From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Improved statistical methodology for high-throughput omics data analysis

Wenjiang Deng

Karolinska Institutet

Stockholm 2021

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB 2021.

ISBN 978-91-8016-279-1

Improved statistical methodology for high-throughput omics data analysis

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Wenjiang Deng

Time and location: September 17th, 2021, kl 09.00 in the lecture hall Atrium, Nobels väg 12 B, Karolinska Institutet, Solna

Principle supervisor:
Professor Yudi Pawitan
Karolinska Institutet
Dept. of Medical Epidemiology and Biostatistics

Co-supervisors:
Assistant Professor Trung Nghia Vu
Karolinska Institutet
Dept. of Medical Epidemiology and Biostatistics

Assistant Professor Xia Shen
Karolinska Institutet
Dept. of Medical Epidemiology and Biostatistics

Opponent:
Associate Professor Raffaele Adolfo Calogero
University of Turin
Dept. of Molecular Biotechnology and Health Sciences

Examination board:
Professor Rebecka Jörnsten
University of Gothenburg and Chalmers University of Technology
Division of Applied Mathematics and Statistics

Docent Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Docent Fredrik Wiklund
Karolinska Institutet
Dept. of Medical Epidemiology and Biostatistics

To all who I love

To be proud of whom I've tried to be
And this life I choose to live

Abstract

Over the last two decades, the advent of high-throughput omics technology has substantially revolutionized biological and biomedical research. A large volume of omics data has been produced with the rapid development of sequencing techniques. Meanwhile, researchers have developed a wide range of computational tools to manage and analyze the omics data. Although the implementation of these tools generates significant discoveries, processing and interpreting the omics data efficiently and accurately is still a big challenge.

In this thesis, we aim to develop novel statistical methodologies and algorithms for omics data analysis. We implement the methods for both simulated and real data from different types of cancers. Based on the evaluation and comparison with existing tools, we find that our methods achieve higher accuracy and better performance in analyzing different types of omics data.

In **Study I**, we build an analysis pipeline to integrate multiple levels of omics data and identify potential driver genes in neuroblastoma. The pipeline employs gene expression profile, microarray-based comparative genomic hybridization data, and functional gene interaction network to detect cancer-related driver genes. We identify a total of 66 patient-specific and four common driver genes. The genes are summarized into a driver-gene score (DGscore) for each patient. We find that the patients with a low DGscore have better survival than those with a high DGscore (p-value=0.006).

In **Study II**, we develop a novel method named XAEM to quantify isoform-level expression using RNA sequencing data. There are two major components in this method. First, we construct a design matrix X as the starting parameter in the quantification model. Second, we utilize an alternating Expectation Maximization algorithm to estimate the design matrix X and isoform expression β iteratively. We compare XAEM with several quantification methods using both simulated and real data. The result shows that XAEM achieves higher accuracy in multiple-isoform genes and obtains substantially better rediscovery rates in the differential-expression analysis.

In **Study III**, we extend the algorithm from Study II and develop an approach named MAX to quantify mutant-allele expression at the isoform level. For a given gene and a list of mutations, we first generate the mutant reference by incorporating all possible mutant isoforms from the wild-type isoform. The

alternating Expectation Maximization algorithm is then applied to estimate the isoform abundance. We implement MAX to a real dataset of acute myeloid leukemia. Using the mutant-allele expression, we discover a subgroup of NPM1-mutated patients that has better drug response to a kinase inhibitor.

In **Study IV**, we build a pipeline to detect fusion genes at DNA level using whole-exome sequencing data. The pipeline is utilized to three comprehensive datasets of acute myeloid leukemia and prostate cancer patients. Compared with the detection results from RNA sequencing data, we find that several major fusion events in these two cancer types are validated in some of the patients. However, the overall results indicate that it is challenging to identify chimeric genes using exome sequencing data due to its inherent limitations.

Altogether, we have developed several statistical and bioinformatics tools to analyze different types of omics data, which demonstrate higher accuracy and better performance than other competing approaches. The results in this thesis will provide novel insights into omics data analysis and facilitate significant discoveries in cancer research.

List of scientific papers

I. Chen Suo*, **Wenjiang Deng***, Trung Nghia Vu, Mingrui Li, Leming Shi, Yudi Pawitan. Accumulation of potential driver genes with genomic alterations predicts survival of high-risk neuroblastoma patients. *Biology Direct,* 2018; 16;13(1):14. doi:10.1186/s13062-018-0218-5.
(*Contributed equally)

II. **Wenjiang Deng**, Tian Mou, Krishna R Kalari, Nifang Niu, Liewei Wang, Yudi Pawitan and Trung Nghia Vu. Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. *Bioinformatics*, 2020; 1;36(3):805-812. doi:10.1093/bioinformatics/btz640.

III. **Wenjiang Deng**, Tian Mou, Yudi Pawitan and Trung Nghia Vu. Quantification of mutant-allele expression at isoform level in cancer from RNA-seq data. (*Submitted*)

IV. **Wenjiang Deng**, Sarath Murugan, Johan Lindberg, Venkatesh Chellappa, Xia Shen, Yudi Pawitan and Trung Nghia Vu. Fusion gene detection using whole-exome sequencing data in cancer patients. (*Manuscript*)

Publications not included in the thesis

I. Tian Mou, **Wenjiang Deng**, Fengyun Gu, Yudi Pawitan and Trung Nghia Vu. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing.
Frontiers in Genetics, 2020; doi:10.3389/fgene.2019.01331.

II. Trung Nghia Vu, **Wenjiang Deng**, Quang Thinh Trac, Stefano Calza, Woochang Hwang and Yudi Pawitan. A fast detection of fusion genes from paired-end RNA-seq data.
BMC Genomics, 2018; doi: 10.1186/s12864-018-5156-1.

III. Tian Mou, Yudi Pawitan, Matthias Stahl, Mattias Vesterlund, **Wenjiang Deng**, Rozbeh Jafari, Anna Bohlin, Albin Österroos, loannis Siavelis, Helena Bäckvall, Tom Erkers, Santeri Kiviluoto, Brinton Seashore-Ludlow, Päivi Östling, Lukas M Orre, Olli Kallioniemi, Sören Lehmann, Janne Lehtiö and Vu Trung Nghia.
The transcriptome-wide landscape of molecular subtype-specific mRNA expression profiles in acute myeloid leukemia. *American Journal of Hematology*, 2021; doi: 10.1002/ajh.26141.

Contents

List of abbreviations

AEM	Alternating expectation-maximization
AGS	Altered gene set
AML	Acute myeloid leukemia
APE	Absolute proportional error
AUC	Area under the curve
BAM	Binary Alignment/Map
ChIP	Chromatin immunoprecipitation
CNA	Copy number alteration
CNV	Copy number variation
ddNTPs	di-deoxynucleotide triphosphates
DE	Differentially expressed
DGscore	Driver-gene score
DNA	Deoxyribonucleic acid
dNTPs	deoxynucleotides triphosphates
EFS	Event-free survival
ELN	European Leukemia Net
Eqclass	Equivalence class
FEQ	Fusion equivalence class
FGS	Functional gene set
GB	Giga-bases
GWAS	Genome-wide association study
HGP	Human Genome Project
HMP	Human Microbiome Project
HTS	High-throughput sequencing
IC50	Inhibitory concentration
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
iHMP	integrative Human Microbiome Project
indel	Small insertion/deletion
ITD	Internal tandem duplication
MNA	MYCN amplification
mRNA	Messenger RNA

NEA	Network enrichment analysis
NGS	Next-generation sequencing
ONT	Oxford Nanopore Technology
PacBio	Pacific Biosciences
PE	Paired-end
RDR	Rediscovery rate
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
SAM	Sequence Alignment/Map
SEQC	Sequencing Quality Control Consortium
SMRT	Single-molecule real-time
SNP	Single-nucleotide polymorphism
SOLiD	Sequencing by oligonucleotide ligation and detection
SRA	Sequence Read Archive
SV	Structural variation
TC	Transcription cluster
TCGA	The Cancer Genome Atlas
TE	TMPRSS2-ERG
TNs	Transcript neighbors
tRNA	Transfer RNA
TRP	Transcript response profile
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WT	Wild-type
ZMW	Zero-mode waveguide

1 Background

Cancer is a leading cause of death across the world. In 2019, 18 million new cases were diagnosed and about 9 million deaths occurred due to cancer [1]. The increasing morbidity and mortality highlight the urgent need to characterize the pathophysiologic mechanism of different cancers. Over the last two decades, the advancement of sequencing technology, especially the next-generation sequencing approach, has allowed researchers to interrogate cancers using multiple omics data [2]. It is well accepted that cancer is closely associated with genetic abnormalities, such as structure variation, copy number alteration, single nucleotide variation, and fusion gene [3]. The sequencing technologies provide a fast, comprehensive and cost-effective way to capture these genetic aberrations simultaneously. In recent years, a wide range of studies employing sequencing approaches have been conducted. Specifically, two of the largest studies, The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have led the way to produce a huge amount of sequencing data in diverse cancer types [4, 5]. It can be expected that cancer-related research will continue to benefit from the sequencing approaches and yield more valuable findings for the prevention, diagnosis, treatment, and prognosis of cancer patients.

1.1 Sequencing technologies

Sequencing is the biochemical process to measure the order of nucleotides in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). A DNA molecule has four constituent bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Different orders and combinations of bases are implicated in unique functional impacts on disease occurrence; thus, it is of great importance to determine the sequence of nucleotides in research samples.

1.1.1 First-generation sequencing

Great efforts have been made to characterize the sequences of DNA, RNA, and protein in living organisms [7]. As Figure 1.1 shows, sequencing technologies have been developed and improved continuously since the 1970s. The Sanger

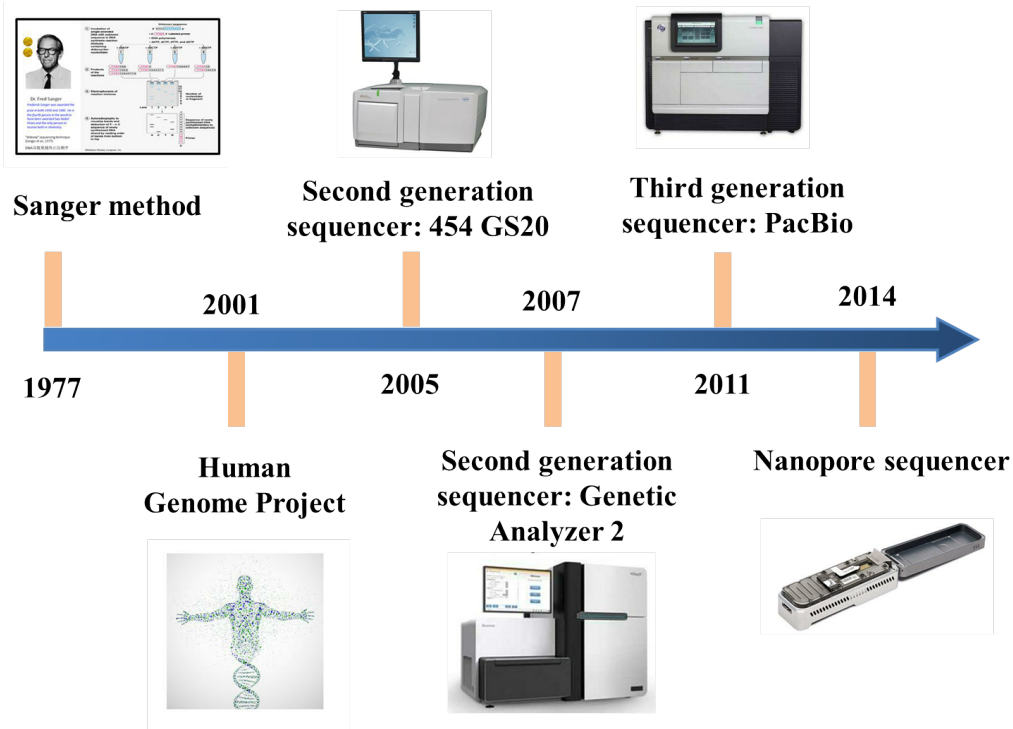

Figure 1.1: The history of sequencing technologies since 1977. *The figure is edited based on the work of Yang et al. [6] and reprinted with permission from Frontiers Media Group.*

method, which marks the wide implementation of first-generation sequencing, was developed by Frederick Sanger in 1977 [8]. The key concept of the Sanger method is the utilization of di-deoxynucleotide triphosphates (ddNTPs). The ddNTPs lack the 3'-OH group, which is required for the bonding between successive nucleotides. Since the absence of the 3'-OH group will terminate the growth of the DNA chain, the Sanger method is also called the chain-termination method. In the Sanger sequencing protocol, DNA samples are first divided into four separate reaction vessels, which contain four normal deoxynucleotides triphosphates (dNTPs) and DNA polymerase. Next, only one of the four ddNTPs (ddATP, ddTTP, ddCTP and ddGTP) is added to each vessel, so that the DNA strand will be terminated in selective positions. After the reaction, DNA fragments with different lengths are produced; the gel electrophoresis is then used to sequence the DNA fragments. Since the shorter and lighter fragments

will migrate further to the bottom of the electrophoresis plate, the DNA sequences are measured according to the pattern of DNA bands.

The Sanger method is further improved in automated DNA sequencing instruments, where the four ddNTPs terminators are labeled with fluorescent dyes. Each of the ddNTPs can be detected by the dye fluorescence using the capillary electrophoresis and laser detector device. The first commercialized sequencer using the Sanger method was manufactured by Applied Biosystems in 1986, which allows the generation of an individual read with the length at 1000 bases. Since then, the automated version of Sanger sequencing became the most widely used method until the middle 2000s. Notably, it was the essential sequencing method for the Human Genome Project (HGP), which published the first draft of human genome in 2001 (Figure 1.1).

1.1.2 Next-generation sequencing

First-generation sequencing provides an unprecedented technique to determine the DNA sequence in a wide variety of organisms. However, its usage is limited due to the low sequencing volume and high cost for large number of DNA targets. To address this problem, second-generation sequencing, or next-generation sequencing (NGS), is developed to sequence millions of fragments simultaneously.

454 pyrosequencing

The 454 pyrosequencing is designed by 454 Life Sciences, which released the first next-generation sequencer, 454 GS20, to the sequencing market in 2005. In the pyrosequencing procedure, DNA sequences are sheared into small fragments and then amplified inside water droplets with an oil solution [9]. The single strand of DNA is kept as a template and the complementary strand will be synthesized using the four types of dNTP and DNA polymerase. When a dNTP is incorporated onto the template, a pyrophosphate will be released and then catalyzed into light signal by luciferase. The emitted light is captured by camera and analyzed as corresponding nucleotide in complementary strand. The pyrosequencing approach can produce about 500 million base pair (bp) per run with read length at 450 bp.

Sequencing by synthesis (Illumina)

Sequencing by synthesis is a major strategy employed in many sequencing machines manufactured by Illumina. Literally, it means a nucleotide is measured when incorporated to the template fragment in the synthesis of the complementary strand. There are three main steps in this method [10, 11]. The first step is sample preparation, where the DNA is cut into smaller fragments with the size of 100–500 bp. The fragment is then ligated with adapters on both ends. Each adapter contains three different parts: (1) sample index, (2) binding site for sequencing primer and (3) the sequence complementary for oligonucleotides (oligo) on flow cell. The bottom of the flow cell is coated with millions of oligos; each DNA fragment is then attached to the flow cell lane.

The second step is DNA cluster generation, also known as bridge amplification. In this step, a DNA strand will bend over and attach to an oligo to form a bridge-like shape. The DNA polymerase binds to the strand and generates a complementary strand. The original strand (forward strand) is then washed away and only the reverse strand is retained. The reverse strand will attach to the oligo again and generate a new forward strand. Both strands are then denatured and the bridge amplification is repeated to produce hundreds of thousands of DNA copies.

In the third step, the sequencing starts by adding the fluorescently labeled dNTPs to the flow cell. When a dNTP incorporates to the template DNA strand, a unique fluorescent light is emitted and captured by camera. The sequencing machine records the light signal and interprets it as a corresponding nucleotide. In this process, only one dNTP can be incorporated to the DNA strand because the fluorophore blocks the binding of next nucleotide. However, this blocking is reversible; when a dNTP is recorded, the fluorophore will be washed away so that the next nucleotide can attach to the DNA strand. In this step, millions of clusters on the flow cell are sequenced simultaneously, which generate huge amount of sequencing reads and outputs.

As shown in Figure 1.1, in 2007, Illumina released the Genetic Analyzer 2 sequencer, which produces one giga-bases (GB) per run with read length at 100 bp. After that, several sequencing machines with improved performance are released, e.g. the HiSeq, MiSeq and HiScanSQ sequencing platform series. The Hiseq 2000 can generate 600 GB reads per run in $\sim$eight days, which becomes

one of the most powerful sequencers in the market.

Sequencing by oligonucleotide ligation and detection

Unlike the pyrosequencing and sequencing by synthesis, the sequencing by oligonucleotide ligation and detection (SOLiD) leverages oligonucleotide probes instead of DNA polymerse [12]. The SOLiD sequencing employs 16 eigth-mer oligonucleotide probes, where the first two bases in the probe use a two-base encoding scheme. Each pair of bases has corresponding fluorescent dye in 5 prime. In the process of sequencing, each probe is ligated to the target strand and the color is recorded by the sequencer. The last three bases in the probe are then cleaved together with the dye to allow next probe to ligate. For each target fragment, seven different probes will be ligated in separate round and five rounds are performed using different primers. Due to the implementation of the two-base encoding strategy, the SOLiD approach achieves a high sequencing accuracy at 99.94%. However, the major disadvantage is the relatively short read length at 50 bp. The first commercialized SOLiD sequencer was produced by Applied Biosystems in 2008, which can generate up to 60 GB reads per run.

Ion Torrent sequencing

When a dNTP is incorporated to the target strand, a hydrogen ion will be released and change the pH of the solution. The pH change is detected and analyzed by an ion-sensitive ion sensor, which is a micro semiconductor chip underneath the reaction wells. In the process of sequencing, the four types of dNTP are added to the reaction wells together with the DNA polymerase. If the introduced dNTP is complementary to the template strand, the hydrogen ion is released and detected by the sensor, and the unattached dNTPs will be washed away. In this way, the sequencer can measure the template DNA when the process repeats. Ion Torrent sequencing is also called Ion semiconductor sequencing and pH mediated sequencing. The first commercial Ion Torrent sequencer was released by Life Technologies in 2010. It provides a rapid sequencing in two hours and generates sequencing reads at 200–400 bp with an accuracy at 99%.

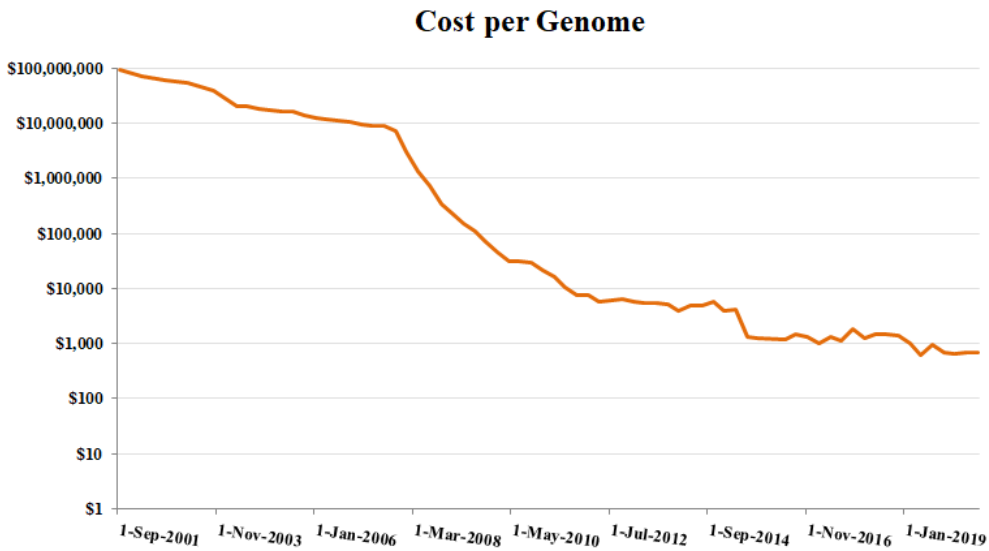

Figure 1.2: The cost of sequencing per genome from 2001 to 2019. Source data are obtained from National Human Genome Research Institute.

Sequencing price overview

The massively parallel capability and high throughput of next-generation sequencing have led to a substantial decrease in sequencing price. Figure 1.2 shows the trend of costs per genome over the last two decades. In 2001, the price to sequence a human genome was about 100 million dollars, while in 2019, the cost decreased strikingly to about 1000. In particular, the first draft of human genome was released in 2001 by the Human Genome Project. The HGP was the first and biggest project at the time to measure the complete base pairs in human genome [13]. It was started in 1990 and officially completed in April 2003, involving a huge amount of collaborations and efforts between researchers worldwide. The cost for the draft genome was about 300 million dollars and the later refinement cost another 150 million. In Figure 1.2, there is a noteworthy decrease near 2008 due to the transition from the usage of first-generation sequencing (Sanger method) to next-generation sequencing. With the continuous improvement of sequencing technologies, we anticipate that the sequencing will be conducted with an even lower price and higher throughput in the near future.

1.1.3 Third-generation sequencing

The next-generation sequencing is high-throughput, efficient, and cost-effective. However, a major drawback of NGS is the relatively short read length, which ranges from 50 bp to 700 bp maximum. The short read length often complicates the downstream bioinformatics analysis, such as isoform quantification, de novo assembly and structural variation detection. To tackle this problem, third-generation sequencing, or long-read sequencing was developed, which can produce reads with tens or hundreds of kilo bases [14, 15].

Single-molecule real-time (SMRT) sequencing

The SMRT method employs a zero-mode waveguide (ZMW), a structure of 70 nm in diameter. A single molecule of DNA fragment and a DNA polymerase are attached in the bottom of the ZMW hole. Four types of DNA bases labeled with fluorescent dyes are added to the reaction. When a nucleotide is incorporated onto the template strand, the fluorescence is observed by the detector and interpreted as a corresponding base. The SMRT is commercialized by Pacific Biosciences (PacBio), which released the first SMRT sequencer in 2011. The average read length using the SMRT method is 10–15 kilo bases, and 500 million bases can be generated per SMRT run [16, 17].

Nanopore sequencing

Nanopore is a nano-scale pore embedded in electrically resistant polymer membrane [18]. When a DNA strand passes through the Nanopore, voltage changes will be triggered and recorded; thus, the DNA bases can be measured from the current signal. The Nanopore method utilizes helicase to unwind the target DNA into two strands; one strand is translocated and passing through the Nanopore for sequencing. Unlike all methods mentioned above, the Nanopore approach does not need PCR amplification or chemical labeling of nucleotide. This feature makes Nanopore sequencing independent of expensive equipment and reagents, which allows it to be used in remote places with limited laboratory resources. Nanopore sequencing was developed and released by Oxford Nanopore Technology (ONT), with the first Nanopore sequencer manufactured in 2014. The average read length of Nanopore sequencer is 20 kilo bases, and the maximum read length can reach to 2.3 million bases [19].

Comparison of sequencing technologies

Table 1.1: A brief comparison between Sanger sequencing, Sequencing by synthesis (Illumina) and Nanopre sequencing. Note that the throughput and times per run in the table are approximate numbers depending on respective platform and equipment.

	Sanger method	Sequencing by synthesis (Illumina)	Nanopore sequencing
Read length	400-900 bp	75-300 bp	up to 2.3 million bp
Accuracy	99.9%	99.9%	95%
Throughput	900 bp	1000 Gb	42 Gb
Times per run	2 hours	2-5 days	one or two days
Advantages	long read length	high throughput	longest read length
Disadvantages	low throughput	short read length	low accuracy

The diverse features of each sequencing method provide researchers with a wide range of choices in their research projects. Table 1.1 shows a brief comparison of performance between three major types of sequencing technologies. The Sanger method has been utilized for half a century and still remains popular. The most important advantage of the Sanger method is the long read length (up to 900 bp) and fast experiment procedure. Although its low throughput limits the large scale usage, it is one of the best choices for sequencing small numbers of gene targets. Next-generation sequencing has became the dominant approach due to the ultra high sequencing volume. For example, the Illumina Hiseq 4000 platform generates more then 1000 Gb reads in five days, which guarantees a high sensitivity and power in many biomedical studies. However, the read length from the Illumina sequencer is only 75–150 bp, which makes the data analysis complicated and challenging [20, 21]. The third-generation sequencing, e.g. Nanopore method, is designed to tackle the short read length problem. It can produce sequencing reads up to 2.3 million bases, facilitating the application in de novo assembly and novel transcript detection. The throughput of the Nanopore method is relatively high, which can be 42 Gb per run. The next and third generation sequencing are also named as high-throughput sequencing (HTS). Although the utilization of third-generation sequencing has become more general in these years, its usage is still limited due to high error rate and high sequencing cost per base [22].

1.2 Omics data and applications

Omics data are defined as the comprehensive dataset of the same type of molecule generated using the high-throughput sequencing methods [23]. For example, genomics measures entire DNA sequence, transcriptomics quantifies all transcripts, and proteomics profiles the complete set of proteins. Each type of omics data provides significant and unique insights into biological mechanisms underlying human disease. In the last ten years, omics studies have been thriving tremendously due to the advancement in sequencing technologies [24, 25].

1.2.1 Genomics

Genomics aims to elaborate the structure, component, function, and modification of the whole genome [26]. In human research, a primary task is to identify genomic mutations associated with different phenotypes and diseases. The most common types of mutations include single-nucleotide polymorphism (SNP), structural variation (SV), copy number variation (CNV), and small insertion/deletion (indel). Different sequencing strategies such as whole-genome sequencing (WGS) and whole-exome sequencing (WES) are widely used to delineate specific mutations [27, 28].

SNP and genome-wide association study (GWAS)

SNP is a genetic variant when a single nucleotide differs between a group of samples. In human genome, there are about 4–5 million SNPs, in that SNP occurs in every 1000 bases. Most SNPs have no functional effects, but some are closely associated with human traits and the increased risk of diseases [29]. In the last two decades, GWAS has identified a wide range of candidate loci related to complex diseases [30]. For example, Fachal et al. have summarized a total of 83 susceptibility loci in breast cancer from GWAS. These loci explain $\sim$14% of breast cancer heredity and provide significant insights into cancer risk stratification [31].

Structural variation (SV)

Structural variation occurs in the structure of chromosome that affects $>$100 bases (referred to as indel if $<$100 bp). Compared with SNP, which only involves

the substitution of single nucleotide, SV alters much longer sequences in DNA. SVs can be categorized into two groups: (1) balanced SVs, i.e. inversion and translocation, which do not change the total number of genomic bases; (2) unbalanced SVs such as deletion, insertion and copy number variation (deletion and duplication), which will add or remove nucleotides from the genome [32].

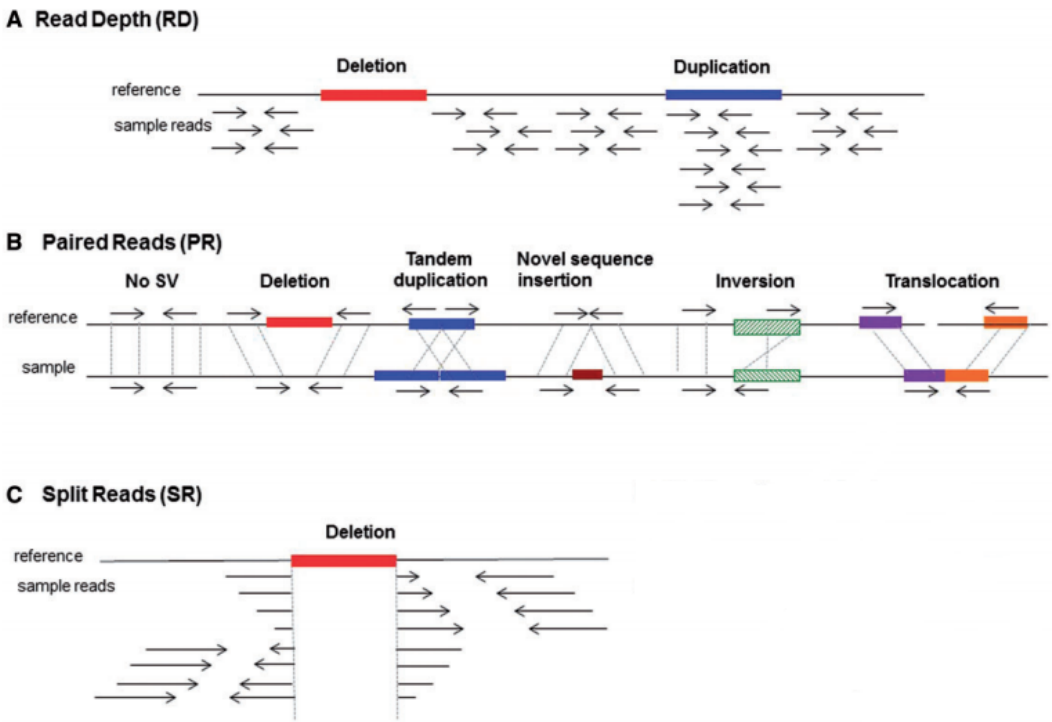

Figure 1.3: Strategies to detect SVs using sequencing reads. The solid black arrows are reads with orientations. *The figure is re-edited from the work of Docampo et al. [33] and reprinted with permission from Oxford University Press.*

Whole-genome sequencing and whole-exome sequencing data are frequently used to identify SVs [34]. Detection of SVs often involves three steps: (1) DNA sequencing using high-throughput sequencing approaches; (2) reads mapping using a known reference genome; (3) variation calling and functional annotation of disease-related mutations. Figure 1.3 shows different strategies to detect SVs from sequencing reads [33]. Copy number variation is a type of SV with the length ranging from 1000 to three million bases. As shown in Figure 1.3(A), a

CNV deletion is identified when the read depth is substantially lower than the average depth of the genome segment, while a CNV duplication is detected if the read depth is higher than reference regions.

Paired-end (PE) sequencing generates read pairs with a fixed length of insert size in between. This feature is often used to detect SVs. Figure 1.3(B) shows that if a pair of reads map to reference genome with correct position, concordant orientation and exact length of insert size, it means there is no SV event occurring. However, any violation of these criteria can indicate the presence of variations. For example, deletion is identified if a read pair has a bigger distance than insert size when mapped to reference genome, and an insertion is detected if the alignment distance is smaller than the insert size. Inversion can be found when two reads are mapping to the reference genome with the same orientation. Translocation is determined if the read pair is mapped to two different chromosomes or having a large distance. Besides, a deletion can be identified when a single read is split at the position of breakpoints (Figure 1.3(C)).

1.2.2 Epigenomics

Epigenomics aims to study the reversible modifications on DNA that affect the gene expression level without changing the DNA nucleotides. Histone modification and DNA methylation are two of the most common epigenomic variants [35]. Studies have shown that these modifications are closely related to biological mechanisms underlying cancer and other human diseases [35, 36]. For example, histone mutations significantly contribute to the formation of paediatric gliomas [37]; methylations of DNA-repair genes are frequently observed in stomach cancer [38]. Several consortia have been launched to characterize human epigenome, such as the Roadmap Epigenomics Mapping Consortium and International Human Epigenome Consortium [39, 40]. To measure the genome-wide epigenomic mutations, sequencing based approaches such as chromatin immunoprecipitation (ChIP) sequencing and bisulfite sequencing are widely used. Recently, due to the great advances in third-generation sequencing technologies, the epigenetic modifications have been detected directly when the DNA bases are sequenced. The application of the Nanopore and SMRT sequencing provides a more accurate and rapid identification of epigenomic aberrations [41].

1.2.3 Transcriptomics

Transcriptomics is the study to characterize entire RNA transcripts. RNA participates in a wide variety of biological processes, including protein synthesis, regulation of gene expression and communicating cellular signals [42]. RNAs can be categorized into coding and non-coding groups. Coding RNA, i.e. messenger RNA (mRNA), is served as template to synthesize proteins in the process of translation. In human transcriptome, mRNA only accounts for 3% of all RNAs and the rest 97% are non-coding RNAs. Transfer RNA (tRNA) and ribosomal RNA (rRNA) represent two of the most common non-coding RNAs, both involved in the synthesis of proteins. Non-coding RNAs are mostly constant regardless of cellular or disease status, while the expression and type of mRNA are dynamically affected by healthy/cancerous conditions in living organism [43]. Hence, it is of great interest to quantify mRNA between different experimental conditions. Previously, the hybridization-based microarrays were widely used to measure the expression of mRNA. However, the microarray method has several limitations; for example, prior knowledge of DNA/RNA sequences are needed to design probes.

RNA sequencing (RNA-seq)

With the tremendous advancement in sequencing technologies, the RNA quantification has embraced substantial improvement in the last ten years. RNA-seq can apply various sequencing approaches, e.g. next-generation sequencing to investigate the presence and abundance of RNA molecules [21]. The overall procedure of RNA-seq is similar with DNA sequencing as introduced above; a major difference is the step for complementary DNA (cDNA) synthesis. In the library preparation step, RNA is isolated from genomic DNA using enzymes such as deoxyribonuclease (DNase). Next, the mRNA is selected or kept by removing non-coding RNAs. The mRNA is then reverse transcribed to cDNA for amplification and sequencing. Several sequencing platforms can be used to perform RNA-seq, such as Illumina and SOLiD. Also, mRNAs can be sequenced directly using Nanopore sequencing without cDNA synthesis nor amplification steps [44].

Quantification of isoform expression using RNA-seq data

Isoforms are different transcripts produced by the same gene with the alternative splicing mechanism. Isoforms have highly similar sequences; however, their functional effects can be distinct or even opposite. For example, full-length p53β isoform from TP53 induces the apoptosis of cancer cell while the Δ133p53 isoform inhibits the cell death process [45, 46]. In this case, it is essential to quantify expression at the isoform level instead of the traditional gene level. In the last decade, a large number of tools have been developed to estimate isoform abundance. These tools can be classified into alignment-based and alignment-free groups. The alignment-based methods include Cufflinks [47], RSEM [48] and eXpress [49]. The first step before running these methods is to align RNA-seq reads to a genome/transcriptome reference. Several aligners can be used for this purpose, such as BWA [50] and Bowtie2 [51]. Most recently, a group of alignment-free methods have been introduced to leverage the idea that precise alignment is not necessary to distribute reads to their original isoforms. The alignment-free methods include Sailfish [52], Salmon [53] and Kallisto [54]. Sailfish and Salmon employ a quasi-mapping concept that maps the k-mers of a read rapidly to a predefined reference index [53]. Kallisto utilizes a de bruijn graph to check the compatibility between reads and transcript segments [54]. All three methods provide an ultra-fast speed in the processing of RNA-seq data and an accurate estimation of isoform expression compared with alignment-based approaches [55].

1.2.4 Proteomics

Proteomics aims to explore the complete set of proteins in terms of structure, function, abundance and interaction. According to the central dogma of molecular biology, protein is the last layer of genetic information flow, thus indicating ultimate consequences from mutations at DNA and RNA level. One of the major applications of proteomics is to develop potential drugs for the treatment of cancers. Many efforts have been made to predict the three-dimensional (3D) structure of disease related proteins using experimental or computational methods. Based on the 3D profile, a new drug could be designed to interfere with protein/enzyme and potentially inactivate the function of protein [56]. Apart from the mutational effects from DNA and RNA level,

proteins can undergo a wide range of post-translational modifications such as phosphorylation and ubiquitination. These chemical modifications are often implicated in enzyme activity and cell structure maintenance, which can be used to monitor cancer formation and progression [57]. Several databases are constructed to store protein sequences and annotated functional information. The UniProt [58] and PROSITE [59] represent two of the largest proteomic repositories, which provide rich resources for protein research.

1.2.5 Microbiomics

Microbiomics is the study to investigate the entire micro-organisms or microbiota such as bacteria, viruses, and fungi. These organisms reside on all parts of the human body, including skin, gastrointestinal tract, uterus, and lung [60]. Microbiomic and epidemiological studies have demonstrated that human microbiome have crucial impacts on inflammatory bowel disease, type II diabetes, obesity, and neurodevelopmental disorders [61, 62]. A recent study shows that the gut microbiome play an important role in body metabolism and contribute to the increasing prevalence of diabetes and obesity [63]. In another study of autism, researchers found that the composition of gut microbiome is significantly different between individuals with autism and those without [64]. The high-throughput sequencing technologies have been commonly used to elucidate the genetic landscape of microbiome. In 2007, the first phase of the Human Microbiome Project (HMP) was launched to characterize the microbial types and components from 300 healthy participants [65]. The second phase, known as integrative Human Microbiome Project (iHMP), was conducted in 2014 to investigate the functional impacts of microbiome on human physiology and disease development [66].

1.3 Cancer research and overview

Cancer is a disease where cells grow uncontrollably with the potential to spread to other parts of the body. A malignant cancer is defined when tumor cells invade other tissues or organisms, while a benign tumor is localized and does not spread. It is well recognized that cancer is closely related to heritable or somatic mutations, which result in abnormal cellular growth, exceptional angiogenesis and suppression of normal cell signaling [67]. In the last two decades, the

sequencing technologies and omics data have been widely used to elucidate the genetic and pathological mechanisms in various cancers. The characterization of genetic mutation provides significant insights into diagnosis, stratification, treatment, and prognosis for cancer patients. In this thesis, we utilize the sequencing methods and omics data from several cancer types for isoform quantification and mutation detection.

1.3.1 Neuroblastoma

Neuroblastoma is the most common extracranial solid tumor in children under the age of five. It ranks third among the most prevalent pediatric cancers after leukemia and brain cancer [68]. The tumor emerges from the sympathetic nervous system, and develops mostly in the adrenal glands, abdomen, chest, or neck. A neuroblastoma is highly heterogeneous with clinical behaviors such as spontaneous regression or aggressive progression despite intensive therapy [24]. The patients can be classified into low, intermediate, and high risk groups. The low- and intermediate-risk patients have a favorable outcome with 90% event-free survival (EFS) rate in three years. However, the high-risk group shows a <50% EFS rate [69]. A wide variety of genetic mutations have been observed in neuroblastoma patients, which are implicated in the tumorigenesis and cancer progression. One of the major objectives of this thesis is to identify potential driver genes in neuroblastoma and provide useful guidance for individual prognosis and treatment.

MYCN amplification

Figure 1.4 shows a comprehensive collection of genetic variations detected to date, which include gene amplification, chromosomal alteration, and polymorphism. MYCN is a protein coding gene and a member of the MYC gene family of transcription factors. The MYCN proteins regulate several cellular processes such as cell proliferation, differentiation, and apoptosis. MYCN amplification (MNA), which contains >10 copies of the gene, is observed in 25% of neuroblastoma patients [24]. MNA is a significant predictor of poor survival in neuroblastoma patients; the amplification and over-expression of MYCN gene can be found in 40% of high-risk cases [69]. MYCN status (amplification versus non-amplification) is frequently used in neuroblastoma risk classification [24].

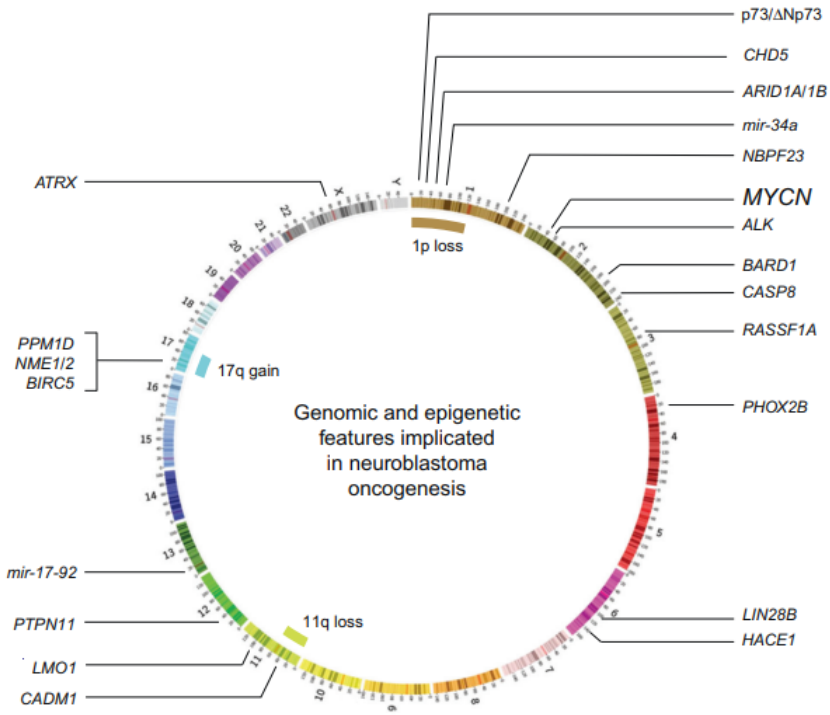

Figure 1.4: A collection of genetic mutations identified in neuroblastoma patients that contribute to the formation of the tumor. The mutations comprise gene amplifications, mutations, deletions and epigenetic modifications such as DNA methylation. *The figure is edited from Morgenstern et al. [70] and reprinted with permission from Elsevier Publishing Group.*

Chromosomal abnormalities

Neuroblastoma has a great cytogenetic heterogeneity in terms of the wide range of genomic aberrations. As shown in Figure 1.4, apart from MYCN amplification, neuroblastoma patients harbor several chromosomal abnormalities, including 17q gain, 1p loss and 11q loss. Gaining the long arm of chromosome 17 is identified in 50% of all cases and ∼90% of high-risk tumors, while the deletion of chromosome 1 short arm is detected in 33% of patients [70]. Both 17q gain and 1p loss can be co-occuring with MYCN amplification and consequently associates with adverse outcomes. Loss of 11q is observed in one-third of high-risk cases, but rarely related to the occurrence of MNA. Neuroblastoma can also be categorized into three subtypes, i.e. Type 1, Type 2A,

and Type 2B, based on the three genomic abnormalities. Type 1 are patients without MYCN amplification or any of the three chromosomal aberrations, usually having a favorable outcome. Type 2A tumor contains 17q gain or 11q loss but without MNA. This category has an intermediate risk and survival compared with Type 1. Type 2B is defined as MNA together with 1p loss or 17q gain, which has the highest risk and worst outcomes [71].

1.3.2 Acute myeloid leukemia

Acute myeloid leukemia (AML) is a hematological malignancy with excessive number of abnormal myeloid stem cells. Several risk factors are related to the formation of AML, such as old age, smoking, chemotherapy treatment, radiation, and genetic abnormalities. In the last 15 years, the advances of high-throughput sequencing technologies have greatly facilitated the detection of genetic mutations in AML [72]. In 2013, the cancer genome atlas (TCGA) research group conducted a comprehensive study of 200 adult AML patients using WGS, WES, and RNA-seq approaches [73]. A total of 23 genes with recurrent mutations were detected, including NPM1 (27% frequency), FLT3 (28%), TP53 (8%), et al. In 2017, the European Leukemia Net (ELN) published the latest recommendation for diagnosis and classification based on AML mutations [74]. As shown in Table 1.2, the patients can be classified into three risk groups, which are favorable, intermediate, and adverse. The favorable risk group is characterized by a patient carrying RUNX1-RUNX1T1 fusion gene, CBFB-MYH11 fusion, or mutated NPM1 without FLT3 internal tandem duplication (ITD) (NPM1+/FLT3-ITD-). The intermediate category can be defined with NPM1+/FLT3-ITD+, NPM1-/FLT3-ITD-, or the chimeric gene of MLLT3-KMT2A. The BCR-ABL1 fusion or NPM1-/FLT3-ITD+ represent the adverse category, which mark the subgroup with inferior outcome and poor survival. Continuous efforts have been made to identify novel mutations to provide further guidance for AML diagnosis and therapy [75].

1.3.3 Breast cancer

Breast cancer is one of the most diagnosed cancers in women worldwide. In America, over 280,000 new cases are detected and 44,000 related deaths are found every year [76]. Many risk factors are involved in the formation of breast

Table 1.2: 2017 ELN recommendation for AML risk stratification based on genetic alterations

Risk group	Genetic alterations
Favorable	t(8;21)(q22;q22.1), RUNX1-RUNX1T1
	inv(16)(p13.1q22), CBFB-MYH11
	NPM1 mutated and FLT3-ITD non-mutated, NPM1+/FLT3-ITD-
	CEBPA biallelic mutated
Intermediate	NPM1+/FLT3-ITD+
	NPM1-/FLT3-ITD-
	t(9;11)(p21.3;q23.3), MLLT3-KMT2A
Adverse	t(9;22)(q34.1;q11.2), BCR-ABL1
	NPM1-/FLT3-ITD+
	Mutated RUNX1, ASXL1, TP53

cancer, including being female, obesity, older age, alcoholism, and genetic mutations. Breast cancer can be classified into three categories according to the presence and absence of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 (HER2): the ER/PR positive and HER2 negative (70%), the HER2 positive (15%), and all three molecular markers negative (triple-negative, 15%). The three subgroups indicate different clinical features and respond to distinct treatment strategies. Genetic variations have triggered $\sim$10% of all breast cancer cases. BRCA1 and BRCA2, two of the most well-characterized susceptibility genes, account for 60% of the total genetic influence on breast cancer [77]. BRCA1 and BRCA2 are tumor suppressor genes with the cellular function to repair damaged DNA fragments. Genetic mutations on these two genes often lead to a high risk of breast cancer [78].

1.3.4 Prostate cancer

Prostate cancer is the second most common cancer in men globally, which causes 1.2 million new cases every year. In Sweden, 10,000 men are newly diagnosed and around 2,500 deaths occur each year, which makes prostate cancer the most frequent and deadliest tumor in the country [79]. The risk factors of prostate cancer include age, obesity, race, family history, and genetic alterations. The prostate-specific antigen (PSA) testing has been widely used for cancer

screening. However, its accuracy and efficacy are still controversial. In recent years, artificial intelligence (AI) approaches have been applied to diagnose and stratify prostate cancer using biopsy images. Results show that AI methods achieve a high accuracy and provide clinically useful aids to urological clinicians for the analysis of prostate biopsy samples [80]. Genetic mutations have been implicated in the formation and progression of prostate cancer. For example, using whole-exome sequencing data, researchers have identified deletions in PTEN (10q23) and NKX (8p21) as recurrent genomic alterations associated with prostate tumorigenesis [81]. The fusion gene represents another major type of mutation identified in prostate cancer. The chimeric gene between TMPRSS2 and ETS gene family, especially TMPRSS2-ERG and TMPRSS2-ETV1/4, are frequently detected in cancer patients. The TMPRSS2-ERG fusion originates from an interstitial deletion in chromosome 21, which is the most frequent fusion event observed in more than 55% of cases [82].

2 Aims of this thesis

Although the sequencing platforms have yielded tremendous amount of omics data, it remains a major challenge to analyze these data with high accuracy and efficiency. The overall aim of this thesis is to develop novel statistical methods to analyze the high-throughput omics data and make biologically meaningful interpretations in cancer studies. The specific aims of the four studies are as follows:

- ◇ To integrate multiple omics data from neuroblastoma patients and identify potential driver genes contributing to the formation and progression of the disease. The datasets utilized in this study include microarray comparative genomic hybridization data, gene expression profile, gene interaction network, and clinical records.
- ◇ To develop a novel statistical method for the quantification of gene expression at the isoform level using RNA-seq data. The method also aims to correct all potential biases in the sequencing data.
- ◇ To build a new approach to quantify mutant-allele expression at the isoform level using RNA-seq data and investigate the association between isoform-level expression and drug response in cancer patients.
- ◇ To develop an analysis pipeline for the detection of fusion genes using whole-exome sequencing data and re-targeted sequencing data from acute myeloid leukemia and prostate cancer samples.

3 Materials and methods

In this thesis, we have utilized a wide range of omics data from several cancer types. Both real and simulated sequencing data have been employed to test the performance of the methods developed in this doctoral project.

3.1 Integrative analysis of neuroblastoma omics data

From the Critical Assessment of Massive Data Analysis (CAMDA 2017) challenge (http://camda.info/), we obtain the RNA-seq data for gene expression quantification, array-based comparative genomic hybridization (aCGH) data for copy number alteration (CNA) detection and an external functional gene interaction network dataset for network enrichment analysis (NEA). The challenge provides a total of 498 neuroblastoma patients, among which 145 cases have matched RNA-seq and aCGH data. A subset of 48 out of the 145 patients are clinically classified as high-risk cases.

Figure 3.1 shows the workflow to integrate multiple data types and detect potential driver genes in neuroblastoma patients. In the first step, the gene expression profile including 60,776 genes are quantified using the Magic-AceView (MAV) method [83]. The expression level for each gene is ranked across all 498 patients. Within each patient, we take the 100 highest ranked and 100 lowest ranked genes as patient-specific extremely expressed genes, also denoted as expression altered gene set (AGS). Secondly, we employ two computational tools, MPSS and cnvpack, to detect recurrent CNAs using aCGH data. We annotate the genes harboring CNAs and keep those with consistent functional impact on gene expression. For example, if a gene carries a duplication, the expression of CNA altered samples is expected to be significantly higher than the non-altered cases (one-sided Welch's t-test, p-value<0.05). The genes with corresponding effects on gene expression are named functional gene set (FGS) as indicated in 3.1. Thirdly, we apply a network enrichment analysis to identify potential driver genes using FGS and AGS results. The key point of NEA analysis is that the functional effects of each FGS gene can be evaluated by the number of links with AGS genes in the interaction network. A driver gene is defined as CNA altered gene with consistent expression pattern and functionally significant in network enrichment analysis.

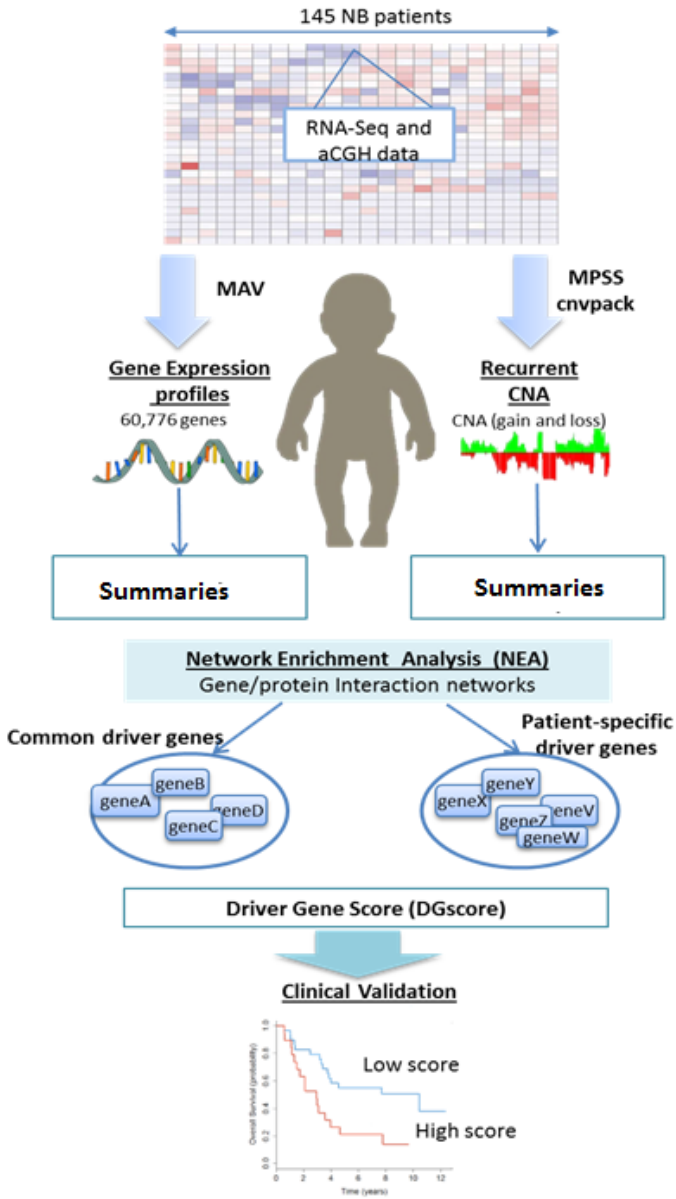

Figure 3.1: Integrative analysis pipeline to identify driver genes in neuroblastoma patients and subsequent clinical validation. *The figure is from Study I and reprinted with permission from BioMed Central Ltd [24].*

Next, we summarize the total number of driver genes in each patient as the driver gene score (DGscore). We then assess the prognostic significance of DGscore by

comparing the patients' survival in high and low DGscore groups.

3.2 Isoform quantification using RNA-seq data

3.2.1 Isoform quantification model

In this project, we use the concept of 'equivalence class' (eqclass) introduced in a recent study [84]. An eqclass defines exon(s) shared by several isoforms and the reads mapped to the shared exon(s). Note that the eqclass does not have to be biologically meaningful exons; it refers to any sequence that causes a sequence sharing problem. We summarize the number of reads aligned to each eqclass using a mapper named Rapmap from the input RNA-seq data [84]. We define y_j as the number of reads (read count) mapped to eqclass j. For a specific eqclass J with T isoforms, we denote β_t to be the expression level of isoform t. The major task is to estimate isoform abundances β_t from the read count data $y = \{y_j, j = 1,...,J\}$. By adding up the read counts of multiple isoforms, we model the expected number of reads in eqclass j as

$$\mu_j = w\sum_t a_{jt}L_t p_{jt}\beta_t \equiv \sum_t x_{jt}\beta_t, \tag{3.1}$$

which can also be written as

$$\mu = X\beta, \tag{3.2}$$

where $x_{jt} \equiv wa_{jt}L_t p_{jt}$. Here w is the total number of mapped reads normalized by isoform length and library size, a_{jt} is the isoform-specific bias or non-uniformity effect, L_t is effective length and p_{jt} is the proportion of reads in eqclass j under uniform distribution. For each isoform t we have $\sum_j x_{jt} \equiv 1$. It is conventionally assumed that y_j has Poisson distribution with mean μ_j. In general, both X and β in equation (3.2) are unknown parameters, so we have a bilinear model with two variables to estimate. Under the uniform read distribution assumption, we have $a_{jt} \equiv 1$, so (3.1) becomes

$$\mu_j \equiv w\sum_t L_t p_{jt}\beta_t \tag{3.3}$$

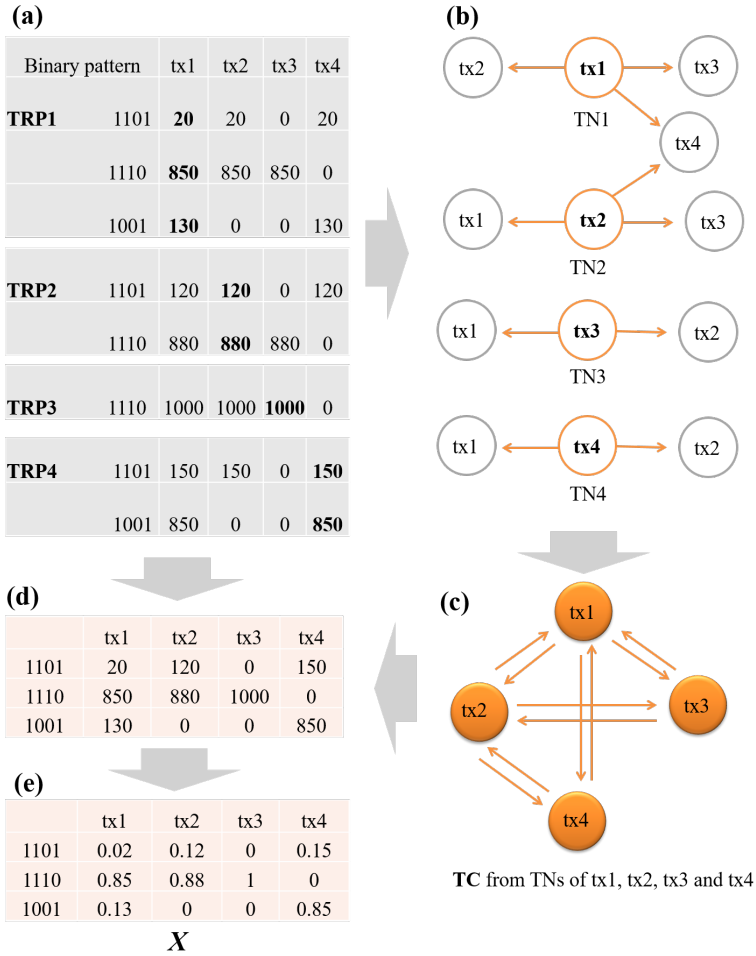

Figure 3.2: Construction of the initial matrix X using simulated RNA-seq data. *The figure is from Study II and reprinted with permission from Oxford University Press [21].*

3.2.2 Construction of X matrix

According to model (3.3) and the definition, X matrix should contain three components: (1) a group of isoforms sharing multiple exons; (2) a list of eqclasses that define exons shared between isoforms; (3) the proportion of read counts from each eqclass that contributes to the total expression of each isoform. Figure 3.2 shows the steps to construct the initial X matrix using a simulation

scheme. For each transcript, we simulate the corresponding RNA-seq sample using the R package Polyester [85]. For example, sample1 is simulated for tx1, where we assign read counts to tx1 only and other transcripts should not be expressed. We then utilize Rapmap for read alignment and read quantification in each eqclass. The result is summarized in the transcript response profile (TRP) matrix as illustrated in Figure 3.2(a). In TRP1, each row represents an eqclass and the number of reads mapped into this eqclass. The binary pattern indicates if a transcript has reads originating from the specific eqclass. Each TRP also defines transcript neighbors (TNs), which are isoforms associated with each other due to exon sharing. As shown in Figure 3.2(b) and (c), we continue to generate TRPs for other transcripts and summarize the associations into a transcription cluster (TC). The unique set of binary patterns and original read counts from each TRP are recorded in Figure 3.2(d). In Figure 3.2(e), the read counts in each transcript are standardized with the sum of one to generate the initial X matrix.

3.2.3 Alternating expectation-maximization algorithm

The starting X matrix is served as an input in equation 3.2. Figure 3.3 shows the workflow of our quantification method XAEM to estimate the isoform abundance. In step (a) we generate the Y matrix using RNA-seq data. The matrix records the number of reads mapped to each eqclass in multiple samples. Step (b) involves the X matrix constructed as mentioned above. In step (c), we estimate both X and β using an alternating expectation-maximization (AEM) algorithm. The estimation is conducted iteratively until X and β have less than 1% difference between successive iterations. In this estimation process, a potential issue could be caused by paralogs, which are transcripts with extremely similar sequences. Paralogs in X matrix will make the X matrix singular and the β non-identifiable. To deal with this issue, we use the k-means clustering to combine paralogs into one transcript.

3.2.4 Simulated and real RNA-seq data

Simulated data are commonly used for benchmarking the quantification approaches. We implement Polyester to simulate RNA-seq reads based on the expression values from a human colon cancer cell line HCT116 [86]. Polyester

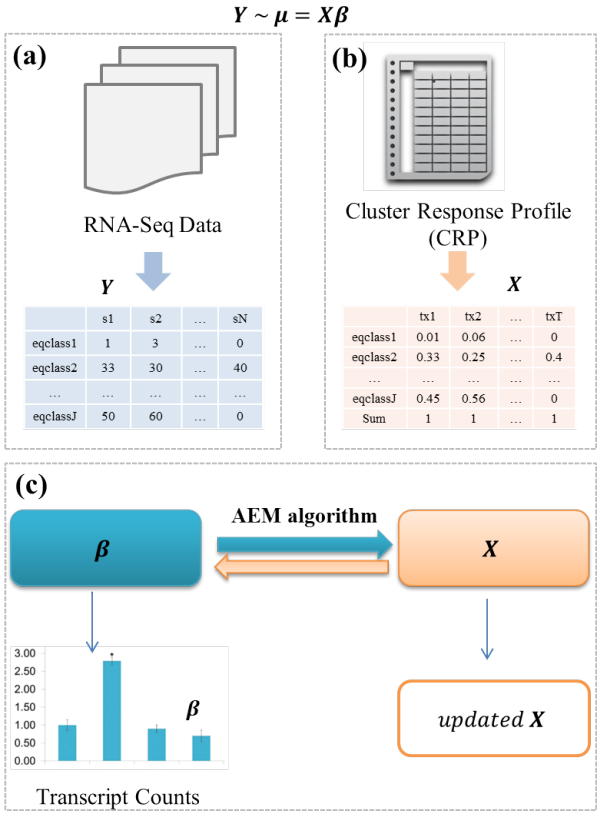

Figure 3.3: The workflow of XAEM to quantify isoform level expression. *The figure is from Study II and reprinted with permission from Oxford University Press [21].*

can generate sequencing reads under uniform and non-uniform distribution, so that we simulate 100 RNA-seq samples under uniform condition and another 100 samples with non-uniform read distribution. Paired-end reads are generated with read length of 100 bp and fragment length at 250 bp. We obtain two real RNA-seq datasets in this project. The first comprises 384 cells from a triple negative breast cancer cell line (MDA-MB-231). The dataset includes two batches and 50% of cells in each batch are treated with metformin. The second real dataset is downloaded from the Sequencing Quality Control Consortium (SEQC) project [87]. The dataset contains two unique RNA samples and hundreds of replicates sequenced in several laboratory sites. We select four replicates for each sample and obtain the RNA-seq data from the Sequence Read

Archive (SRA) repository. A qPCR validated expression profile is also acquired for the eight replicates.

We compare the quantification performance of XAEM with other existing approaches such as Cufflinks [47], Sailfish [84], Kallisto [54], and Salmon [53]. An absolute proportional error (APE) is calculated using the equation 3.4, where E is the estimated expression value and T is the ground truth.

$$APE = |E - T|/(T + 1). \tag{3.4}$$

3.3 Estimation of mutation-allele expression

It is well recognized that DNA mutations play crucial roles in cancer initiation and progression [88]. However, traditional quantification methods often ignore mutant status and alleles. To address this issue, we extend the idea of X matrix and AEM algorithm to estimate mutation-allele expression at the isoform level. We use a more flexible strategy by estimating the sum of all mutant isoforms originating from the same wild-type isoform. For instance, two mutant isoforms, isoform_mut1 and isoform_mut2, are associated with the wild-type version isoform_wt. In the process of X matrix construction, we rename both isoform_mut1 and isoform_mut2 as isoform_mut. We then recode the binary pattern in respective eqclass and merge those with the same pattern. This processing will generate only one mutant version for each wild-type isoform, and the number of total isoforms in the X matrix will be up to $M * 2$, where M is the number of wild-type isoforms. The X matrix including both wild-type and mutant isoforms is then served as input in equation 3.2.

We simulate two RNA-seq datasets to evaluate the accuracy of our method MAX and another quantification method Salmon [53]. The first dataset comprise 100 non-mutated samples where we only assign read counts to wild-type isoforms. The second dataset contain 100 mutated samples with equal read counts to both wild-type and mutant isoforms. For the real RNA-seq data, we obtain a total of 461 RNA-seq samples from the BeatAML study [89]. The dataset also includes whole-exome sequencing data, clinical records, and drug response data. The BeatAML project provides a detailed list of genetic variations detected using variation callers such as Mutect [90] and Varscan2 [91].

3.4 Fusion gene detection at DNA level

In this study, we build a pipeline to detect fusion genes using paired-end whole-exome sequencing and targeted sequencing data. We first align reads to genome reference using aligners such as BWA [50] or Bowtie2 [51]. The output is in the Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM) format, which records the mapping position, flag, mapping quality, CIGAR string and other alignment results. Based on these information, we extract (1) the discordant reads, where the two reads are mapping to different genes and (2) split reads, where a single read is partially mapping to more than one gene. The fusion gene identification from split reads can be straightforward since split reads spanning the fusion break point directly. For discordant reads, the idea of equivalence class mentioned in section 3.2.1 is utilized to construct a fusion equivalence class (FEQ). Each FEQ comprises the constituent genes and the number of reads supporting the fusion event. We then apply multiple filters to exclude the fusion candidates that are false positives. We test the performance of our method on three large cancer datasets including BeatAML data, TCGA-AML data, and the Prostate Biomarkers cohort. The BeatAML cohort provides a total of 531 samples with WES data and 411 patients with RNA-seq data. The TCGA-AML project performs whole-exome sequencing on 150 samples and RNA-seq on 179 samples. From the Prostate Biomarkers cohort, we obtain a total of 65 patients with targeted deep-sequencing data.

4 Main results

4.1 Study I

We apply the analysis pipeline as shown in Figure 3.1 to a subset of 48 high-risk neuroblastoma patients. A total of 274 genes with recurrent copy number alteration and parallel impact on gene expression are identified. We apply the network enrichment analysis (NEA) to detect patient-specific driver genes, where the input AGS are the top 200 extremely expressed genes from each patient and the input FGS are those CNA-altered genes in each patient (subset of the 274 recurrent genes). The enrichment analysis identifies 66 patient-specific driver genes; the full list is given in Additional File 4 attached with Study I. Next, we detect the common driver genes where the input FGS and AGS are genes present in at least five patients (10% of 48 samples). We detect four common drivers: ERCC6, HECTD2, KIAA1279, and EMX2.

We summarize the total number of common and patient-specific drivers in each sample as DGscore and evaluate its clinical relevance in patients' survival. The 48 patients are divided into high and low DGscore groups based on the median value of the score. Figure 4.1(a) shows that the low DGscore group has a significantly better outcome than the high DGscore group (p-value=0.006). In Figure 4.1(b), we only use the 274 CNA altered genes without the NEA step; the result indicates that it cannot distinguish the survival between high and low DGscore groups (p-value=0.492). In Figure 4.1(c) and (d), the DGscore is calculated only using patient-specific or common driver genes. The results indicate that either type of driver genes is insufficient to predict the patients' survival (p-value>0.2).

4.2 Study II

In study II, we develop a novel method named XAEM for the quantification of isoform abundance. We first apply XAEM to quantify the isoform expression using simulated data and compare the accuracy with existing methods such as Salmon [53], Sailfish [84], Kallisto [54], and Cufflinks [47]. Table 4.1 summarizes the median APE for each method using 100 uniform and non-uniform samples. The isoforms are divided into three categories: (1)

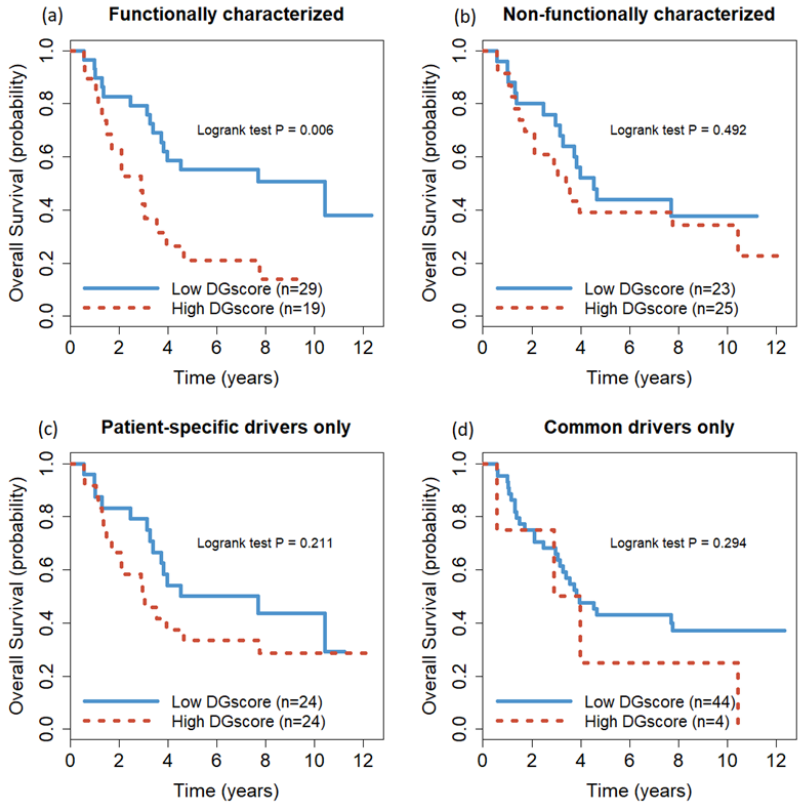

Figure 4.1: Survival analysis of 48 high-risk patients under different driver gene conditions. *The figure is from Study I and reprinted with permission from BioMed Central Ltd [24].*

singletons, which originate from genes that produce only single isoform; (2) non-paralogs, where a gene generates multiple isoforms but not paralogs; (3) paralogs, which are extremely similar isoforms as described in section 3.2.3. It can be seen that the estimation of singletons is accurate in most methods. The median APEs for XAEM, Salmon, Kallisto, and Sailfish are 0, while Cufflinks has a median APE at 0.28 and 0.45 under uniform and non-uniform setting, respectively. The quantification of multiple isoforms is more challenging than singletons. Table 4.1(A) shows that the APEs of XAEM, Salmon, Kallisto, and Sailfish increase to 0.18, 0.18, 0.20, 0.20 in the uniform samples. Table 4.1(B) indicates that XAEM achieves higher accuracy under the non-uniform scenario with the median APE at 0.37, while the APE for Salmon is 0.42, Kallisto 0.44,

Table 4.1: Comparison of the quantification accuracy of XAEM, Salmon, Kallisto, Sailfish, and Cufflinks. The isoforms are divided into singletons, non-paralogs, and paralogs. The median APE is calculated for each method using (A) 100 uniform and (B) 100 non-uniform simulated samples.

		Multiple isoforms	
Methods	Singletons	Non-paralogs	Paralogs
(A) Uniform	(*N*=14,446)	(*N*=25,838)	(*N*=6,112)
XAEM	0	0.18	0.12
Salmon	0	0.18	0.45
Kallisto	0	0.20	0.47
Sailfish	0	0.20	0.47
Cufflinks	0.28	0.36	0.54
(B) Non-uniform	(*N*=14,446)	(*N*=18,597)	(*N*=13,353)
XAEM	0	0.37	0.15
Salmon	0	0.42	0.966
Kallisto	0	0.44	0.969
Sailfish	0	0.45	0.968
Cufflinks	0.45	0.69	0.970

and Sailfish 0.45. Cufflinks has the worst performance with the APE at 0.69. The quantification of paralogs is a major challenge for all methods. For this category, we calculate the APE using the original output in each method; hence, in XAEM, the paralogs remain merged but are separated in other methods. We can see that the APEs of XAEM are 0.12 in uniform samples and 0.15 in non-uniform samples, which are substantially smaller than other approaches. The APEs for Salmon, Kallisto, Sailfish, and Cufflinks are close to one in non-uniform samples, indicating that these methods have inferior estimations for paralog isoforms.

Differential expression analysis is commonly used to identify differentially expressed (DE) genes between cancerous and healthy samples. The breast cancer cell line MDA-MB-231 contains two batches of cells; each batch has 96 metformin treated cells and 96 control cells. To evaluate the performance of XAEM in real RNA-seq data, we generate a training set and validation set. The training set comprises 40 randomly selected treated cells and 40 control cells.

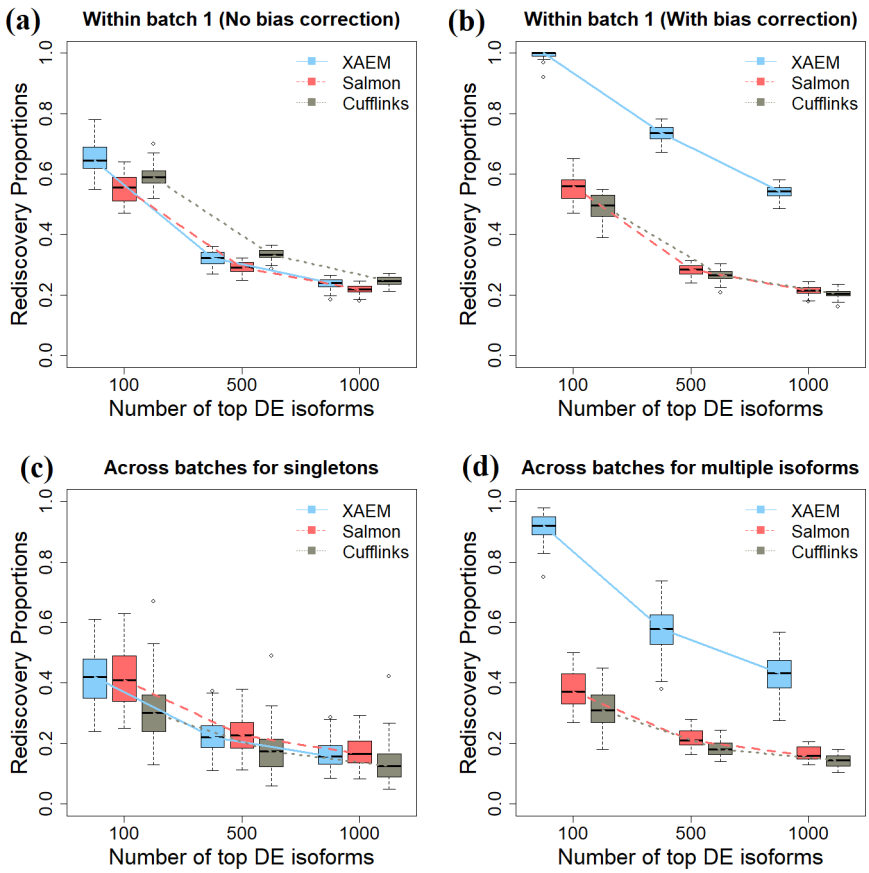

Figure 4.2: Detection and validation of differentially expressed isoforms using breast cancer cell line RNA-seq data. The comparison is between XAEM, Salmon, and Cufflinks. *The figure is from Study II and reprinted with permission from Oxford University Press [21].*

The validation set contains another set of 40 treated and 40 control cells. We calculate a rediscovery rate (RDR) that indicates the number of significant DE isoforms from the training set that are validated in the validation set. Figure 4.2 shows the comparison between XAEM, Salmon and Cufflinks. In Figure 4.2(a), all three methods are implemented on batch 1 without bias correction. Therefore, in XAEM, we do not run the AEM step to correct biases from RNA-seq data. It can be seen that the RDR is similar between the three methods. In Figure 4.2(b), the bias correction step is added back to each method, in that the AEM step is

used in XAEM's estimation. We can see a notable improvement of XAEM, where the RDRs of top 100, 500, 1000 DE isoforms are 1.0, 0.56, and 0.50, respectively, which is substantially higher than those in Salmon and Cufflinks. Figure 4.2(c) shows the comparison across batches for singleton, indicating that there is no significant difference in RDR among the three methods. The finding is in agreement with Table 4.1 since the quantification of singletons is trivial. Figure 4.2(d) shows that XAEM achieves higher RDR for multiple isoforms (non-paralogs) across batches. The overall RDRs for XAEM, Salmon, and Cufflinks are 0.77, 0.26, and 0.22, respectively.

4.3 Study III

Table 4.2: Comparison of estimation accuracy between MAX and Salmon using 100 non-mutated and 100 mutated samples for FLT3, NPM1 and TP53.

	FLT3		**NPM1**		**TP53**	
	WT	Mut	WT	Mut	WT	Mut
Non-Mutated samples						
MAX	0.05	0	0.06	0	0.24	0
Salmon	0.22	0	0.29	0	0.99	0
Mutated samples						
MAX	0.03	0.04	0.07	0.07	0.22	0.25
Salmon	0.87	0.85	0.65	0.87	0.99	0.99

In this study, we develop a method named MAX to quantify the mutant-allele expression from both simulated and real RNA-seq data. The simulated data include 100 mutated samples and 100 non-mutated samples for FTL3, NPM1, and TP53. Table 4.2 shows the median APE for wild-type (WT) and mutant (Mut) isoforms from MAX and Salmon. In non-mutated samples, where only wild-type isoforms should be expressed, MAX achieves a higher accuracy than Salmon. The median APEs for MAX in FLT3, NPM1 and TP53 are 0.05, 0.06, and 0.24, while the APEs from Salmon are 0.22, 0.29, and 0.99, respectively. The median APEs of mutant isoforms from MAX and Salmon are all zero, indicating that both methods have little false positives in non-mutated scenario. The estimation in mutated RNA-seq data can be more challenging. As Table 4.2 shows, for TP53 gene, MAX has median APEs at 0.22 and 0.25 for WT and Mut isoforms. This could be because TP53 has 15 WT isoforms and 49 distinct

mutations, which results in a total of 522 MT isoforms. The accuracy of Salmon becomes extremely low, where the median errors across three genes are >0.65. For example, the median APEs from Salmon for WT and Mut isoforms in TP53 are close to one. Overall, compared with Salmon, MAX achieves a much higher accuracy for the quantification of wild-type and mutant isoforms under non-mutated and mutated conditions.

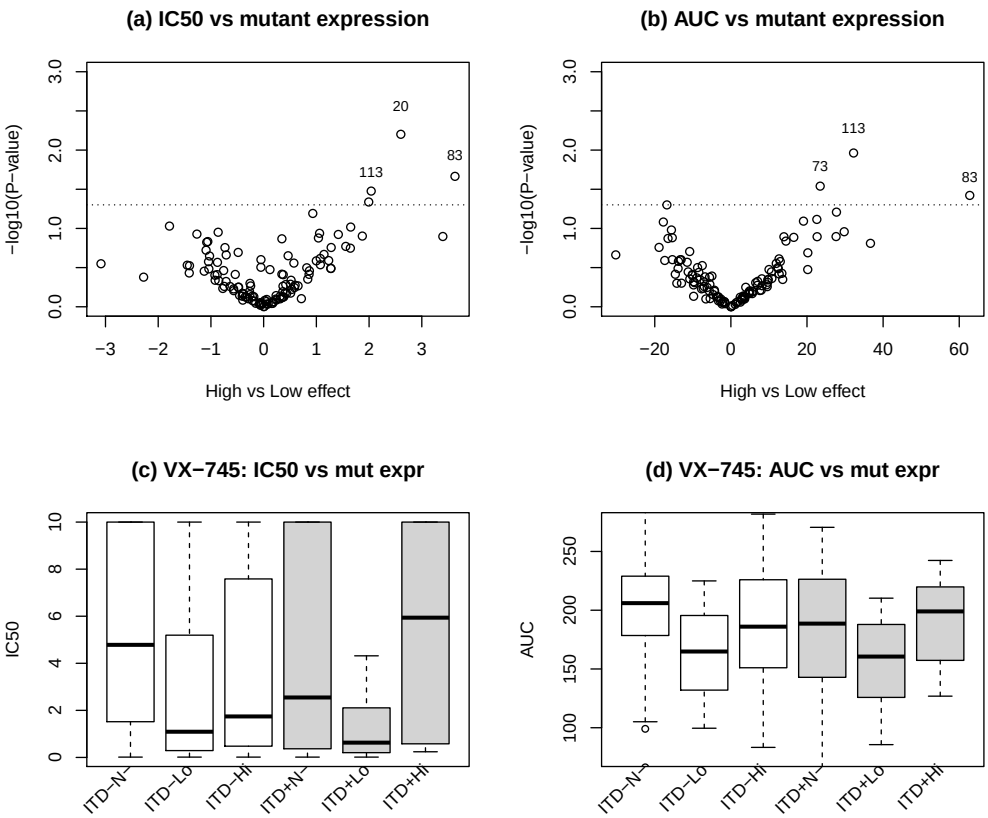

Figure 4.3: Differential analysis of drug response in NPM1 high/low mutant expression and ITD positive/negative groups: (a) drug response is measured in IC50; (b) AUC; (c) boxplots of IC50 in six subgroups based on FLT3-ITD and NPM1 mutant expression; (d) boxplots using AUC values.

We implement MAX to BeatAML RNA-seq data and investigate the clinical impact of the mutant-allele expression on drug response. The BeatAML

conducted drug screening experiments on 122 small-molecule inhibitors and the drug sensitivities are measured in inhibitory concentration (IC50) and area under the curve (AUC). We focus on 82 samples carrying the NPM1 mutation and calculate a Mut/WT allele expression ratio for each isoform. Based on the expression ratio of the dominant NM_002520 isoform, we split the 82 patients into high and low mutant expression groups. Figures 4.3(a) and (b) show the volcano plots using IC50 and AUC metrics. It can be seen that drug 83 (panobinostat) and drug 113 (VX-745) have significantly differential responses between high and low ratio groups. We check the expression of VX-745 further by dividing the patients into six subgroups based on the presence of FLT3-ITD alteration and the high/low expression ratio of NPM1. Figures 4.3(c) and (d) illustrate the drug response in IC50 and AUC, respectively. In each panel, the three left-most boxplots represent FLT3-ITD negative patients and the three right-most boxplots are FLT3-ITD positive patients. Among the FLT3-ITD negative samples, both low and high NPM1 mutated groups have better drug response than NPM1 non-mutated samples (p-value=0.04 for IC50 and 0.004 for AUC). However, in the FLT3-positive patients, samples with low NPM1 expression are the only group to have good drug response (p-value=0.001 for IC50 and 0.03 for AUC). The results indicate that, based on the mutant-allele expression profile, we identify a subgroup of patients having better drug response to a kinase inhibitor.

4.4 Study IV

In study IV, we build a method named Fuseq-WES to identify fusion genes from whole-exome and targeted sequencing data. We apply the method to three large cancer datasets and focus on the validation of several well-established fusion genes in AML and prostate cancer. The PML-RARA, CBFB-MYH11, and RUNX1-RUNX1T1 fusions are among the most common fusion genes in AML [89]. Table 4.3 shows that in BeatAML dataset, PML-RARA is detected in 16 patients using the RNA-seq data. Eleven out of the 16 patients have matched WES data, and four are validated by Fuseq-WES with a validation rate of 36%. Fuseq-WES identifies CBFB-MYH11 in 15 samples using WES data, indicating a much higher validation rate of 63%. For the RUNX1-RUNX1T1 fusion, none of the six samples with WES data have been validated. In the TCGA data, three

cases are detected to carry PML-RARA and two are harboring the RUNX1-RUNX1T1 fusion. In contrast, no cases are identified with CBFB-MYH11, which indicates a validation rate at 0.

Table 4.3: The number of patients detected with fusion genes using RNA-seq data; number of patients with matched WES data and number of patients carrying the fusion events validated using WES data.

	BeatAML RNA-seq data	**WES data**	**Fuseq-WES**
PML-RARA	16	11	4
CBFB-MYH11	25	24	15
RUNX1-RUNX1T1	9	6	0
	TCGA RNA-seq data	**WES data**	**Fuseq-WES**
PML-RARA	16	6	3
CBFB-MYH11	11	6	0
RUNX1-RUNX1T1	7	4	2

Table 4.4: Fusion detection results for TMPRSS2-ERG in Prostate Biomarker dataset

	Positive IGV	**Negative IGV**	**Total**
Postive Fuseq-WES	36	5	41
Negative Fuseq-WES	1	23	24
Total	37	28	65

TMPRSS2-ERG (TE) is a predominant fusion in prostate cancer, which can be observed in >55% of patients. From the Prostate Biomarker project we obtain 65 samples with targeted sequencing data. We first implement four individual tools for fusion detection, including SvABA [92], GRIDSS [93], LUMPY [94], and a python-based tool named SVcaller. We then verify the fusion genes using the Integrative Genomics Viewer (IGV) and keep those successfully verified by IGV. Table 4.4 shows the comparison between IGV and Fuseq-WES detection. TMPRSS2-ERG is identified in 41 and absent in 24 patients using Fuseq-WES, while the IGV method detects the fusion in 37 samples. The comparison shows that the results are concordant in 36 TE positive and 23 TE negative samples, indicating an overall concordance at 91%.

5 Discussion and conclusion

Integrative analysis of multi-omics data has been widely used in the era of high-throughput sequencing. In **Study I**, we present an integrative analysis pipeline combining RNA-seq, copy number alteration and network enrichment profile to detect driver genes in neuroblastoma patients. A total of 66 patient-specific and four common driver genes are detected from 48 high-risk cases. We calculate a DGscore based on the driver genes and evaluate its clinical impact in survival analysis. Results show that patients with a low DGscore have significantly better outcomes than those with a high DGscore.

A key feature of the integrative pipeline is that it combines signals from multiple omic sources. A driver gene is defined when it fulfills three criteria: altered copy number, having a consistent impact on gene expression and enriched in gene functional network. The result in Figure 4.1(b) shows that without functional characterization in the gene interaction network, the DGscore cannot distinguish the survival of high and low DGscore groups. Besides, the DGscore takes both patient-specific and common drivers into account. As shown in Figures 4.1(c) and (d), using patient-specific or common driver genes only is not sufficient to predict the patients' survival. In a Cox regression analysis of high-risk neuroblastoma patients, DGscore emerged as the strongest prognostic factor with the adjustment of age, tumor stage and MYCN amplification.

Notably, MYCN is a well-established oncogene and a significant predictor for survival in neuroblastoma patients [69]; however, it is not necessarily effective for high-risk groups. Unlike DGscore, which integrates several levels of signals, MYCN amplification alone is insufficient to predict the outcome of high-risk patients. The result shows the importance to consider multiple attributes ranging from mutation status to functional impacts in identifying candidate driver genes.

The quantification of isoform level expression is a fundamental task in RNA-seq data analysis. Compared with gene level quantification, which simply adds up all reads mapped to a single gene, the estimation of isoform abundance is trickier due to the alternative splicing mechanism and exon sharing. A major problem is how to distribute reads mapping to exons shared by different isoforms.

In **Study II**, we develop a method named XAEM to quantify the isoform expression from RNA-seq data. Many existing methods utilize a linear model

$Y = X\beta$ with a possibly known X and estimate only the β. In contrast, XAEM leverages a more flexible bi-linear model where both X and β are unknown. We construct the initial X matrix using a simulation scheme and divide the whole transcriptome into small and feasible units as isoform clusters. The X matrix is then served as an input variable in the model where β and X are estimated using an AEM algorithm. In the process of updating the X matrix, the AEM algorithm automatically corrects all potential biases observed from multiple input RNA-seq samples.

We utilize simulated and real RNA-seq data to evaluate the performance of XAEM and compare it with other approaches. The comparison shows that XAEM achieves higher accuracy in multiple isoforms and better rediscovery rate in differential expression analysis. Paralogs are isoforms with extremely similar sequences. The quantification of paralogs using short read-length sequencing data remains a big challenge in existing approaches. In the XAEM model, we employ a special strategy to merge paralogs into a combined isoform group, which leads to a more accurate estimation of paralog abundance.

In **Study III**, we extend the concept of AEM algorithm and develop a method named MAX to quantify mutant-allele expression at isoform level. The major obstacle in this analysis is the highly similar sequences between wild-type isoform and potentially large number of mutant isoforms. For example, if there are M wild-type isoforms and N distinct mutations detected in a gene, the total number of wild-type and mutant isoforms can be $M \times 2^N$. The large number of isoforms will make the X matrix sparse and result in an indeterminate solution in the equation 3.2.

To address this issue, we merge the mutant isoforms from the same wild-type isoform into a single mutant set. In this case, each wild-type isoform only has one mutant version, thus reducing the dimension of the X and making the quantification feasible. We assess the accuracy of MAX using mutated and non-mutated RNA-seq data and compare it with a standard quantification method Salmon. The results indicate that MAX achieves substantially better performance than Salmon under both scenarios.

One advantage of MAX over other methods is the utilization of explicit X matrix, which integrates both wild-type and mutant isoforms, providing an effective way to deal with the huge amount of sequence similarities. We apply MAX to a real RNA-seq dataset from BeatAML project; the analysis shows little

false positive estimates in non-mutated samples, which can be considered as a validation of MAX's accuracy in real-world data. To evaluate the clinical significance of mutant-allele expression, we investigate the expression pattern in the drug response data comprising 122 drugs. We find that a subgroup of NPM1 mutated patients has a better drug response than other groups. The results demonstrate that mutant-allele expression can provide significant information for patient stratification and individualized treatment.

Whole-exome sequencing data have been frequently used to investigate the genetic landscape in numerous diseases. The large number of WES data provide rich resources for further exploitation. In **Study IV**, we build an analysis pipeline named Fuseq-WES to detect fusion genes from WES and targeted sequencing data. We apply the method to AML and prostate cancer cohorts, validating several well-known fusion genes in these two cancers. In the BeatAML data, we detect 15 patients carrying the CBFB-MYH11 fusion, which indicates a validation rate of 63%. The PML-RARA fusion is validated in 36% patients. We check the gene structure of MYH11 and notice that it contains 43 exons with the gene length at 154,000 nucleotides. Both the PML and RARA gene have nine exons with a gene length of $\sim$50,000 bases. Accordingly, we speculate that the density of exons could be a key factor for validation rate, where a high density can facilitate the detection of fusion events. However, in the TCGA dataset, the CBFB-MYH11 is not validated in any samples. From the WES data, we find that the read depth of TCGA data is 15x while in BeatAML data it is 40x. The result suggests that read depth could also affect the fusion gene detection using WES data.

This is confirmed when we use targeted sequencing data to detect TMPRSS2-ERG fusion in prostate cancer patients. The average read depth of targeted sequencing data is 1500x, which provides an ultra deep coverage for fusion detection. The result shows that Fuseq-WES achieves a 97% concordance in TMPRSS2-ERG positive cases and 88% concordance in negative cases. Although we validate several fusion genes using exome sequencing data, the overall detection result indicates that it is difficult to replicate the fusion genes detected in the RNA-seq dataset. An inherent disadvantage of WES data is that only exonic regions are sequenced. Thus, if a fusion gene occurs in the intron or intergenic region, the WES is unable to capture the breakpoint.

6 Future perspectives

In this thesis, we develop several methods to analyze a wide variety of omics data from cancer patients. Each method aims to tackle specific questions in omics data analysis and subsequently facilitates the investigation of pathophysiologic mechanisms underlying different cancer types. Our studies have several limitations. For example, in Study I, we detect driver genes on a small number (48) of high-risk neuroblastoma patients. The same problem exists in Study III and IV, where the number of NPM1 mutated samples or patients with WES data is limited. Ideally, if sufficient omics data had been available, we would be able to identify candidate driver genes with higher confidence. Besides, we could validate the efficacy and performance of our methods in a separate validation set.

The second limitation comes from the completeness of genome/transcriptome reference used for read alignment. The reference is served as the basis to detect genetic mutations and quantify isoform abundance. However, since the release of the first draft of human genome, many complex regions remain unfinished or erroneous. These incomplete and incorrect segments can have negative effects on gene annotation and other downstream analyses. According to a latest pre-print study [95], a research consortium has almost filled in the unfinished gaps and added $\sim$115 protein-coding genes. We anticipate that a complete and refined human genome would generally improve the performance of related analyses.

Another limitation regarding the XAEM approach is the applicability to single-cell (sc) RNA-seq data. The current version of XAEM achieves the best performance using bulk RNA-seq data, while the scRNA-seq data usually have lowly expressed isoforms and produce a sparse read-count matrix. Therefore, improving the accuracy of XAEM using scRNA-seq data is worth further investigation.

7 Acknowledgements

Finally I am reaching this point here, with all the memories reappearing in my head. During my last several years of PhD study, there are so many of you who helped and supported me warmly. Here I would like to take the chance to express my sincere gratitude to all of you.

First of all, to my main supervisor, **Yudi Pawitan**. Thank you for giving me the chance to conduct the PhD study in your group. I learned so much from you and I believe I have improved myself a lot with your help. Also, thank you for taking us to so many different restaurants for the delicious food, I surely miss those happy times.

My co-supervisor, **Trung Nghia Vu**. Thank you for being so patient with me all the time. I feel that you are my "elder brother", who is always there and ready to help when I have problems in my project. I am so grateful that you help me debug and improve my codes and analysis. For so many times I was impressed by your intelligence and efficiency to solve the problems that I got stuck.

My co-supervisor, **Xia Shen**. It is really pleasant and inspiring every time when I see you and hear your talk. You are very interesting person with a lot of brilliant ideas in your head. I feel like you are my "younger brother" because you told me which video game is the most popular, which restaurant serves the most authentic food, which country has the best view, which mountain is the best for skiing and so on. I am happy that you are here because it is joyful to talk with you.

Thanks to my mentor, **Di Wu**, who is also my landlord, roommate, "part-time boss", collaborator and a true friend. We spend a lot of times together and talk a wide range of topics every day. I have learned so much from you when we discuss about academy, industry, technology and sometimes philosophy. Special thanks to **Yanling Cai**, for providing me the accommodation in the last four yeas, I wish you a big success in your career and life journey.

Thanks to everyone in the group: **Zheng Ning**, for helping me since the first day when I arrived in Sweden and giving me advises to prepare each important

seminar; to **Tian Mou**, **Stefano Calza**, **Sophie Debonneville**, **Lu Pan**, **Quang Thinh Trac**, **Sarath Kumar Murugan**, **Dat T Nguyen** and **Tingyou Zhou**, thanks for the inspring discussions during our weekly meeting.

Thanks to my office mates: **Linda Abrahamsson**, **Xingrong Liu**, for helping me at the very beginning of my PhD study, and our everyday fika within the office. **Anastasia Lam**, **Nurgul Batyrbekova** and **Henrik Olsson**, thanks for your accompany, I have learned so many interesting things from your study and project.

I would like to express my gratitude to everyone in **MEB**, past and present. Thanks to **Chen Suo**, it is pleasant and enjoyable to have every talk and discussion with you, and I wish you a big success in your academic career. **Chuen Seng Tan**, thanks for hosting me warmly in Singapore. To **Jingru Yu**, **Shihua Sun**, **Weiwei Bian**, **Yun Du**, **Shuan Hao**, **Can Cui**, **Xinhe Mao**, **Yang Xu**, **Erwei Zeng**, **Shengxin Liu**, **Qian Yang**, **Jiangwei Sun**, **Lin Li**, **Yunzhang Wang**, **Yinxi Wang**, **Xia Li**, **Shuyang Yao**, **Venkatesh Chellappa**, **Tong Gong**, **Jie Song**, **Yiqiang Zhan**, **Yi Lu**, **Zheng Chang**, **Donghao Lv**, **Xu Chen**, **Jiangrong Wang**, **Haomin Yang**, **Tingting Huang**, **Xiaoying Kang**, **Jiayao Lei**, **Qing Shen**, **Le Zhang**, **Ruyue Zhang**, **Chen Wang**, **Wei He** and **Yufeng Chen**. Each of you are so fabulous and talented. I feel lucky to be your friend and colleague.

To the **Biostatistics Group** colleagues: **Marie Reilly**, **Marie Jansson**, **Gabriel Isheden**, **Andreas Karlsson**, **Peter Ström**, **Keith Humphreys**, **Maya Alsheh Ali**, **Rickard Strandberg**, **Lili Meng**, **Julien Bryois**, **Hannah Bower**, **Elisabeth Dahlqwist**, **Alexander Ploner** and so on. Thank you for creating such a nice and brilliant group.

Thanks to the IT, HR and Administration group in MEB, for helping me deal with the document and contract, especially thanks to **Alessandra Nanni**, **Gunilla Nilsson Roos**, **Jenny Lindgren**, **Lina Werner** and **Gunilla Sonnebring**.

To my friends in Sweden: **Xiaoyan Qian**, **Xingwu Zhou**, **Xiang Jiao**, **Dan Bai**, it is a lot of fun to hang out with you guys. To my previous colleagues in China,

Lusheng Huang, **Bin Yang**, **Zhiyan Zhang**, **Huashui Ai**, **Shijun Xiao**, **Jun Ren**, **Aleksei Traspov** and **Wanbo Li**. Thank you for your help during my master's study.

To **Tat Wah**, thank you for all the joyful times together. I hope we can travel to more destinations in the near future.

To my parents and my family, thank you for supporting me as always. And to my four adorable nephews and niece, which are **Zhong En Jun**, **Zhong Xin Yue**, **Lan Ying Hao** and **Lan Xi Jie**, I am lucky to be your uncle, and watching all you little guys growing up. Let us hope that someday in the future, when you are able to read these words, you would be proud of your uncle and consider him as your role model.

The PhD study is a unique experience in my life. It is filled with happiness, sorrow, challenges and achievements. I will cherish this special journey and continue to improve myself as a better person.

References

[1] Sciacovelli, M., Schmidt, C., Maher, E.R., Frezza, C.: Metabolic drivers in hereditary cancer syndromes. Annual Review of Cancer Biology **4**, 77–97 (2020) 1

[2] Nones, K., Patch, A.-M.: The Impact of Next Generation Sequencing in Cancer Research. Multidisciplinary Digital Publishing Institute (2020) 1

[3] Rusch, M., Nakitandwe, J., Shurtleff, S., Newman, S., Zhang, Z., Edmonson, M.N., Parker, M., Jiao, Y., Ma, X., Liu, Y., *et al.*: Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nature communications **9**(1), 1–13 (2018) 1

[4] Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary oncology **19**(1A), 68 (2015) 1

[5] Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al.: International cancer genome consortium data portal—a one-stop shop for cancer genomics data. Database **2011** (2011) 1

[6] Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., Zhang, L.: Review on the application of machine learning algorithms in the sequence data mining of dna. Frontiers in Bioengineering and Biotechnology **8**, 1032 (2020) 2

[7] Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., Yu, J.: The next-generation sequencing technology: a technology review and future perspective. Science China Life Sciences **53**(1), 44–57 (2010) 1

[8] Sanger, F., Nicklen, S., Coulson, A.R.: Dna sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences **74**(12), 5463–5467 (1977) 2

[9] Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, W.T.: Accurate determination of microbial diversity from 454 pyrosequencing data. Nature methods **6**(9), 639–641 (2009) 3

[10] Bronner, I.F., Quail, M.A., Turner, D.J., Swerdlow, H.: Improved protocols for illumina sequencing. Current protocols in human genetics **79**(1), 18–2 (2013) 4

[11] McElhoe, J.A., Holland, M.M., Makova, K.D., Su, M.S.-W., Paul, I.M., Baker, C.H., Faith, S.A., Young, B.: Development and assessment of an optimized next-generation dna sequencing approach for the mtgenome using the illumina miseq. Forensic Science International: Genetics **13**, 20–29 (2014) 4

[12] Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., Shi, L.: Next-generation sequencing and its applications in molecular diagnostics. Expert review of molecular diagnostics **11**(3), 333–343 (2011) 5

[13] Olson, M.V.: The human genome project. Proceedings of the National Academy of Sciences **90**(10), 4338–4344 (1993) 6

[14] Bleidorn, C.: Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and biodiversity **14**(1), 1–8 (2016) 7

[15] Schadt, E.E., Turner, S., Kasarskis, A.: A window into third-generation sequencing. Human molecular genetics **19**(R2), 227–240 (2010) 7

[16] Roberts, R.J., Carneiro, M.O., Schatz, M.C.: The advantages of smrt sequencing. Genome biology **14**(7), 1–4 (2013) 7

[17] Ardui, S., Ameur, A., Vermeesch, J.R., Hestand, M.S.: Single molecule real-time (smrt) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic acids research **46**(5), 2159–2168 (2018) 7

[18] Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al.: The potential and challenges of nanopore sequencing. Nanoscience and technology: A collection of reviews from Nature Journals, 261–268 (2010) 7

[19] Jain, M., Olsen, H.E., Paten, B., Akeson, M.: The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. Genome biology **17**(1), 1–11 (2016) 7

[20] Ari, Ş., Arikan, M.: Next-generation sequencing: advantages, disadvantages, and future. In: Plant Omics: Trends and Applications, pp. 109–135. Springer, ??? (2016) 8

[21] Deng, W., Mou, T., Kalari, K.R., Niu, N., Wang, L., Pawitan, Y., Vu, T.N.: Alternating em algorithm for a bilinear model in isoform quantification from rna-seq data. Bioinformatics **36**(3), 805–812 (2020) 8, 12, 26, 28, 34

[22] Petersen, L.M., Martin, I.W., Moschetti, W.E., Kershaw, C.M., Tsongalis, G.J.: Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. Journal of Clinical Microbiology **58**(1), 01315–19 (2019) 8

[23] Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., Ma, S.: A selective review of multi-level omics data integration using variable selection. High-throughput **8**(1), 4 (2019) 9

[24] Suo, C., Deng, W., Vu, T.N., Li, M., Shi, L., Pawitan, Y.: Accumulation of potential driver genes with genomic alterations predicts survival of high-risk neuroblastoma patients. Biology direct **13**(1), 1–11 (2018) 9, 15, 24, 32

[25] Huang, S., Chaudhary, K., Garmire, L.X.: More is better: recent progress in multi-omics data integration methods. Frontiers in genetics **8**, 84 (2017) 9

[26] Bhardwaj, R., Sethi, A., Nambiar, R.: Big data in genomics: An overview. In: 2014 IEEE International Conference on Big Data (Big Data), pp. 45–49 (2014). IEEE 9

[27] Zuryn, S., Le Gras, S., Jamet, K., Jarriault, S.: A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics **186**(1), 427–430 (2010) 9

[28] Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., Ding, L.: Somaticsniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics **28**(3), 311–317 (2012) 9

[29] Li, P., Guo, M., Wang, C., Liu, X., Zou, Q.: An overview of snp interactions in genome-wide association studies. Briefings in functional genomics **14**(2), 143–155 (2015) 9

[30] Rosenthal, S.L., Barmada, M.M., Wang, X., Demirci, F.Y., Kamboh, M.I.: Connecting the dots: potential of data integration to identify regulatory snps in late-onset alzheimer's disease gwas findings. PloS one **9**(4), 95152 (2014) 9

[31] Fachal, L., Dunning, A.M.: From candidate gene studies to gwas and post-gwas analyses in breast cancer. Current opinion in genetics & development **30**, 32–41 (2015) 9

[32] Guan, P., Sung, W.-K.: Structural variation detection using next-generation sequencing data: a comparative technical review. Methods **102**, 36–49 (2016) 10

[33] Escaramís, G., Docampo, E., Rabionet, R.: A decade of structural variants: description, history and methods to detect structural variation. Briefings in functional genomics **14**(5), 305–314 (2015) 10

[34] Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., Kamatani, Y.: Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome biology **20**(1), 1–18 (2019) 10

[35] Jones, P.A., Baylin, S.B.: The epigenomics of cancer. Cell **128**(4), 683–692 (2007) 11

[36] Baylin, S.B., Esteller, M., Rountree, M.R., Bachman, K.E., Schuebel, K., Herman, J.G.: Aberrant patterns of dna methylation, chromatin formation and gene expression in cancer. Human molecular genetics **10**(7), 687–692 (2001) 11

[37] Diaz, A.K., Baker, S.J.: The genetic signatures of pediatric high-grade glioma: no longer a one-act play. In: Seminars in Radiation Oncology, vol. 24, pp. 240–247 (2014). Elsevier 11

[38] Jin, B., Robertson, K.D.: Dna methyltransferases, dna damage repair, and cancer. Epigenetic alterations in oncogenesis, 3–29 (2013) 11

[39] Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., *et al.*: The nih roadmap epigenomics mapping consortium. Nature biotechnology **28**(10), 1045–1048 (2010) 11

[40] Stunnenberg, H.G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., *et al.*: The international human epigenome consortium: a blueprint for scientific collaboration and discovery. Cell **167**(5), 1145–1149 (2016) 11

[41] Simpson, J.T., Workman, R.E., Zuzarte, P., David, M., Dursi, L., Timp, W.: Detecting dna cytosine methylation using nanopore sequencing. Nature methods **14**(4), 407–410 (2017) 11

[42] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**(1), 57–63 (2009) 12

[43] Heegaard, N.H., Schetter, A.J., Welsh, J.A., Yoneda, M., Bowman, E.D., Harris, C.C.: Circulating micro-rna expression profiles in early stage nonsmall cell lung cancer. International journal of cancer **130**(6), 1378–1386 (2012) 12

[44] Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., Lacroix, V., Aury, J.-M.: Transcriptome profiling of mouse samples using nanopore sequencing of cdna and rna molecules. Scientific reports **9**(1), 1–12 (2019) 12

[45] Aoubala, M., Murray-Zmijewski, F., Khoury, M.P., Fernandes, K., Perrier, S., Bernard, H., Prats, A.-C., Lane, D.P., Bourdon, J.-C.: p53 directly transactivates δ133p53 α, regulating cell fate outcome in response to dna damage. Cell Death & Differentiation **18**(2), 248–258 (2011) 13

[46] Mondal, A.M., Horikawa, I., Pine, S.R., Fujita, K., Morgan, K.M., Vera, E., Mazur, S.J., Appella, E., Vojtesek, B., Blasco, M.A., *et al.*: p53 isoforms regulate aging-and tumor-associated replicative senescence in t lymphocytes. The Journal of clinical investigation **123**(12), 5247–5257 (2013) 13

[47] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols **7**(3), 562 (2012) 13, 29, 31

[48] Li, B., Dewey, C.N.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics **12**(1), 1–16 (2011) 13

[49] Roberts, A., Pachter, L.: Streaming fragment assignment for real-time analysis of sequencing experiments. Nature methods **10**(1), 71–73 (2013) 13

[50] Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:1303.3997 (2013) 13, 30

[51] Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357 (2012) 13, 30

[52] Patro, R., Mount, S.M., Kingsford, C.: Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. Nature biotechnology **32**(5), 462–464 (2014) 13

[53] Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**(4), 417–419 (2017) 13, 29, 31

[54] Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic rna-seq quantification. Nature biotechnology **34**(5), 525–527 (2016) 13, 29, 31

[55] Zhang, C., Zhang, B., Lin, L.-L., Zhao, S.: Evaluation and comparison of computational tools for rna-seq isoform quantification. BMC genomics **18**(1), 1–11 (2017) 13

[56] Klebe, G.: Protein modeling and structure-based drug design. Springer (2013) 13

[57] Woolfrey, K.M., Dell'Acqua, M.L.: Coordination of protein phosphorylation and dephosphorylation in synaptic plasticity. Journal of Biological Chemistry **290**(48), 28604–28612 (2015) 14

[58] Consortium, U.: Uniprot: a hub for protein information. Nucleic acids research **43**(D1), 204–212 (2015) 14

[59] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., Sigrist, C.J.: The prosite database. Nucleic acids research **34**(suppl_1), 227–230 (2006) 14

[60] Rajendhran, J., Gunasekaran, P.: Human microbiomics. Indian journal of microbiology **50**(1), 109–112 (2010) 14

[61] Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D.B., Morgun, A., Shulzhenko, N.: Role of gut microbiota in type 2 diabetes pathophysiology. EBioMedicine **51**, 102590 (2020) 14

[62] Li, Q., Han, Y., Dy, A.B.C., Hagerman, R.J.: The gut microbiota and autism spectrum disorders. Frontiers in cellular neuroscience **11**, 120 (2017) 14

[63] Singer-Englar, T., Barlow, G., Mathur, R.: Obesity, diabetes, and the gut microbiome: an updated review. Expert review of gastroenterology & hepatology **13**(1), 3–15 (2019) 14

[64] Devaraj, S., Hemarajata, P., Versalovic, J.: The human gut microbiome and body metabolism: implications for obesity and diabetes. Clinical chemistry **59**(4), 617–628 (2013) 14

[65] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. Nature **449**(7164), 804–810 (2007) 14

[66] lita. proctor@ nih. gov Lita Proctor Jonathan LoTempio Aron Marquitz Phil Daschner Dan Xi Roberto Flores Liliana Brown Ryan Ranallo Padma Maruvada Karen Regan R. Dwayne Lunsford Michael Reddy Lis Caler, N.H.M.P.A.T.: A review of 10 years of human microbiome research activities at the us national institutes of health, fiscal years 2007-2016. Microbiome **7**, 1–19 (2019) 14

[67] Chin, L., Andersen, J.N., Futreal, P.A.: Cancer genomics: from discovery science to personalized medicine. Nature medicine **17**(3), 297 (2011) 14

[68] Simon, T., Hero, B., Schulte, J.H., Deubzer, H., Hundsdoerfer, P., von Schweinitz, D., Fuchs, J., Schmidt, M., Prasad, V., Krug, B., *et al.*: 2017 gpoh guidelines for diagnosis and treatment of patients with neuroblastic tumors. Klinische Pädiatrie **229**(03), 147–167 (2017) 15

[69] Huang, M., Weiss, W.A.: Neuroblastoma and mycn. Cold Spring Harbor perspectives in medicine **3**(10), 014415 (2013) 15, 39

[70] Dellaire, G., Berman, J.N., Arceci, R.J.: Cancer Genomics: from Bench to Personalized Medicine. Academic Press, ??? (2013) 16

[71] Costa, R.A., Seuánez, H.N.: Investigation of major genetic alterations in neuroblastoma. Molecular biology reports **45**(3), 287–295 (2018) 17

[72] Daver, N., Schlenk, R.F., Russell, N.H., Levis, M.J.: Targeting flt3 mutations in aml: review of current knowledge and evidence. Leukemia **33**(2), 299–312 (2019) 17

[73] Network, C.G.A.R.: Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. New England Journal of Medicine **368**(22), 2059–2074 (2013) 17

[74] Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F.R., Büchner, T., Dombret, H., Ebert, B.L., Fenaux, P., Larson, R.A., *et al.*: Diagnosis and management of aml in adults: 2017 eln recommendations from an international expert panel. Blood **129**(4), 424–447 (2017) 17

[75] Saultz, J.N., Garzon, R.: Acute myeloid leukemia: a concise review. Journal of clinical medicine **5**(3), 33 (2016) 17

[76] Waks, A.G., Winer, E.P.: Breast cancer treatment: a review. Jama **321**(3), 288–300 (2019) 17

[77] Sun, J., Meng, H., Yao, L., Lv, M., Bai, J., Zhang, J., Wang, L., Ouyang, T., Li, J., Wang, T., *et al.*: Germline mutations in cancer susceptibility genes in a large series of unselected breast cancer patients. Clinical Cancer Research **23**(20), 6113–6119 (2017) 18

[78] Abdel-Razeq, H., Al-Omari, A., Zahran, F., Arun, B.: Germline brca1/brca2 mutations among high risk breast cancer patients in jordan. BMC cancer **18**(1), 1–11 (2018) 18

[79] Loeb, S., Folkvaljon, Y., Curnyn, C., Robinson, D., Bratt, O., Stattin, P.: Uptake of active surveillance for very-low-risk prostate cancer in sweden. JAMA oncology **3**(10), 1393–1398 (2017) 18

[80] Goldenberg, S.L., Nir, G., Salcudean, S.E.: A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology **16**(7), 391–403 (2019) 19

[81] Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.-P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., *et al.*: Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. Nature genetics **44**(6), 685–689 (2012) 19

[82] Zhou, F., Gao, S., Han, D., Han, W., Chen, S., Patalano, S., Macoska, J.A., He, H.H., Cai, C.: Tmprss2-erg activates no-cgmp signaling in prostate cancer cells. Oncogene **38**(22), 4397–4411 (2019) 19

[83] Thierry-Mieg, D., Thierry-Mieg, J.: Aceview: a comprehensive cdna-supported gene and transcripts annotation. Genome biology **7**(1), 1–14 (2006) 23

[84] Srivastava, A., Sarkar, H., Gupta, N., Patro, R.: Rapmap: a rapid, sensitive and accurate tool for mapping rna-seq reads to transcriptomes. Bioinformatics **32**(12), 192–200 (2016) 25, 29, 31

[85] Frazee, A.C., Jaffe, A.E., Langmead, B., Leek, J.T.: Polyester: simulating rna-seq datasets with differential transcript expression. Bioinformatics **31**(17), 2778–2784 (2015) 27

[86] Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., *et al.*: Quantitative assessment of single-cell rna-sequencing methods. Nature methods **11**(1), 41 (2014) 27

[87] Su, Z., Mason, C.: Seqc/maqc-iii consortium a comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat. Biotechnol **32**(9), 903–914 (2014) 28

[88] Halazonetis, T.D., Gorgoulis, V.G., Bartek, J.: An oncogene-induced dna damage model for cancer development. science **319**(5868), 1352–1355 (2008) 29

[89] Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M., *et al.*: Functional genomic

landscape of acute myeloid leukaemia. Nature **562**(7728), 526–531 (2018) 29, 37

[90] Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology **31**(3), 213–219 (2013) 29

[91] Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K.: Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research **22**(3), 568–576 (2012) 29

[92] Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., *et al.*: Svaba: genome-wide detection of structural variants and indels by local assembly. Genome research **28**(4), 581–591 (2018) 38

[93] Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., Papenfuss, A.T.: Gridss: sensitive and specific genomic rearrangement detection using positional de bruijn graph assembly. Genome research **27**(12), 2050–2060 (2017) 38

[94] Layer, R.M., Chiang, C., Quinlan, A.R., Hall, I.M.: Lumpy: a probabilistic framework for structural variant discovery. Genome biology **15**(6), 1–19 (2014) 38

[95] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al.: The complete sequence of a human genome. bioRxiv (2021) 43