From DEPARTMENT OF MEDICAL EPIDEMIOLOGY
AND BIOSTATISTICS
Karolinska Institutet, Stockholm, Sweden

STATISTICAL METHODS FOR THE DETECTION, ANALYSES AND INTEGRATION OF BIOMARKERS IN THE HUMAN GENOME AND TRANSCRIPTOME

Chen Suo

Stockholm 2014

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice AB

ISBN 978-91-7549-552-1

Statistical Methods for the Detection, Analyses and Integration of Biomarkers in the Human Genome and Transcriptom

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Chen Suo

Principal Supervisor:
Professor Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor(s):
Doctor Stefano Calza
University of Brescia
Department of Molecular and Translational Medicine

Doctor Agus Salim
La Trobe University
Department of Mathematics and Statistics

Opponent:
Professor Wolfgang Huber
European Molecular Biology Laboratory
Genome Biology Unit

Examination Board:
Professor Mats Gustafsson
Uppsala University
Department of Medical Sciences
Cancer Pharmacology and Computational Medicine

Docent Keith Humphreys
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Docent Erik Kristiansson
Chalmers University of technology
Department of Mathematical Science

For my Family. For my Love.

ABSTRACT

Most human diseases have been shown to have a genetic basis that is linked to regulation of gene expression at the transcriptional or post-transcriptional level. In the central dogma of biology, deoxyribonucleic acid (DNA) is transcribed to messenger ribonucleic acid (mRNA), and then translated into proteins; dysfunction in any of these processes may contribute to the development of disease. Sources of such potential irregularities include, but not limited to, the following: point mutations in DNA sequences, copy number alterations (CNAs) and abnormal mRNA and microRNAs (miRNAs) expression. MiRNAs are a type of non-coding RNA that inhibit the transcription and/or translation of specific target mRNAs. Current technologies allow the identification of biomarkers and study of the complex interplay between DNA, mRNA, miRNA and phenotypic variation. This thesis aims to tackle the statistical challenges that have arisen with the application of these technologies to investigate various genomic and transcriptomic alterations.

In study I, modified least-variant set normalization for miRNA microarray, a new algorithm and software were developed for microRNA array data normalization. The algorithm selects miRNAs with the least array-to-array variation as the reference set for normalization. The selection process was refined by accounting for the considerable differences in variances between probes. Data are provided to show that this algorithm results in better operating characteristics than other methods.

In study II, joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-Seq data, a joint model and software were developed to estimate isoform-specific read distribution and gene isoform expression, using RNA-sequencing data from multiple samples. Observation of similarities in the shape of the read distributions solves the problem that the non-uniform read intensity pattern is not identifiable from the data provided by one sample.

In study III, integrated molecular portrait of non-small cell lung cancers, molecular markers at the DNA, mRNA and miRNA level that can distinguish between different histopathological subtypes of non-small cell lung cancer were identified. Additionally, using integrated genomic data including CNAs and mRNA and miRNA expression data, three potential driver genes were identified in non-small cell lung cancer, namely *MRPS22, NDRG1* and *RNF7*. Furthermore, a potential driver miRNA, *hsa-miR-944*, was identified.

In study IV, integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. An analytic pipeline to process large-scale whole-genome and transcriptome sequencing data was created, and an integrative approach based on network enrichment analyses to combine information across different types of omics data was proposed to identify putative cancer driver genes. Analysis of 60 patients with breast cancer provided evidence that patients carrying more mutated potential driver genes had poorer survival.

LIST OF SCIENTIFIC PAPERS

This thesis is based on the following original articles which will be referred to in the text by their Roman numerals.

I. **Suo C**, Salim A, Chia KS, Pawitan Y and Calza S: Modified least-variant set normalization for miRNA microarray. RNA 16(12):2293-303, 2010.

II. **Suo C**, Calza S, Salim A and Pawitan Y: Joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-Seq data. Bioinformatics 30(4):506-13, 2014.

III. Lazar V, **Suo C**, Orear C, van den Oord J, Balogh Z, Guegan J, Job B, Meurice G, Ripoche H, Calza S, Hasmats J, Lundeberg J, Lacroix L, Vielh P, Dufour F, Lehtiö J, Napieralski R, Eggermont A, Schmitt M, Cadranel J, Besse B, Girard P, Blackhall F, Validire P, Soria J, Dessen P, Hansson J and Pawitan Y: Integrated molecular portrait of non-small cell lung cancers. BMC Medical Genomics 6:53, 2013.

IV. **Suo C**, Lee D, Pramana S, Saputra D, Joshi H, Calza S and Pawitan Y: Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. Manuscript.

CONTENTS

LIST OF ABBREVIATIONS

The abbreviations below have been used in this thesis and in the four original publications attached at the end.

AC	Adenocarcinoma
aCGH	Array-based comparative genomic hybridization
AGS	Altered gene set
ASTD	Alternative Splicing and Transcript Diversity
AUC	Area under curve
BAM	Binary aligned format
BWA	Burrows-Wheeler alignment tool
CCA	Canonical correlation analysis
CGH	Comparative genomic hybridization
CHEMORES	Chemotherapy resistance consortium
CNA	Copy number alteration
CNV	Copy number variation
DE	Differentially expressed
DGS	Driver-Gene search
DGscore	Driver gene score
DNA	Deoxyribonucleic acid
FC	Fold change
FDR	False discovery rate
FGS	Functional gene set
GEA or GSEA	Gene-set enrichment analysis
GEO	Gene Expression Omnibus
GLM	Generalized linear model
GWAS	Genome-wide association studies
Inv-P	Invariant-set normalization

IWLS	Iterative weighted least square
LCC	Large-cell carcinoma
lowess	Locally weighted scatter plot smoothing
LVS	Least variant set
MAQC	Microarray Quality Control project
miRNA/miR	MicroRNA
mRNA	Messenger ribonucleic acid
NEA	Network enrichment analysis
NGS	Next generation sequencing
NSCLC	Non-small cell lung cancer
OC	Operating characteristic
PC	Principal component
qPCR	Quantitative TaqMan real-time Polymerase chain reaction
RLM	Robust linear model
RNA	Ribonucleic acid
RPKM	Reads per kilobase per million reads
SCC	Squamous cell carcinoma
SCCA	Sparse canonical correlation analysis
SCS	Single-cell sequencing
SD	Standard deviation
SNP	Single nucleotide morphism
TGS	Third-generation sequencing
T/N	Tumor/normal
TCGA	The Cancer Genome Atlas
VSN	Variance stabilizing normalization

LIST OF TABLES

LIST OF FIGURES

1 CHAPTER 1 - INTRODUCTION

It has been found that genetic variations and patterns of gene expression variability are associated with most complex human diseases, such as cancer, diabetes and Alzheimer's disease (Stratton *et al*., 2009). Deoxyribonucleic acid (DNA) and messenger ribonucleic acid (mRNA) play fundamental roles in the central dogma of biology (Figure 1.1), in which DNA is transcribed to mRNA and then translated to protein. Irregularities at the DNA and RNA level would potentially disturb normal biological processes, and lead to a transformed phenotype.

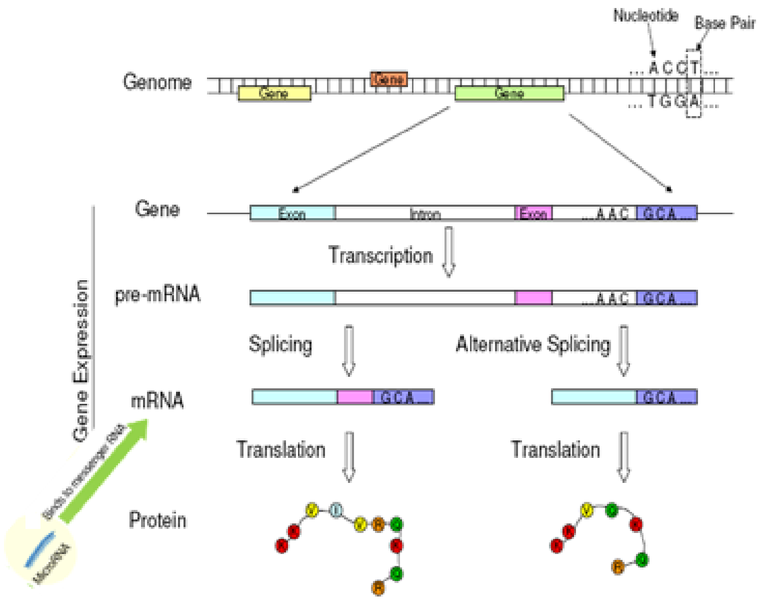


Figure 1.1 DNA is transcribed to mRNA and then translated to protein. In the transcription process, isoforms are generated due to alternative splicing. MicroRNAs, another type of RNA, bind to target mRNAs and negatively control their expression through both the transcriptional and post-transcriptional regulation. (Adapted from Jiang, 2009).

The genome and transcriptome can 'go wrong' in many aspects. Variation at the DNA level can occur in a variety of ways, including a single substitution of a nucleotide, duplication or deletion of a large segment of nucleotide sequences, and other structural variations. DNA contains genes, which are templates to produce proteins. How much protein is produced is regulated by the level of gene expression, which is a measure of gene activity. An abnormal level of gene expression is thus an indicator of abnormal production of proteins, and subsequent phenotypic changes. In recent years, the substructures of genes – isoforms – have attracted much attention in disease association studies. As shown in Figure 1.1, during the transcription process, introns are removed and exons are joined together to form the template to produce proteins. Due to alternative splicing, different combination of exons is used to form the isoform. So for a gene containing multiple isoforms, its expression level is in fact the sum of individual isoforms expression. Studying the isoform level expression is useful for identifying the exact transcript associated with the phenotype. During the transcription process, microRNAs (miRNA; another type of RNA) are also transcribed. MicroRNAs have been found to negatively regulate gene expression, and are widely recognized as potential biomarkers of disease (Pasquinelli *et al*., 2005; Fabbri *et al*., 2008; Guarnieri and DiLeone, 2008).

Microarray technology has been widely used to simultaneously measure relative mRNA expression levels of tens of thousands of genes. The development of microarray technology has provided researchers and clinicians with a great opportunity to identify potential genetic features that show differential expression across subjects with different phenotypes. In order to accurately estimate the expression levels from raw intensity values, one of the most important steps is the selection of appropriate pre-processing methods which commonly entail background correction, summarization and normalization. While a huge amount of work has been done for well-known array platforms such as mRNA and comparative genomic hybridization (CGH) array data, miRNA platforms have only recently been investigated. High-throughput RNA sequencing is another advanced tool at the forefront of expression level measurement. However, there are also problems regarding the accuracy of absolute expression estimation using RNA-sequencing data.

Currently, genetic molecules are identified through separate studies conducted on different platforms. Genetic profiles identified in a single study tend to be of the same class (for example, DNA, mRNA or miRNA). Different platforms are supposed to add different information; however, methods that utilize biological knowledge to combine data from different measurements are lacking. When integrated properly, different types of signal detected from the increasing amount of genomic and molecular data could lead to investigations which may result in better understanding of the regulatory network of disease-causing pathways. Ideally, this could be used to study the way genes function and interact with each other.

While microarray and sequencing technology has developed rapidly in recent years, comprehensively extracting information from different platforms and analysing data efficiently still pose statistical challenges. In projects I and II, we aimed to address the issues by developing a normalization method and an expression estimation method which adapts to the underlying data type and relies on few assumptions. Despite the great success of genome-wide mRNA expression analysis in biomedical research, current methods have focused on analysis of a single type of marker that accounts for only a small proportion of risk variants. We anticipated that the new statistical methodologies, proposed in projects III and IV, to utilize data from multiple platforms and to identify biologically plausible pathways would provide the tool and knowledge required to facilitate the joint discovery of genetic and molecular risk factors for various diseases.

The development of new statistical methods and software is much needed, in order to detect, analyse and integrate genetic data. The main focus of this thesis was the development and application of computational analysis (1) to quantify expression levels accurately from raw array and sequencing data and (2) to distinguish between critical genetic alterations, which are potentially driving tumorigenesis, and functionally neutral mutational events.

This thesis consists of six chapters. In Chapter 2 we provide a review of the genetic and transcriptional features studied in the four projects, explaining commonly used terms as well as discussing the platforms used to generate the data and potential statistical problems arisen from the technology. The aim of the four studies is described briefly in Chapter 3. In Chapter 4 and 5, the motivation, methods and results of each study are outlined and discussed. Finally, in Chapter 6, our contribution to the field is summarized with concluding remarks.

2 CHAPTER 2 - BACKGROUND

In this chapter, I will introduce the most commonly used terms and concepts in whole-genome and transcriptome analysis, and provide an overview of existing methods and technologies used in data generation, pre-processing, analysis and integration. This may be helpful in understanding the following chapters.

2.1 GENETIC VARIATION

In humans, there are 22 pairs of chromosomes, termed autosomes, plus the X and Y sex chromosomes. Chromosomes carry genetic coding information, encoded as sequences of nucleotides A, T, C and G. The entire genetic information of an organism is termed the *genome*. There are more than 3 billion bases in the human genome (International Human Genome Sequencing Consortium, 2004).

Genetic variations are differences in the DNA sequence of the genome relative to a reference genome. A genetic variation can occur in single or multiple nucleotides, and the latter can occur on a small or large scale. Small-scale variations are known as insertions or deletions while variations on a large scale, more than 1 kb, are known as copy number variations (CNVs) (Freeman *et al*., 2006). Healthy individuals can carry genetic variations; it is important to distinguish between variations among normal and abnormal biological samples. Conventionally, the term single nucleotide polymorphism (SNP) refers to a substitution in a single base pair at the general population level with a common frequency of at least 1%, whereas rare variants with a <1% frequency are often considered to be mutations. There are two types of mutations, *germline* and *somatic*. Germline mutations are inherited and passed from parent to child, whereas somatic mutations are not inherited; for example, mutations in tumour tissues are not passed to the next generation. Similarly, CNV refers to variations in the number of copies of a section of DNA that occur in the general population, and copy number alteration (CNA) usually refers to potentially harmful CNVs in diseased persons. In this thesis, we focus on somatic mutations and CNAs, to investigate their relationship with cancer. Currently, microarray and sequencing technologies can simultaneously measure DNA changes in thousands to millions of loci in the genome (this is reviewed in section 2.4).

2.1.1 Somatic mutation

In human genetics, a mutation is an alteration in the nucleotide sequence of the genome, which may or may not cause phenotypic changes. Mutations can result in several different types of consequences in the subsequent biological process; mutations in genes can have no effect, prevent the gene from functioning properly if not functioning at all. Results from a study of genetic variations between different species of *Drosophila* suggest that 70% of protein changes produced by a gene mutation are likely to be harmful (Sawyer *et al*., 2007). Human cancer is caused at least in part by such mutations (Stratton *et al*., 2009).

The occurrence and rate of somatic mutations are related to many factors. Mutations can occur during DNA replication and may be repaired by a biological pathway known as DNA repair. Thus if the function of DNA repair is damaged due to mutation, intuitively we know that the mutation rate may be increased. In fact, it has been established that DNA repair is an important pathway in the development of cancer. The number of somatic mutations can differ considerably between individuals, because some have a high background mutation rate (Conrad *et al*., 2011). Mutations may also occur because of external environmental factors, such as radiation or extreme heat.

With rapid advances in DNA sequencing technology, a huge amount of sequencing data is generated and can be used for identification of mutations. The chi-squared test is probably the simplest method for detection of somatic mutations, by examining the allelic proportions in tumour and normal tissues. Various programs, such as MuTect (Cibulskis *et al*., 2013), have been developed to identify somatic point mutations in next-generation sequencing (NGS) data of cancer genomes.

2.1.2 Copy number alteration

CNA is a form of large-scale alteration in the number of copies of one or more sections of DNA. In contrast to point mutations, which affect only a single nucleotide base, CNA affects a relatively large region ranging from 1 kb to several megabases in the DNA sequence. CNA either increases or decreases the normal number of nucleotides on the chromosomes, potentially resulting in abnormal cell function and thus a transformed phenotype, especially when the altered region contains certain genes (Hastings *et al*., 2009). CNA may also have an impact on expression level. Over-expressed genes in amplified or duplicated regions and under-expressed genes in regions of deletion are likely to contribute to cancer development (Santarius *et al.*, 2010; Shridhar *et al*., 2002).

Approximately 12% of human genomic DNA may contain CNVs (Stankiewicz and Lupski, 2010). Comprehensive characterization of this type of genetic defect is necessary for understanding the molecular aetiology of cancer and contributing to the realization of targeted treatment. CNA profiling is mainly array-based to generate raw intensity data of DNA aberrations in tumour samples; recently sequencing has also become popular despite problems due to non-uniform sequencing coverage in CNA detection. In this thesis, only array-based comparative genomic hybridization (aCGH) data have been used. For detection we have used a combination of two R packages, MPSS (Teo *et al*., 2011) and CNVpack (Teo *et al*., 2010). Firstly, the former takes a robust smooth segmentation approach to identify whether a segment is a true CNA, then the latter identifies recurrent CNA regions that are found in at least 10% of individuals.

Similar to mutations, it has been found that CNAs are also associated with the occurrence of cancer. Gene copy number can be increased in cancer cells. For instance, the *EGFR* copy number was found to be increased in non-small cell lung cancer (Sebat *et al*., 2004). By contrast, a deletion in exons 24 and 25 of the tumour suppressor gene *RB* can cause low-penetrance retinoblastoma (Bremner *et al.*, 1997).

The CNA regions most likely to contain genes central to disease initiation and progression are those that recur among diseased individuals (Pinkel *et al.*, 2005). Such regions probably contain the so-called "driver" alterations, which are functionally important changes, rather than "passenger" alterations, which do not have pathological relevance.

2.2 TRANSCRIPTIONAL BIOMARKERS

2.2.1 Genes and isoforms

The transcriptome is the entire repertoire of all transcripts in a species, and is a key link between information encoded in DNA and proteins. It is a challenge to fully quantify the large number of transcripts in the transcriptome. In humans, there are over 20,000 genes in total (International Human Genome Sequencing Consortium, 2004). Generated by the mechanism of alternative splicing, isoforms are mRNAs that are produced from the same gene but are different in their protein sequence thus potentially altering function. Many isoforms are known to be implicated in a wide range of human diseases and functional roles (Nagao *et al.,* 2005; Wang *et al*., 2010); for example, aberrant splicings of the *PTCH* gene have been detected in patients with autosomal dominant nevoid basal cell carcinoma syndrome (Nagao *et al.,* 2005).

As the importance of alternative splicing (which greatly diversifies the transcriptome) becomes clear, whole-transcriptome shotgun sequencing (RNA-Seq) is rapidly gaining popularity as it offers the possibility of detecting isoform expression. But the use of the technology requires much more effort in terms of statistical modelling in order to make accurate estimation of expression.

During RNA sequencing, millions of reads are generated. The number of nucleotides in a read is termed the "read length". These reads will be aligned to their genomic position on a reference genome. It should be evident that more reads will be mapped to a gene if its transcript is longer and the sequencing is deeper. So for gene expression estimation, we only need to sum up the number of reads falling into a gene, as the number of reads coming from a gene is proportional to the number of copies of transcripts produced by a gene. However, isoform expression estimation is not that straightforward because which isoform a given read comes from is unknown. For example, Figure 2.1 shows a three-exon gene with two isoforms: a read mapped to exon 1 may come from either isoform 1 or 2. Thus, statistical modelling is required to estimate isoform-specific expression. Note that a read can also be mapped to a region that lies on the exon–exon boundary within an isoform; this is known as a junction read. A junction read may provide more information to quantify isoform expression; for example in the gene shown in Figure 2.1, a junction read between exons 1 and 3 can only come from isoform 2.

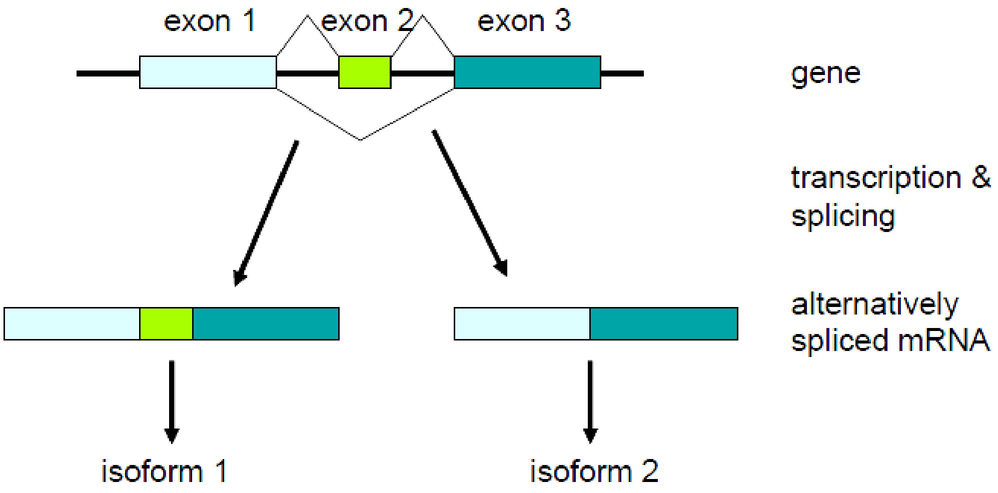


Figure 2.1 A simplified diagram of alterative splicing events. (Adapted from http://en.wikipedia.org/wiki/Alternative_splicing).

Additional difficulties may arise for example from annotation accuracy and counting duplicate reads. The complexity of the analysis may prevent many researchers from using sequencing techniques and benefiting from isoform expression quantification. Therefore reliable and user-friendly statistical tools need to be developed to advance the use of RNA-sequencing techniques. In paper II, we address the problem of non-uniform read distribution (Suo *et al*., 2014).

2.2.2 MicroRNA

MicroRNAs (miRNAs) are short (~18 to 24-nucleotide) non-coding RNAs. Although relatively short in terms of nucleotide length, miRNAs play an important role in gene regulation as they negatively regulate mRNA expression by binding to the 3’ untranslated region of their target mRNAs. Both the transcriptional and post-transcriptional regulation of gene expression is mediated by miRNAs (Calin and Croce, 2006). The miRBase database (release 21) reported 1,881 human miRNAs, each of which may potentially regulate many target genes (Kozomara and Griffiths-Jones, 2014).

It has been recognized that miRNAs have a significant role in human cancer (Pasquinelli *et al*., 2005; Fabbri *et al*., 2008; Guarnieri and DiLeone, 2008). Volinia *et al.* (2006) found that cancer cells showed distinct miRNA profiles compared with normal cells. In addition, recent evidence suggests that miRNAs might also function as tumour suppressors and oncogenes, damage of which may be selected for in cancer (Shenouda and Alahari, 2009).

Much emphasis has been placed on studying the impact of genetic alterations and patterns of gene expression variability related to cancer (Akavia *et al*., 2010). Genomic aberrations and miRNA expression should also be studied simultaneously to obtain a comprehensive understanding of tumour formation. In paper III (Lazar *et al*., 2013), we proposed an integrative model that identifies the potential driving role of miRNAs and mRNAs in cancer.

2.3 PATHWAYS AND FUNCTIONAL NETWORKS

Gene-set enrichment analysis (GEA or GSEA) is commonly used to characterize experimentally derived altered gene sets (AGSs); for instance, features identified as having the top significant *P*-value in association studies, or a list of genes containing high-impact mutations in coding regions. Based on known functional databases, such as Gene Ontology or KEGG, GEA identifies previously known functional gene sets (FGSs) that are over-represented in AGSs, using a hypergeometric test. Although this is easy to understand and simple to compute, the fact that the majority of genes have not been assigned to a biologically informative category is problematic. The sensitivity of these analyses is related to the size of the AGS, but increasing the number of genes in the AGS, e.g. including genes that are less differentially expressed, is not always biologically meaningful, as it may also increase the number of false-positive results in the AGS. Moreover, it is not feasible to perform GEA if there is only one gene in the AGS.

To overcome this problem, several network-based methods have been proposed to reveal network patterns that are enriched compared to those expected by chance, based on biological knowledge of gene and protein interaction (Huttenhower *et al.,* 2009; Shojaie and Michailidis, 2010). To integrate the topological information in the gene network and the functional information about biological processes, Alexeyenko *et al*. (2012) presented a method of network enrichment analysis (NEA) that systematically implements the network approach to describe novel gene sets with biologically meaningful functional categories. The method integrates two types of biological data: functional information and network connectivity of nearly all protein-coding genes. In contrast to traditional GEA, the NEA method quantifies the over- and under-representation of the functional group members among the neighbours in the gene network rather than simply counting the number of AGSs in a known pathway or FGS.

NEA provides the possibility of combining the results of molecular data from different experiments; for example, to investigate the relationship between genome-wide mutation analysis and gene isoform expression in order to identify whether mutations have a functional impact on the protein network. This method has been utilized in papers III and IV. It is a critical step to detect putative cancer drivers (Suo *et al.*, manuscript) and to characterize identified potential driving genes (Lazar *et al*, 2013).

2.4 TECHNOLOGIES

2.4.1 Microarray

Microarray has been the most commonly used method to measure targeted loci or genes of interests for almost 20 years (Lashkari *et al*., 1997). It allows simultaneous profiling of thousands of genetic features, such as SNPs, CNVs, mRNAs and miRNAs. Typically an mRNA array measures the expression of over 20,000 genes, and an miRNA array measures the expression of ~1,000 microRNAs.

However there are still problems in the preprocessing of the data. In addition to the true signal, raw microarray data may exhibit systematic differences between samples due to bias introduced by technical factors. Proper normalization is one of the critical steps in order to ensure downstream suitable comparative data analysis in terms of minimizing false negative and false positive results.

Because array technology is based on the mutual and specific affinity of DNA strands, it relies on a known reference genome and transcriptome before the microarray platform can be designed. The technology also has limited probe density, especially in detecting SNPs or CNVs where the number of targeted loci can reach a few million; however this number is very low considering there are 3 billion base pairs in the human genome.

2.4.2 First-generation sequencing

First-generation sequencing was first developed by Frederick Sanger in 1977 (Sanger *et al*., 1977) and is usually referred to as “Sanger sequencing”. Typically, read lengths are ~800-1000 bases in Sanger sequencing (Hert *et al*., 2008). Sanger sequencing was the major sequencing method in general use for almost 30 years, until NGS was introduced about 10 years ago.

Although long reads reduce mapping error, Sanger sequencing is tedious and expensive; it is not able to process more than 96 sequence reads in a single run, limiting its application to large-scale genome-wide sequencing efforts for many individuals (Mardis, 2008). For instance, using Sanger sequencing it would take more than 400 years to completely sequence the *Drosophila melanogaster* genome, which with only 120 million bases, is much smaller than the human genome (3 billion bases) (Schadt *et al.*, 2010).

2.4.3 Next-generation sequencing

The use of NGS has grown rapidly during the last decade. This technology permits global measurement of the whole genome or transcriptome to produce large amounts of sequencing reads in a single run within a short time frame, in a cost-effective manner relative to traditional Sanger sequencing (Metzker, 2010). NGS allows a DNA fragment to be repeatedly sequenced (a procedure known as *deep sequencing*), delivers greatly increased sensitivity and accuracy, and has revolutionized the world of genomics. The technique has most recently been extended to the analysis of the transcriptome by what is known as RNA-Seq. Current commercial NGS systems include Illumina, Applied Biosystems Supported Oligonucletide Ligation Detection System (SOLiD), the Roche 454 and so on. NGS has the potential to measure all known mutations, structure variants in the genome, genes and isoforms and miRNAs in the transcriptome and, furthermore, to discover novel variants. To facilitate and accelerate the process of identifying genetic variations at the population level, whole-genome sequencing of a large number of individuals was performed at great effort by the 1000 Genomes Project (http://www.1000genomes.org). To characterize disease-specific alterations in cancer genomes, the International Cancer Genome Consortium (ICGC; http://www.icgc.org/home) and The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) sequenced over 20,000 cancer genome in at least 50 types of cancer.

In sequencing experiments, millions of reads are generated and stored in FASTA format. The aligned reads are usually saved as Binary Aligned Format (BAM), from which read count information can be computed for downstream estimation and analysis. However, there are some technical problems in processing and analysing NGS data. Firstly, due to polymerase chain reaction (PCR) amplification bias and sequencing error, initial raw reads must be pre-processed and filtered properly. Secondly, annotation is still incomplete. Inaccurate annotation on gene-isoform structure may cause bias in estimating isoform-level expression. Thirdly, non-uniform read coverage is an important issue especially in RNA-sequencing experiments, because the coverage not only ensures adequate information but is also related to the expression level to be estimated. Several issues may further complicate the use of sequencing technology, for example counting reads that span more than one region, multiple mapped reads and the challenge of dealing with paired-end as compared to single-end reads.

In this thesis, whole-genome sequencing data have not been analysed; instead, we analysed Exome-seq data, which is a less expensive alternative approach that only sequences the exon regions of the genome. It is known that exons comprise roughly 1% of the genome (Gilissen *et al*., 2011), so the compromised approach reduces the sequenced region by 99%, while the most informative sources of genetic variation remain. An important project to identify genetic variants in coding regions is the Exome Sequencing Project (ESP; http://evs.gs.washington.edu/EVS/). This is a multi-cohort project on heart, lung and blood disorders, to discover novel genes and mechanisms contributing to the diverse phenotypes.

2.5 NORMALIZATION ALGORITHMS FOR MIRNA ARRAY

Microarrays measure relative expression level as intensity value for each arrayed feature, such as mRNAs and miRNAs. Subsequently these intensity values can be used to identify biologically relevant patterns of expression by comparing the intensities between biological conditions on a feature-by-feature basis. For example, comparing the expression level of particular genes between a diseased and a normal tissue can provide useful information about early diagnostic biomarkers. However, in addition to the true signals, observed expression levels usually include technical variation that is introduced at all stages of the experiment, therefore direct comparisons of the raw intensity data are inappropriate. Obscuring variation may arise as a result of a number of potential sources (Hartemink *et al*., 2001), including dye bias, hybridization bias and batch bias. Hence, to reduce these systematic technical sources of bias, a normalization step is needed before the expression levels can be appropriately compared. Different choices of normalization methods exist (all previously developed for mRNA arrays), but there is no consensus on their relative performance for miRNAs. Below, several commonly used methods for mRNA array normalization are reviewed.

Global normalization

The simplest method for mRNA array normalization is probably global normalization, which is applied to each array independently. The basic idea of this method is to centre the mean or median values based on total array intensities, resulting in the mean or median intensity among chips equal to an arbitrary target value T. For example, in Affymetrix Microarray Suite 5.0 (MAS5), the target value is 500. The following equation is used for normalization:

$$X_i^{norm} = \frac{T}{\acute{X}_i} \times X_i = k \times X_i,$$

where X_i is the raw intensity, $\acute{X}_i$ is the mean intensity of array *i*, X_i^{norm} is the intensity after normalization and *k* is the normalization factor.

Lowess normalization

Global mean normalization described above utilizes a linear scaling technique. However, due to different background intensities or dye effects in labelling, data from cDNA microarray studies may exhibit a banana-shape in an MA plot, where the x-axis is the average log intensity and the y-axis is the log intensity ratio. Yang *et al*. (2002) proposed that lowess smoothing, also known as locally weighted regression (Cleveland, 1979) and smoothing scatterplots, can correct for non-linear array-to-array variation. This method assumes that changes of intensities between arrays are roughly symmetrical across all intensities instead of only around the mean or median. For two-colour arrays, the data points are commonly displayed in an MA plot. For single-colour platform, lowess normalization is also applicable in the sense that it corrects for pairs of arrays to be normalized to each other. The normalized intensity ratio is computed using the following equation:

$$log\left(\frac{X_r}{X_g}\right)^{norm} = log\left(\frac{X_r}{X_g}\right) - c(A)$$

where X_r and $X_g$are intensity data points from two arrays and c(A) is the lowess fit to the MA plot.

The procedure cycles through all pairwise combinations of arrays until convergence. An alternative to lowess is smooth spline normalization, which is also an intensity-dependent method. Thus, non-linear normalization correction compensates for intensity-dependent bias, and the results of a simulation study showed that it performs better than global normalization methods in terms of standard deviation between samples (Park *et al.*, 2003).

Quantile normalization

The aim of quantile normalization is to equalize the distribution of intensities for all arrays (Bolstad *et al*., 2003). Simply forcing equal mean or median intensity for the arrays using global normalization, and likewise equalizing means at all intensities using the lowess technique, might not be sufficient because the entire distribution may vary. As explained by Bolstad *et al*. (2003), the algorithm for quantile normalization follows these steps:

1. For *n* arrays of length *m*, a matrix with intensities as rows and samples as columns is formed.
2. Each column is sorted.
3. The mean intensity is computed in each rank across chips.

4. Each intensity value is replaced by the mean intensity of its rank and finally the new intensities are brought back to the original order.

Using only the observation ranks and thus no particular distribution assumed, the algorithm is also able to manage quite nasty non-linear trends. It reduces variance slightly better than lowess, runs relatively fast and is easy to implement.

Inv-P and housekeeping gene normalization

A common feature of the above methods is that they use the whole set of genes with the assumption that the majority of genes do not vary across samples. But this assumption is not always satisfied, either in spike-in experiments or in many studies with real data where imbalanced regulation might occur (Porter *et al.*, 2002; Haslett *et al*., 2003; Timmons *et al*., 2005). To overcome this problem, a data-driven procedure to select genes that do not vary across arrays has been proposed (Li and Wong, 2001), thus providing a good subset of reference genes for normalization. To select the reference set, the procedure attempts to identify genes that are expressed at similar levels in the compared samples based on ranks, following the idea of using ranks in quantile normalization. An iterative procedure is used to select the so-called invariant set of probes:

1. A reference array is selected, for example, a mean or median array, i.e. for each probe, an average among chips is computed to become the probe expression on the reference chip.
2. Within each chip, each probe is ranked according to expression level and compared to the corresponding value in the reference array. If the change in ranks divided by the total number of probes on the array is smaller than a cutoff value, the probe is selected for the invariant set and excluded from the ranking list.
3. Steps 1 and 2 are repeated until the number of invariant probes is sufficient.

4. All the arrays with the invariant set of genes are normalized towards the reference array created in step 1, based on a lowess or smoothing spline. The fitted curve is then used to map intensities of the non-invariant set of genes in each array to be normalized.

The advantage of invariant set normalization is that it relaxes the assumption on the balanced proportion of over- and under-regulated genes.

Similarly, housekeeping gene normalization chooses reference genes that express approximately the same across samples based on prior biological knowledge, instead of selection of the subset of genes that does not vary based on the dataset itself. But there are hardly any housekeeping genes practically, so this approach is generally deprecated

Variance stabilizing normalization

Briefly, the variance stabilizing normalization (VSN; Huber *et al*., 2002) procedure first makes sample-to-sample linear calibration so that data are on a common scale and have a common distribution. This step assumes that the data of all genes on an array are subject to the same systematic effects. More complex intensity-dependent calibration can also be used to correct for deviations from a linear line.

Next, VSN based on a parametric arcsinh (inverse of hyperbolic sine) transformation is performed to address the dependence of the variance on the mean intensity. Under the assumption of a quadratic relationship between mean expression and variance, affine-linear transformation will transform the data to a scale where the variance of the data is almost independent of the mean. Overall, the total transformation is represented by the equation:

$$h(x) = arcsinh((s_0 + \lambda_0(\lambda_i S_{ij} + s_i))$$

where λ_i and s_i are the scaling and shifting parameters, respectively, in linear transformation, and λ_0 and s_0 are the parameters in the variance stabilization step. The parameter h can be estimated with a robust variant of maximum-likelihood estimation.

One advantage of VSN transformation is that the arcsinh function is continuous across zero and coincides with log-ratio values at high intensity. Compared to the more traditional log transformation, VSN is able to manage initially negative intensities.

2.6 EXPRESSION QUANTIFICATION ALGORITHMS

Advanced computational methods are required for expression quantification with RNA-Seq due to sequencing bias and the large number of parameters that need to be estimated for all genes and their isoforms. A number of models have been derived and some typical examples are described below.

In the score proposed by Mortazavi *et al*. (2008), the expression level of a transcript is first measured in reads per kilobase of the transcript per million mapped reads (RPKM), which is a normalized measure of counts against the sequencing depth and transcript length. This expression score allows a relatively fair comparison between measurements across genes and samples, compared to raw counts. However it is not possible to compute isoform-level RPKM because we cannot distinguish reads mapped to an exon shared by more than one isoform.

Jiang *et al*. (2009) developed a Poisson regression-based approach to model the relationship between read counts mapped to exons and isoform-specific expression. This simple model has been developed and used in several subsequent RNA-Seq studies (Wang *et al*., 2010; Li *et al*., 2010) and is referred to as the *standard method* in this thesis. However, the method relies on the key assumption of a uniform sampling of reads across transcripts.

To correct for the non-uniform read distribution, Cufflinks (Trapnell *et al*., 2010), one of the most commonly used tools, accounts for sequence-specific bias problems in isoform expression estimation and assumes uniform read distribution in its basic model, and providing an ad hoc correction of the bias step (Roberts *et al*., 2011). The model also estimates positional bias, which determines whether fragments are preferentially located towards either end of the transcripts. Unlike the base-level bias correction method, NURD models both the read distribution and expression jointly. A global bias curve is estimated for all genes, and an approximate local bias curve for each gene is estimated using non-parametric models (Ma and Zhang, 2013).

In chapter 4.2, we describe in detail the method and software we propose for isoform-specific read distribution and gene isoform expression estimation (Suo *et al*., 2014). In brief, the method assume the same read intensity distribution at isoform level across different samples, and the isoform expressions are estimated using Poisson model where the Poisson rates reflects the non-uniform read intensity, expression level and isoform length.

3 CHAPTER 3 - AIMS

The overall aim of this thesis is to develop and apply statistical and bioinformatics tools to improve the accuracy of estimating expression levels when using microarray and sequencing technologies, in order to analyse and integrate biomarkers associated with cancer. The thesis is divided into four studies, as described below:

I. A method was developed and implemented as an R package LVSmiRNA to normalize miRNA expression microarray data. In downstream analysis, normalized intensity values by LVSmiRNA resulted in a better performance in terms of sensitivity and specificity to detect differentially expressed miNRAs, compared to the values normalized by other methods.

II. A method and software were developed to jointly estimate gene and isoform expression levels, as well as read intensity patterns, in RNA-sequencing experiments. This was challenging, considering the large number of parameters estimated, noise in RNA-sequencing data and the time consuming nature of dealing with large-scale datasets.

III. The molecular profiles between adenocarcinoma and squamous cell carcinoma samples from non-small cell lung cancer were characterized. Potential driver genes were identified by integration of genomic data on copy-number alterations, mRNA expression and microRNA expression.

IV. A pipeline for processing, analyzing and integrating genomic and transcriptomic sequencing data was designed. The pipeline incorporates existing bioinformatics tools, as well as a novel proposed method for integrating mutation and expression profiles based on network enrichment analyses. Real data from The Cancer Genome Atlas (TCGA) project was used to demonstrate application of the pipeline and investigate the prognostic value of the identified potential driver genes for the overall survival of patients with breast cancer.

4 CHAPTER 4 - PAPER SUMMARIES

4.1 STUDY I: MODIFIED LEAST-VARIANT SET NORMALIZATION FOR MIRNA MICROARRAY

4.1.1 Motivation

Traditional normalization algorithms for gene expression arrays rely on the expression levels of a large number of genes/features not varying across samples. However, there is a substantially smaller total number of miRNAs and large numbers of miRNAs are differentially expressed (DE) between samples.

4.1.2 Methods

The main novelty of this algorithm is that it accommodates for the potential heterogeneous variance between probes, in order to select the most appropriate subset of miRNAs that have the least variation across arrays.

The algorithm was built from the least variant set (LVS) normalization method, which was initially developed for RNA microarrays. The so-called modified LVS normalization method comprises three steps:

1. Fitting of a robust linear model on the background-corrected raw probe-level data, where the mean and variance are modelled jointly.
2. Selection of a subset of miRNAs that have the least variation across arrays, on the basis of the plot of the array-effect test statistics versus the residual standard deviations (SDs) provided by the model in Step 1.
3. Normalization of the raw data at either the miRNA level or probe level, where the miRNA level normalization requires the data to be initially summarized.

4.1.3 Results

Comparison to other algorithms using spike-in data

To assess the performance of the modified LVS normalization method, four samples, for which 697 miRNAs were spiked-in and thus their fold-changes (FC) were known, were analysed. A set of 173 miRNAs having a true constant level of 1 provided the ideal reference set for normalization; however, in practice, the true expression levels of these miRNAs are not available. The four samples were organized into group A and B. The advantages of the LVS normalization method were evaluated in terms of the sensitivity and specificity for detecting differential expression between group A and B compared to various normalization procedures. The LVS normalization method achieved a similar level of sensitivity and specificity compared to the best possible normalization method based on FC1 miRNAs. The LVS normalization method was also superior to other normalization methods, including no normalization, the 75th percentile shift, quantile, inv-P, global median, VSN and locally-weighted scatter-plot smoothing (lowess) methods (Figure 4.1).

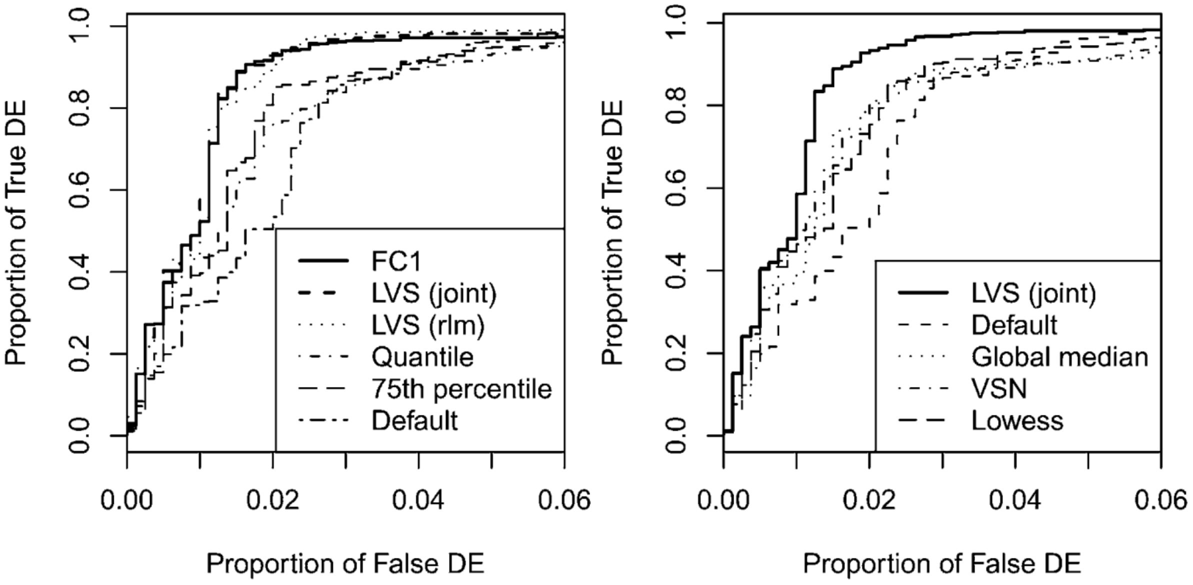


Figure 4.1 Sensitivity and specificity of the normalization methods for spike-in data. Proportions of true discoveries are plotted against the proportion of false discoveries. Positives are defined as miRNAs both present and with FC not equal to 1. (Figure from Suo *et al.*, 2010).

Comparison to other algorithms using RT-PCR data

The expression patterns of brain and heart samples, which are considered to be very distinct tissues, and skeletal muscle and heart, which are expected to have a lower level of variation, were compared. True expression fold changes between the tissues were computed from qPCR data as an independent "gold standard" method. To assess the effect of different normalization algorithms, their ability to detect the fold changes computed from the normalized expression levels was evaluated in terms of sensitivity and specificity. The LVS normalization method performed better than the other procedures tested, regardless of whether relatively large FC were detected (brain vs. heart) or the expression levels in the tissues were similar, making detection potentially more difficult (skeletal muscle vs. heart; Figure 4.2).

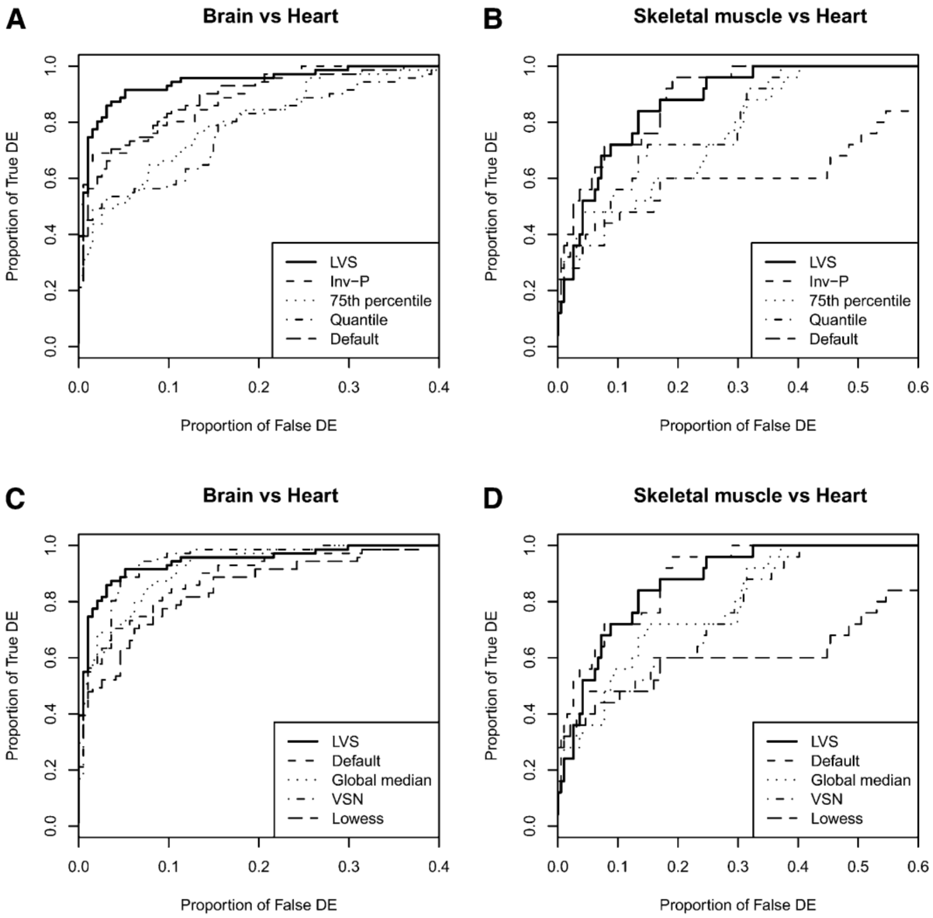


Figure 4.2 Sensitivity and specificity analysis of the normalization methods both in two extremely different tissues (brain and heart) and in two similar tissues (skeletal muscle and heart). Proportions of true discoveries are plotted against the proportion of false

discoveries. Positives are defined as miRNAs with a FC (FC = brain or skeletal muscle/heart) >3, either over- or underexpression. Panels (A) and (C) show OC curves for brain vs. heart comparisons for all the different methods considered. Similarly, panels (B) and (D) show OC curves for skeletal muscle vs. heart comparisons. LVS has the advantage of being flexible enough to successfully adapt to either situation. (Figure from Suo *et al.*, 2010).

Implementation

The method is implemented in an R package called LVSmiRNA. The package can be downloaded from the Bioconductor website (http://www.bioconductor.org/packages/release/bioc/html/LVSmiRNA.html).

4.2 STUDY II: JOINT ESTIMATION OF ISOFORM EXPRESSION AND ISOFORM-SPECIFIC READ DISTRIBUTION USING MULTI-SAMPLE RNA-SEQ DATA

4.2.1 Motivation

RNA-sequencing technologies provide a powerful tool for quantification of expression, especially for distinguishing between the expression levels of different isoforms of the same gene. Typically, to estimate isoform expression in a biological sample, the number of reads falling into a transcript unit with multiple isoforms is modelled as a Poisson process with uniform sampling across each isoform. However, the uniform sampling assumption is often violated due to, for example, the local nucleotide composition effect and 5` or 3` bias. The main challenge when estimating non-uniform read intensity patterns is that the pattern cannot be identified from data provided by a single sample. Additionally, the expression pattern could be isoform-specific (Kozarewa *et al.*, 2009), which increases the number of parameters to be estimated. In our study, the interesting observation of similarities in the shape of the read distributions across samples makes it possible to estimate the read distributions.

On this basis, a novel method for jointly accounting for non-uniform isoform-specific read distribution and gene isoform expression estimation was developed. Regularization via a smoothing penalty was imposed to control for the number of parameters when estimating the read distribution.

4.2.2 Methods

A widely accepted model under uniform read distribution assumption is

$$\lambda_{ri} = w_i \sum_{j=1}^{J} L_j X_{rj} \theta_{ji}.$$

Our joint model incorporating isoform-specific read intensity c_{rj} is

$$\lambda_{ri} = w_i \sum_{j=1}^{J} (c_{rj} L_j X_{rj}) \theta_{ji}.$$

The joint estimation of c_{rj} and θ_{ji} can be performed iteratively via a block Gauss-Seidel method. In practice, to ensure a robust and high speed computational procedure, the estimation of θ_{ji} given c_{rj} is realized via a generalized linear model with an identity link function, where iterative-weighted least squares with robust modification is employed against potential outliers (Pawitan, 2001; Chapter 6.7). In addition, when estimating c_{rj} given θ_{ji}, we consider a model with smoothness penalty to allow the possibility of smooth transition between neighboring regions. This is done using a generalized linear mixed model with isoform-specific read intensity as correlated random effects (Pawitan, 2001; Chapter 18).

4.2.3 Results

Implementation

The method is implemented in an R package called Sequgio for fast processing of RNA-Seq data and expression quantification. The package is freely available on the web at http://www.meb.ki.se/~yudpaw.

Comparison to other methods with simulation

Sequgio was compared to three existing methods: Standard method, Cufflinks (Trapnell *et al.*, 2010) and NURD (Ma and Zhang, 2013) using simulated reads for 10 samples from two simulators: 1) a model-based simulator where the parameters are based on real data, and 2) a simulator called RNASeqReadSimulator that is fully independent of our model (http://www.cs.ucr.edu/_liw/rnaseqreadsimulator.html). For each simulator, both the aligned and unaligned reads were generated and examined to determine whether expression estimation was affected by the alignment procedure.

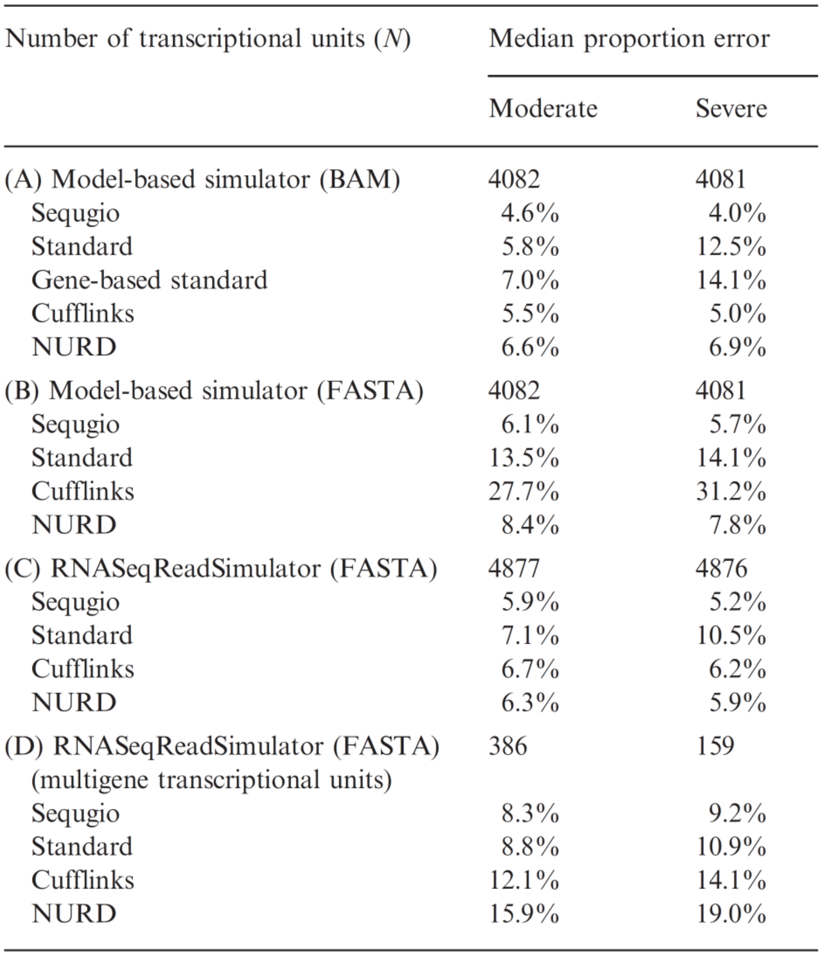
The median proportion error between the predetermined true expression values and the expression estimates are presented in Table 4.1. Overall, Sequgio performed better than the other methods, and provided consistent good performance across various simulation settings.

Sensitivity in differential expression analysis

Using simulated data based on information from a typical multi-isoform transcriptional unit, power analysis showed that larger numbers of true DE transcripts could be identified from the Sequgio estimates than the standard estimates. For genes with a fold change of approximately 1.2, the gain in power was as much as 20%.

Application on a real RNA-Seq dataset

Sequgio and the standard method were applied to publicly available mouse tissue RNA-Seq data, comprising six samples from skeletal muscle, brain and liver tissue (Mortazavi *et al*., 2008). After correction for multiple testing, 68.5% of the standard models had *P*-values < 0.05 in the goodness-of-fit test, while using Sequgio, model fitting improved for 70.3% of the poorly-fitted standard models. Differential analysis of the tissues showed that analysis at the gene-level and isoform-level may lead to different conclusions. Taking brain versus liver tissues for example, 18.7% of the 30,140 transcripts exhibited differential expression at the isoform-level. Among genes containing these DE isoforms, 20.4% did not show a differential expression pattern when analysed at the gene-level, indicating that isoform-level analysis may reveal distinct patterns of expression.



Number of transcriptional units (*N*)	Median proportion error	
	Moderate	Severe
(A) Model-based simulator (BAM)	4082	4081
Sequgio	4.6%	4.0%
Standard	5.8%	12.5%
Gene-based standard	7.0%	14.1%
Cufflinks	5.5%	5.0%
NURD	6.6%	6.9%
(B) Model-based simulator (FASTA)	4082	4081
Sequgio	6.1%	5.7%
Standard	13.5%	14.1%
Cufflinks	27.7%	31.2%
NURD	8.4%	7.8%
(C) RNASeqReadSimulator (FASTA)	4877	4876
Sequgio	5.9%	5.2%
Standard	7.1%	10.5%
Cufflinks	6.7%	6.2%
NURD	6.3%	5.9%
(D) RNASeqReadSimulator (FASTA) (multigene transcriptional units)	386	159
Sequgio	8.3%	9.2%
Standard	8.8%	10.9%
Cufflinks	12.1%	14.1%
NURD	15.9%	19.0%

Table 4.1 Results of comparing Sequgio, Cufflinks, NURD, the transcriptional-unit and gene-based standard method from four simulation settings. (Table from Suo *et al*., 2014).

Implementation

The algorithm is implemented in the R package Sequgio that can be freely downloaded from http://www.meb.ki.se/~yudpaw. The package prepares an annotation file and design matrix for all transcripts given a specific version of annotation, for example hg19 and GRCH37. The main inputs are the annotation, design matrix and reads in BAM format mapped by an alignment program, such as Bowtie (Langmead *et al*., 2009), Tophat (Trapnell *et al*., 2009) or BWA (Li and Durbin, 2009).

4.3 STUDY III: INTEGRATED MOLECULAR PORTRAIT OF NON-SMALL CELL LUNG CANCERS

4.3.1 Motivation

Lung cancer accounted for 13% of all cancer cases and 20% of all deaths from cancer in 2012 (Ferlay *et al.*, 2012), and thus represents a significant social and economic burden. The histological subtype of lung cancer affects prognosis and is used to determine treatment planning and patient management. The two major subtypes of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which together represent ~80% of all primary lung cancers. Currently, classification relies on surgical specimens, but in reality, small biopsies or cytology specimens can be obtained in only 70% of cases.

4.3.2 Methods

The samples used in this study were obtained from the CHEMORES initiative (Chemotherapy Resistance Consortium), which includes 19 academic centres, organizations for cancer research, and research-oriented biotechnology companies. Paired snap-frozen tumour and adjacent normal lung tissue samples were obtained from a total of 123 patients who diagnosed with NSCLC and underwent surgery. Copy number alteration profiling was obtained using an Agilent G2505C DNA Microarray scanner. Gene expression profiling was performed using a dual-colour 244K Human exon array from Agilent. MicroRNA expression was obtained using an Agilent G2565C DNA microarray scanner. Candidate gene sequencing reactions were performed using a 48-capillary 3730 DNA Analyzer®. Sequence analysis and alignment was performed using SeqScape® software (Applied Biosystems).

4.3.3 Results

Using aCGH data, 34 genomic clusters were identified, of which several contained genes exhibiting a different profile of alterations between adenocarcinoma (AC) and squamous cell carcinoma (SCC), including the genes *PIK3CA*, *SOX2*, *THPO*, *TP63* and *PDGFB*. Principal component analysis of the mRNA expression data revealed that AC could be separated from

SCC based on transcriptomic variability. Indeed, a 15-gene classifier achieved a cross-validated area under the curve (AUC) of 96% for separating the two histological subtypes. Furthermore, gene expression profiling analysis identified *SPP1*, *CTHRC1* and *GREM1* as potential biomarkers for the early diagnosis of lung cancer, and *SPINK1* and *BMP7* for distinguishing between AC and SCC using small biopsies or blood samples.

Using an integrated genomics approach, three potential driver genes: *MRPS22, NDRG1* and *RNF7* were identified in frequently altered regions, and correlated with as many as ~800 genes across the genome, and also had a high predictive value for discriminating between the histological subtypes. Using the same procedure, the potential driver microRNA hsa-miR-944 was found to frequently undergo copy number gains, and, on average, was also correspondingly overexpressed in tumor samples which showed copy number gains for this miRNA. The potential driver miRNA had a significant AUC of 88% and median AUC of 78% for predicting AC and SCC in the validation dataset.

4.4 STUDY IV: INTEGRATION OF SOMATIC MUTATION, EXPRESSION AND FUNCTIONAL DATA REVEALS POTENTIAL DRIVER GENES PREDICTIVE OF BREAST CANCER SURVIVAL

4.4.1 Motivation

Analysis of whole genome and transcriptome sequencing experiments provides a useful tool for comprehensively exploring human cancer, and may help to identify the genetic alterations that drive cancer development. However, no widely accepted standard protocols exist to integrate and fully utilize the complex information across the different types of omics data. Several algorithms for integrative analysis have been developed to distinguish driving genetic alterations from the vast number of ‘passengers’ that have neutral or less deleterious effects. However, many challenges remain in this new field, such as identification of the patient-specific mutation events that may partially account for tumour heterogeneity.

4.4.2 Methods

An analytic pipeline was created using existing bioinformatics tools, including GATK (McKenna *et al*., 2010), SnpEff (Cingolani *et al*., 2012), Sequgio (Suo *et al*., 2014) and NEA (Alexeyenko *et al*., 2012), and a novel method was proposed to integrate genomic and transcriptomic profiles based on network enrichment analyses. This pipeline provides statistical evidence for the functional implications of the mutated potential driver genes identified within and between patients, termed common driver genes and patient-specific driver genes, respectively. A so-called driver gene score (DGscore) was developed to reflect the cumulative effect of such genes. To contribute to the score, a driver gene has to be frequently mutated, have a high or moderate mutational impact, and exhibit extreme expression and functional changes linked to a large number of DE neighbours in the functional network.

The samples used in this study are part of The Cancer Genome Atlas breast cancer project, which provided sixty matched tumor and normal samples from the same patients. Exome sequencing with a read length 100 bp was performed on Illumina at the Genome Institute at Washington University and the sequences were aligned to the human genome GRCh 37 using BWA (Li and Durbin, 2009). The RNA samples were assayed via 50 bp HiSeq Illumina 2000 paired-end sequencing at the University of North Carolina.

4.4.3 Results

From analysis of the 60 patients whose samples were available from the TCGA, 17 common driver genes were identified, which together with the identified patient-specific driver genes were summarized into a DGscore for each patient. A high DGscore, defined as larger than the median, was associated with poor survival (p = 0.001). The good performance of the DGscore for predicting patient survival is the result of the integration of mutation, isoform-level expression and functional data, a properly designed weighting scheme for putative driver genes, and a mechanism for identifying driver genes in a generalized and patient-specific manner. Failing to incorporate any of these components decreases the *P*-values: using the crude number of non-functionally characterized mutations would not be able to predict the patients' survival (*p* = 0.25), demonstrating the importance of the frequency of mutation pattern, expression level and the functional characterization by NEA; DGscore calculated at gene-level instead of isoform-level is not a significant prognostic factor (*p* = 0.12); an un-weighted DGscore predicts patient survival, but yielding a slightly less significant *P*-value of 0.005; an incomplete DGscore based on either mutation or extreme expression pattern only is not associated with patient survival (*p* = 0.72 and *p* = 0.38, respectively); DGscore that summarizes either common driver genes or patient-specific driver genes alone cannot predict patient survival well (*p* = 0.08 and *p* = 0.04, respectively). Therefore, purposefully ignoring parts of the informative data demonstrated that the performance of the DGscore method is dependent on each of the components assessed.

DGscore is compared to two existing signatures, the MammaPrint 70-gene signature (van 't Veer *et al*., 2002) and PAM 50-gene signature (Parker et al., 2009). It remains the most significant predictor, whereas MammaPrint and PAM50 have a *P*-value of 0.40 and 0.15, respectively, in predicting patient survival.

5 CHAPTER 5 – DISCUSSION

5.1 CHARACTERISTICS OF A GOOD NORMALIZATION METHOD

In order to compare data from different genomic expression arrays, it is inevitable that some biological information will be lost during the process of normalization, especially if the normalization method is not chosen carefully. To effectively separate the intrinsic biological variation in reported expression levels, a good statistical method should retain interesting variation information while at the same time account for systematic errors. A major assumption of most normalization procedures employed in mRNA pre-processing is that most genes are not differentially expressed, and that there is an approximately balanced proportion of over- and under-expressed genes. While this is generally acceptable for mRNAs, this assumption is unrealistic for miRNAs both biologically, as we expect most miRNAs to be differentially expressed, and technically, as the small number of features available on miRNA array chips makes the standard normalization algorithms highly unstable (Davison *et al*. 2006).

In general, the modified LVS normalization method proposed in paper I of the thesis will have widespread utility for other platforms with replicated-probe design, such as Platform miRCURY that has just fewer numbers of probes and replicates for each miRNA compared to Agilent, and miRXplore that has two channels instead of one-colour design, where each miRNA is targeted by four repetitions; in such case, the colour effect can be included to correct for dye bias.

The advantage of the LVS algorithm over other invariant-set based procedures is that it can extract more information as it operates on the raw signal prior to any processing, such as summarization, and does not rely on additional external data (Wang *et al* 2010). The basic concept of LVS algorithm is to simply compute a measure of between-sample variability accounting for heterogeneity in between-probe variances within a miRNA, thus exploiting all of the information content in probe-level data. The result is a more sophisticated version of a variance filtering procedure, where low-variance features are used as a reference set for normalization. This rather straightforward method does not depend on any specific assumptions, such as the existence of mixing distributions, so it is applicable in most situations.

An optimal pre-processing procedure maximizes the ability of any statistical test to identify a true signature and minimizes the burden of false discoveries. In the operating characteristic curves, the proposed LVS algorithm improved on existing normalization procedures in terms of sensitivity and specificity, especially for datasets with a relatively high number of differentially expressed features. Moreover, the LVS algorithm is flexible enough to successfully adapt to various scenarios.

5.2 CHALLENGES IN DEVELOPING EXPRESSION QUANTIFICATION METHOD FOR RNA-SEQUENCING

In paper II of the thesis, Sequgio is proposed for expression quantification in RNA-sequencing experiments. In order to demonstrate its superior performance, a new method should be compared against a gold standard method. But gold standards for transcript-level expression are difficult to obtain experimentally. The improvements in model fitting offered by Sequgio are mainly demonstrated by empirical means via the goodness-of-fit χ^2 statistics. Simulations and limited isoform-level RT-PCR data were also used to assess the accuracy of the results. In the simulations, both non-uniform distributions and slight deviations from uniformity were considered, and all parameter values were estimated from the real data, so the testing procedure was fair. One limitation of many current methods, including Sequgio is that it is assumed that all isoforms of a gene are known. However, the current annotation is incomplete due to the huge amount of information in different isoform-level annotation databases and the complex structure of the transcriptome; we suspect that these issues may lead to discrepancies during RT-PCR validation. Most biological annotation databases may be updated almost every week, whereas other databases will be closed and merged with others, e.g. ASTD was integrated into the Ensembl database. Therefore, there is a need to develop a reliable and comprehensive mega annotation database.

The main assumption of the joint Sequgio model is that non-uniform read distributions can be identified using information across samples, assuming that the observed read distribution is consistent across samples. In developing Sequgio, evidence has been presented to demonstrate that the sample-to-sample similarity generally holds across the genome, even between different tissues. In practice, it is recommended that users to check consistency of read distribution across samples, especially when pooling information from two biological groups, e.g. diseased versus healthy tissues. If the read distribution is inconsistent, then estimation should be performed separately.

When applied to human brain tissue data, Sequgio performed fairly well based on the correlation with Cufflinks and RT-PCR estimates, although only three samples were available. When there are fewer than ten samples, it is recommended using all of the samples in the estimation. On the other hand, if a large number of samples are available and the computational system is limited, it would be useful to consider a two-staged procedure: (i) in the first stage, the read intensities are estimated from a random subsample, and (ii) in the second stage these intensities are fixed, so only the expression levels need to be estimated.

5.3 VALIDATION OF INTEGRATIVE ANALYSIS

In practice, it is never easy to find a perfect method to validate findings especially for integrative analysis. It is difficult to obtain a validation dataset that has the exact same types of molecular data profiled as the discovery dataset, and raw data files are usually not accessible to public. Therefore, validated results should be interpreted carefully.

In paper III, to validate the three candidate driver genes predictive of adenocarcinoma and squamous cell carcinoma of the lungs, an independent dataset published by Chitale *et al*. (2009) are used to get a bioinformatics validation. Agilent 44K CGH arrays are used to assess copy-number alteration profiles; these arrays are much less dense than the 244K arrays used in Chemores which is the discovery dataset. As the sensitivity of CNV detection algorithms is limited by the resolution of the array, it was decided to directly validate the frequency of copy number gains for the candidate driver genes, as well as their properties including the number of correlated genes and the relationship between the copy number status and gene expression. Significant copy-number gains were observed for the identified driver genes. Using a threshold *P*-value of < 0.001, the frequency of copy number gains profiled by the Agilent 44K CGH arrays in Chitale *et al.* was 11.6%, 28.1% and 7.5% for *MRPS22*, *NDRG1* and *RNF7*; these values are similar or exceed the values reported for patients in Chemores. A one-sided Welch *t*-test was performed to compare the gene expression levels in patients with copy number gains vs. the patients without mutations in these genes; The *P*-values were 0.07, 7.5×10^{-6} and 0.2 for *MRPS22*, *NDRG1* and *RNF7*, respectively. If a *P*-value threshold of 0.05 was used to define copy number gains, all three candidate driver genes exhibited significantly upregulated gene expression in the samples with amplifications, with corresponding *P*-values of 0.002, 6.7×10^{-7} and 0.0009, respectively, suggesting that expression of the three potential driver genes exhibited the expected positive correlations between copy number gains and up-regulated gene expression.

In paper IV, the proposed integrative algorithm was applied to identify potential driver genes in a validation dataset of 671 samples from the TCGA, and the association between the derived DGscore and the overall survival of the patients was tested. The resulting insignificant *P*-value was consistent with the observations in the discovery dataset, in which we intentionally tested at the gene-level. Therefore, the insignificant association between the

DGscore and overall survival is probably due to a lack of isoform-level resolution. In addition, a completely independent microarray dataset based on 17 common driver genes was tested; the negative result (*p*=0.75) was also in line with expectations as the mutation status and isoform-level information was not available.

Some differences were observed in the survival curves when MammaPrint and PAM50 were applied to the TCGA expression profiles, although none of the *P*-values for these methods were as significant as those of the DGscore method. The reasonable performance of these two previously established signatures also implies that the TCGA data was pre-processed in an appropriate manner using the proposed integrative pipeline. It is noted that 67% of the 70 gene identifiers in MammaPrint mapped to RefSeq gene names in the TCGA dataset, whereas the unmapped gene identifiers are genes not annotated in ResSeq. It is possible that the lower number of genes mapped to RefSeq relative to the original study of MammaPrint (van 't Veer *et al*., 2002) may have resulted in the suboptimal performance of MammaPrint.

5.4 NEW CHALLENGES OF NEW TECHNOLOGIES

Third-generation sequencing (TGS), also referred to as single-molecule sequencing, is a new generation of technology with the aims of producing longer read lengths of potentially more than 1,000 bp, and of reduced cost and time relative to first- and second-/next-generation techniques (Schadt *et al*., 2010). The major innovation of TGS is that it does not rely on PCR to amplify a specific DNA template; instead, it examines single molecules of DNA, therefore problems due to PCR amplification can be avoided such that duplicate sequencing reads are largely reduced (Whiteford *et al*., 2009). However TGS can still be much improved, and whether data generated using this sophisticated methodology are superior to those from previous technologies remains to be determined. For example, the raw read error rate is generally in excess of 5% for the first commercially available sequencing instrument, the Helicos Genetic Analysis Platform (Harris *et al*., 2008).

Another cutting-edge technique is single-cell sequencing (SCS; Macaulay *et al.*, 2014). As the name implies, SCS amplifies DNA from single cells, thereby enabling identification of the heterogeneous pattern of genomic or transcriptomic profiling between cells. Utilizing SCS, it has been demonstrated that the tumour cells in bladder cancer are derived from a single ancestral cell, but subsequent evolution leads to two distinct tumour cell subpopulations (Li *et al*., 2012). This finding is important to study specific genetic mutations that are critical in different aspects of tumour development. The technology is also particularly useful for profiling individual circulating tumour cells which are scarce even in cancer patients.

Technology has played a major role in revolutionizing research, and will continue to motivate biostatisticians to develop and apply statistical methodologies that have resulted from the application of technological advances for additional improvement in data acquisition, processing, detection, analysis and integration of biomarkers of interests.

6 CHAPTER 6 – CONCLUSIONS

- Copy number alterations, somatic mutations, and differential microRNA, gene and isoform expression are important sources of variation in the human genome and transcriptome, and are associated with many complex phenotypes.
- Novel statistical methods and data processing pipelines are required for the detection and analyses of these variants.
- In study I, a normalization method that relies on fewer assumptions and applies joint modeling was developed, enabling optimal performance in the downstream analysis of miRNA expression levels.
- In study II, a novel method was introduced that allow users to use RNA-Seq data from multiple samples to estimate isoform expression levels as well as non-uniform read distributions.
- Histopathological classification of NSCLC using small tissue samples is difficult. Identification of differentially-expressed sets of secreted and non-secreted genes may help in the diagnosis and classification of NSCLC using serum or tissue samples. These issues were addressed in study III.
- The novel driver-gene search algorithm for integrating genomic data, mRNA and miRNA expression was used to identify potential driver genes, which may be useful for follow-up experimental validation.
- A practical pipeline was developed to perform somatic variant calling, quantification of gene and isoform expression, and integrate genomic and transcriptomic profiles based on known biological networks and the functional impact on protein coding.
- Putative driver genes that were frequently mutated across multiple patients and patient-specific putative driver genes were identified on a basis of a network-based enrichment method.

- It was demonstrated that patients with breast cancer who carry more mutated potential driver genes with functional implications and extreme expression patterns had poorer survival than patients with lower numbers of mutated potential driver genes.

7 ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to many people who have contributed to this work in different ways.

First, I wish to thank my supervisors Yudi Pawitan, Stefano Calza and Agus Salim, who have always been generous with their guidance and encouragement, without which this thesis would not have been possible. Thank you for your valuable input to the design and conception of the studies, statistics and programming advice, and comments on academic writing. Thank you also for weekly meetings and Skype conference calls, for fruitful discussions and troubleshooting.

To my mentor Marie Reilly, thank you for your care and guidance, for all the kind invitations to dinner at your lovely home, and for drafting ‘angry’ emails to the student housing administrator because of a ‘war’ with bed bugs! My PhD journey was wonderful because of you.

Thank you to my Master’s thesis advisor Kee Seng Chia, who introduced me to Yudi, for encouraging me to pursue a PhD at Karolinska Institutet.

To my advisor Jurg Ott who hosted me as a visiting student at the Institute of Psychology, Chinese Academy of Sciences.

And to my undergraduate supervisor Yap Von Bing, who first introduced me to the field of Genome Wide Association Studies.

To my wonderful co-authors and colleagues, Donghwan Lee, Woojoo Lee, Setia Pramana, Fredrik Jonsson, Dhany Saputra, Jinseub Hwang and Himanshu Joshi, thank you for your help and for many fruitful discussions.

To Qian Yang and Shumei Teo, thank you for accompanying me on many academic and holiday trips, and for your patience when I whined about my programming code.

To my wonderful officemates, Benedict Delcoigne, Andreas Karlsson and Myeongjee Lee, I enjoyed our lunches together and every minute we shared in the office. And thank you for the difficult task of keeping my plants alive when I was away.

To the Biostat group at MEB, as a person who is not usually very talkative when speaking English, I treasure our conversations during classes, along the corridor, beside the fruit basket or at the Monday/Thursday Biostat fika. Special thanks to Therese Andersson, Sandra Eloranta, Caroline Weibull, Robert Karlsson, Alexander Ploner, Anna Johansson, Mark Clements, Keith Humphreys, Cecilia Lundholm, Johan Zetterqvist, Peter Ström, Xingrong Liu, Flaminia Chieasa, Annika Tillander, Arvid Siölander, Paul Dickman, Alessandra Grotta, Robert Szulkin, Sven Sandin, Henrik Olsson, Alexandra Jauhiainen, Abbas Cheddad, Hannah Bower, Li Yin and Rino Belloco.

To all the MEB administrators, especially Camilla Ahlqvist and Marie Jansson, my life at MEB would have been much more difficult without you.

To all present and former doctoral students at MEB, a special thanks for your friendly companionship, particularly Hatef Darabi, Andrea Ganna, Daniela Mariosa, Iffat Rahman, Ida Mattsson, Linda Abrahamsson, Sara Ekberg, Edoardo Colzani, He Gao and Jingmei Li.

To my wonderful Chinese friends at MEB, particularly Huan Song, Jiaqi Huang, Tong Gong, Donghao Lv, Ci Song, Zhiwei Liu, Zheng Chang, Jie Song, Shuyang Yao, Fei Yang, Xu Chen, Jianwei Zhu, Jiangrong Wang, Qi Chen, Haiyun Wang, Xingdong Chen, Ruoqing Chen, Yiqiang Zhan, Ruijingfang Jiang, Bojing Liu and other friends elsewhere in Stockholm, Ting Zhuang, Rui Wang, Qiang Zhang, Yang Xuan, Tiantian Liu, Jianren Song, Na Guan, Yixin Wang, Dong Yang, Yuning Zhang, Tianwei Gu, Kai Du, Qin Xiao, Xiaoyuan Ren, Mi Wang, Meiqiongzi Zhang and Chengjun Sun. Special thanks to Weimin Ye and Fang Fang.

To my friends Yang Wang, Hui Gong, Zhuo Chi, Yin Lu, Li Wang, Xiaohua Zhang, Boya Xie, Yajing Zhan, Shu Xu and Jingxia Yang.

I would like to thank my colleagues from Singapore, Yik Ying Teo, Xueling Sim, Kaavya Narasimhalu, Chung Seng Tan, Chee Seng Ku, Haiyan Xu, Hui Miao, E Shyong Tai, Sze Ling Cha, Katherine Kasiman, Peng Chen, Xu Wang, Jundong Jiang, Sharon Wee and Rick Ong and many others for their friendship.

Last but not the least, I am grateful to my family, for supporting me and for their endless love. To my husband Li Kan, who makes me believe in love, for all the treasurable moments we have no matter whether we are together or apart.

8 REFERENCES

1. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA and Pe'er D: An integrated approach to uncover drivers of cancer. Cell 143:1005–1017, 2010.

2. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J and Pawitan Y: Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformaics 13:226, 2012.

3. Bremner R, Du DC, Connolly-Wilson MJ, Bridge P, Ahmad KF, Mostachfi H, Rushlow D, Dunn JM and Gallie BL: Deletion of RB exons 24 and 25 causes low-penetrance retinoblastoma. American Journal of Human Genetics 61:556–570, 1997.

4. Bolstad BM, Irizarry RA, Astrand M and Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2), 185-193, 2003.

5. Calin GA and Croce CM: MicroRNA signarures in human cancer. Nature Reviews Cancer 6(11):857-66, 2006.

6. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, Golas B, Wang L, Motoi N, Szoke J, Reinersman JM, Major J, Sander C, Seshan VE, ZakowskiMF, Rusch V, Pao W, Gerald W and LadanyiM: An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene 28(31):2773–2783, 2009.

7. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES and Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 31, 213-219, 2013.

8. Cingolani P, Platts A, Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM and Lu X: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Landes Bioscience 6(2):80-92, 2012.

9. Cleveland WS: Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74. 836-859, 1979.

10. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P and 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. Nature Genetics 12;43(7):712-4, 2011.

11. Davison TS, Johnson CD and Andruss BF: Analyzing micro-RNA expression using microarrays. Methods Enzymol 411: 14–34, 2006.

12. Fabbri M, Croce CM and Calin GA: MicroRNAs. Cancer Journal 14: 1–6, 2008.

13. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW and Lee C: Copy number variation: new insights in genome diversity. Genome Research 16(8):949-61, 2006.

14. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 21/08/2014.

15. Gilissen C, Hoischen A, Brunner HG and Veltman JA: Unlocking Mendelin disease using exome sequencing. Genome Biology 12:228, 2011.

16. Guarnieri D and DiLeone R: MicroRNAs: A new class of gene regulators. Annual Medicine 40: 197–208, 2008.

17. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H and Xie Z: Single-molecule DNA sequencing of a viral genome. Science 4;320(5872):106-9, 2008.

18. Hartemink A, Gifford D, Jaakkola T and Young R: Maximum likelihood estimation of optimal scaling factors for expression array normalizations. SPIE Bios, San Jose, California, 2001.

19. Haslett JN, Sanoudou D, Kho AT, Han M, Bennett RR, Kohane IS, Beggs AH and Kunkel LM: Gene expression profiling of Duchenne muscular dystrophy skeletal muscle. Neurogenetics 4(4), 163-171, 2003.

20. Hastings PJ, Lupski JR, Rosenberg SM and Ira G: Mechanisms of change in gene copy number. Nature Review Genetics 10: 551-564, 2009.

21. Hert DG, Fredlake CP and Barron AE: Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29: 4618-4626, 2008.

22. Huber W, von Heydebreck A, Sultmann H, Poustka A and Vingron M: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18: S96–S104, 2002.

23. Huttenhower C, Haley E, Hibbs M, Dumeaux V, Barrett D, Coller H and Troyanskaya O: Exploring the human genome with functional maps. Genome Research, 19:1093–1106, 2009.

24. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature 21;431(7011):931-45, 2004.

25. Jiang H: Computational and statistical approaches in RNA sequencing analysis. Doctoral dissertation, 2009.

26. Jiang H and Wong WH: Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25, 1026–1032, 2009.

27. Kiss T: Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions. Cell 109: 145–148, 2002.

28. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M and Turner DJ: Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (GþC)-biased genomes. Nature Methods 6, 291–295, 2009.

29. Kozomara A and Griffiths-Jones S: miRBase: annotating high confidence microRNAs using deep-sequencing data. Nucleic Acids Research 42(Database issue):D68-73, 2014.

30. Langmead B, Trapnell C, Pop M and Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25, 2009.

31. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO and Davis RW: . Proceedings of the National Academy of Sciences U.S.A. 94(24): 13057–13062, 1997.

32. Lazar V, Suo C, Orear C, van den Oord J, Balogh Z, Guegan J, Job B, Meurice G, Ripoche H, Calza S, Hasmats J, Lundeberg J, Lacroix L, Vielh P, Dufour F, Lehtiö J, Napieralski R, Eggermont A, Schmitt M, Cadranel J, Besse B, Girard P, Blackhall F, Validire P, Soria J, Dessen P, Hansson J and Pawitan Y: Integrated molecular portrait of non-small cell lung cancers. BMC Medical Genomics 6:53, 2013.

33. Li H and Durbin R: Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25, 1754–1760, 2009.

34. Li C and Wong WH: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Sciences 98:31-36, 2001.

35. Li J, Jiang H and Wong WH: Modelling non-uniformity in short-read rates in RNA-Seq data. Genome Biology 11(5):R50, 2010.

36. Li Y, Xu X, Song L, Hou Y, Li Z, Tsang S, Li F, Im KM, Wu K, Wu H, Ye X, Li G, Wang L, Zhang B, Liang J, Xie W, Wu R, Jiang H, Liu X,Yu C, Zheng H, Jian M, Nie L, Wan L, Shi M, Sun X, Tang A, Guo G, Gui Y, Cai Z, Li J, Wang W, Lu Z, Zhang X, Bolund L, Kristiansen K,Wang J, Yang H, Dean M, Wang J: Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. Gigascience 14;1(1):12, 2012.

37. Ma X and Zhang X: NURD: an implementation of a new method to estimate isoform expression from non-uniform RNA-Seq data. BMC Bioinformatics 14, 220, 2013.

38. Macaulay IC and Voet T: Single cell genomics: advances and future perspectives. PLoS Genetics 10(1): e1004126, 2014.

39. Mardis ER: The impact of next-generation sequencing technology on genetics. Trends Genetics 24:133–141, 2008.

40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-303, 2010.

41. Metzker ML: Sequencing technologies – the next generation. *Nature Reviews 2010;* **11**:31-46.

42. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 621–628, 2008.

43. Nagao K, Togawa N, Fujii K, Uchikawa H, Kohno Y, Yamada M and Miyashita T : Detecting tissue-specific alternative splicing and disease associated aberrant splicing of the PTCH gene with exon junction microarrays. Humam Molecular Genetics 14, 3379–3388, 2005.

44. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ,Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM and Bernard PS: Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 10;27(8):1160-7, 2009.

45. Park T, Yi SG, Kang SH, Lee SY, Lee YS and Simon R: Evaluation of normalization methods for microarray data. BMC Bioinformatics 4:33, 2003.

46. Pasquinelli AE, Hunter S and Bracht J: MicroRNAs: A developing story. Current Opinion in Genetics and Development 15: 200–205, 2005.

47. Pawitan Y: In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press, the United States, 2001.

48. Pinkel D and Albertson DG: Array comparative genomic hybridization and its applications in cancer. Nature Genetics 37 (Suppl): S11-S17, 2005.

49. Porter JD, Khanna S, Kaminski HJ, Rao JS, Merriam AP, Richmonds CR, Leahy P, Li J, Guo W and Andrade FH: A chronic inammatory response dominates the skeletal muscle molecular signature in dystrophin-deffcient mdx mice. Human Molecular Genetics 11(3), 263C272, 2002.

50. Roberts A, Trapnell C, Donaghey J, Rinn JL and Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 12, r22, 2011.

51. Sanger F, Nicklen S and Coulson AR. DNA sequencing with chain-terminating inhibitors: Proceedings of the National Academy of Sciences U.S.A. 74: 5463–5467, 1977.

52. Santarius T, Shipley J, Brewer D, Stratton MR and Cooper CS. A census of amplified and overexpressed human cancer genes. Cancer 10(1):59-64, 2010.

53. Sawyer SA, Parsch J, Zhang Z and Hartl DL: Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. Proceedings of the National Academy of Sciences U.S.A. 104 (16): 6504–10, 2007.

54. Schadt EE, Turner S and Kasarskis A: A window into third-generation sequencing. Human Molecular Genetics 19: R227-240, 2010.

55. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J,Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A and Wigler M. Large-scale copy number polymorphism in the human genome. Science 305(5683):525-8, 2004.

56. Shenouda SK and Alahari SK: MicroRNA function in cancer: oncogene or a tumor suppressor? Cancer Metastasis Review 28(3-4):369-78, 2009.

57. Shojaie A and Michailidis G: Network enrichment analysis in complex experiments. Statistical Applications in Genetic and Molecular Biology 9(1):Article22, 2010.

58. Shridhar V, Sen A, Chien J, Staub J, Avula R, Kovats S, Lee J, Lillie J and Smith DI. Identification of underexpressed genes in early- and late-stage primary ovarian tumors by suppression subtraction hybridization. Cancer Reseach 1;62(1):262-70, 2002.

59. Stankiewicz P and Lupski JR. Structure variation in the human genome and its role in disease. Annual Review of Medicine 61:437-55, 2010

60. Stratton MR, Campbell PJ and Futreal PA: The cancer genome. Nature Vol 458|9, 2009.

61. Suo C, Calza S, Salim A and Pawitan Y: Joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-seq data. Bioinformatics 15;30(4):506-13, 2014.

62. Suo C, Salim A, Chia KS, Pawitan Y and Calza S: Modified least-variant set normalization for miRNA microarray. RNA 16(12):2293-303, 2010.

63. Suo C, Lee D, Pramana S, Saputra D, Joshi H, Calza S and Pawitan Y: Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. Manuscript.

64. Teo SM, Pawitan Y, Kumar V, Thalamuthu A, Seielstad M, Chia KS and Salim A: Multi-platform segmentation for joint detection of copy number variants. Bioinformatics 27(11):1555–1561, 2011.

65. Teo SM, Salim A, Calza S, Ku CS, Chia KS and Pawitan Y: Identification of recurrent regions of copy-number variants across multiple individuals. BMC Bioinformaics 2010, 11:147.

66. Timmons JA, Jansson E, Fischer H, Gustafsson T, Greenhaff PL, Ridden J, Rachman J and Sundberg CJ: Modulation of extra-cellular matrix genes reects the magnitude of physiological adaptation to aerobic exercise training in humans. BMC Biology, 3, 19, 2005.

67. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L: Transcript assembly and quantification by RNA-Seq reveal unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28, 511–515, 2010.

68. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536, 2002.

69. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC and Croce CM: A microRNA expression signature of human solid tumors defines cancer gene targets. PNAS Vol. 103, no.7, 2257-2261, 2006.

70. Wang B, Wang XF, Howell P, Qian X, Huang K, Riker AI, Ju J, Xi Y: A personalized microRNA microarray normalization method using a logistic regression model. Bioinformatics 26: 228–234, 2010.

71. Wang X, Wu Z and Zhang X: Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. Journal of Bioinformatics and Computational Biology 8 Suppl 1:177-92, 2010.

72. Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, Zaranek AW, Abnizova I and Brown C: Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics volume 25, issue 17, pp.2194-2199, 2009.

73. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30(4), e15, 2002.