Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

COMPUTATIONAL METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA

by
Wei Wang

Recently, next generation sequencing (NGS) technology has emerged as a powerful approach and dramatically transformed biomedical research in an unprecedented scale. NGS is expected to replace the traditional hybridization-based microarray technology because of its affordable cost and high digital resolution. Although NGS has significantly extended the ability to study the human genome and to better understand the biology of genomes, the new technology has required profound changes to the data analysis. There is a substantial need for computational methods that allow a convenient analysis of these overwhelmingly high-throughput data sets and address an increasing number of compelling biological questions which are now approachable by NGS technology.

This dissertation focuses on the development of computational methods for NGS data analyses. First, two methods are developed and implemented for detecting variants in analysis of individual or pooled DNA sequencing data. SNVer formulates variant calling as a hypothesis testing problem and employs a binomial-binomial model to test the significance of observed allele frequency by taking account of sequencing error. SNVerGUI is a GUI-based desktop tool that is built upon the SNVer model to facilitate the main users of NGS data, such as biologists, geneticists and clinicians who often lack of the programming expertise. Second, collapsing singletons strategy is explored for associating rare variants in a DNA sequencing study. Specifically, a gene-based genome-wide scan based on singleton collapsing is performed to analyze a whole genome

sequencing data set, suggesting that collapsing singletons may boost signals for association studies of rare variants in sequencing study. Third, two approaches are proposed to address the 3'UTR switching problem. PolyASeeker is a novel bioinformatics pipeline for identifying polyadenylation cleavage sites from RNA sequencing data, which helps to enhance the knowledge of alternative polyadenylation mechanisms and their roles in gene regulation. A change-point model based on a likelihood ratio test is also proposed to solve such problem in analysis of RNA sequencing data. To date, this is the first method for detecting 3'UTR switching without relying on any prior knowledge of polyadenylation cleavage sites.

COMPUTATIONAL METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA

by
Wei Wang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2014

Copyright © 2014 by Wei Wang

ALL RIGHTS RESERVED

APPROVAL PAGE

COMPUTATIONAL METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA

Wei Wang

Dr. Zhi Wei, Dissertation Advisor Associate Professor of Computer Science, NJIT	Date
Dr. Jason T.L. Wang, Committee Member Professor of Bioinformatics and Computer Science, NJIT	Date
Dr. Usman Roshan, Committee Member Associate Professor of Computer Science, NJIT	Date
Dr. Vincent Oria, Committee Member Associate Professor of Computer Science, NJIT	Date
Dr. Zhigen Zhao, Committee Member Assistant Professor of Statistics, Temple University	Date

BIOGRAPHICAL SKETCH

Author: Wei Wang

Degree: Doctor of Philosophy

Date: May 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2014

- Bachelor of Engineering in Computer Science and Technology,
  Nankai University, Tianjin, P. R. China, 2008

Major: Computer Science

Publications:

Wang, W., *et al*. A change-point model for identifying 3'UTR switching by next generation RNA sequencing. *Bioinformatics*, in press.

Wang, W., *et al*. Collapsing singletons may boost signal for associating rare variants in sequencing study. *BMC Proceedings*, in press.

Zhang, Z., *et al*. Dysregulation of synaptogenesis genes antecedes motor neuron pathology in spinal muscular atrophy. *Proceedings of the National Academy of Sciences of the United States of America* 2013;110(48):19348-19353.

Zhao, Z., *et al*. An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *The Annals of Applied Statistics* 2013;7(4): 2229-2248.

Younis, I., *et al*. Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. *Elife* 2013;2: e00780.

Zaidi, S., *et al*. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 2013;498(7453):220-223.

Wei, Z., *et al*. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *American journal of human genetics* 2013;92(6):1008-1012.

O'Rawe, J., *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013;5(3):28.

Wang, W., *et al*. SNVerGUI: a desktop tool for variant analysis of next-generation sequencing data. *J Med Genet* 2012;49(12)753-5.

Lyon, G.J., *et al*. Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications. *Discov Med* 2011;12(62):41-55.

Wei, Z., *et al*. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* 2011;39(19):e132.

Wang, W., *et al*. Simultaneous set-wise testing under dependence, with applications to genome-wide association studies. *Statistics and Its Interface* 2010;3(4):501-512.

Dedicated to my beloved family:
(谨此献给我挚爱的家人:)

Wang, Shihe (王世和), father (父亲)
Liu, Meixuan (刘美宣), mother (母亲)
Tan, Ludan (谭露丹), wife (妻子)

"Where there's a will, there's a way!"
"有志者，事竟成!"

ACKNOWLEDGMENT

First, I would like to take this opportunity to express my heartfelt appreciation to my dissertation advisor, Dr. Zhi Wei, for his invaluable advice, infinite patience, technical guidance, and his confidence in me which enabled me to bring this dissertation to its culmination. Over the past few years, Dr. Wei has been a constant source of encouragement and a valued mentor to me. I will always be indebted to Dr. Wei for encouraging and supporting me to present our research achievements to reputed journals, which significantly improved my confidence and gave me great opportunities to exchange insights with leading researchers in the world.

Second, I am extremely grateful to Dr. Usman Roshan, Dr. Jason T.L. Wang, Dr. Vincent Oria and Dr. Zhigen Zhao for serving on my committee. In addition, I would like to extent special thanks to my collaborators, Dr. Hakon Hakonarson from The Center for Applied Genomics at The Children's Hospital of Philadelphia, Dr. Gideon Dreyfuss and Dr. Hongzhe Li from University of Pennsylvania, Dr. Lyon Gholson from Cold Spring Habor Laboratory, Dr. Pingzhao Hu from The Hospital for Sick Children, Canada. They have provided me with academic advice inside and outside my research field. This dissertation would not have been possible without their invaluable guidance and generous help. I would also like to thank my fellow graduate students for their assistance and support, in particular Jichao Sun and Xiguo Ma. Their friendship and wit have been tremendous source of emotional support.

Third, I heartily appreciate my family for being my emotional anchor throughout the roughness of Ph.D. study. I thank my parents, Mr. Shihe Wang and Mrs. Meixuan Liu, for their faith in me and allowing me to be as ambitious as I wanted to pursue my

Ph.D. degree abroad. I also would like to thank my lovely wife, Ludan Tan, who has supported me for all the time that I gained so much strength and ability to tackle challenges head-on.

Last, but not the least, I would like to thank the people who have stayed by my side for the past five years, giving me advice on non-research related matters and keeping me sane. In particular, I would like to thank my friends Jichao Sun, Xiguo Ma, Shuangyi Zhang and Bing Li.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

The past few years have seen a dramatic development in sequencing technology, which has made the per-base cost of DNA sequencing plummet by ~100,000-fold over the past decade, far outpacing Moore's law of technological advance in the semiconductor industry (Lander, 2011). Because of affordable cost and high digital resolution, the new or "next generation" sequencing (NGS) technology is replacing the traditional hybridization-based microarray technology. And this new engine has been 'turbo-charged' by several orders of magnitude compared to its predecessor (Figure 1.1) since the basic mechanisms for data generation have been changed radically, producing far more sequence reads per instrument run and at a significantly lower expense (Mardis, 2011).

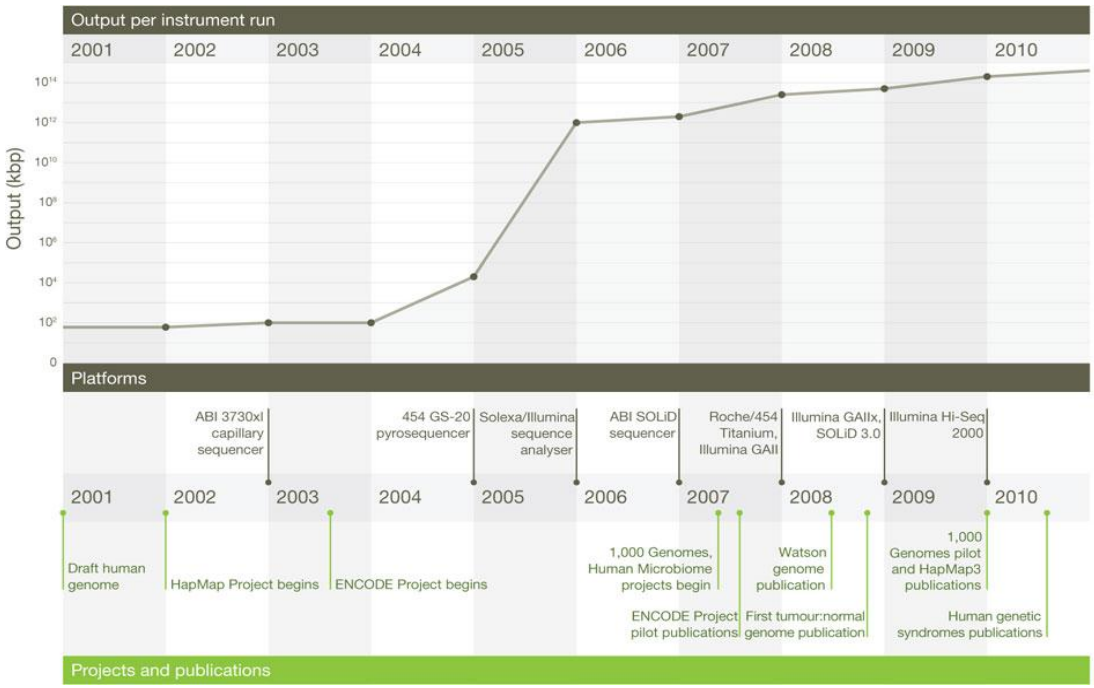

Figure 1.1 Changes in instrument capacity over the past decade.
Source: (Mardis, 2011)

Table 1.1 Sequencing Platform Comparison

	Roche/454	**Life Technologies SOLiD**	**Illumina Hi-Seq 2000**	**Pacific Biosciences RS**
Library amplification method	emPCR on bead surface	emPCR on bead surface	Enzymatic amplification on glass surface	NA (single molecule detection)
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase-mediated incorporation of end-blocked fluorescent nucleotides	Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3′ end of oligonucleotide	Chemical cleavage of fluorescent dye and 3′ blocking group	NA (fluorescent dyes are removed as part of PPi release on nucleotide incorporation)
Error model	Substitution errors rare, insertion/deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/deletion errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp	>1,000 bp

Source: (Mardis, 2011)

Several platforms have been developed using the so called “massively parallel” sequencing technology. Although each instrument is distinctly different in its specifics, as showed in Table 1.1 (Mardis, 2011), all massively parallel devices share certain attributes: First, the initial preparatory steps are fewer and simpler to perform than for Sanger sequencing. Instead of a bacterial cloning step followed by DNA isolation, massively

parallel sequencing begins with the production of a library formed by ligating platform-specific synthetic DNAs (adapters) onto the ends of the fragment population to be sequenced. Second, all platforms require the library fragments to be amplified on a solid surface (either a glass slide or a microbead) by a polymerase-mediated reaction that produces many copies of each single library fragment. Amplification is needed so that the ensuing sequencing reactions produce sufficient signal for detection by the instrument's optical system. However, this step also provides a source of sequencing error that is perpetuated through the downstream processes, because polymerases are never 100% accurate. Third, these instruments perform sequencing reactions as an orchestrated series of repeating steps that are performed and detected automatically. The specifics of the DNA sequencing reaction are different for each platform, emphasizing the amazing range of innovation in chemistry, molecular biology and engineering required to produce sequence information from hundreds of thousands to hundreds of millions of DNA molecules simultaneously. For example, the Roche/454 instrument detects each polymerase-catalysed nucleotide incorporation event by a downstream series of reactions that produce light ('pyrosequencing'), initiated by the pyrophosphate molecules released on nucleotide incorporation. The Life Technologies SOLiD uses a unique DNA ligase-mediated process that, through multiple rounds of template-directed ligation, sequences each nucleotide twice. The Illumina sequencer incorporates fluorescently labelled nucleotides that are chemically blocked such that only one nucleotide incorporation event occurs per fragment population per sequencing cycle. Regardless of the details, massively parallel sequencing reactions are distinguished by the fact that they occur in a nucleotide-by-nucleotide stepwise fashion, rather than by discrete separation and

detection (in a 96-at-a-time fashion) of already produced Sanger sequencing reaction products on a capillary instrument. The fourth shared feature of these systems is the ability to obtain sequence information from both the ends of the DNA fragments comprising the sequencing library. Depending on the instrument system and the library construction approach used, one can either sequence at both ends of linear fragments ('paired end sequencing') or from both ends of previously circularized fragments ('mate pair sequencing').

There are many applications have been conducted by taking the unprecedented advantages of NGS. According to (Lander, 2011), an early application of massively parallel sequencing was to create 'epigenomic maps', showing the locations of specific DNA modifications, chromatin modifications and protein-binding events across the human genome. Chromatin modification and protein binding can be mapped by chromatin immunoprecipitation-sequencing (ChIP-Seq) (Barski, et al., 2007; Mikkelsen, et al., 2007), and the sites of DNA methylation can be found by sequencing DNA in which the methylated cytosines have been chemically modified (Methyl-Seq) (Meissner, et al., 2008). Next, as the technology has improved, the focus has turned to re-sequencing human samples to study inherited variation or somatic mutations. One can re-sequence the whole genome (Bentley, et al., 2008) to varying degrees of coverage or use hybridization-capture techniques (Okou, et al., 2007) to re-sequence a targeted subset, such as the protein-coding sequences (referred to as the 'exome'). Furthermore, sequencing is also being extensively applied to RNA transcripts (RNA-Seq), to count their abundance, identify novel splice forms or spot mutations (Mortazavi, et al., 2008). These applications have great impact on genomics research and allow to not only understand the structure of genomes and the

biology of genomes, but also further understand the biology of disease, advance the science of medicine in the near future, in addition to improve the effectiveness of healthcare ultimately (Green and Guyer, 2011), which is represented in Figure 1.2.

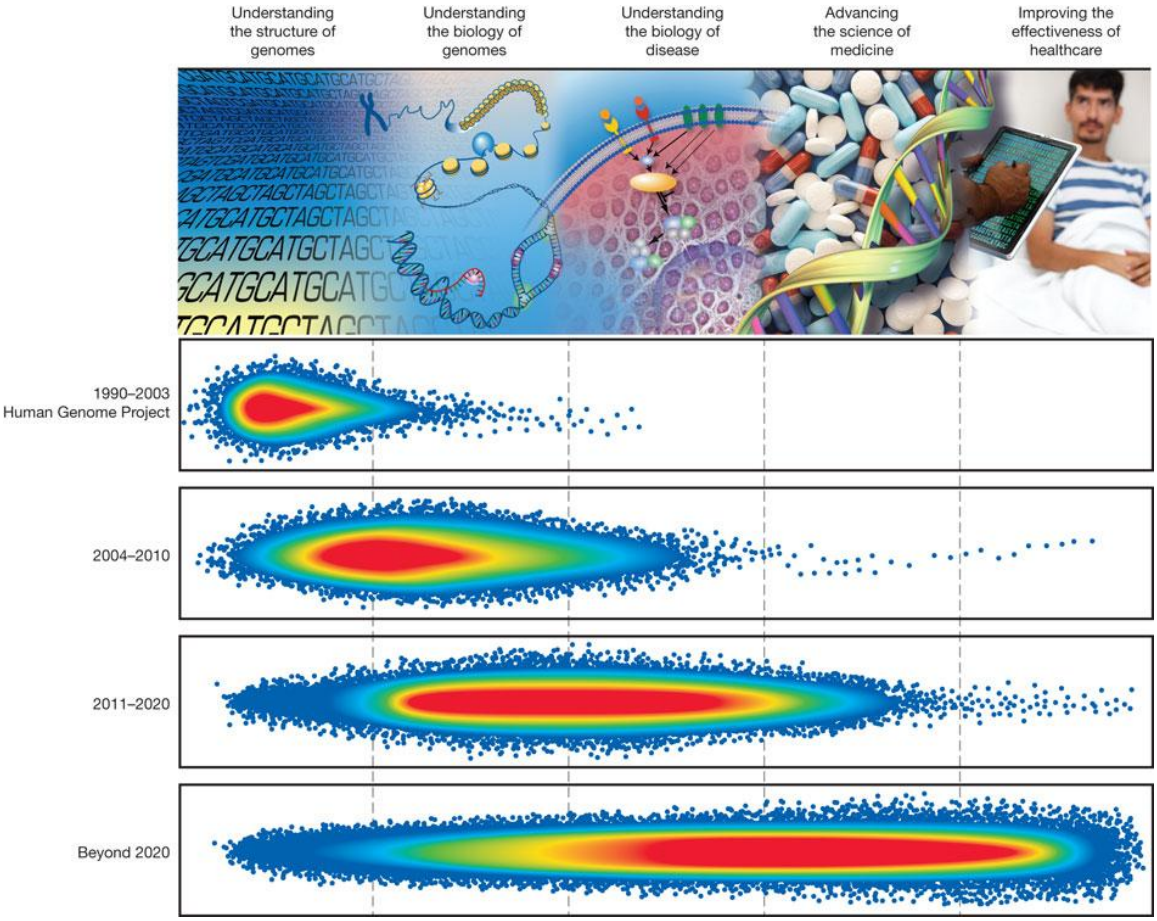

Figure 1.2 Schematic representations of accomplishments across five domains of genomics research.
Source: (Green and Guyer, 2011)

Although next generation sequencing have significantly extended the ability to study the human genome and to better understand the biology of genomes, the new technology has required profound changes to the data analysis pipelines than those of previous technology. For example, how to handle the huge amount of data, how to deal with the error profiles of the sequencing platforms and how to model the significant decrease in the read length become more challenging. These challenges have resulted in a

revitalization of the bioinformatics-based pursuit for the analysis of next generation sequencing data at all levels, in order to address an increasing number of compelling biological questions that are now approachable by NGS technology.

This dissertation focuses on the development of computational methods for next generation sequencing analysis. First, two methods are developed and implemented for detecting variants in analysis of individual or pooled NGS data. SNVer formulates variant calling as a hypothesis testing problem and employs a binomial-binomial model to test the significance of observed allele frequency by taking account of sequencing error in NGS. SNVerGUI is a GUI-based desktop tool that is built upon the SNVer model in order to facilitate the main users of NGS data, such as biologists, geneticists and clinicians who often lack of the programming expertise. Second, collapsing singletons strategy is explored for associating rare variants in a NGS study. Specifically, a gene-based genome-wide scan based on singleton collapsing is performed to analyze a whole genome sequencing data, suggesting that collapsing singletons may boost signals for association studies in sequencing data. Third, two approaches are proposed to solve the problem of identification of 3'UTR length changes in RNA-Seq data. On one hand, PolyASeeker, is a novel bioinformatics pipeline for identifying polyadenylation cleavage sites from RNA-Seq data. Followed by the conventional method using Fisher's exact test, the 3'UTR switching can be detected. On the other hand, a change-point model based on a likelihood ratio test has been proposed, which is the first available method to allow alternative cleavage and polyadenylation (APA) analysis without relying on any PolyA information and hence is more powerful and accurate in the APA studies than the traditional method.

This dissertation is organized in the following manner. Chapter 2 discusses the background and related work of variant calling and association study in DNA-Seq analysis, together with the identification of 3'UTR length changes from RNA-Seq data. Chapters 3 and 4 introduce two proposed methods, SNVer and SNVerGUI, for variant detection in analysis of individual or pooled sequencing data. Chapter 5 demonstrates that collapsing singletons may boost signal for associating rare variants in sequencing study. Chapters 6 and 7 propose two methods, PolyASeeker and a change-point model-based approach, for identifying 3'UTR length changes in RNA-Seq studies. Finally, Chapter 8 summarizes the contribution of this dissertation and discusses future directions for research.

CHAPTER 2

BACKGROUD

2.1 Variant Detection by DNA-Seq

For genetics studies, NGS holds the promise to revolutionize genome-wide association studies (GWAS). The recently completed phase of GWAS mainly addresses common SNPs with minor allele frequency > 5%, based upon the common disease/common variant (CD/CV) hypothesis (Manolio, et al., 2009). However, the identified common variants explain only a small proportion of heritability (Hindorff, et al., 2009). Rare variants therefore have been hypothesized to account for the missing heritability (Dickson, et al., 2010; Wang, et al., 2010). To identify rare variants, a direct and more powerful approach is to sequence a large number of individuals (Li and Leal, 2009). This line of thought also implicitly motivates the recent 1000 Genomes Project, which will sequence the genomes of 1,200 individuals of various ethnicities by NGS (Hayden, 2008). It is expected to extend the catalogue of known human variants down to a frequency near 1%.

Although the cost of whole-genome or exome sequencing of all enrolled subjects is prohibitively high now, such studies will eventually be carried out in a manner similar to GWAS with very large sample sizes (Cirulli and Goldstein, 2010). While the cost is being brought down to as low as $1000 for sequencing a whole genome (Service, 2006), in the interim, a cost-effective strategy has to be taken in order to take the full advantage of NGS. Such issues with cost and labor are not new as similar problems were confronted in the early expensive stage of GWAS and were circumvented by focusing on small candidate regions and the use of pooling of genomic DNAs (Norton, et al., 2004; Sham, et al., 2002). Borrowing the same idea, many targeted re-sequencing applications utilizing pooling have

been seen in the past few years (Calvo, et al., 2010; Momozawa, et al., 2011; Nejentsev, et al., 2009; Out, et al., 2009).

The first-step analysis of NGS data for genetics study is often to identify genomic variants among sequenced samples. Quite a few SNP calling tools have been implemented to identify SNPs from sequencing of individual genomes. SNP calling is a relatively straightforward problem in analysis of sequencing data of individual genomes, because the frequency of a candidate allele can be only 0 (non-variant), 0.5 (heterozygous) or 1 (alternate homozygous) for a diploid genome. Despite (high) sequencing error of NGS, a reliable call can be easily made given a high depth of coverage, say 20X to 30X. Consequently, statistical models for SNP calling have been developed and integrated as one simple functional module in many NGS short reads analysis tools such as SAMtools (Li, et al., 2009) , MAQ (Li, et al., 2008) and VarScan (Koboldt, et al., 2009). SAMtools and MAQ use a Bayesian statistical model to compute the posterior probabilities of the three possible genotypes. Specifically, for the likelihood part, they employ a binomial distribution to characterize sampling of the two haplotypes, and the prior probability, like other Bayesian approaches, is pre-specified. SAMtools and MAQ empirically set the prior probability of observing a heterozygote to be 0.001 for the discovery of new SNPs, and 0.2 for inferring genotypes at known SNP sites. Such Bayesian approaches may not be ideal for multiplicity control because of the subjectivity of assigning the prior probability. VarScan implements a heuristic/statistical method. For each candidate site, it applies several heuristic filters such as having a minimum number of supporting reads and allele frequency reaching a minimum threshold. It also conducts a Fisher's exact test for testing the deviation of the read counts supporting variant alleles from being generated because of

sequencing error. Those heuristic filters overlap with the Fisher's exact test in terms of reducing false positives. When not systematically considered, they may distort the statistics distribution under null and thus void the resultant p values for multiplicity control.

Identifying SNPs from pooled NGS data is more challenging in that pooled DNA are sampled from a number of individuals, which consequently will give rise to variant allele frequencies other than simply 0, 0.5 or 1. Driven by the need for analysis of increasing amount of pooled NGS data, several programs/methods for the detection of variants from the pooled data have been developed. SNPSeeker employs the large deviation theory for SNP detection (Druley, et al., 2009). It compares observed allele frequencies against the distribution of sequencing errors as measured by the Kullback Leibler (KL) distance (Kullback and Leibler, 1951). One limitation of this approach is that its error model has to be estimated from negative control data. SNPSeeker was recently extended to SPLINTER with two main improvements (Vallania, et al., 2010). First, it is capable of detecting rare short indels. Second, it provides a good cutoff after ranking all candidate variants to balance power and type I error rate, which, however, requires an additional positive control data. CRISP (Bansal, 2010) models the number of reads of the reference and alternate alleles at a particular position across all pools as a contingency table, which is then tested by the Fisher's exact test. Its working hypothesis is that, due to rareness, presence of rare variants in all pools will be sporadic and then results in an excess of reads with the alternate allele as compared with the other pools, which is expected to be captured by the Fisher's exact test. CRISP then conducts a complementary test for the overabundance of alternate alleles within each pool against the sequencing error rate.

Although it is shown that CRISP outperforms SNPSeeker, MAQ, and VarScan (Bansal, 2010), it has the following limitations. First, its working hypothesis does not hold well for common variants. When the minor allele frequency is large and/or the number of individuals in each pool is large, sporadic presence will disappear and result in no prominent excess of reads that can be captured by the Fisher's exact test. Second, their method is not applicable for single pool data. Third, rareness and overabundance of alternate alleles are related but are captured separately using two different models, which may not be an efficient approach. In addition, these two separate tests make it hard to obtain an overall multiplicity control. Finally, its computational efficiency makes scalability an issue and may prevent its application in analysis of whole-exome or genome sequencing data. The main bottleneck comes from computing the p-value of a large number of contingency tables in the Fisher's exact test.

In addition to the above direct SNP calling programs, there are also other relevant studies for analysis of pooled NGS data, including estimating allele frequencies from pooled sequencing (Ingman and Gyllensten, 2009), evaluating the ability to detect rare SNPs (Out, et al., 2009), and investigating the power of variant detection in pooled DNA for NGS and the optimal pooling designs (Lee, et al., 2011), among others.

2.2 Association Studies for Rare Variants

The limitations of genome-wide association (GWA) studies that focus on the phenotypic influence of common genetic variants have motivated human geneticists to consider the contribution of rare variants to phenotypic expression. Recent advances in next-generation sequencing (NGS) technology have made it technically and economically feasible to capture the full spectrum of genomic variation. NGS provides a powerful tool for

systematic exploration of common and rare variants in the entire genome, even in large population-scale studies (1000 Genomes Project Consortium, 2010). However, pinpointing causal variants remains a major challenge, particularly for associating rare variants with complex traits (Cooper and Shendure, 2011).

An illustration of scenarios in DNA-Seq studies in Figure 2.1, where the blue lines indicate genomic regions and red boxes indicate variants. A) Variants at a single locus with common alleles are more frequent in cases then controls. B) Multiple rare variations contribute to the phenotype such that the collective frequency of these variations is greater in cases. This would create a greater diversity of haplotypes or DNA sequences among the cases. C) Multiple rare variations contribute to the phenotype but act in a synergistic fashion, such that cases are likely to have more similar DNA sequences compared to controls. D) Multiple rare variations contribute to a phenotype but the variations contributing to the phenotype reside in specific genomic regions. This situation would create greater sequence diversity among the cases, as in part b, but only in the relevant genomic regions.

There is a substantial need for computational methods that allow for efficient association analysis of rare variants. Several powerful approaches tailored for rare-variant association studies have been proposed recently (Daye, et al., 2012; Li and Leal, 2008; Neale, et al., 2011; Wu, et al., 2011). Although these tests offer the powerful tool to investigate rare variants in the entire genome, resulted from the increasing availability of high-throughput sequencing technologies, these methods may not be sufficient for their success as appropriate analytical methods are also needed (Bansal, et al., 2010).

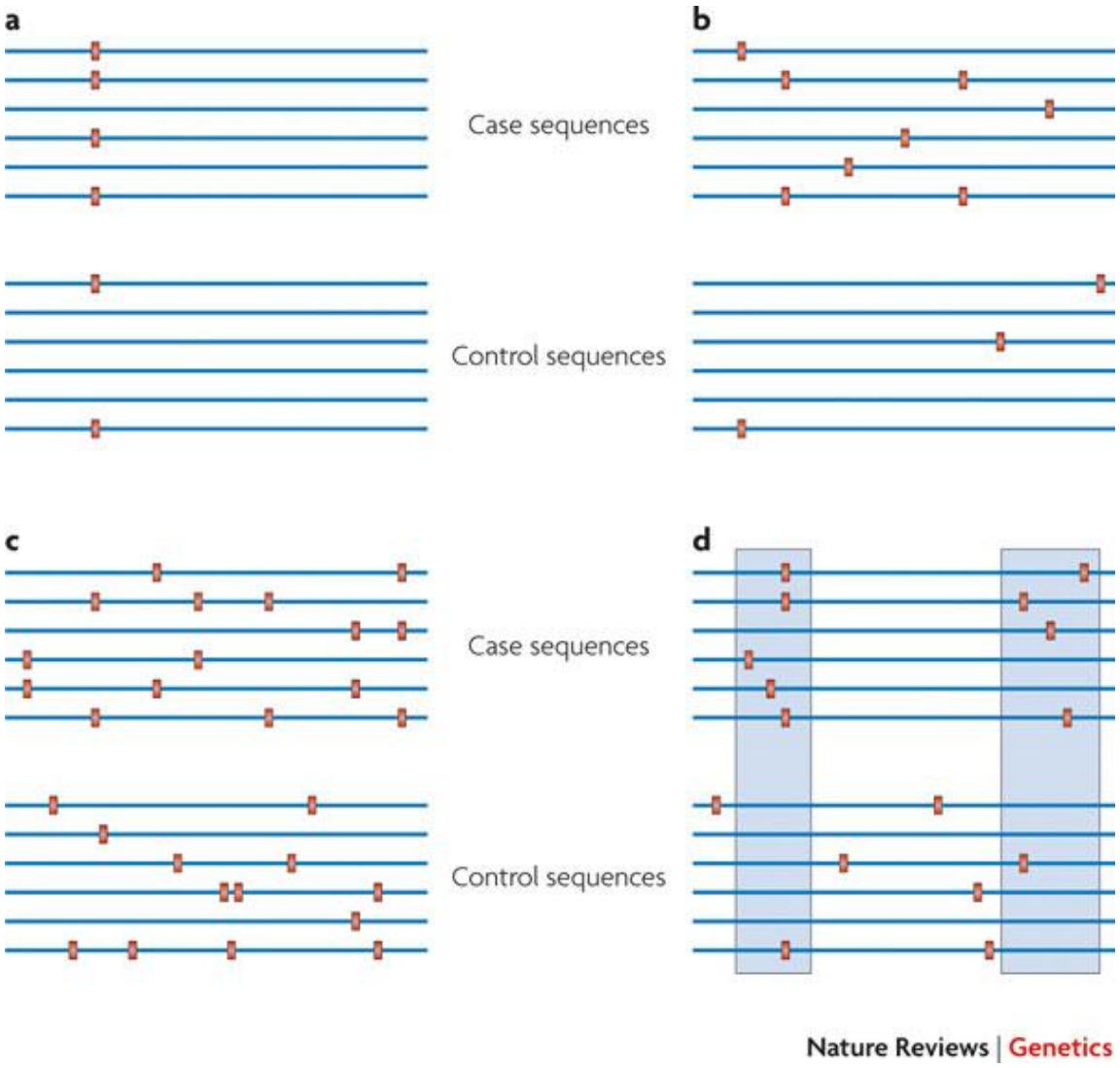

Figure 2.1 Scenarios in which DNA sequence variants distinguish cases and controls. Source: (Bansal, et al., 2010)

2.3 Identification of 3'UTR Switching from RNA-Seq

For transcriptome study, the introduction of RNA-Seq technology along with new analytic methods makes it possible to address an increasing number of compelling biological questions that may not be possible using microarray technology. In particular, alternative RNA splicing and processing, common phenomena in eukaryotes, play so critical a role in

gene function regulation that they receive much attention in RNA-Seq analysis (Keren, et al., 2010) and motivate quite a few methodological developments. For example, MISO employs a probabilistic mixture model to quantify alternative splicing and processing, then test the equality of transcript isoform ratios between samples (Katz, et al., 2010); MATS by using a Bayesian statistical framework offers the flexibility to identify differential alternative splicing and processing events that match a given user-defined pattern (Shen, et al., 2012); DEXSeq employs generalized linear models to test for differential usage of exons and provides reliable control of false discoveries by taking biological variation into account (Anders, et al., 2012); other developments include (Griffith, et al., 2010; Rogers, et al., 2012; Trapnell, et al., 2013). Despite the success of these methods, detecting 3' untranslated regions (3'UTR) switching remains challenging. Very few, if any, methods and tools are available for directly analyzing this special alternative RNA processing event.

3' end processing plays a crucial role in eukaryotic mRNA maturation (Colgan and Manley, 1997). Through cis elements in the 3' translated regions (3'UTR) of mRNAs, post-transcriptional gene regulation frequently occurs and determines the stability, localization and translation of mRNA (Martin and Ephrussi, 2009; Moore, 2005). These roles are mediated by interactions with RNA-binding proteins (RBPs) and microRNAs (miRNAs) (Licatalosi and Darnell, 2010). Over half of mammalian genes contain alternative cleavage and polyadenylation sites, which lead to various mRNA isoforms differing in their 3'UTRs (Zhang, et al., 2005). The analysis of alternative cleavage and polyadenylation in 3'UTR, including shortening and lengthening, has recently been appreciated as a global phenomenon under different cell conditions (Flavell, et al., 2008; Ji, et al., 2009; Mayr and Bartel, 2009; Sandberg, et al., 2008) and different species

(Sherstnev, et al., 2012; Smibert, et al., 2012; Ulitsky, et al., 2012). And this phenomenon has received particular attention in cancer studies (Fu, et al., 2011; Lembo, et al., 2012; Lin, et al., 2012; Mayr and Bartel, 2009).

In contrast with the increasingly recognized importance of APA, computational methods and tools for the APA analysis using RNA-Seq are underdeveloped. In unraveling APA regulation, Ji and colleagues scored relative expressions by taking the ratio of short reads density in extended and common regions, as defined by distal and proximal PolyA sites, respectively (Ji, et al., 2011). A higher score, therefore, indicated higher abundance of long 3'UTR isoform. A similar approach was taken in a recent tandem 3'UTR analysis, where the statistical significance was assessed by Fisher's exact test for the switch-score under different conditions (Wang, et al., 2008). The same group further improved the approach and implemented a new computational tool, MISO (Katz, et al., 2010). Specifically, tandem 3'UTR was treated as special alternative processing, and thus the quantification of expression level for each isoform can be estimated by computing PSI (Percent Spliced Isoform). These existing methods, however, have one critical drawback; namely, they rely on prior knowledge of annotated PolyA sites. For example, MISO constructs 3'UTR isoform based on PolyA sites information collected from the PolyA site database (Lee, et al., 2007; Zhang, et al., 2005). It is noted that the PolyA sites from the current database are computationally inferred from cDNA/EST sequences. It is far from complete and may also contain false positives. Therefore, these approaches that depend on PolyA information may not be precise or powerful due to incomplete information of all potential cleavage sites on 3'UTR.

This motivates the analysis of 3'UTR switching without relying on any PolyA annotations, one major limitation of existing methods. In addition, existing tools also have the same limitations, such as not capable of handling sample replicates, not supporting multiple isoforms and no confidence interval estimates for the change-point. These limitations warrant development of new bioinformatics methods.

CHAPTER 3

SNVER: A STATISTICAL TOOL FOR VARIANT CALLING

3.1 Introduction

This chapter proposes a statistical tool, SNVer (Single Nucleotide Variant caller/seeker), for calling common and rare variants in analysis of pooled or individual next-generation sequencing (NGS) data. It formulates variant calling as a hypothesis testing problem and employs a binomial-binomial model to test the significance of observed allele frequency against sequencing error. SNVer reports one single overall p-value for evaluating the significance of a candidate locus being a variant, based on which multiplicity control can be obtained. This is particularly desirable because tens of thousands loci are simultaneously examined in typical NGS experiments. Each user can choose the false positive error rate threshold he or she considers appropriate, instead of just the dichotomous decisions of whether to "accept or reject the candidates" provided by most existing methods. Both simulated data and real data demonstrate the superior performance of the program in comparison with existing methods. SNVer runs very fast and can complete testing 300K loci within an hour. This excellent scalability makes it feasible for analysis of whole-exome sequencing data, or even whole-genome sequencing data using high performance computing cluster.

3.2 Methods

3.2.1 Statistical Models for Single Pool Data

For a genomic locus, let θ be its minor allele frequency (MAF) in a population. If θ is larger than a threshold θ_0 ($\theta > \theta_0$), then it is a single nucleotide polymorphism (SNP). Suppose that sample *N* individuals (haploids) from this population for pooled sequencing. It can be assumed that the number of individuals (*n*) carrying the minor allele follows a binomial distribution *b(N, θ),* namely,

$$n \sim b(N, \theta)$$

with

$$\text{Prob}(n; \theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$$

Re-sequence this genomic region and suppose that *K* short reads cover this locus. If no sequencing error, given n individuals carrying the minor allele, the number of minor alleles *X* that observed from the *K* short sequence reads follows also a binomial distribution *b(K, n/N)*, namely,

$$X \sim b(K, n/N)$$

with

$$\text{Prob}(X|n) = \binom{K}{X} (\frac{n}{N})^X \left(1 - \frac{n}{N}\right)^{K-X}$$

Now assume sequencing error rate to be ε, under which the minor allele will be flipped to one of the other three alternate alleles, and vice versa. So the observed *X* follows a binomial distribution $b\left(K, \frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)$, namely,

$$X \sim b\left(K, \frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)$$

with

$$\text{Prob}(X|n) = \binom{K}{X}\left(\frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)^{X}\left(1 - \left(\frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)\right)^{K-X}.$$

Since *n* is not observable, sum it out and obtain the statistical model for *X* as

$$\begin{aligned}\text{Prob}(X;\theta) &= \sum_{n=0}^{N} \text{Prob}(X|n)\text{Prob}(n;\theta) \\ &= \sum_{n=0}^{N} \binom{K}{X}\left(\frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)^{X}\left(1 - \left(\frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)\right)^{K-X} \\ &* \binom{K}{X}\left(\frac{n}{N}\right)^{X}\left(1-\frac{n}{N}\right)^{K-X}.\end{aligned}$$

Now consider the hypothesis test of whether this locus is a (rare) variant ($\theta > \theta_0$)

$$H_0: \theta \le \theta_0 \; versus \; H_1: \theta > \theta_0$$

Its significance *p* value will be

$$p = \text{Prob}(X \ge x; \theta = \theta_0) = 1 - \text{Prob}(X < x;\ \theta = \theta_0)$$

3.2.2 Partial Conjunction Test for Multiple Pool Data

The above statistical model is for testing a locus in one single pool data. If M pools, the test is performed in each pool separately. Therefore a set of *M* hypotheses for each candidate variant can be obtained. The problem of making a variant call at one specific locus involves the simultaneous testing of hypotheses at the set level. Typical questions considered in the multiple-testing framework include: (i) Are all *M* hypotheses in the set true? (ii) Are all *M* hypotheses in the set false? (iii) Are at least *u* out of *M* hypotheses in the set false? These questions are referred to as conjunction test, disjunction test and partial conjunction test, respectively (Benjamini and Heller, 2008). Testing whether a locus is a variant based on multiple pool data is equivalent to the partial conjunction test that at least *u* =1 out of the *M* hypotheses for that locus is false. Let $P_{(1)},P_{(2),\dots,}P_{(M)}$ be the ordered *p*-values obtained from each single pool test. Following (Benjamini and Heller, 2008), the Simes method is employed to calculate the pooled p-value for the partial conjunction test as

$$p^{1/M} = min\left\{\frac{M}{j}P_{(j)}, j = 1, \dots, M\right\}$$

If the set of *M* null p-values at the tested locus are independent, Benjamini and Heller show that $p^{1/M}$ is a valid p-value for testing the partial conjunction null (Benjamini and Heller, 2008). The Benjamini Hochberg (BH) procedure (Benjamini and Hochberg, 1995) and other multiple-test adjustments can then be applied to the pooled Simes' p-values for multiplicity control when testing a large number of loci. It has been shown that this Simes-BH procedure controls the false discovery rate (FDR) at the pre-specified nominal level (Benjamini and Heller, 2008).

3.3 Data Sets

3.3.1 Simulated Data

This section simulates synthetic data to investigate the numerical performances of the proposed approach. For the single pool scenario, a total of 10,000 datasets are generated under each combination of several conditions:

- Sequencing coverage: low (10X) and high (30X)
- Sequencing error: low (0.01) and high (0.05)
- Minor allele frequency (MAF): rare variants with $\theta \sim U(0.001, 0.01)$, less common variants with $\theta \sim U(0.01, 0.05)$, and very common variants $\theta \sim U(0.05, 0.5)$
- The number of sequenced individuals from low to high with N = 10, 20, 50, 100, 200, 500, 1000, 1500, 2000

For each MAF setting $\theta \sim U(\theta_{min}, \theta_{max})$, the power of the proposed approach is computed for detecting variants by testing the null hypothesis $H_0: \theta < \theta_{min}$. Meanwhile, it is demonstrated that type I error is controlled at the nominal level by the proposed test, by simulating $\theta \sim U(0, \theta_{min})$ and evaluating how likely the same null hypothesis $H_0: \theta < \theta_{min}$ will be rejected by mistake. For both power and type I error evaluations, a variant is called at the nominal level 0.05.

For the multiple-pool scenario, the simulation follows the above single pool simulation settings except that simulates five pools with the same number of individuals in each pool and the total N = 10, 20, 50, 100, 200, 500, 1000, 1500, 2000.

3.3.2 Real Data

This section assesses the performance of the proposed method in analysis of two pooled and one individual real NGS datasets as summarized in Table 3.1. The first one is an in-house Autism dataset generated using ABI SOLiD platform from sequencing three

genomic regions, denoted as Core, CDH9 and CDH10, of size 187Kb, 158Kb and 158Kb, respectively, on chromosome 5 of the human genome. 24 pools with six individuals in each were made, totaling 144 samples. There are 12 pools for Autism case samples and the other half 12 pools for control samples. One case pool experiment failed and therefore 23 pools left in total for analysis. Short sequence reads were aligned by the Bioscope software from ABI SOLiD with default parameters. The mapped short sequence reads covered >96% of the three target regions with average 90X depth of coverage per individual. Meanwhile, individual genotyping data was collected for each sample, which were generated from Illumina HumanHap550v3 SNP arrays with ~550,000 markers. With individual genotyping data, the concordance of identified variants was calculated between pooled sequencing data and individual genotyping data for evaluating variant call quality.

Table 3.1 Summary of T1D and Autism Pooled Sequencing and ADHD Individual Sequencing Datasets

Disease	Platform	Total Reads	Reads Length	#Pool		#Individual per pool	Region	Coverage per individual
				Case	Ctrl			
Autism	SOLiD	~402M	50bp	11	12	6	~502 Kb	~90X
T1D	454	~9.4M	~250bp	10	10	48	~31 Kb	~80X
ADHD	Illumina	~57M	76bp x 2	3 individuals			~38 Mb	~20X

The second dataset was collected in a recent study of causative Type 1 Diabetes (T1D) variants (Nejentsev, et al., 2009). Exons and splice sites of 10 candidate genes were resequenced by the 454 sequencing system. Ten pooled samples each comprising equal amounts of DNA from 48 T1D patients and ten pooled samples each comprising equal

amounts of DNA from 48 healthy controls were made, totaling 480 T1D patients and 480 healthy controls from Great Britain. For each of the 20 pooled DNA samples, the numbers of produced short reads range from 281,270 to 579,102, with average length of 250 bases and 9,416,365 reads in total. These reads were mapped by BWA-SW (Li and Durbin, 2010) with default parameters and the average depth of coverage is 80X per individual.

The third one was an in-house individual sequencing dataset. Paired-end exome sequencing was performed on three members affected with attention deficit/hyperactivity disorder (ADHD) in a pedigree, using the Illumina Genome Analyzer IIx platform with read lengths of 76 base pairs. It targets all human exonic regions totaling approximately 38 Mb. The short reads were aligned by BWA with default parameters and removed duplicates by PICARD (http://sourceforge.net/projects/picard/). These mapped and cleaned short reads were then re-aligned locally by the GATK IndelRealigner tool (DePristo, et al., 2011). The average depth of coverage is about 20X for each patient. Meanwhile, the genotyping data were collected of these three patients, generated from the Illumina Human610-Quad version 1 SNP arrays with ~610,000 markers (including ~20,000 non-polymorphic markers).

For pooled sequencing data, CRISP has been shown to outperform other existing methods (Bansal, 2010), the comparison between SNVer and CRISP was performed for evaluation. SAMtools was also included for comparison although it was not designed for pooled sequencing data. For the ADHD individual data, SNVer was compared with SAMtools and GATK. Variant positions were called and filtered by SAMtools with all default settings plus using awk '($3 = = "*" &$6> = 50) || ($3! ="*" &$6> = 20) ', as suggested by the SAMtools website. For the ADHD data, SAMtools with the suggested

setting returned so many variants that SAMtools results were reported with an additional filtering -d20 to remove variant calls with sequencing coverage less than 20, for getting comparable numbers of variant calls as SNVer. Variants were also called using the GATK UnifiedGenotyper, followed by further filtering based on the latest recommendations from the GATK. CRISP has its own pileup procedure integrated in its analysis pipeline. To make a fair comparison, following CRISP (Bansal, 2010) similar quality control was performed and set the same processing parameters such as mapping quality and base quality filtering thresholds.

3.4 Results

3.4.1 Power and Type I Error Evaluations

The single pool results are shown in Figure 3.1. It can be seen that the proposed method can control type I error rate at the nominal level 0.05 in all settings. The number of sampled individuals (sample size) and the depth of coverage are both shown to be helpful in improving power. The largest improvement of ~10% attributed to depth of coverage (from 10X to 30X) is observed in the rare variants and high sequencing error (up-right panel). The improvement contributed by larger sample size keeps increasing at a decreasing rate until saturated. These power improvement curves would be helpful for pooling experiment design and provide guidance as to how to balance sample size (cost) and desired power. As expected, rare variants are much harder to be detected than common variants. A large sample is required for achieving high power to detect them. Finally, higher sequencing error (0.05 vs 0.01) puts a small dent to power.

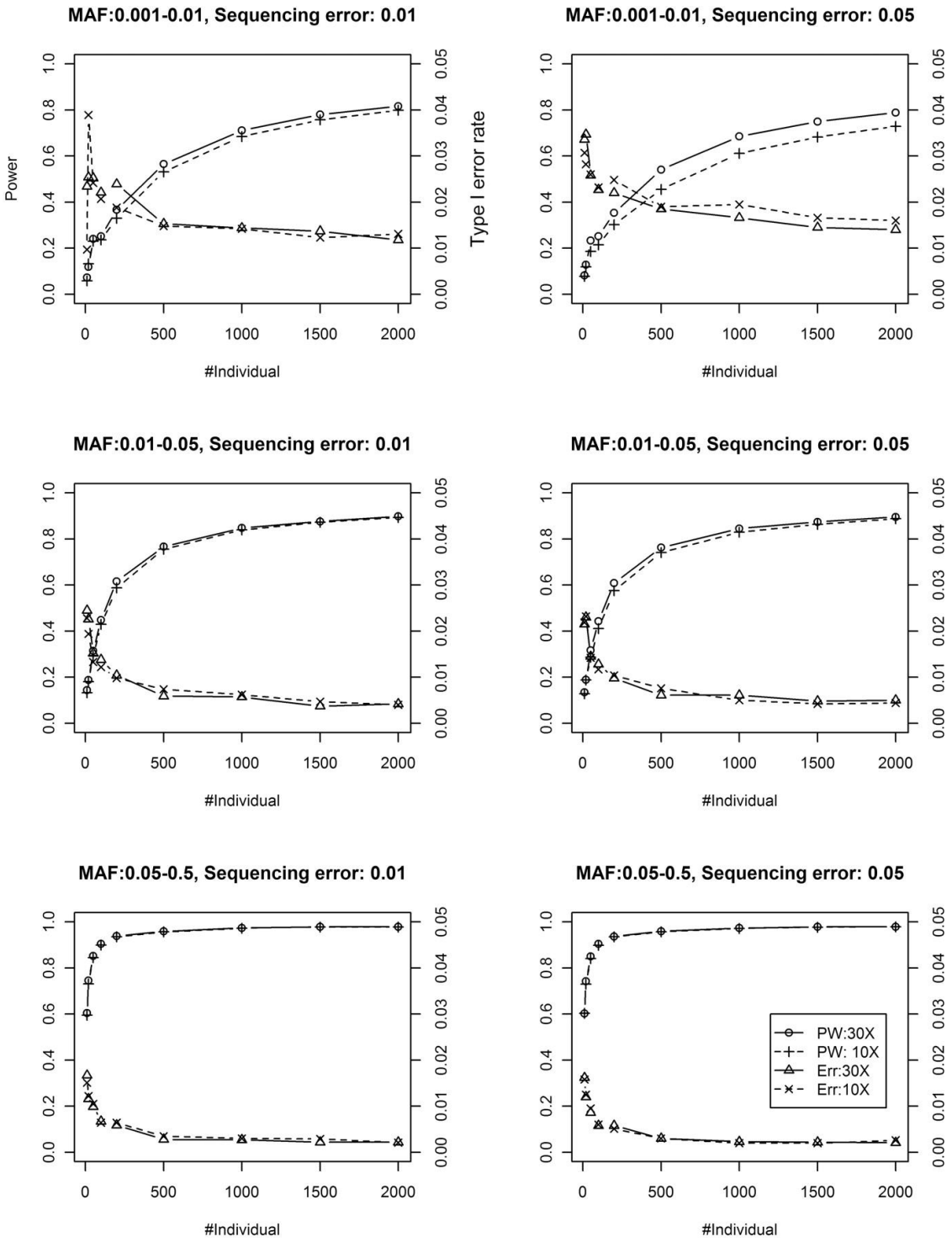

Figure 3.1 Power (PW) and Type I error rate (Err) of SNVer using single pool data at low (10X) and high (30X) coverage.

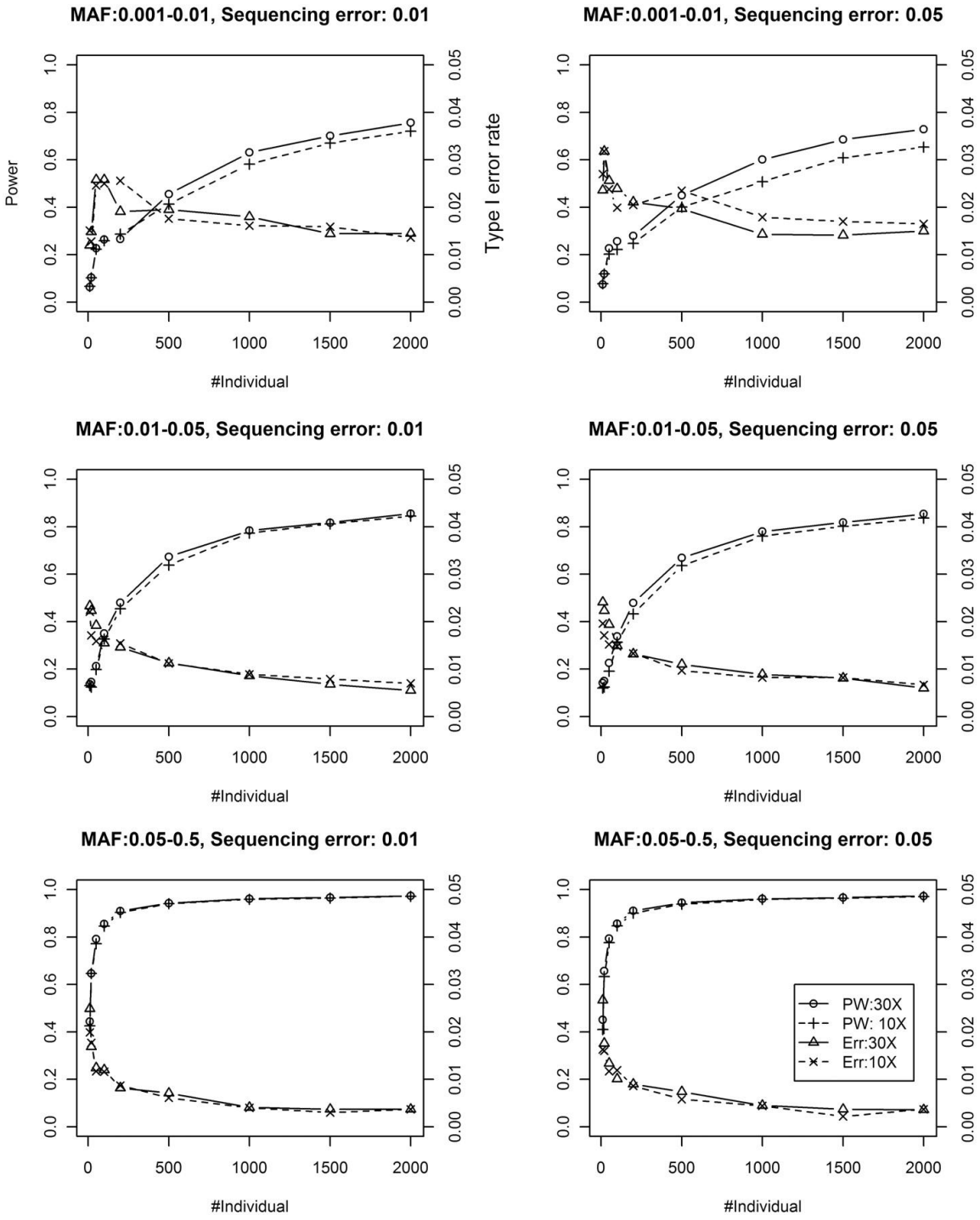

Figure 3.2 Power (PW) and Type I error rate (Err) of SNVer using multiple pool data at low (10X) and high (30X) coverage.

Figure 3.2 shows similar results for the multiple-pool scenario. Again, type I error rate is controlled at the nominal level 0.05. It is also observed that given the same number of sequenced individuals, single pool design yields a bit higher power with lower type I error rate in comparison with multiple pool design, for example, 1000 individuals using one single pool vs five pools with 200 individuals in each. CRISP selects candidate SNPs by the Fisher's exact test, which is then followed by additional filtering steps. In the multiple-pool scenario, it is shown that the rankings of candidates SNPs by the proposed test is superior to those by the Fisher's exact test employed by CRISP. To compare the efficiencies of these two rankings, the 10,000 positives with $\theta \sim U(\theta_{min}, \theta_{max})$ and 10,000 negatives with $\theta \sim U(0, \theta_{min})$ are divided into 100 groups, each with 100 positives and 100 negatives. These 200 loci are then ranked by their significance levels of testing the null H_0: $\theta < \theta_{min}$ using the statistical models. Rankings based the Fisher's exact test are also generated. The area under the curve (AUC) score averaged over 100 groups is used to evaluate these two rankings as shown in Figure 3.3 for the typical scenario of 30X coverage and 0.05 sequencing error. It can be seen that the Fisher's exact test is very inefficient for detecting common and less common variants. CRISP therefore has to rely on additional sequencing error models to complement the Fisher's exact test for detecting common variants. The BH procedure is applied to control FDR at the nominal level of 0.1 and 0.05. The number of sequenced individuals is modeled in the test and is shown to be helpful. This information is not explicitly utilized by CRISP in its Fisher's exact test and therefore contributes very little for detecting common and less common variants, although CRISP models it at the later filtering step.

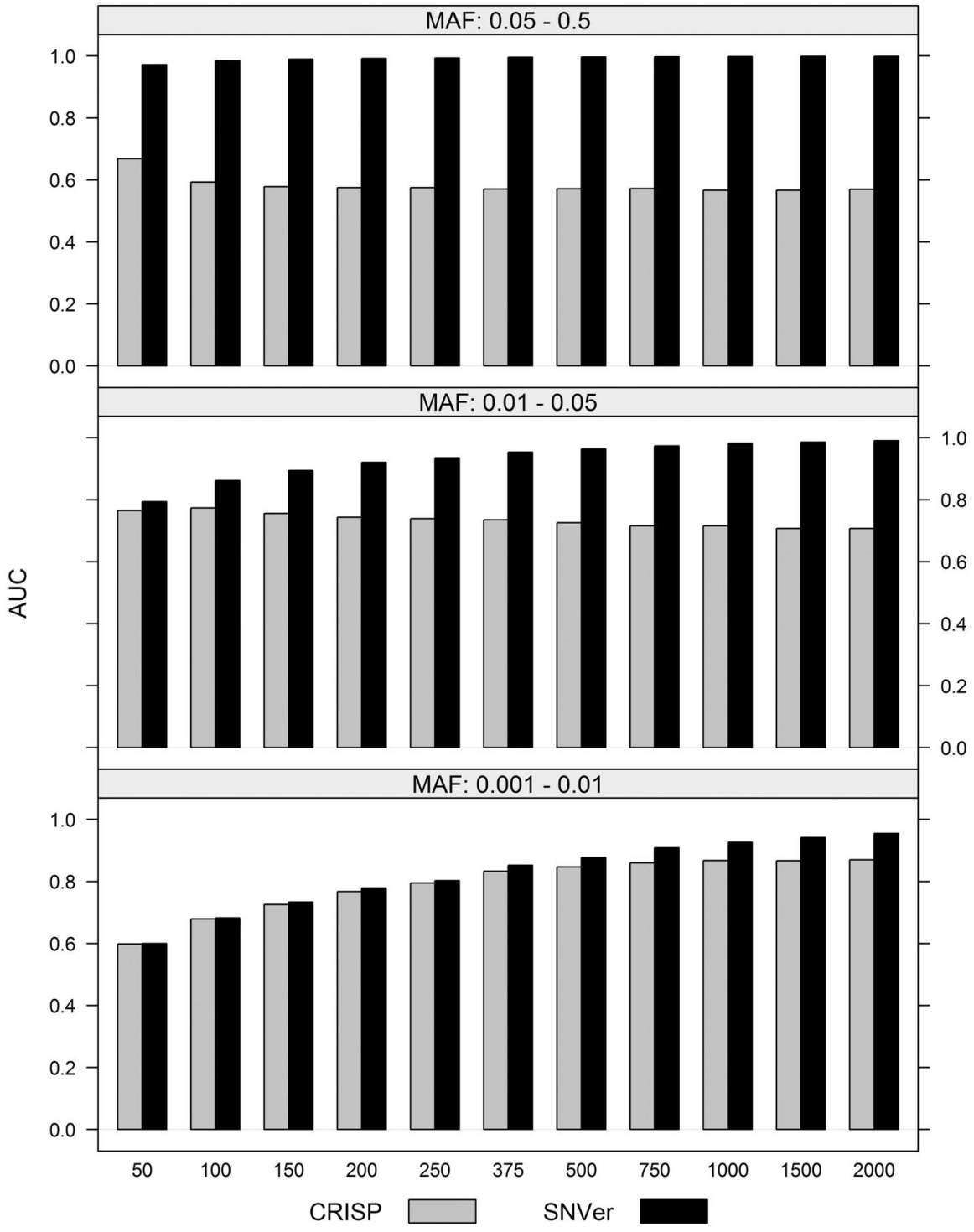

Figure 3.3 Ranking efficiency of the binomial models employed by SNVer vs the Fisher's exact test employed by CRISP.

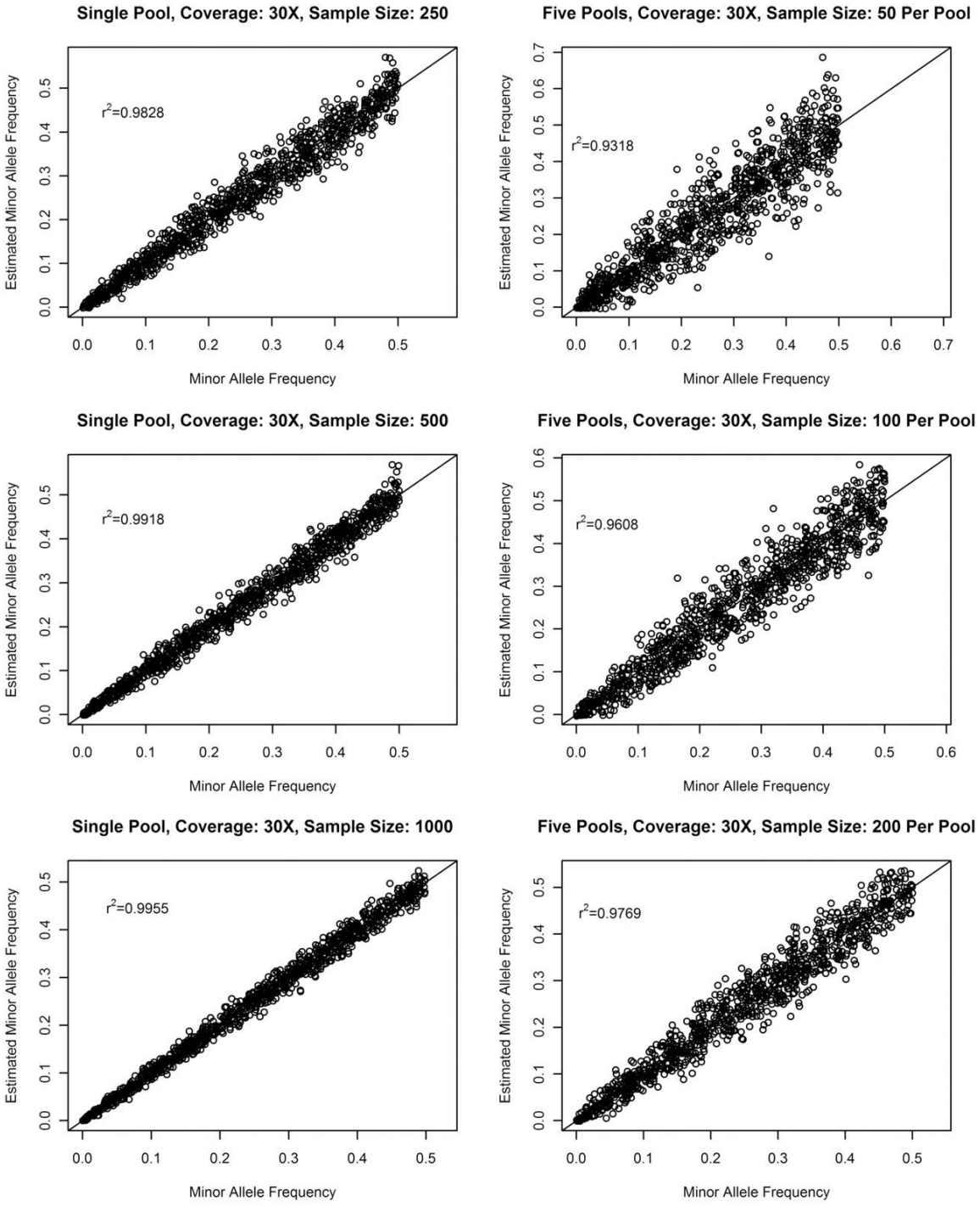

Figure 3.4 Correlation between the minor allele frequencies and its estimates in pooled sequencing.

The accuracy of allele frequency estimation has an impact on variant call, and is more critical for establishing association in genetics studies. Therefore the estimated minor

allele frequency (MAF) against the actual MAF when $\varepsilon = 0.01$ is plotted in Figure 3.4. For a moderate sample size of 250, good concordance are observed with correlation coefficients r^2=0.9828 and r^2=0.9318 for the single-pool design and the multiple-pool design, respectively. When the sample size increases to 1000, the concordance improves to r^2=0.9955 and r^2=0.9769 for the single-pool design and the multiple-pool design, respectively. The lower concordance of the multiple-design may be attributed to its additional between-pool variance. It also explains why single pool analysis yields fewer false positives than the multiple pools design for the same set of samples.

3.4.2 Better Performance

The user of SNVer only needs to set the sequencing error rate ε and the variant threshold θ_0. SNVer will then report the significance p values of the tested loci of how likely their MAF $\theta < \theta_0$. Assume $\varepsilon = 0.01$ for all real datasets. CRISP calls both rare and common variants, so θ_0=0 is set for SNVer to compare their performance in calling variants. CRISP will output the variants it calls, while SNVer will report overall significance p values for each locus, based on which the user can choose a threshold he/she feels appropriate and make variant calls. To make a comparison, loci are ranked by their p values output by SNVer and the significance threshold is taken that gives the same number of variants called by CRISP. The loci identified as variants by these two programs are then annotated by SeattleSeq (http://gvs.gs.washington.edu/SeattleSeqAnnotation/), and count how many of them have been confirmed as variants in dbSNP. Following (DePristo, et al., 2011), variant call quality is evaluated by examining dbSNP rate, transition/transversion (Ti/Tv) ratio and concordance of sequencing and individual genotyping calls. A higher Ti/Tv ratio generally indicates a higher accuracy; this metrics is particularly helpful for assessing novel single

nucleotide variant calls (DePristo, et al., 2011). The variant call results are summarized in Table 3.2. For the Autism and T1D pooled sequencing datasets, SNVer has the higher dbSNP rates, the higher overall Ti/Tv ratios, and the higher Ti/Tv ratios for new sites, in comparison with CRISP. It indicates the better quality of the call sets SNVer produced. In contrast, SAMtools made much fewer SNP calls which led to much lower sensitivities, despite its higher Ti/Tv ratios. Out of the 110 SNPs that have been genotyped by SNP arrays in the Autism dataset, SAMtools identified only 16 SNPs with 100% genotyping concordance, while both SNVer and CRISP called about 100 SNPs with 100% genotyping concordance. This confirms that SAMtools may not be appropriate for pooled sequencing data. The correlation between alternate allele frequencies in individually genotyped DNA samples and frequency estimates in the sequenced DNA pools is plotted in Figure 3.5, with r^2=0.92 and r^2=0.94 for the Autism case and control, respectively. The achieved 100% genotype concordance with less perfect frequency estimates is not surprising because accurate estimate of allele frequency θ is only critical for rare variants when testing $\theta>0$.

As shown in Table 3.2 , for the ADHD individual sequencing data, under family-wise error rate 0.05 level, SNVer also obtained the variant call sets with good quality. This is evidenced by the ~97% dbSNP rates, the ~2.9 overall Ti/Tv ratios, the 2.22 to 2.73 Ti/Tv ratios for novel sites, and the 99% genotype concordance. SAMtools with suggested parameters/filters made 2+ times more variant calls than SNVer (e.g. ~49K vs ~18K). The lower Ti/Tv ratios and genotype concordance suggest poorer quality for these larger call sets made by SAMtools. When applied with an additional filtering of sequencing depth >=20X, SAMtools identified fewer SNPs than SNVer. But it still has lower quality as indicated by the lower Ti/Tv ratios and genotype concordance. Compared

with GATK, SNVer has similar performance, while with the higher Ti/Tv ratios for novel variants in all three individuals.

Table 3.2 Comparison of SNP Calling by CRISP, SAMtools, GATK and SNVer

Data		Method	No. of SNP				Ti/Tv[&]			Concord.[+]
			All	Known	Novel	dbSNP%	All	Known	Novel	TP/P (%)
Autism (Pooled)	Case	CRISP	2182	1791	391	82.1	1.68	1.79	1.26	101/101 (100%)
		SNVer	2182	1795	387	82.3	1.71	1.81	1.35	102/102 (100%)
		SAMtools	261	260	1	99.6	2.26	2.29	0/1	16/16 (100%)
	Control	CRISP	2063	1610	453	78.0	1.68	1.83	1.27	96/96 (100%)
		SNVer	2063	1617	446	78.4	1.78	1.89	1.45	95/95 (100%)
		SAMtools	239	238	1	99.6	2.06	2.05	1/0	16/16 (100%)
T1D (Pooled)	Case	CRISP	306	93	213	30.3	0.95	2.58	0.63	N/A
		SNVer	306	126	180	41.2	1.71	2.15	1.47	
		SAMtools	14	9	5	64.3	10/4	8/1	2/3	
	Control	CRISP	167	110	57	65.9	1.49	2.93	0.46	
		SNVer	167	120	47	71.9	2.34	3.00	1.35	
		SAMtools	18	12	6	66.7	14/4	11/1	3/3	
ADHD (Individual)	84060	SNVer	18001	17535	466	97.4	2.89	2.89	2.73	4158/4183 (99%)
		SAMtools	48988	47513	1475	97.0	2.66	2.68	2.16	4437/8116 (55%)
		SAMtools[20X]	15038	14538	500	96.7	2.70	2.72	2.11	2034/3158 (64%)
		GATK	19655	19713	482	97.6	2.91	2.94	2.15	4649/4657(100%)
	84615	SNVer	17436	16914	522	97.0	2.85	2.87	2.22	4032/4063 (99%)
		SAMtools	46037	44489	1548	96.6	2.64	2.67	1.94	4173/7643 (54%)
		SAMtools[20X]	15510	14942	568	96.3	2.74	2.77	2.02	2062/3247 (64%)
		GATK	18892	18419	473	97.5	2.89	2.92	2.03	4537/4566(99%)
	92157	SNVer	18676	18208	468	97.5	2.90	2.92	2.37	4192/4224 (99%)
		SAMtools	49729	47693	2036	95.9	2.69	2.73	2.03	4251/7996 (53%)
		SAMtools[20X]	15881	15370	511	96.8	2.80	2.83	1.99	2028/3259 (62%)
		GATK	20100	19631	469	97.7	2.98	3.00	2.35	4700/4710(100%)

&: Transition and transversion ratio for the identified variants. When the number of variants is small it just reports the numbers but not calculate the ratio, e.g., 10/4 for all variants in T1D case by SAMtools means 10 transitions and 4 transversions.
+: genotype concordance. P represents the number of variants called by each program and also genotyped. TP represents the number of variant calls concordant between sequencing data and individual genotyping data.
20X: Additional filtering of sequencing depth >=20 is applied.

It is noted that the Ti/Tv ratios for the pooled sequencing data are low for both programs. This suggests that they may not perform well if estimating the false positive

rates using the Ti/Tv ratios (DePristo, et al., 2011) and confirms that variant calling is more challenging for pooled sequencing. Meanwhile, estimating false positive rates using this summary statistic should be cautious for pooled sequencing. First, Ti/Tv estimate for pooled samples is not as accurate as for individual samples. Second, targeted resequencing regions are usually small, e.g., 31 Kb for the T1D data and 500Kb for the Autism data, and therefore will exhibit higher genomic and statistical variances. For example, the ADHD individual 84060 has an exome-wide Ti/Tv ratio of 2.89 for all variants; if calculating Ti/Tv ratios based on only 500Kb regions, then the smallest Ti/Tv ratio obtained is 1.31, and the largest 7.00 with SD=1.53 (consider only 500Kb regions with at least 30 variants for having stable Ti/Tv ratio estimates).

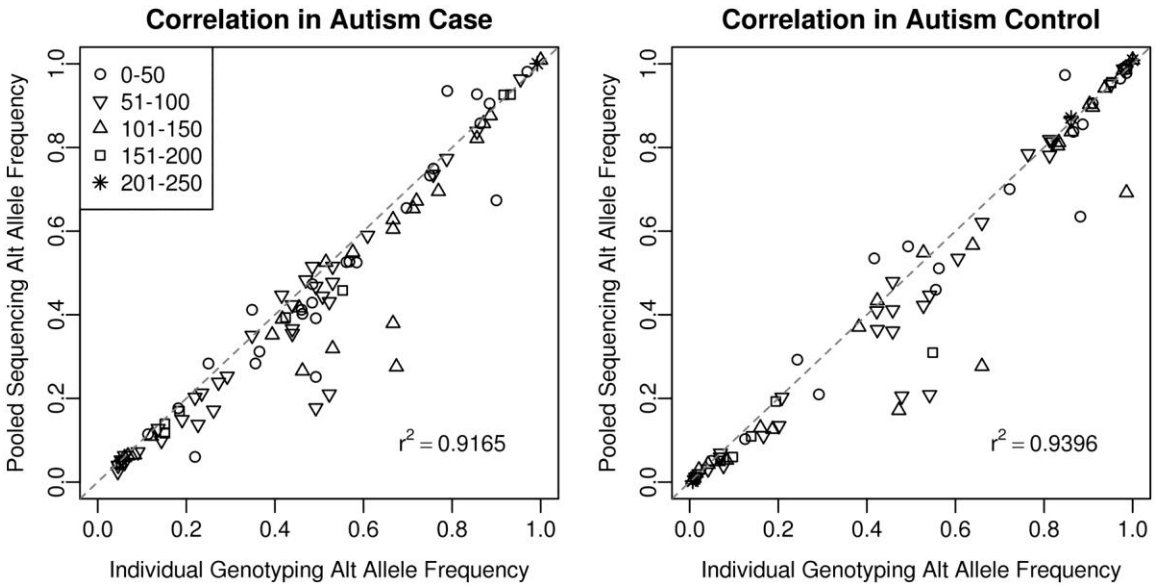

Figure 3.5 Correlation between alternate allele frequencies in individually genotyped DNA samples and its estimates in the sequenced DNA pools for the Autism dataset. Different symbols represent different depth of coverage ranges as shown in the legend.

3.4.3 Better Scalability

SNVer and SAMtools exhibit similar efficiency in terms of running time. The running time of SNVer and CRISP in analysis of the T1D and Autism datasets is given in Figure 3.6. The main bottleneck of CRISP comes from computing the p-value of a large number of contingency tables in the Fisher's exact test. Therefore, in additional to the number of tests, its time efficiency is also largely dependent on the number of pools and the depth of coverage. In contrast, these two factors have little impact on SNVer and its running time is roughly linear with the region size (the number of tests). For example, SNVer spends 0.1 hr on 31Kb and 1.5 hr on 502Kb for the two datasets, respectively. SNVer is much faster than CRISP. Taking the T1D case for example, SNVer is about 500-fold faster than CRISP and achieves 300Kb/hr. Such efficiency makes feasible the application of SNVer to analysis of whole-exome sequencing data, or even whole-genome sequencing data using high performance computing cluster, both of which, however, will take prohibitively longer time for CRISP.

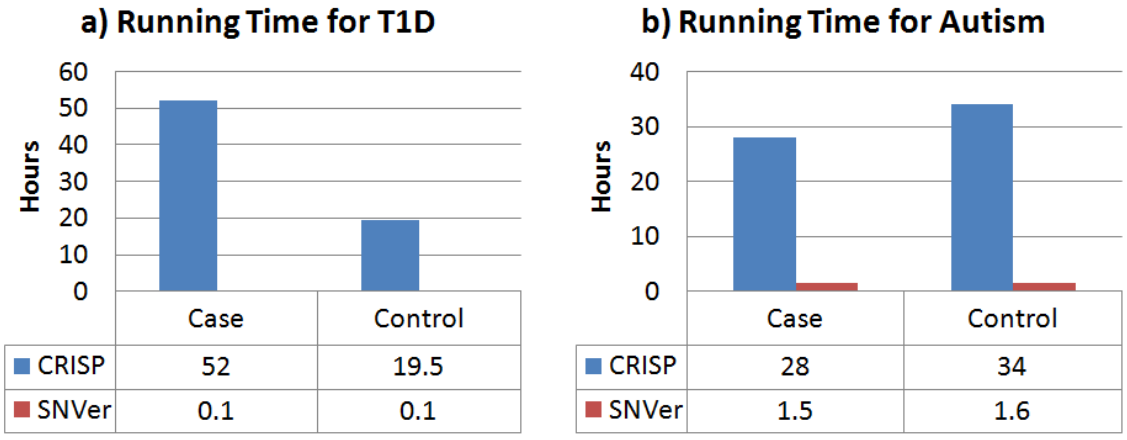

Figure 3.6 Comparison of running time of SNVer and CRISP for testing the T1D ~30Kb region and the Autism ~500Kb region. Running time of SNVer is mainly determined by the region size (the number of tests), while larger pool numbers and sequencing depth will take additional time for CRISP.

3.4.4 Informative Ranking and Multiplicity Control

SNVer reports one single overall significance p-value for each locus, based on which the rankings of all tested loci can be produced. Such rankings are more informative and accurate than the dichotomous decision of whether to "accept or reject the candidate as a variant" provided by CRISP and most other existing methods. For example, four rare variants have been found to be associated with T1D based on the T1D dataset by comparing the estimated MAF in cases and controls (Nejentsev, et al., 2009). If use SNVer to call these four variants by testing the null hypothesis $\theta \leq \theta_0$=0.01, the rankings of them by SNVer can be obained in Table 3.3, as well as the dichotomous decisions made by CRISP. For SNVer, very significant ranking changes of these four SNPs can be observed, which are consistent with their MAFs (relative to the threshold 0.01) and the MAF differences. CRISP identifies three of them, rs35337543, ss107794688 and ss107794687, as variants in both cases and controls, exhibiting no informative differential changes. It should be noted that the ranking difference may only reflect frequency difference. Large frequency difference between case and control of those variants may suggest their potential association with the phenotype, but their functional importance to the phenotype is yet to be assessed by further experiments.

In addition to ranking, valid p values given by SNVer also make multiplicity control possible. Tens of thousands or millions loci are usually simultaneously examined in typical NGS experiments. It is particularly desirable to have multiplicity control, which gives the user an idea of the chance of making any errors and/or the proportion of false positives among the variant calls they make. Each user can choose the type I error rate

threshold he or she considers appropriate, instead of just the dichotomous decisions of whether to "accept or reject the candidates" provided by most existing methods.

Table 3.3 Informative Rankings of Four Rare Variants with the Null Hypothesis $\theta \leq \theta_0$=0.01

SNP	T1D Case			T1D Control		
	Estimated MAF	**SNVer Ranking**	**CRISP CALL**	**Estimated MAF**	**SNVer Ranking**	**CRISP CALL**
rs35337543	0.36%	17557	Y	2.51%	45	Y
rs35667974	0.72%	17557	N	2.42%	59	Y
ss107794688	0.50%	17557	Y	1.79%	56	Y
ss107794687	1.07%	145	Y	2.45%	51	Y

3.5 Conclusion and Discussion

This chapter has developed a novel statistical tool SNVer for calling SNPs in analysis of pooled NGS data. Different from the previous models employed by CRISP, it analyzes common and rare variants in one integrated model, which considers and models all relevant factors including variant distribution and sequencing errors simultaneously. As a result, the user does not need to specify several filter cutoffs as required by CRISP. Some variant calling methods simply discard loci with low depth of coverage to achieve reliable variant calls. The statistical model does not discriminate against poorly covered loci. Loci with any (low) coverage can be tested and depth of coverage will be quantitatively factored into the final significance calculation. SNVer reports one single overall significance p-value for evaluating the significance of a candidate being a variant. An advantage of

reporting results on a more continuous scale, instead of just the dichotomous decision of whether to "accept or reject the candidate as a variant" as most existing methods do, is that the user can choose the alpha threshold he or she considers appropriate. In this chapter, both simulated data and real data are used to demonstrate the superior performance of the program in comparison with pre-existing methods. Although SNVer is motivated by the need for analysis of pooled NGS data, it can also be applied to individual NGS data as a special case (N=2 for diploid species), as shown in the ADHD dataset.

The current program can be improved and extended in several ways. First, small indels are not supported. Indels impose a great challenge for NGS including DNA amplification and reads mapping which are under fast development. When those techniques become mature in handling indels, it is worth investigation of their distribution and work out a proper calling strategy. Second, sequencing quality scores can be utilized to estimate site-specific sequencing error. Third, the majority loci of sequenced segments are known to carry no variants. The density of SNP is estimated to be around 1 out of 1000 bases. Such prior percentage of non-nulls information may help obtain more precise multiplicity control. Fourth, the dependency among tests will also be informative in increasing testing efficiency. It has shown that the LD dependency information is very informative in increasing the efficiency of conducting genome-wide association tests in analysis of GWAS data (Wei, et al., 2009). It is also found recently that dependency information is helpful for increasing the efficiency of testing hypotheses at the set level (Sun and Wei, 2011). For NGS data, one non-null (variant) is expected from every 1000 consecutive genomic bases. Such dependency patterns, if appropriately modeled, may help further improve testing efficiency. Finally, the current program focuses on calling variants,

namely, testing whether θ is larger than a threshold. Under the same framework, the models can be naturally extended for case-control association studies by testing whether $\theta_{case} = \theta_{control}$.

CHAPTER 4

SNVERGUI: A DESKTOP TOOL FOR VARIANT ANALYSIS

4.1 Introduction

Advances in next-generation sequencing (NGS) technology have made it possible to comprehensively interrogate genome-wide genetic variations. However, most existing tools for variation detection are based on command-line interface, which discourages the main end users of NGS data, such as biologists, geneticists and clinicians, from utilizing the software. This chapter develops the SNVerGUI, a graphical user interface (GUI)-based tool for variant detection and analysis. Compared with other methods for variant calling, the proposed approach is unique in that it is applicable to both individual and pooled sequencing data. With friendly GUI, end users can easily adjust running parameters to optimize variant calling for their specific needs. SNVerGUI supports commonly used input and output file formats that allows SNVerGUI to be seamlessly integrated into common NGS data analysis pipelines. SNVerGUI is implemented in Java, which is platform-independent and therefore easy to install and run on the commonly used operating systems, such as Linux, Mac, and Windows.

4.2 Results

Most existing NGS variant calling tools (Bansal, 2010; DePristo, et al., 2011) provide only command-line interfaces. Typically, users must execute these tools and sometimes apply additional filters from the command line. This may discourage biologists, geneticists, clinicians, and other end users who often lack the programming expertise to allow them to easily apply non-GUI tools. Quite a few GUI-based variant calling softwares have been

developed for addressing this concern (Bao, et al., 2009; Hou, et al., 2010; Qi, et al., 2010). However, they are not adapted to detecting variants from pooled sequencing data, which account for a sizable proportion of current NGS studies (Calvo, et al., 2010; Nejentsev, et al., 2009; Out, et al., 2009). This is the motivation of the SNVerGUI, a graphical user interface (GUI)-based desktop environment, in order to exploit the unique merits of the recently developed statistical tool SNVer (Wei, et al., 2011) for detecting SNVs and indels from both pooled and individual NGS data. The pipeline of SNVerGUI is illustrated in Figure 4.1. Its new and key merits are highlighted as follows.

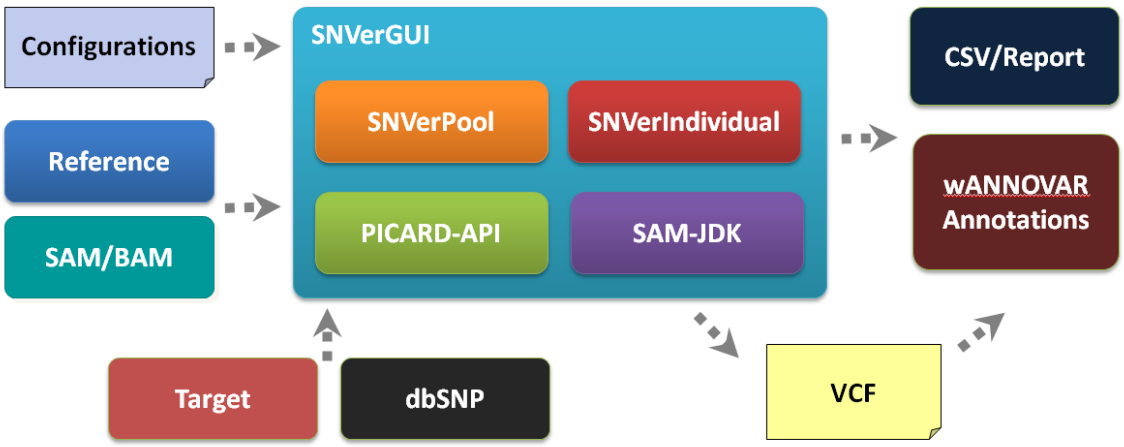

Figure 4.1. Pipeline of SNVerGUI. SNVerGUI employs PICARD-API and SAM-JDK for processing alignments, and utilizes SNVerPool and SNVerIndividual for calling variants (both SNV and indel) in analysis of pooled or individual NGS data.

First, compared with its previous command-line version (Wei, et al., 2011), SNVerGUI adds three new features. 1) It can estimate locus-specific sequencing error from data, and thus users don't need to specify this critical parameter. 2) SNVerGUI can call indel variants. 3) Variant outputs in VCF can be directed to the user-friendly web version

of the popular annotation tool wANNOVAR (Chang and Wang, 2012) for delineating their functional consequences.

Second, SNVerGUI is applicable to both individual and pooled NGS data by using a unified binomial-binomial statistical model. It can handle single-pool NGS data, which cannot be processed by most, if not all, existing state-of-the-art tools. Its computational efficiency makes it feasible to analyze whole-exome or whole-genome NGS data.

Third, SNVerGUI supports widely used input and output file formats. SNVerGUI accepts aligned read data and reference sequence data in popular file formats, such as .fasta, .sam and .bam files. Variant detection results are outputted in the CSV format that can be directly opened by Excel. They are also outputted in the standard VCF (Variant Call Format) (Danecek, et al., 2011) that can be accepted by other powerful tools as input, e.g., VarSifter (Teer, et al., 2011) for filtering and ANNOVAR (Wang, et al., 2010) for annotation.

Fourth, SNVerGUI provides flexible interactive post-call processing. Analysis results are displayed in easy-to-analyze table views that support sorting variants by p-value, sequencing depth, allele frequency, etc. Users can easily customize the output according to their needs, based on their experience and/or desired multiplicity control by setting different cutoffs.

Finally, SNVerGUI is platform-independent. As a Java-based software, SNVerGUI inherits Java's trademark property to allow to "write once, run anywhere". The program is wrapped in "one-click" easy-to-install package and runs on the commonly used operating systems, such as Linux, Mac, and Windows.

The superiority of the statistical model for variant detection, including accuracy, sensitivity, and computational efficiency, has been extensively documented in (Wei, et al., 2011). Table 4.1 briefly describes how to apply the GUI tool to analyze two real datasets in order to illustrate the analysis pipeline. Specifically, one pooled targeted resequencing Type 1 Diabetes (T1D) (Nejentsev, et al., 2009) dataset and one individual exome sequencing attention deficit/hyperactivity disorder (ADHD) data (Lyon, et al., 2011) were analyzed. The reads of the two datasets were mapped using BWA-SW(Li and Durbin, 2010) and BWA(Li and Durbin, 2009), respectively, with default parameters. Then mapping results were outputted in BAM format.

The two data sets were tested on Windows 7 Professional 64-bit OS, with Intel(R) Core(TM) i7 3.07GHz processor and 12.0 GB installed memory (RAM). The aligned BAM files, together with target regions, reference genome, and dbSNP file, were specified in SNVerGUI's graphical user interface. Variant calling was then executed using default parameters. It took SNVerGUI ~1.45h to analyze the T1D pooled data and ~1.75h to analyze the ADHD individual data (Table 4.1). This efficiency demonstrates that SNVerGUI is capable of analyzing high volume NGS data within very reasonable time. Notably, the analyses were performed using a fixed memory of only 1GB, which shows that SNVerGUI is so memory-efficient that it can be used as a desktop tool to analyze even whole-genome NGS data. Moreover, one can increase the Java Virtual Machine heap size by simply modifying the configuration file before launching SNVerGUI. The running time for analyzing such data sets would be shortened with larger memory, which highlights the flexibility of SNVerGUI in the aspect of memory management. Therefore, as a desktop

tool, it is feasible to analyze a myriad of NGS data at different scales if enough memory and CPU are provided.

Table 4.1 Summary and Performance on T1D Pooled Sequencing and ADHD Individual Sequencing Data

Data	Platform	Total reads	Read Length	#pool	Pool Size	Target Region	Coverage	Time
T1D	454	~4.9M	~250bp	10	48	~31kb	~80X	~1.45h
ADHD	Illumina	~19M	76bp PE	-	-	~38Mb	~20X	~1.75h

4.3 Summary

In summary, this chapter has developed the SNVerGUI, a user-friendly desktop tool to call variants for the analysis of pooled sequencing and individual sequencing data. Using this software, users can perform sophisticated variant detection by simply configuring several parameters in a friendly graphical user interface and annotate variants in wANNOVAR (Chang and Wang, 2012) conveniently. Using two real datasets, it has been shown that SNVerGUI tool is capable of analyzing very high volume NGS data in feasible time. Hence, SNVerGUI is a fast and easy GUI tool for identification of genomic variants. As a desktop tool, it has demonstrated the feasibility of conducting variant detection analysis on personal computers. It makes bioinformatic analysis as simple and effortless as possible, as needed for clinical genetics and personalized medicine. It should be expected its popularity among geneticists, clinicians, and biologists, as well as small labs which cannot afford costly bioinformatics personnel and infrastructure as is currently required for analyzing NGS datasets (Sboner, et al., 2011). With the advent of desktop sequencers (Loman, et al., 2012), such a desktop bioinformatics tool would become much demanded for NGS data analysis.

CHAPTER 5

COLLAPSING SINGLETONS FOR ASSOCIATION STUDY

5.1 Introduction

Advances in next-generation sequencing (NGS) technology have made it possible to comprehensively interrogate the entire spectrum of genomic variations including rare variants. They may help capture the remaining genetic heritability which has not been fully explained by previous genome-wide association studies (GWAS). This chapter performs a gene-based genome-wide scan to identify hypertension susceptibility loci in analysis of a whole genome sequencing cohort of 103 unrelated individuals. It has found that collapsing singletons may boost signals for associating rare variants and identified SETX statistically significant by a genome-wide gene-based threshold (P value $< 5.0 \times 10^{-6}$). The function of SETX in hypertension may be worthy of further investigation.

5.2 Data

The Genetic Analysis Workshop 18 (GAW18) provided a whole genome sequencing dataset for a hypertension cohort of 483 individuals. These samples were sequenced by Complete Genomics with ~60x coverage, and odd numbered autosomes data were made available for analysis. After quality control, 464 individuals and 24 million SNPs remained. Of those SNPs, over 51% had MAFs < 1%, which was the focus of the analysis in this chapter. The longitudinal hypertension phenotypes were provided for up to four time points. Since the analysis was focused on binary traits, it treated individuals diagnosed with hypertension in any of the four times as cases. 103 genetically unrelated individuals

were extracted with both phenotype data and sequencing data, where 39 unaffected controls and 64 cases affected by hypertension were found.

5.3 Methods

The variants were stored in VCF files. The pre-processing includes the following. 1) Filtered out SNPs that were present in dbSNP132 or MAFs > 1%, for getting rare variants. 2) Filtered out SNPs with genotype missing rate > 5%. 3) The remaining missing genotypes were resampled from non-missing individuals. 4) Collapsed singletons, a variant being observed only once among all the samples, as one indicator variable by

$$X_i = \begin{cases} 1 & if\ any\ x_{i,j} = 1 \\ 0 & otherwise \end{cases}.$$

The rationale was that the distribution of singletons as a group may reflect the association between target region and phenotype. 5) Grouped variants into sets based on RefSeq gene annotations (Pruitt, et al.), requiring SNPs lie between the RefSeq transcript start site (TSS) and transcript end site (TES). SNPs outside gene boundary were not analyzed. In total, 10,148 genes from odd numbered chromosomes were used.

This chapter employed three recently published rare association tests, qMSAT (Daye, et al.), C-alpha (Neale, et al., 2011) and CMC (Li and Leal). The qMSAT was a quality-weighted multivariate score association test that can utilize genotype quality information. However, genotype quality score information was not available in the GAW18 raw VCF files. Without utilizing quality information, the qMSAT test was equivalent to the linear SKAT (Wu, et al.), SSU (Pan, 2009) and C-alpha. The C-alpha test

compared the assumed binomial distribution of rare variants in cases versus controls via a homogeneity test. CMC, a combined Multivariate and Collapsing Method, collapsed variants in subgroups according to allele frequencies and combines these subgroups using Hotelling's T^2 test. For all these three tests, permutations were used to evaluate association significance. Because permutation was computationally expensive, a two-step strategy in searching and testing candidate loci was used. Specifically, it first used 1000 permutations, from which it can identify candidates with estimated P value < 0.001. Then for these candidates it conducted 10^6 permutations so as to know if any loci were significant at a genome-wide gene-based threshold ($0.05/10000=5.0\times10^{-6}$) using a Bonferroni assumption.

Table 5.1 Genes with P < 0.001 from at Least One Method Using 10^3 Permutations

Chr	Gene	#Variants(Singletons)	qMSAT	C-alpha	CMC
chr1	NUP210L	674(221)	Y		
chr1	USP1	51(19)	Y	Y	
chr7	CUL1	348(114)	Y		
chr9	RAB14	88(32)	Y		
chr9	**SETX**	**380(135)**	**Y**	**Y**	**Y**
chr11	FLJ39051	44(11)	Y		
chr11	GDPD5	338(104)	Y		
chr19	GRIN3B	24(8)	Y		
chr19	LOC100505495	249(78)	Y	Y	
chr19	PSG5	111(26)	Y		Y
chr5	CXXC5	96(35)		Y	
chr15	RCCD1	32(13)		Y	
chr17	WSCD1	183(70)		Y	
chr17	MLLT6	92(33)			Y
chr1	ATF6	576(173)			Y
chr7	ZNF775	69(21)			Y
chr19	LOC100128252	45(13)			Y
chr5	LOC728342	495(146)			Y

5.4 Results

After the pre-processing step, it obtained ~2.2 million rare variants, which were assigned to 10148 genes for testing. Then the three tests were performed, qMAST, C-alpha and CMC, using R (www.r-project.org). The qMSAT, C-alpha and CMC identified 10, 6 and 7 genes with estimated P value < 0.001, respectively (Table 5.1). Only SETX revealed significance for all of the three methods. Using 10^6 permutations, qMSAT, C-alpha and CMC yielded more precise p-values of 2.0×10^{-6}, 1.0×10^{-6}, and 6.0×10^{-6}, respectively, for SETX (Table 2). The CMC p-value was slightly higher than the genome-wide gene-based threshold, which was possibly due to its lower power compared to qMSAT and C-alpha (Daye, et al.).

SETX locates in chr9:135,136,827-135,230,372 and is a relatively large gene among all the human genes. The length of SETX (93,545bp) is far greater than the median number (17,970bp) of all the genes (P-value$<2.2\times10^{-16}$, one-sided Wilcoxon signed rank test). Although it contains 26 exons, the total length of coding regions is only 8,034, suggesting that SETX includes large intronic regions. In order to pinpoint causal regions, the 380 variants of SETX into three groups were divided based on its functional annotations. Specifically, ANNOVAR was applied (Wang, et al.) to annotate the variants of SETX and grouped them into coding sequence regions (CDS), untranslated regions (UTR) and intronic regions (INTRON) (Table 5.2). It is observed that majority of rare variants were, indeed, from intronic region. These three regions were tested using the same tests with 10^6 permutations. The UTR group and the CDS group were far from being significant, suggesting that they may be irrelevant. Another possible reason may be that there are very few variants in these categories. Due to the fact that the INTRON group became more significant than the whole gene-based tests after excluding the variants from

these two groups, and therefore it may be concluded that causal variants were located in the intronic region of SETX.

The next was to elucidate why and where the signal came from. To this end, several in-depth analyses for SETX were performed. First, the Fisher's Exact test was conducted on the super feature created by collapsing singletons. It has been found that, by collapsing all the 135 singletons on SETX, it achieved a very significant p-value (3.7×10^{-6}), together with OR=8.8 and 95% CI = [3.12, 27.43] (Table 5.2). This explained why SETX could be detectable under such a small sample size. It obtained more significance when testing the super feature with only singletons within the intronic regions (P-value=8.8×10^{-7}, OR=9.5 and 95% CI=[3.43,28.70]), which was consistent with the results from three rare variant association tests. Second, each rare variant and singleton were checked individually by performing the same test. It turned out that none of them were significant, where the minimum p-value was merely 0.14. This demonstrated that the significance of SETX was very unlikely due to technical artifact, such as systematic sequencing error or imputation bias, because the new feature was a combination of hundreds of singletons. It also highlighted that collapsing singletons may increase power when studying association of rare variants using a relatively small sample size. Third, a closer examination was took for allele frequencies of the 380 rare variants located in SETX. 92 of them could be found in 1000 Genomes Project (2012 Feb release, http://www.1000genomes.org/). It can be seen that their allele frequencies in general population were extremely low (mean frequency=0.0004 for 92 rare variants), indicating that these variants were so rare that they may collectively have a composite effect of OR=8.8 while missed in previous studies.

Table 5.2 Functional Annotation and Test of the Rare Variants in SETX

		P-value | OR | 95%CI			P-value*		
Regions	**#Variants (Singletons)**	**Fisher's Exact Test on Super Variant[&]**			**qMSAT**	**C-alpha**	**CMC**
SETX (GENE)	380(135)	3.7×10^{-6}	8.8	[3.12,27.43]	2.0×10^{-6}	1.0×10^{-6}	6.0×10^{-6}
CDS	14(8)	1.000	1.0	[0.18, 6.94]	1.000	0.544	0.837
UTR	6(4)	0.632	0.6	[0.04, 8.60]	1.000	0.662	0.990
INTRON	360(123)	8.8×10^{-7}	9.5	[3.43,28.70]	$<1.0\times10^{-6}$	$<1.0\times10^{-6}$	$<1.0\times10^{-6}$

*: P-values were calculated using 10^6 permutations.

&: Super variant was defined by collapsing all the singletons.

Finally, to further remove possible cofounding effect of population stratification, a principle component analysis (PCA) was performed on a set of randomly selected 100k common variants with no missing value and MAF>0.1. Logistic regression test was then conducted on the created super feature for SETX, together with the first ten principle components as covariates. It can be found that the super feature remained significant with a P-value=6.7×10^{-5} while the ten principle components were not.

The protein encoded by SETX contains a DNA/RNA helicase domain at its C-terminal, suggesting its involvement in both DNA and RNA processing. Mutations in SETX have been reported to be associated with ataxia-ocular apraxia-2 (AOA2) (Arning, et al., 2008) and an autosomal dominant form of juvenile amyotrophic lateral sclerosis (ALS4) (Chen, et al., 2004; Pruitt, et al.). However, the function of SETX and its role in hypertension remains unclear and may be worthy of further investigation.

5.5 Summary

This chapter performs three rare-variant association tests for the analysis of a whole genome sequencing dataset to identify susceptibility genes in hypertension. It groups and collapsed rare variants in a gene-based manner for two reasons: 1) the deleteriousness of variants could come from protein-coding sequence changes or non-coding intronic regions that contain regulatory elements. 2) Based on the previous simulation study (Daye, et al.), the power of the analysis could be as low as 0.2 (sample size<200). By collapsing singletons, one may benefit from increasing power. This idea is essentially similar as those burden tests for rare CNV in GWAS and de novo mutations in sequencing study. Indeed, the signal is identified mainly from the intronic regions of SETX in a collective manner of those singletons.

CHAPTER 6

POLYASEEKER: A PIPELINE FOR IDENTIFYING POLYA SITES

6.1 Introduction

Alternative polyadenylation (APA) of mRNA plays a crucial role in post-transcriptional gene regulation. To date, however, no bioinformatics tools exist for identifying polyadenylation cleavage sites from increasingly popular RNA-Seq data. This chapter proposes a bioinformatics pipeline, PolyASeeker, to fill this void. The novelties of this work include a probabilistic scoring scheme that takes sequencing quality into account to select polyA containing reads, and utilizing the mating information in paired-end reads. Simulation studies and applications to real data demonstrate that the proposed tool can efficiently and precisely identify PolyA sites for the analysis of RNA sequencing data. It may be expected that the knowledge of APA mechanisms and their roles in gene regulation will be greatly enhanced with the aid of the tool as increasingly more RNA-Seq data become available.

6.2 Methods

6.2.1 Score PolyA Reads by Incorporating Sequencing Quality

This section proposes a novel scoring method that takes sequencing quality into account to select PolyA containing reads. For a sequence, different weights are given to the bases of A, T, C and G as follows:

$$S(base) = \begin{cases} 1, & if\ base = A \\ -1, & if\ base = \bar{A} \end{cases}$$

When considering sequencing error ε, the expected score for each base is computed as:

$$E(A) = (1-\varepsilon) \times S(A) + \frac{\varepsilon}{3} \times S(\bar{A}),$$

$$E(\bar{A}) = (1-\varepsilon) \times S(\bar{A}) + \frac{\varepsilon}{3} \times S(A),$$

where ε can be obtained from the Phred-scaled base quality score, Q, in FASTQ file, i.e.,

$$\varepsilon = 10^{(Q/-10)}.$$

Each aligned read is scored by evaluating its unmapped region. Specifically, let L be the length of the unmapped region, and then compute the summation of the expected score of each base in the unmapped region as follows:

$$Score = \sum_{j=1}^{L} \pi_j \times E_j(A) + \left(1-\pi_j\right) \times E_j(\bar{A}),$$

where π_j is an indicator of whether a base is A or not,

$$\pi_j = \begin{cases} 1, & if\ base = A \\ 0, & if\ base = \bar{A} \end{cases}.$$

6.2.2 Novel Method for Filtering Internal Priming

This section develops a novel method for filtering internal priming. It extends 20 bp from each side of a candidate cleavage site, and examine this 40-bp genomic region in the reference genome. If this region is A-rich then it suggests internal priming. The above scoring scheme for evaluating PolyA reads can be similarly applied by setting $\varepsilon = 0$. Now let $X = \{e_1, e_2, \ldots, e_{40}\}$ be a one-dimensional array with e_i being the expected score of the i^{th} base in the sequence. It formulates the problem of seeking an A-rich segment in the 40-bp sequence as solving the maximum-subarray problem, namely, to find a nonempty and contiguous subarray of X whose values give the largest sum (Bentley, 1984). It scores the internal priming for each alignment in terms of the sum of its maximum-subarray.

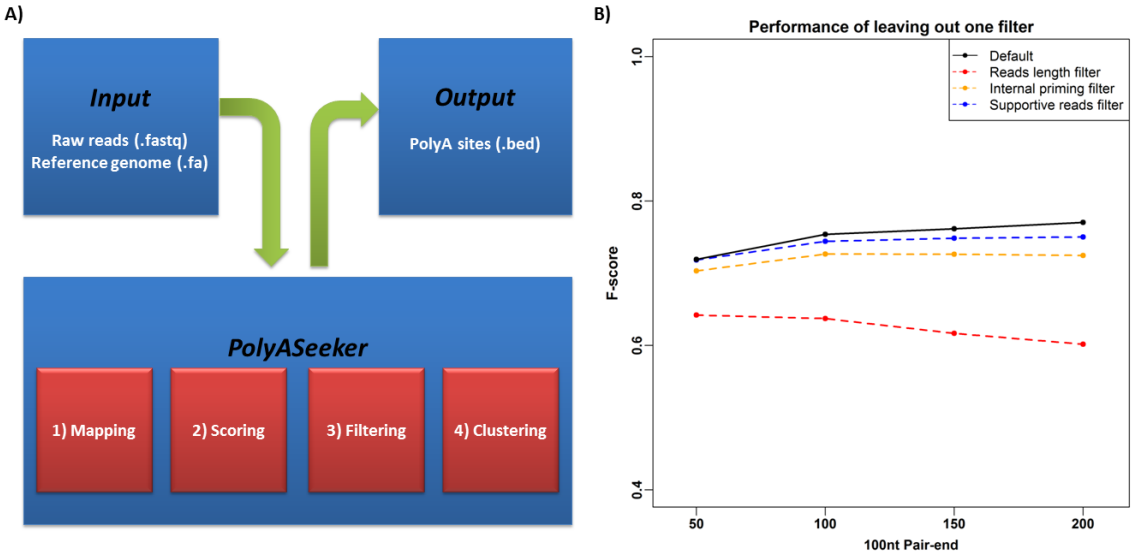

Figure 6.1 Illustration of PolyASeeker. A) Pipeline: PolyASeeker supports widely used input and output file formats and integrates four steps from mapping to clustering, making the tool easy to use. B) Filter contribution: the performances of leaving one filter out (dashed lines) are worse than that of using all three filters (solid line), suggesting the individual contribution to the improvement of PolyA site predictions made by each filter.

6.2.3 Pipeline of Identifying PolyA Sites from RNA-Seq

The PolyASeeker pipeline is shown in Figure 6.1. PolyASeeker accepts raw reads and a reference genome in their popular formats, fastq and fasta, respectively, as input. The short reads are first mapped using Bowtie2 (Langmead and Salzberg, 2012) local model, which does not require reads to align end-to-end. This feature is particularly suitable for aligning PolyA containing reads. Next, for each alignment, PolyASeeker scores its unmapped region by taking into account sequencing quality and selects candidate reads above a pre-specified threshold (7.8 as default). The genomic loci with screened mapped reads are considered as PolyA site candidates. For these candidate loci, PolyASeeker removes potential false positives by applying the following three filters. (1) Read length filter: It requires the length of the mapped region for a PolyA containing read be greater than a threshold, to ensure the mapping accuracy. (2) A novel internal priming filter to reduce false positives introduced by A-rich regions in the genome. (3) Supportive reads filter: PolyASeeker counts how many PolyA reads support a PolyA site candidate and filters out low confident ones. All these filters contribute to enhance PolyA site prediction performance as shown in Figure 6.1. Finally, PolyASeeker clusters PolyA sites within a 24nt window by utilizing the snowball method (Tian, et al., 2005) and outputs a BED file.

PolyASeeker has several key merits. First, it is easy and convenient to use. PolyASeeker integrates all necessary PolyA site analysis steps from mapping to clustering, and provides a convenient one-stop solution. Users just need to provide raw sequencing reads, supporting both pair-end and single-end libraries, and a reference genome to perform the entire process for PolyA site predictions. The output file can be directly uploaded to UCSC genome browser (http://genome.ucsc.edu/) or IGV (Robinson, et al.,

2011) for visualization. Second, PolyASeeker is accurate and powerful, and demonstrates the state-of-the-art performance, as shown in the simulation studies and real data analyses. Third, PolyASeeker takes full advantage of existing powerful NGS tools, for example, Bowtie2 (Langmead and Salzberg, 2012) and BEDTools (Quinlan and Hall, 2010), making the pipeline so efficient that it can process large scale data feasibly.

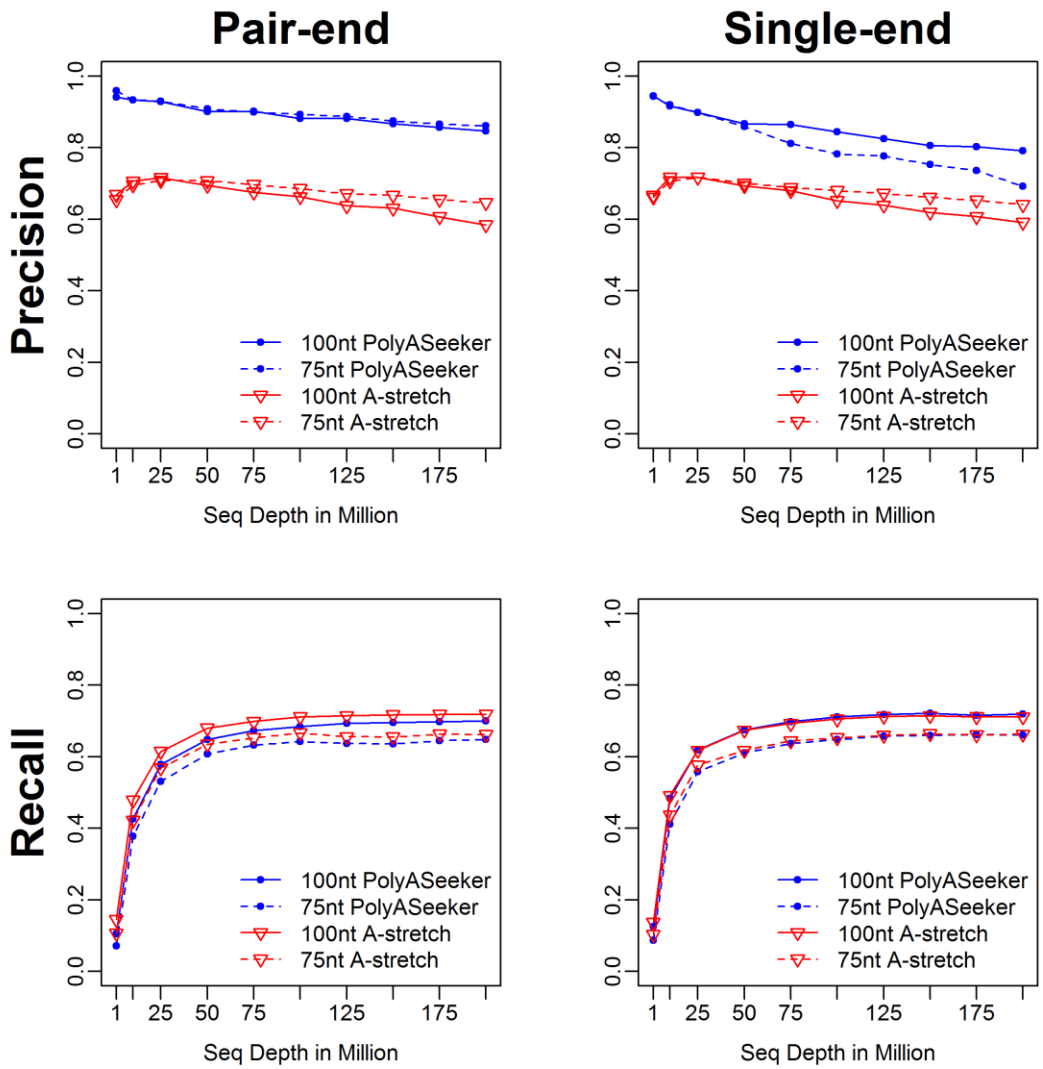

Figure 6.2 Performance of PolyASeeker and the 8A-stretch method for simulated data. With comparable Recall, PolyASeeker outperforms the 8A-stretch method in terms of significantly improved Precision in all the simulation settings.

6.3 Results

6.3.1 Simulation Studies

The simulation covered various RNA-Seq experiments using FluxSimulator (Griebel, et al., 2012) to assess the performance of PolyASeeker. It considered four scenarios: 100bp paired-end, 75bp paired-end, 100bp single-end and 75bp single-end. To simulate more realistic error models, instead of using the default one in FluxSimulator, a custom error model was created from an in-house RNA-Seq dataset from Illumina platform. The model of the polyadenylation process in FluxSimulator was generated by a Weibull-approximation of the normal distribution with shape=2 and scale=300 to sample random lengths of polyA tails of human. The Poly-dT priming RNA-Seq procedure was performed and sequenced different reads depths, from 1M to 200M for all the four scenarios.

The simulation considered only expressed transcripts (molecule>0), which generated ~13,600 out of all the 34,102 transcripts in each simulated dataset. It defined the ends of these expressed transcripts to be true PolyA sites. If a PolyA site prediction fell into +/-5bp of a true site, the prediction was considered as a true positive (TP), and a false positive (FP) otherwise. If a true site was not covered by any prediction correctly, it was considered as a false negative (FN). The performance was evaluated by Precision, Recall and F-score, defined as follows:

$$Precision = \frac{TP}{(TP+FP)},$$

$$Recall = \frac{TP}{(TP+FN)},$$

$$F\text{-}score = \frac{2 \times Precision \times Recall}{(Precision + Recall)}.$$

The simulation compared the performance of PolyASeeker to a simple 8A-stretch method. It found that PolyASeeker achieved comparable recall but much higher precision among all the settings (Figure 6.2). It also found that the best mapping strategy in the proposed pipeline performed better than the uniquely mapping strategy that was commonly used in previous studies (Fu, et al., 2011; Ji, et al., 2011; Pickrell, et al., 2010) (Figure 6.3).

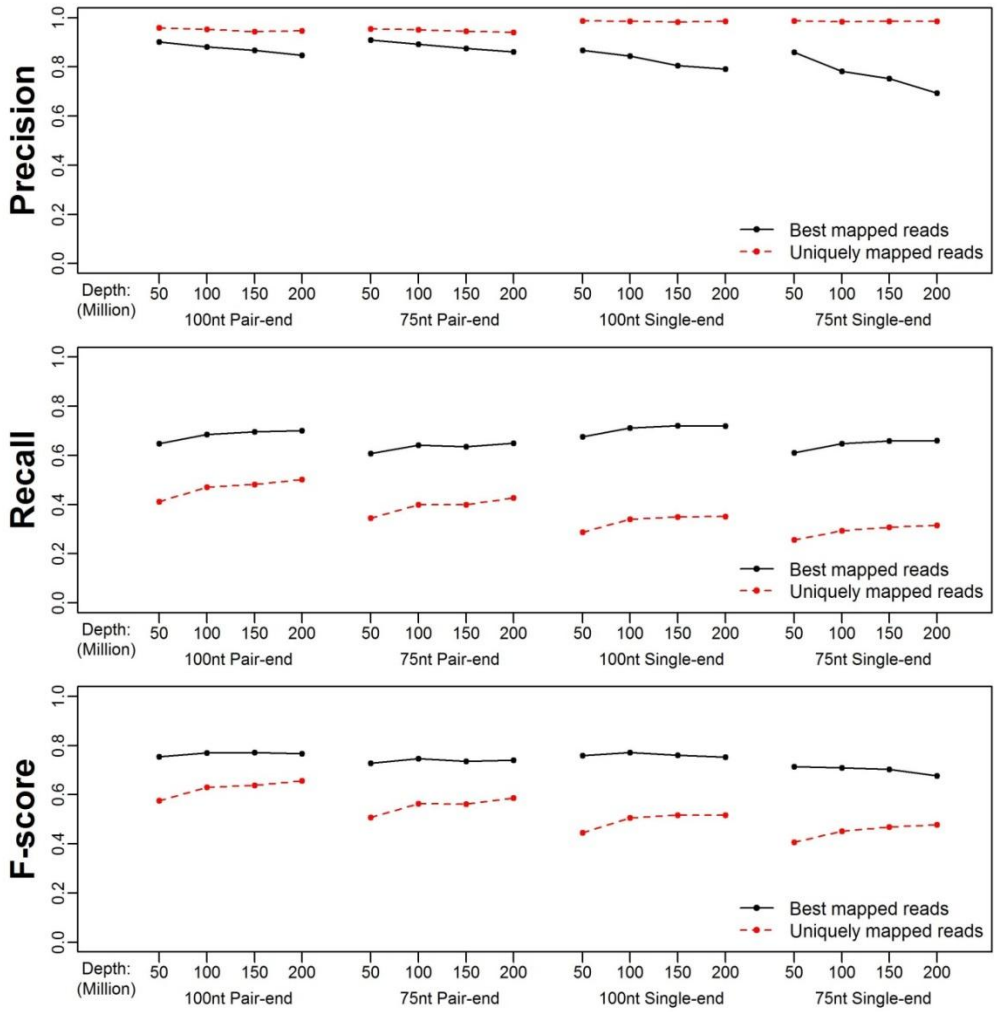

Figure 6.3 The best mapping strategy performs better than the uniquely mapping strategy, especially for single-end data.

Table 6.1 Summaries and Results of Five Real RNA-Seq Datasets

Data	Human Body Map 2.0	Breast Cancer	Arabidopsis Thaliana	Saccharomyces Cerevisiae	Sordaria Macrospora
Library	2x50 paired-end	75 single-end	2x101 paired-end	2x76 paired-end	101 single-end
#Samples	16	3	6	2	8
Total reads	1,278,682,935	31,026,769	69,695,338	23,742,737	465,385,168
PolyA reads	174,418	16,139,436	14,626	118,562	577,897
PolyA sites	12,432	89,475	885	4,991	42,542
Known sites	6,492(52.22%)	56,148(62.75%)	179(20.23%)	497(9.96%)	-
Novel sites	5,940(47.78%)	33,327(37.25%)	706(79.77%)	4,494(90.04%)	-
FDR	5.5%	5.6%	13.5%	13.6%	15.2%
Running time	~118.5h	~4.1h	~9.9h	~3.5h	~47h

6.3.2 Applications to Real NGS Data

PolyASeeker was applied to analyze five RNA sequencing datasets that cover different representative RNA-Seq settings and species for demonstrating its performance: (1) regular paired-end human tissue data, from the Illumina Human Body Map 2.0 project (NCBI GEO: GSE30611), which profiled 16 different human tissues using 2x50 paired-end sequencing; (2) single-end breast cancer data, generated following a special protocol by using modified oligo-d(T) tagged with sequencing primers to sequence polyA containing reads (Fu, et al., 2011); (3) Plant data, using 2x101 paired-end sequencing, from a recent Arabidopsis study (Ren, et al., 2012); (4) Yeast data from expression profiling of Saccharomyces cerevisiae by 2x76 paired-end high throughput sequencing (Levin, et al., 2010); and (5) 101 bp single-end RNA-Seq data for transcriptional landscape of fungal development (Teichert, et al., 2012). The data summaries and analysis results are given in Table 6.1. About 37%, 80% and 90% of the polyA sites identified by the proposed pipeline are novel for human, plant and yeast, respectively, compared to the ones that have been catalogued in the existing public PolyA databases (Human: PolyA-Seq (Derti, et al., 2012) from UCSC genome browser, Arabidopsis: http://www.arabidopsis.org/ and Yeast: PACdb, http://harlequin.jax.org/pacdb/). The high fraction of novel sites in plant and yeast,

compared to human, is mainly due to the fact that the public PolyA databases for plant and yeast are EST-based and have not been enriched by high throughput studies yet. This highlights that the knowledge of PolyA sites among different species can be greatly enhanced by applying the proposed tool to plentifully available RNA-Seq data. It also estimated their false discovery rates (FDRs) following (Pickrell, et al., 2010). PolyASeeker achieved a comparable FDR in the three non-human datasets and much lower FDRs in the two human datasets than the one obtained in (Pickrell, et al., 2010).

The next was to compare the predicted polyA sites with those found by the simple 8A-stretch method (Tian, et al., 2005). To this end, the 8A-stretch method was applied to the same five real datasets. Comparable results were achieved for yeast and fungi data, and much better results for human and plant data in terms of the number of identified PolyA sites and the number of recovered alternative polyadenylated genes (Figure 6.4). It has been found that, for the 2x50 paired-end Human Body Map 2.0 project, ~30% APA genes were recovered by PolyASeeker whereas only ~13% APA genes were identified by the rudimentary 8A-stretch method. The results also show that the sequencing depth of PolyA containing reads would affect the number of identified APA genes. The breast cancer dataset has about 10-fold more PolyA containing reads than the Human Body dataset (Table 6.1). As shown in Figure 6.4, the breast cancer dataset reveals ~75% of genes that produce alternative polyadenylated mRNAs, which is ~2.3 fold more than that of the Human Body dataset. Taken together, these observations indicate that the analysis of relative usages of sites in APA genes could be largely affected if employing different sequencing protocols (leading to different numbers of PolyA containing reads) and choosing different bioinformatics tools.

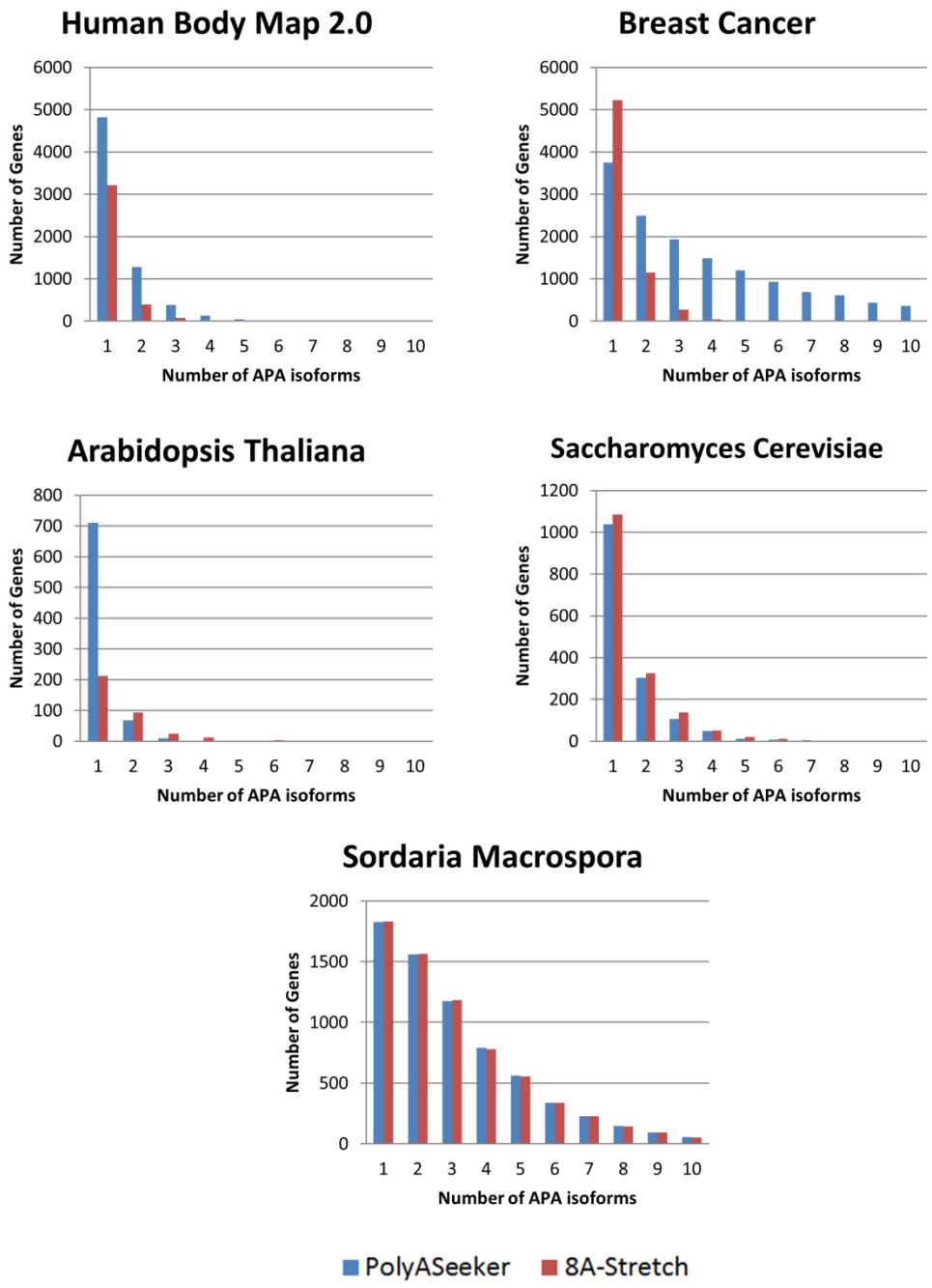

Figure 6.4 Bar plot showing the number of genes with different numbers of PolyA sites detected by PolyASeeker and the simple 8A-stretch method from five real datasets.

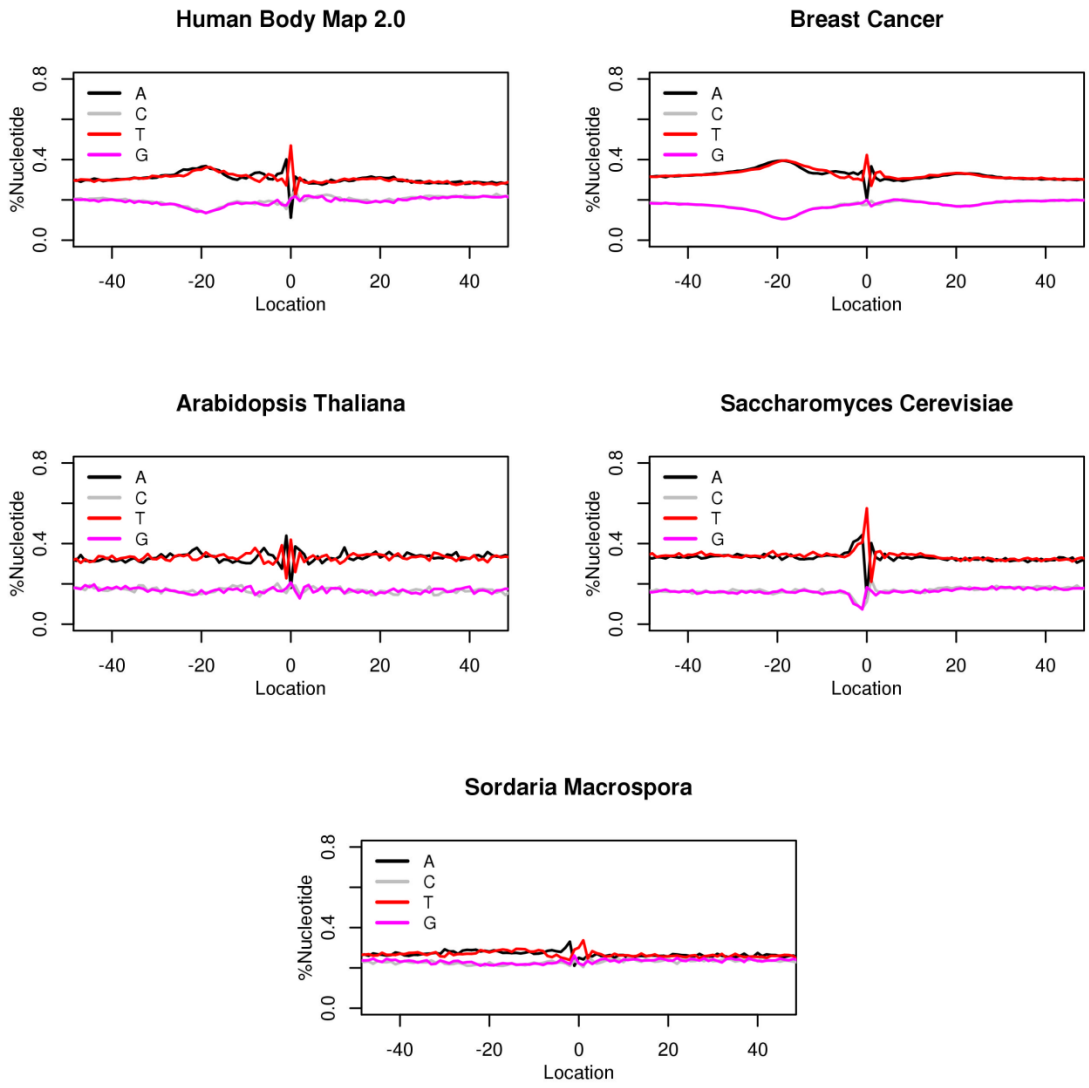

Figure 6.5 The nucleotide composition surrounding polyadenlyation cleavage locations identified by PolyASeeker in five real datasets.

Figure 6.5 shows the nucleotide composition surrounding polyadenylation cleavage locations in each dataset, which is similar to previous studies (Jan, et al., 2011; Ozsolak, et al., 2010; Sherstnev, et al., 2012). These summary statistics and results suggest that the proposed pipeline can be applied to analyze RNA sequencing data accurately and

efficiently. It will facilitate fully exploiting RNA-Seq data for gaining a better understanding of alternative polyadenylation mechanisms.

6.4 Summary

In conclusion, this chapter has developed a useful bioinformatics pipeline, PolyASeeker, to identify PolyA sites from RNA sequencing data. It is the first NGS bioinformatics tool to detect PolyA sites. It may be expected that the knowledge of APA mechanisms and their roles in gene regulation will be greatly enhanced with the aid of the proposed tool as increasingly more RNA-Seq data become available.

CHAPTER 7

A CHANGE-POINT MODEL FOR IDENTIFYING 3'UTR SWITCHING

7.1 Introduction

Next-generation RNA sequencing offers an opportunity to investigate transcriptome in an unprecedented scale. Recent studies have revealed widespread alternative polyadenalytion (APA) in eukaryotes, leading to various mRNA isoforms differing in their 3'UTR, through which, the stability, localization and translation of mRNA can be regulated. However, very few, if any, methods and tools are available for directly analyzing this special alternative RNA processing event. Conventional methods rely on annotation of polyadenylation sites; yet, such knowledge remains incomplete, and identification of PolyA sites is still challenging. The goal of this chapter is to develop methods for detecting 3'UTR switching without any prior knowledge of PolyA annotations.

This chapter proposes using a change-point model for identifying 3'UTR switching, which is the first available method that allows investigators to directly analyze 3'UTR length changes without being dependent on PolyA site information. To determine whether a 3'UTR is shortening or lengthening to a certain extent, it further develops an additional testing procedure to make directional decisions. It has been shown that this directional procedure can control the mixed directional FDR (mdFDR) at a pre-specified nominal level. Simulation studies in various settings and applications to two real NGS data sets have demonstrated that the proposed change-point model and the testing framework are powerful and accurate. This tool will allow investigators to analyze next-generation RNA sequencing data in an effective and efficient way.

7.2 Methods

7.2.1 Change-point Model for 3'UTR Switching

The 3'UTR switching problem and the change-point model are illustrated via a toy example in Figure 7.1. Assume there are two 3'UTR isoforms, isoform 1 and isoform 2, ending with a distal and proximal PolyA site, respectively. These two PolyA sites define the common and extended regions. The expression ratio of the two isoforms across two conditions, treatment and control, can be quantified by the percentage of read counts from the treatment condition (Figure 7.1(C)). It is expected a constant ratio throughout the whole 3'UTR (p_i = C, for i=1,..,T), if the usage of isoforms under these two conditions is identical. A ratio change at a certain position τ implies a ratio change between two isoforms, which is so-called 3'UTR switching. The goal is to test the null hypothesis H_0 that the ratio p_i is constant against the alternative hypothesis that, for some point τ in the 3'UTR, the ratio changes from p_0 to p_τ,

$$H_1: p_i = \begin{cases} p_0, & i = 1, \dots, \tau - 1, \\ p_\tau, & i = \tau, \dots, T. \end{cases}$$

When the change-point location τ is known, e.g., based on isoform knowledge if available, detecting the change is straightforward. However, p_0, p_τ, and, most importantly, τ, are unknown in this problem.

The model starts with a setup for the sequenced reads on 3'UTR with length T. Let $\{X_t \mid t < T\}$ be the number of reads whose first base maps to the left of base location t of a given 3'UTR under the treatment condition. Similarly, let $\{Y_t \mid t < T\}$ be the number of such reads under the control condition. Denote m^x and m^y to be the total number of reads

in the treatment and control conditions, respectively. Let $\boldsymbol{U} = \{U_1,\ U_2,\ \ldots,\ U_{m^x}\}$ and $\boldsymbol{V} = \{V_1,\ V_2,\ \ldots,\ V_{m^y}\}$ be the event locations for processes $\{X_t\}$ and $\{Y_t\}$, namely, U and V are the mapped positions of reads from the treatment and control samples. Let $m = m^x + m^y$ be the total number of reads combined from treatment and control samples, and then obtain combined event locations $\{W_1,\ W_2,\ \ldots,\ W_m\}$. Define an indicator variable Z_i to denote whether an event is a realization of the treatment process or control process as:

$$Z_i = \begin{cases} 1\,, & if\ W_i \in \{U_1,\ U_2, \ldots,\ U_{m^x}\}, \\ 0, & if\ W_i \in \{V_1,\ V_2, \ldots,\ V_{m^y}\}. \end{cases}$$

For any short read i in the combined process, use the term "success" to refer to $Z_i = 1$, that is, the read is from the treatment process. Hence, following (Worsley, 1983) , define a change-point model on the indices $\{1, \ldots,\ T\}$ for read counts by the binomial log-likelihood function. Considering a candidate change point at τ , for $1 < \tau < T$, a generalized likelihood ratio statistic is

$$\Lambda_\tau = \sum_{k\in[1,\tau-1]} \left\{ Z_k \times \log\frac{\hat{p}_0}{\hat{p}} + (1 - Z_k) \times \log\frac{1-\hat{p}_0}{1-\hat{p}} \right\} + \sum_{k\in[\tau,T]} \left\{ Z_k \times \log\frac{\hat{p}_\tau}{\hat{p}} + (1 - Z_k) \times \log\frac{1-\hat{p}_\tau}{1-\hat{p}} \right\},$$

where $\hat{\mathrm{p}}_0$, $\hat{\mathrm{p}}_\tau$ and $\hat{\mathrm{p}}$ are the maximum likelihood estimates of success probabilities:

$$\hat{p}_0 = \frac{\sum_{k=1}^{\tau-1} Z_k}{\tau - 1}\,, \qquad \hat{p}_\tau = \frac{\sum_{k=\tau}^{T} Z_k}{T - \tau + 1},$$

$$\hat{p} = \frac{\sum_{k=1}^{T} Z_k}{T}\,.$$

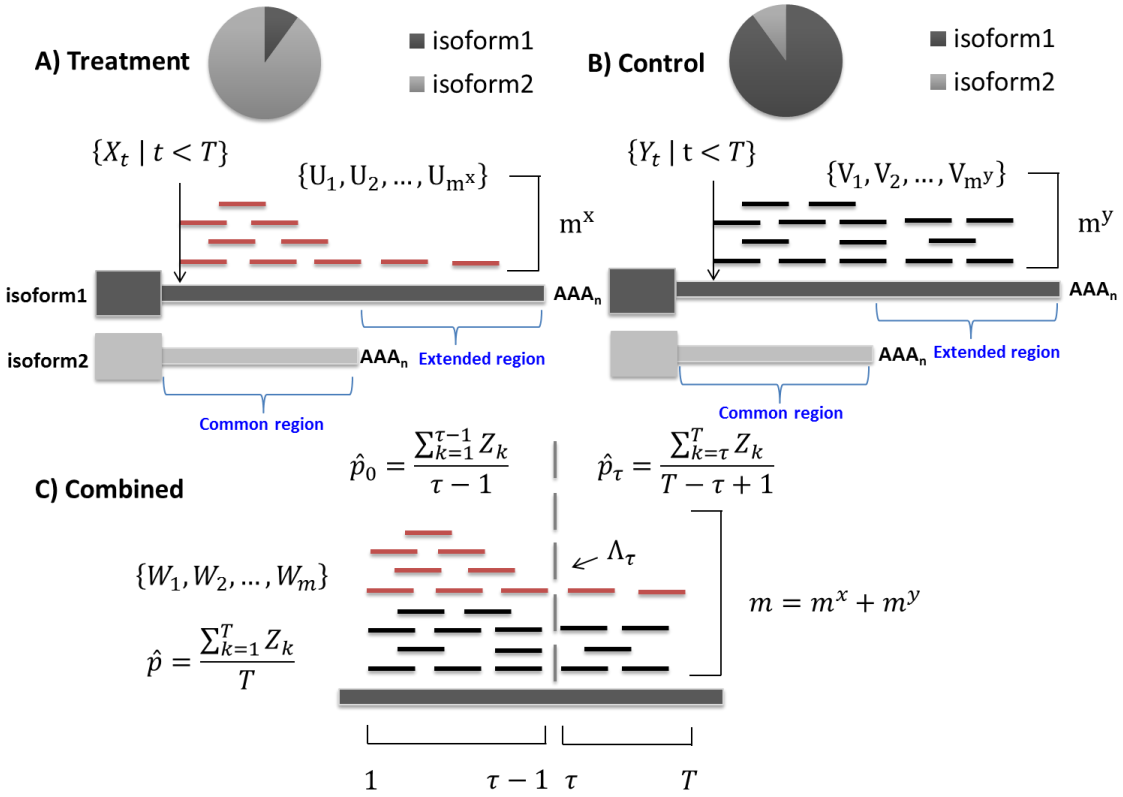

Figure 7.1 Illustration and notations of the change-point model for 3'UTR switching problem.

Note that this is an exact binomial generalized likelihood ratio statistic and can help to quantify the ratio change. Because the change-point location τ is unknown, the statistic for all candidate loci τ=2, …, T-1 can be computed , and find the one yielding the maximal change. The solution is

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \Lambda_\tau.$$

7.2.2 General Iterative Procedure for Calculating P-value

This section seeks to compute the significance p-value for the maximum test statistic. To this end, following (Worsley, 1983) it employs a general iterative procedure to calculate

how likely the maximum likelihood ratio statistic L would be less than $\Lambda_{\hat{\tau}}$, denoted as $\Pr(L < \Lambda_{\hat{\tau}})$, under the null hypothesis.

For the combined process in section 7.2.1, let S_k and $S_k{}'$ be the total numbers of successes (from the treatment process) at intervals [1, k-1] and [k, T], respectively, (k=2,...,T-1). The likelihood ratio test statistic Λ_τ depends only on S_k and $S_k{}'$. Given that $S = S_k + S_k{}'$, and $S = m^x$ is fixed, then Λ_τ depends only on S_k.

Therefore, given $\Lambda_{\hat{\tau}}$ and the test statistics, events of $L_k < \Lambda_{\hat{\tau}}$ can be expressed as events of the form $a_k \le S_k \le b_k$ for suitable choices of $a_k = inf\{S_k : L_k < \Lambda_{\hat{\tau}}\}$ and $b_k = sup\{S_k : L_k < \Lambda_{\hat{\tau}}\}$.

For $k = 1, \dots, T$, define $F_k(v) = Pr\left(\bigcap_{i=1}^{k} events\ of\ L_i < \Lambda_{\hat{\tau}} \,\middle|\, S_k = v\right)$ so that the p-value can be derived as follows:

1) Initially, set $F_1(v) = 1$, for $a_1 \le v \le b_1$.
2) For $2 \le k \le T-1$, find $F_k(v)$ for $a_k \le v \le b_k$ by

$$F_k(v) = \sum_{u=a_{k-1}}^{b_{k-1}} F_{k-1}(u) h_{k-1}(u, v)$$

$$\text{where, for } 0 \le u \le M_{k-1} = \sum_{i=1}^{k-1} m_i\,,\ 0 \le v - u \le m_k,$$

$$h_k(u, v) = \binom{M_{k-1}}{u}\binom{m_k}{v-u} \Big/ \binom{M_k}{v}.$$

3) A final iteration for k=T at v=S will produce $F_T(S)$ equal to $Pr(L < \Lambda_{\hat{\tau}})$.
4) The desired probability will be $Pr(L \ge \Lambda_{\hat{\tau}}) = 1 - Pr(L < \Lambda_{\hat{\tau}})$.

7.2.3 Directional Multiple Testing Procedure

If the usage of the long isoform increases, it is called lengthening, and, if it decreases, shortening. Identifying shortening or lengthening events may be critical for downstream analysis, such as analyzing miRNA target sites. The significance computed in the previous section is for a two-sided test. That is to say, when the null hypothesis is rejected, it can only state that there is a change, either lengthening ($\hat{p}_\tau > \hat{p}_0$) or shortening ($\hat{p}_\tau < \hat{p}_0$). In practice, upon rejecting the null H0, one may often conclude that the change is either lengthening or shortening based on the sign of ($\hat{p}_\tau - \hat{p}_0$). There is a chance that this decision strategy will make a false statement about the sign, which is termed as a directional error, or a type III error (Benjamini, et al., 2005). It is desirable to control this error when making directional conclusions, which may not be negligible when a large number of tests are conducted simultaneously. In the applications of this chapter, it often tests for tens of thousands of genes at a time.

In the multiple-testing field, it is often argued that an exact null hypothesis is never true in reality; instead, more likely only significant differences matter (Benjamini, et al., 2005; Williams, et al., 1999). Here for the 3'UTR switching problem, small change may happen by chance and is irrelevant to the phenotype of interest. Dramatic change may be more robust and easier to replicate. Therefore focusing on dramatic change is particularly meaningful as it often has only one or very few replicates in RNA-Seq experiments.

This section proposes to use the odds ratio (OR) at the estimated change point $\hat{\tau}$ to measure the change direction and magnitude, reasoning that the proposed method essentially chooses the location that gives the strongest association in a 2×2 contingency table among all possible locations. Thus, it performs Fisher's exact test at the estimated

change point $\hat{\tau}$ to make such directional decisions. It formulates this problem as controlling false discoveries within the multiple-testing framework. Using a similar definition as in (Guo, et al., 2010), denote the mixed directional FDR (mdFDR) to be a combination of two parts. One is the false discovery rate (FDR), resulted from the change-point testing procedure. The other is the pure directional FDR (dFDR), derived from Fisher's exact test,

$$mdFDR = FDR + dFDR = E\left\{\frac{C}{R \vee 1}\right\} + E\left\{\frac{F}{R \vee 1}\right\} = E\left\{\frac{C+F}{R \vee 1}\right\},$$

where *C* is the number of falsely rejected true null hypotheses and *R* is the total number of rejected hypotheses among $H_1, \ldots, H_m$. *F* denotes the total number of false null hypotheses among $H_1, \ldots, H_m$ that are correctly rejected while at least one directional error has been made when deciding upon the signs of the components.

To control mdFDR, the expected proportion of Type I and directional errors among all the rejections, this section proposes a directional testing procedure as follows:

1) Apply the BH method at level α to test whether there is a significant change among all the m hypotheses.
2) Let *R* denote the number of hypotheses rejected.
3) For every $i=1,\ldots, R$, perform one-sided Fisher's exact test for testing $OR>d$ $(d\geq 1)$.
4) If Fisher's exact test has a p-value $P^i_{fisher} \leq \frac{R}{m}\alpha$, then reject the null hypothesis.

It is shown that a similar BH procedure using the same two-sided p-value twice can control the mdFDR at level α (Benjamini, et al., 2005). The directional testing procedure proposed here has its novel extension in comparison with the BH procedure in (Benjamini, et al., 2005). Specifically, the same significance p-values are re-used in testing direction

and controlling directional errors in (Benjamini et al. 2005); in contrast, this procedure employs an additional one-sided Fisher's exact test for detecting dramatic change and the rejection is based on these new p-values. It can be seen in the simulation studies that the new testing procedure can control mdFDR at the nominal level. It is noted that when $d=1$, the one-sided test determines the direction of 3'UTR changes. The user may set d to be much larger than 1 to detect genes with more dramatic 3'UTR changes.

7.3 Simulation Studies

This section first presents simulation results to demonstrate the performance of the change-point model. Assume there are two 3'UTR isoforms with different ending PolyA sites as shown in Figure 7.1. The gene expression level ratio before and after the change point (Figure 7.1(C)) will critically influence how difficult the change can be detected. So it generated the 3'UTR with different expression ratios under two conditions. Specifically, under condition 1, the entire 3'UTR has a constant expression level; whereas, under condition 2, the expression level was increased by K-fold in the common region and the extended region remained the same as in condition 1. Gene expression level was measured in RPKM (Reads Per Kilobase per Million mapped reads (Mortazavi, et al., 2008)). It simulated two constant expression levels RPKM=1 and RPKM=2 for condition 1. These two RPKM values are commonly used for determining expressed genes in RNA-Seq real data analyses (Ji, et al., 2011; Zhang, et al., 2013). It assumed that the total number of mapped reads was 100 million/sample and the 3'UTR length was 1000bp. It considered three possible change points at 250bp, 500bp and 750bp of the 3'UTR. It varied the fold change K in the common region from 2 to 4 with increments of 0.5. The null distribution was simulated by setting $K=1$ for estimating type I errors. It simulated 500 3'UTRs with

change (*K>1*) and 500 3'UTRs without change (*K=1*) to estimate the power and FDR of the proposed method, respectively. FDR nominal level=0.05 was set. The simulation was repeated 50 times, and it reported the averaged power and FDR.

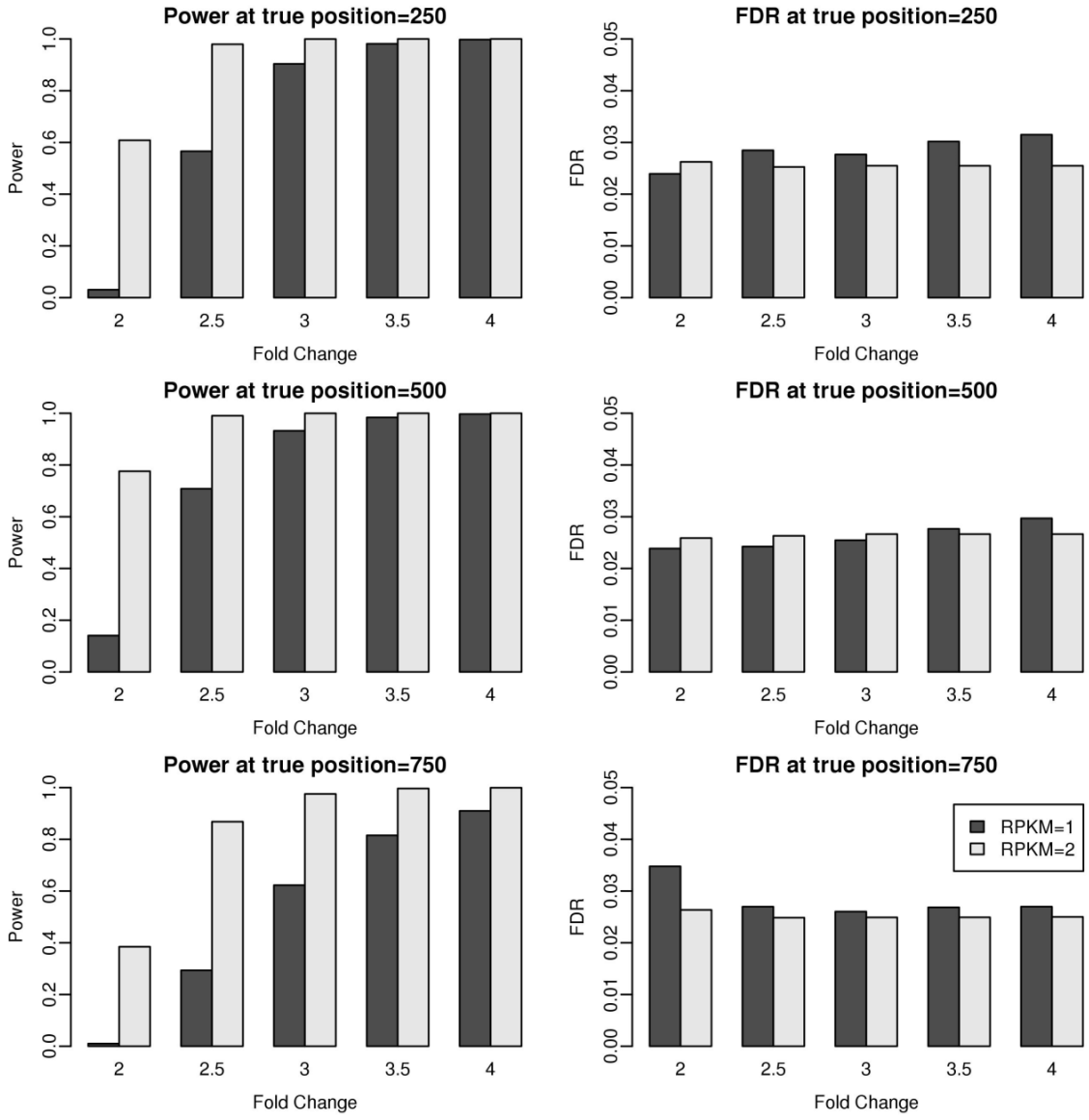

Figure 7.2 Power and FDR evaluation of the change-point model at the nominal level FDR=0.05.

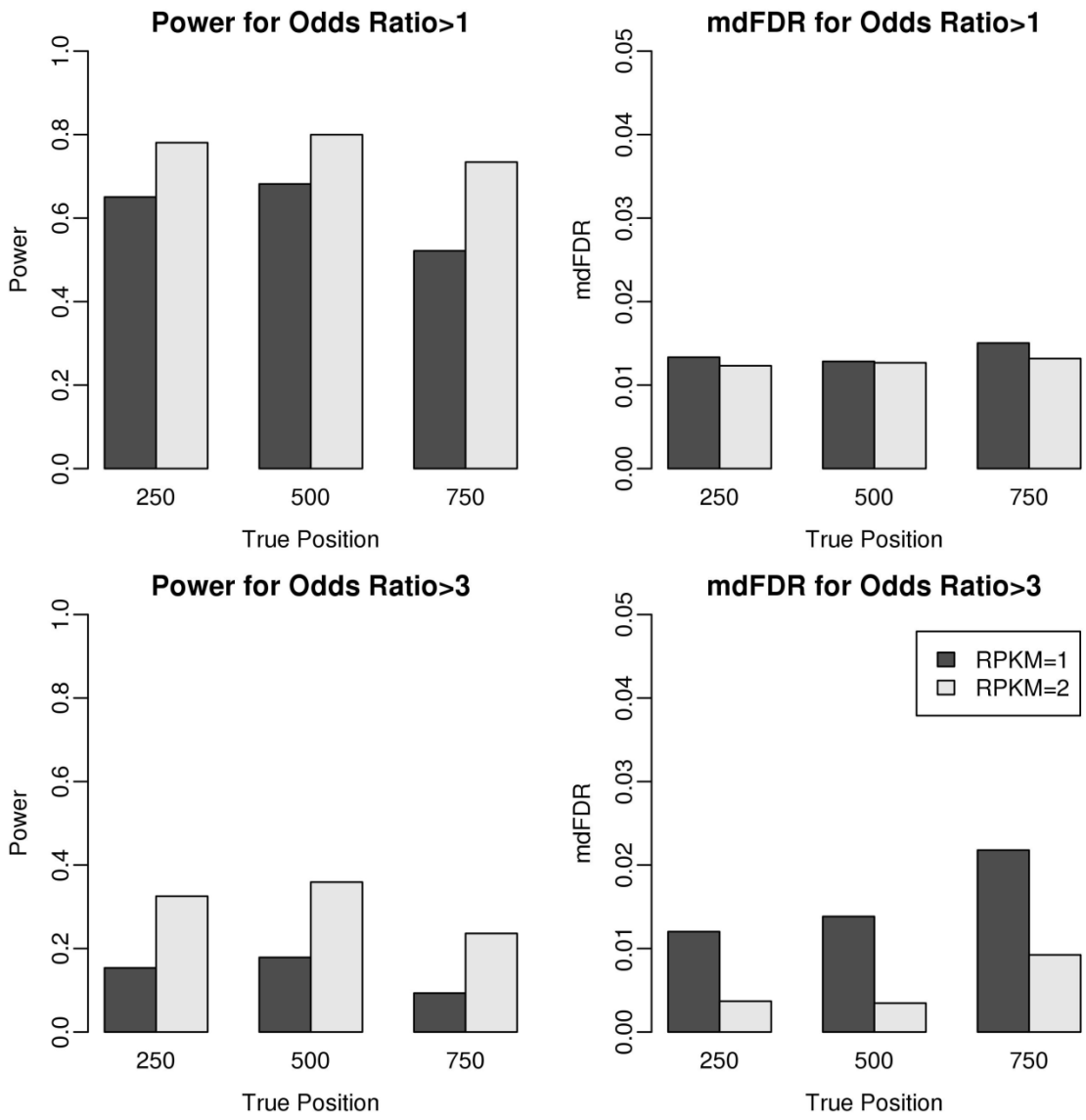

Figure 7.3 Power and mdFDR evaluation of the directional testing procedure at the nominal level mdFDR=0.05.

The simulation results are summarized in Figure 7.2. It can be seen that FDR was controlled at the nominal level=0.05 in all settings, suggesting that the proposed method is a valid testing procedure. Moreover, it can be found that the fold change, expression level and change point position all influence 3'UTR switching detection. First, the power of the proposed method increases with the fold change from small to large. This is expected because the change is more likely to be detected when the signal becomes stronger.

Second, the power increases when the gene expression level increases. Under the same fold change, the power of RPKM=2 is always higher than that of RPKM=1, suggesting that increasing the number of reads that are covered in the 3'UTR will also benefit change detection. Third, the position of the change point has an impact on the performance too. The change point in the middle yielded the highest power, compared to the change points close to the two ends.

Next, the power and mdFDR are evaluated for the proposed two-step testing framework. To simulate alternative hypotheses with mixed odds ratios, it used similar simulations as above but with the following modifications. For the 500 3'UTRs with fold change, they were divided into two groups with 250 each. The fold change for the first group is uniformly distributed from 1 to 3, and the second group is uniformly distributed from 3 to 5. *d=1* and *d=3* to test the changes was set with *OR>1* and *OR>3*, respectively. The proposed directional testing procedure at mdFDR level=0.05 was applied.

As seen in Figure 7.3, the proposed testing framework is able to control mdFDR at the nominal level=0.05 for all the settings. Similarly, the power increases when the expression level doubles from RPKM=1 to RPKM=2, and the change point at the middle position is easier to detect than those closer to the two ends. It is noted that when the hypothesized odds ratio is changed, the results change accordingly. For example, if the interest is detecting the 3'UTRs with *OR>3* by setting *d=3*, the testing procedure then favors the second group of 3'UTRs with $OR \sim unif(3,5)$ and would not reject the 3'UTRs in the first group with $OR \sim unif(1,3)$. When setting *d=1* for testing the 3'UTR changes with *OR>1*, the 3'UTRs in the group *with* $OR \sim unif(3,5)$ are easier to detect because the signal is relatively stronger than that of *d=3*. This explains the power difference between

testing *OR>1* and *OR>3* as shown in Figure 7.3. In summary, it is easier to capture the switching events when the *OR* is higher, the expression level is higher, or the change-point is closer to the middle.

7.4 Real Data Applications

7.4.1 Application to Regular RNA-Seq Data

The proposed method has been applied to analyze regular RNA-Seq data that have been commonly produced to profile transcriptome changes. MYC is a notable transcriptional factor that has been frequently activated in many human cancers with profound cellular influence. Although MYC-binding sites and target genes have been documented extensively in the past decade, thanks to the widespread application of high-throughput technology, the role of MYC and MYC target genes in androgen-controlled breast cancer growth remains unclear. To elucidate MYC regulatory network in molecular apocrine breast cancers, Ni and colleagues employed RNA-Seq to profile transcriptome changes before and after MYC knockdown by siRNA in MDA-MB-453 breast cancer cells with androgen stimulation (Ni, et al., 2013). In summary, they transfected MDA-MB-453 breast cancer cells with control (siCtrl) or MYC siRNA (siMYC) for 48 h, followed by treatment with 10nM DHT (the most potent androgen) or vehicle (veh) for 6 h, resulting in three conditions: siCtrl-veh, siCtrl-DHT and siMYC-DHT. High-throughput 50bp single-end sequencing was performed on Illumina HiSeq 2000 platform for each sample, generating total numbers of short reads ranging from ~26 million to ~39 million. Following the authors, two comparisons was made, siCtrl-DHT vs siCtrl-veh and siMYC-DHT vs

siCtrl-DHT, but to detect 3'UTR shortening events instead of gene expression level changes.

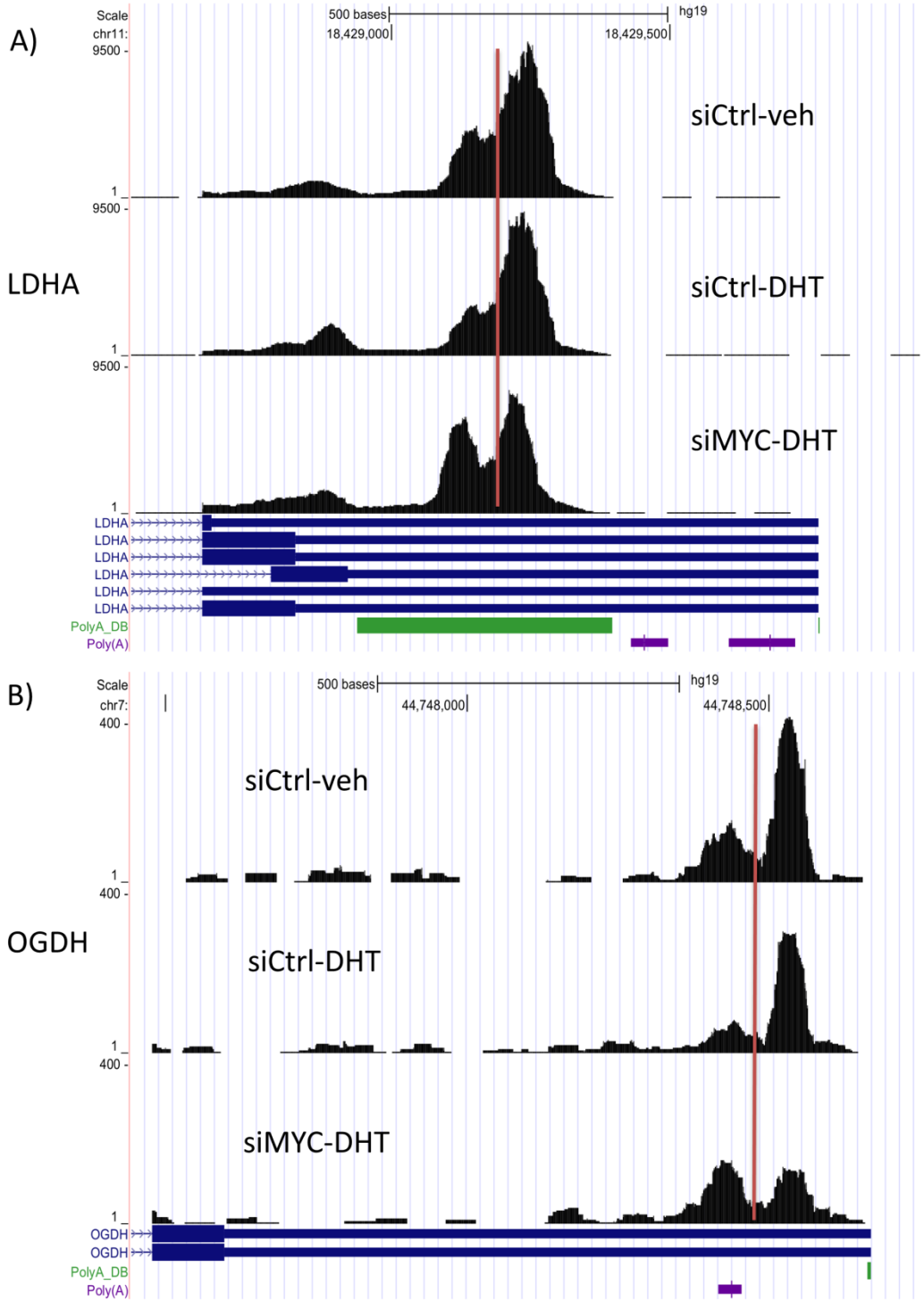

Figure 7.4 Examples of two MYC-dependent 3'UTR shortening events. The vertical lines indicate the estimated change points predicted by the proposed model.

The dataset was downloaded from NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under GSE45202. The raw reads were aligned to hg19 reference genomes using a conventional RNA-Seq aligner Tophat (Trapnell, et al., 2009) v1.3.1 with default parameters. Coverage filter can help to reduce false positives and is a heuristic strategy commonly used in existing RNA sequencing tools and analyses. Following MISO (Katz, et al., 2010), the analysis required that each 3'UTR should have at least 20 supporting reads in both samples, leading to 8052 and 7878 genes in the two comparisons, respectively, for further analysis. The proposed method was applied to detect shortening events with odds ratio>2 at an mdFDR level of 0.05. The analysis identified 947 shortening 3'UTRs in siCtrl-DHT vs siCtrl-veh and 1524 shortening 3'UTRs in siMYC-DHT vs siCtrl-DHT, respectively, with 461 genes in common. The 1063 genes unique in the comparison of siMYC-DHT vs siCtrl-DHT but not siCtrl-DHT vs siCtrl-veh may be associated with MYC knockdown given the DHT treatment. Figure 7.4 shows two examples of significant MYC-dependent shortening events, LDHA and OGDH, on the UCSC genome browser, demonstrating that the proposed method worked well in detecting such 3'UTR switching without relying on any PolyA annotations. It can be observed that a highly non-uniform distribution of data in the 3'UTR, a common phenomenon in RNA-Seq data which may be caused by PolyA mRNA selection bias (Wang, et al., 2009). The PolyA track in the genome browser was included, which showed the annotated PolyA sites from the PolyA_DB. It can be seen that dramatic changes before and after the predicted change-points. Clearly, the two genes LDHA and OGDH tend to use the proximal PolyA site instead of the distal site in siMYC-DHT. These change-points are also consistent and supported by the PolyA sites annotated in the PolyA_DB. Together, these

results suggest that the proposed method works well to detect 3'UTR switching without relying on any PolyA annotations.

LDHA catalyzes the conversion of L-lactate and NAD to pyruvate and NADH in the final step of anaerobic glycolysis. It has been shown to be highly correlated with breast cancer growth (Wang, et al., 2012). OGDH encodes one subunit of the 2-oxoglutarate dehydrogenase complex that catalyzes the overall conversion of 2-oxoglutarate (alpha-ketoglutarate) to succinyl-CoA and CO(2) during the Krebs cycle. It also plays an important role in breast cancer cells (Qattan, et al., 2012). These shortening genes may be worthwhile for further biological study, because, the loss of distal region, if containing miRNA target sites, may help escape degradation destiny or translational repression.

A gene set enrichment analysis (GSEA) of these 1063 MYC-dependent shortening genes was conducted using a hypergeometric test. The canonical pathways definitions were downloaded from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). The results are summarized in Table 7.1. It has been suggested that MYC plays a crucial role in several aspects of cellular function, such as metabolism, growth, replication, differentiation and apoptosis (Ni, et al., 2013). These pathway results suggest very interesting transcription relevant functions of these 1063 MYC-dependent shortening genes, such as splicing, intron processing and transcript elongation. These genes are primarily associated with mRNA processing and gene expression, which are critical in cancer development (David and Manley, 2010; Sotiriou, et al., 2006). The original studies (Ni, et al., 2013) focused on conventional differential expression analysis. Capturing 3'UTR switching from the same RNA-Seq data

set using the proposed method would shed additional light on cancer transcriptome regulations and suggest new roles of MYC.

Table 7.1 Significantly Enriched Canonical Pathways in Analysis of the Breast Cancer Dataset of (Ni, et al., 2013) at FDR=0.05

Canonical Pathway	P Value
REACTOME_MRNA_SPLICING	3.74E-05
REACTOME_GENE_EXPRESSION	4.99E-05
REACTOME_PROCESSING_OF_CAPPED_INTRONCONTAINING_PRE_MRNA	7.74E-05
BIOCARTA_PROTEASOME_PATHWAY	1.06E-04
REACTOME_FORMATION_AND_MATURATION_OF_MRNA_TRANSCRIPT	1.32E-04
REACTOME_METABOLISM_OF_PROTEINS	2.51E-04
REACTOME_ELONGATION_AND_PROCESSING_OF_CAPPED_TRANSCRIPTS	2.55E-04
KEGG_OXIDATIVE_PHOSPHORYLATION	3.35E-04
REACTOME_TRANSLATION	6.15E-04
KEGG_CARDIAC_MUSCLE_CONTRACTION	7.68E-04
REACTOME_INFLUENZA_LIFE_CYCLE	9.18E-04

To compare to existing methods applicable for the 3'UTR switching analysis but relying on PolyA annotations, MISO was also run (Katz, et al., 2010) (version 0.4.1 with default parameters) to analyze this RNA-Seq dataset for identifying 3'UTR shortening events. It filtered tandem 3'UTR events following the MISO manual as follows: (a) at least 1 inclusion read, (b) 1 exclusion read, such that (c) the sum of inclusion and exclusion

reads is at least 10, and (d) the $\Delta \Psi$ is at least 0.2 and (e) the Bayes factor is at least 10, and (a)-(e) are true in one of the two samples. MISO didn't output any tandem 3'UTR events, although it did report other alternative splicing events, such as skipped exons, intron retentions, etc. This shows that methods depending on PolyA annotations may suffer from low power in 3'UTR switching analysis. The capability of the method for detecting 3'UTR switching will fill a void among current alternative splicing and processing analysis tools.

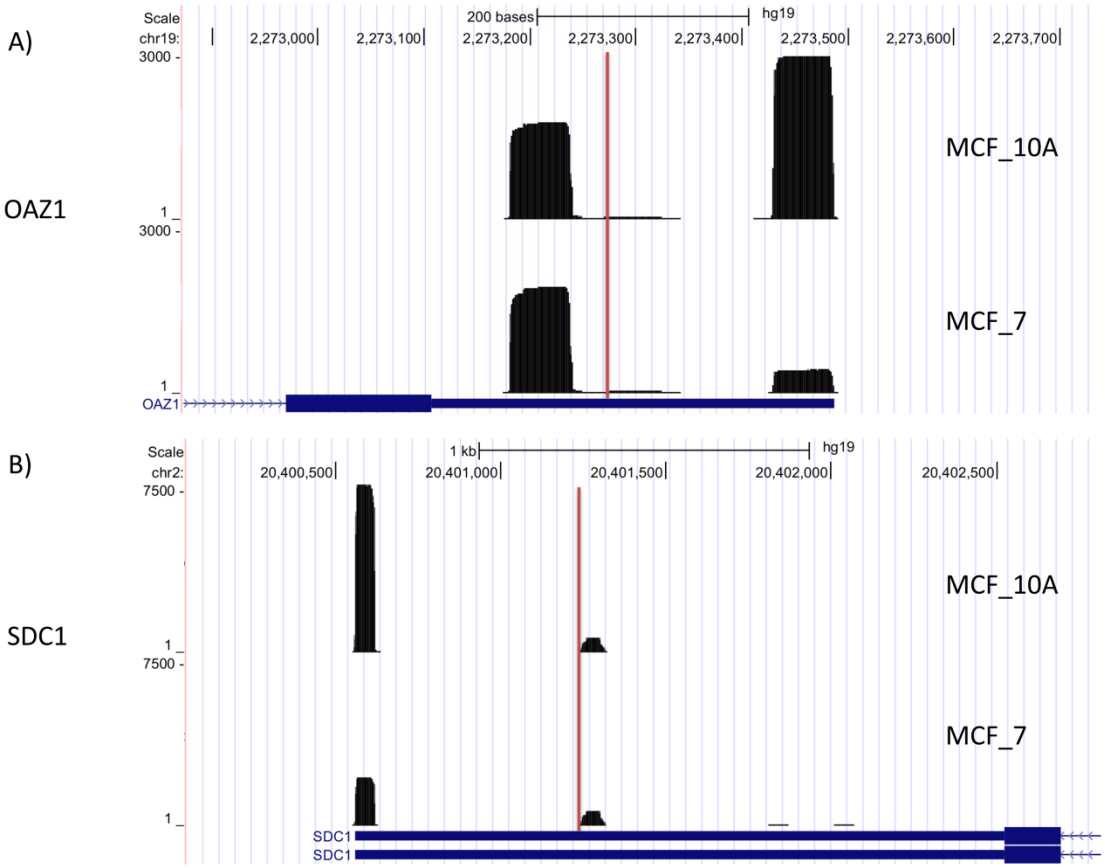

Figure 7.5 Examples of two shortening events that were identified by the method but missed by the linear trend test. The vertical lines indicate the change-points predicted by the proposed model.

7.4.2 Application to Special RNA-Seq Data

Another breast cancer dataset (Fu, et al., 2011) was analyzed to highlight the flexibility of the proposed method to handle special RNA sequencing data. To improve efficiency of capturing APA sites, Fu and colleagues developed a novel strategy to sequence only reads with Poly(A) tails followed by a linear trend test method for analyzing APA site switching (Fu, et al., 2011). Specifically, they modified oligo-d(T) tagged with sequencing primers after PCR to sequence polyadenylated reads. They performed the SAPAS method to profile APA sites of human breast cancer lines and compared with normal cell lines, generating in total ~31 million short reads with 75bp length from the Illumina platform. The dataset was downloaded from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) (accession number SRA023826). PolyA containing reads cannot be mapped to the genome directly. Therefore, Bowtie2 (Langmead and Salzberg, 2012) was used by local model to align those PolyA containing reads, because this model does not require end-to-end mapping. The proposed method was applied to identify 3'UTR shortening events. The authors reported their results at FDR level=0.01. To make a comparison, shortening events at the same mdFDR level of 0.01 for *OR>1* were reported. It has been identified 972 shortening events in the breast cancer cell line (MCF_7) in comparison with the control sample (MCF_10A). Their linear trend test method was conservative according to the authors, and detected only 428 shortened 3'UTRs (Fu, et al., 2011). It has been found that 85% of their shortening genes were also detected as shortening by the proposed method. The larger numbers of 3'UTR shortening events it identified under the same significance level suggest the higher power of the proposed method.

Table 7.2 Significantly Enriched Canonical Pathways in Analysis of the Breast Cancer Dataset of (Fu, et al., 2011) at FDR=0.05

CANONICAL PATHWAY	P Value
REACTOME_MRNA_SPLICING	3.83E-05
REACTOME_ELONGATION_AND_PROCESSING_OF_CAPPED_TRANSCRIPTS	1.78E-04
KEGG_CELL_ADHESION_MOLECULES_CAMS	1.88E-04
KEGG_SPLICEOSOME	2.11E-04
REACTOME_DIABETES_PATHWAYS	2.25E-04
BIOCARTA_EIF_PATHWAY	4.07E-04
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	5.36E-04
REACTOME_FURTHER_PLATELET_RELEASATE	7.13E-04

To demonstrate the accuracy of these findings, the four genes that were validated in their studies were examined. All the four genes, DDX5, SEC61A1, HSBP1 and FAM134A, were detected to be shortening in MCF_7 by the proposed method. The shortenings of these four genes were all experimentally confirmed (see the PCR results in the Supplementary of (Fu, et al., 2011)). Moreover, visualization of the identified shortening events highlighted the accuracy of the prediction. Figure 7.5 shows two genes OAZ1 and SDC1 that were missed by the linear trend method but demonstrated clear shortening patterns. Both genes are known to be related with cancer (Kastl, et al., 2010; Nikolova, et al., 2009). Finally, the GSEA for these 972 shortening genes was conducted. The results are summarized in Table 7.2. Interestingly, it has also found the mRNA splicing pathway to be the most significantly enriched in this breast cancer dataset, as in the

first breast cancer dataset that analyzed in the previous section. In particular, these genes are related to splicesome, a large ribonucleoprotein complex that guides pre-mRNA splicing in eukaryotic cells. Recent studies have demonstrated the contribution of splicesome as a core component in oncology (Quidville, et al., 2013) and its role in determining 3'UTR length (Berg, et al., 2012). Taken together, these results indicate the accuracy of the proposed method in capturing 3'UTR switching. Overall, this real data application highlights the flexibility of the proposed method for analyzing NGS data that are specially generated to sequence and capture polyadenlyation cleavage sites.

7.5 Conclusion and Discussion

This chapter proposes a change-point model based on a generalized likelihood ratio statistic for identifying 3'UTR length change in the analysis of next-generation RNA sequencing data. It develops a directional multiple testing procedure for identifying dramatic shortening or lengthening events. The numerical performances of the approach are investigated using both simulated and real data. The results show that the proposed method is powerful, accurate, and flexible for analyzing various types of next-generation RNA sequencing data.

The proposed method can be improved in several ways. First, one limitation is that the current method cannot handle sample replicates. The extension is to compute joint likelihood over multiple samples, assuming the same change-point across samples but allowing $\hat{p}_0$, $\hat{p}_\tau$, and $\hat{p}$ to vary for different sample comparisons. Second, assume there are only two isoforms with one change point. The extension can be done for multiple isoforms with $K>1$ change points. In principle, it may search similarly for the K points that yield the

most significance with computational complexity of $O(L^K)$, where L is the whole UTR length. It may further assume K is unknown and determine it using model selection (Shen and Zhang, 2012). Third, statistical inference of confidence estimates is as important as point estimates. For example, the confidence intervals on the estimated change points could provide more information as needed for some downstream analyses, such as determining the loss/gain of miRNA target sites. This can be obtained based on the values accepted by a level α of likelihood ratio test (Worsley, 1986).

CHAPTER 8

CONCLUSION AND FUTURE WORK

This dissertation focuses on the development of computational methods for NGS data analyses. The main contributions of this dissertation are as listed below:

First, a statistical tool, SNVer, has been developed for calling common and rare variants in analysis of pooled or individual DNA-Seq data. It formulates variant calling as a hypothesis testing problem and employ a binomial-binomial model to test the significance of observed allele frequency against sequencing error. The reported p-values for candidate loci are particular desirable for multiplicity control. SNVer runs very fast, making it feasible for analyzing high-throughput NGS data.

Second, a graphical user interface(GUI)-based tool, SNVerGUI, has been implemented based upon SNVer model. Compared with other current methods for variant calling, SNVerGUI is unique in that it is applicable to both individual and pooled DNA sequencing data. It allows users who do not have bioinformatics expertise to perform sophisticated variant detection by a simple and user-friendly desktop tool.

Third, a gene-based genome-wide screening method based on collapsing singletons in a whole-genome sequencing dataset has been studied. It has been demonstrated that this strategy can boost signals for associating rare variants in the NGS sequencing analysis.

Fourth, a novel bioinformatics pipeline, PolyASeeker, has been developed to fill the void of identifying polyadenylation cleavage sites from increasingly popular RNA-Seq data. The novelties include a probabilistic scoring scheme to select candidate reads and use the mating information in paired-end data.

Finally, a change-point model based on a likelihood ratio test has been proposed to detecting 3'UTR switching from RNA-Seq data. This is the first available method that allows users to directly analyze 3'UTR length changes without relying on any prior information of PolyA sites. It also allows user to make directional decisions and control mixed directional FDR (mdFDR) at a pre-specified level.

Future work lies in the following directions:

First, it has been shown that there remains significant discrepancy in SNV and indel calling between many of the currently available variant-calling pipelines under near-default software parameterizations (O'Rawe, et al., 2013). This, therefore, demonstrates that further improvement of variant calling algorithm is necessary, especially for indel detection.

Second, the association study in this dissertation focuses on only genic regions using conventional gene annotation, which makes up little more than 1% of the genome. The recent annotation made by the ENCODE consortium has included more than 70,000 "promoter" regions and nearly 400,000 "enhancer" regions that regulate expression of distant genes, which account for roughly 80% of the genome (Dunham, et al., 2012)(Dunham, Kundaje et al. 2012). This new knowledge can be used in future analysis.

Third, in addition to 3'UTR switching analysis, the proposed change-point method can also be extended to other applications. For example, one can merge together the multiple 3'UTRs of a gene, if any, to perform alternative last exon analysis. Moreover, if input vector is the coverage of entire exon regions of a gene, the proposed method can also detect premature cleavage and polyadenylation (PCPA) events, another set of interesting biological phenomena that has received much attention recently (Kaida, et al., 2010).

These phenomena can also be computationally confirmed by identifying PolyA sites using PolyASeeker.

REFERENCES

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467(7319):1061-1073.

Anders, S., Reyes, A. and Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;22(10):2008-2017.

Arning, L., *et al.* Identification and characterisation of a large senataxin (SETX) gene duplication in ataxia with ocular apraxia type 2 (AOA2). *Neurogenetics* 2008;9(4):295-299.

Bansal, V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 2010;26(12):i318-324.

Bansal, V., *et al.* Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010;11(11):773-785.

Bao, H., *et al.* MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* 2009;25(12):1554-1555.

Barski, A., *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* 2007;129(4):823-837.

Benjamini, Y. and Heller, R. Screening for partial conjunction hypotheses. *Biometrics* 2008;64(4):1215-1222.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* 1995;57(1):289-300.

Benjamini, Y., *et al.* False Discovery Rate: Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association* 2005;100(469):71-93.

Bentley, D.R., *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456(7218):53-59.

Bentley, J. Programming pearls: algorithm design techniques. *Commun. ACM* 1984;27(9):865-873.

Berg, M.G., *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 2012;150(1):53-64.

Calvo, S.E., *et al.* High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet* 2010;42(10):851-858.

Chang, X. and Wang, K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 2012;49(7):433-436.

Chen, Y.Z., *et al.* DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am J Hum Genet* 2004;74(6):1128-1135.

Cirulli, E.T. and Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;11(6):415-425.

Colgan, D.F. and Manley, J.L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* 1997;11(21):2755-2766.

Cooper, G.M. and Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;12(9):628-640.

Danecek, P., *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;27(15):2156-2158.

David, C.J. and Manley, J.L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev* 2010;24(21):2343-2364.

Daye, Z.J., Li, H. and Wei, Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res* 2012.

DePristo, M.A., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43(5):491-498.

Derti, A., *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* 2012;22(6):1173-1183.

Dickson, S.P., *et al.* Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8(1):e1000294.

Druley, T.E., *et al.* Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 2009;6(4):263-265.

Dunham, I., *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.

Flavell, S.W., *et al.* Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* 2008;60(6):1022-1038.

Fu, Y., *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* 2011;21(5):741-747.

Green, E.D. and Guyer, M.S. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470(7333):204-213.

Griebel, T., *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* 2012;40(20):10073-10083.

Griffith, M., *et al.* Alternative expression analysis by RNA sequencing. *Nat Methods* 2010;7(10):843-847.

Guo, W., Sarkar, S.K. and Peddada, S.D. Controlling False Discoveries in Multidimensional Directional Decisions, with Applications to Gene Expression Data on Ordered Categories. *Biometrics* 2010;66(2):485-492.

Hayden, E.C. International genome project launched. *Nature* 2008;451(7177):378-379.

Hindorff, L.A., *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106(23):9362-9367.

Hou, H., *et al.* MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Res* 2010;38(Web Server issue):W732-736.

Ingman, M. and Gyllensten, U. SNP frequency estimation using massively parallel sequencing of pooled DNA. *Eur J Hum Genet* 2009;17(3):383-386.

Jan, C.H., *et al.* Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 2011;469(7328):97-101.

Ji, Z., *et al.* Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106(17):7028-7033.

Ji, Z., *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol* 2011;7:534.

Kaida, D., *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 2010;468(7324):664-668.

Kastl, L., Brown, I. and Schofield, A.C. Effects of decitabine on the expression of selected endogenous control genes in human breast cancer cells. *Mol Cell Probes* 2010;24(2):87-92.

Katz, Y., *et al.* Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;7(12):1009-1015.

Keren, H., Lev-Maor, G. and Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 2010;11(5):345-355.

Koboldt, D.C., *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;25(17):2283-2285.

Kullback, S. and Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* 1951;22:79-86.

Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* 2011;470(7333):187-197.

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

Lee, J.S., *et al.* On optimal pooling designs to identify rare variants through massive resequencing. *Genet Epidemiol* 2011;35(3):139-147.

Lee, J.Y., *et al.* PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic acids research* 2007;35(Database issue):D165-168.

Lembo, A., Di Cunto, F. and Provero, P. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. *PLoS One* 2012;7(2):e31129.

Levin, J.Z., *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 2010;7(9):709-715.

Li, B. and Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics* 2008;83(3):311-321.

Li, B. and Leal, S.M. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 2009;5(5):e1000481.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-1760.

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

Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Li, H., Ruan, J. and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;18(11):1851-1858.

Licatalosi, D.D. and Darnell, R.B. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 2010;11(1):75-87.

Lin, Y., *et al.* An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* 2012.

Loman, N.J., *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;30(5):434-439.

Lyon, G.J., *et al.* Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications. *Discov Med* 2011;12(62):41-55.

Manolio, T.A., *et al.* Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-753.

Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* 2011;470(7333):198-203.

Martin, K.C. and Ephrussi, A. mRNA localization: gene expression in the spatial dimension. *Cell* 2009;136(4):719-730.

Mayr, C. and Bartel, D.P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 2009;138(4):673-684.

Meissner, A., *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;454(7205):766-770.

Mikkelsen, T.S., *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;448(7153):553-560.

Momozawa, Y., *et al.* Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat Genet* 2011;43(1):43-47.

Moore, M.J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* 2005;309(5740):1514-1518.

Mortazavi, A., *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;5(7):621-628.

Neale, B.M., *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;7(3):e1001322.

Nejentsev, S., *et al.* Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009;324(5925):387-389.

Ni, M., *et al.* Amplitude modulation of androgen signaling by c-MYC. *Genes Dev* 2013;27(7):734-748.

Nikolova, V., *et al.* Differential roles for membrane-bound and soluble syndecan-1 (CD138) in breast cancer progression. *Carcinogenesis* 2009;30(3):397-407.

Norton, N., *et al.* DNA pooling as a tool for large-scale association studies in complex traits. *Ann Med* 2004;36(2):146-152.

O'Rawe, J., *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013;5(3):28.

Okou, D.T., *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 2007;4(11):907-909.

Out, A.A., *et al.* Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat* 2009;30(12):1703-1712.

Ozsolak, F., *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 2010;143(6):1018-1029.

Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 2009;33(6):497-507.

Pickrell, J.K., *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;464(7289):768-772.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005;33(Database issue):D501-504.

Qattan, A.T., *et al.* Spatial distribution of cellular function: the partitioning of proteins between mitochondria and the nucleus in MCF7 breast cancer cells. *J Proteome Res* 2012;11(12):6080-6101.

Qi, J., *et al.* inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* 2010;26(1):127-129.

Quidville, V., *et al.* Targeting the deregulated spliceosome core machinery in cancer cells triggers mTOR blockade and autophagy. *Cancer Res* 2013;73(7):2247-2258.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-842.

Ren, M., *et al.* Target of rapamycin signaling regulates metabolism, growth, and life span in Arabidopsis. *Plant Cell* 2012;24(12):4850-4874.

Robinson, J.T., *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;29(1):24-26.

Rogers, M.F., *et al.* SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol* 2012;13(1):R4.

Sandberg, R., *et al.* Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 2008;320(5883):1643-1647.

Sboner, A., *et al.* The real cost of sequencing: higher than you think! *Genome Biol* 2011;12(8):125.

Service, R.F. Gene sequencing. The race for the $1000 genome. *Science* 2006;311(5767):1544-1546.

Sham, P., *et al.* DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002;3(11):862-871.

Shen, J.J. and Zhang, N.R. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *arXiv:1206.6627* 2012.

Shen, S., *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 2012;40(8):e61.

Sherstnev, A., *et al.* Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat Struct Mol Biol* 2012;19(8):845-852.

Smibert, P., *et al.* Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* 2012;1(3):277-289.

Sotiriou, C., *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;98(4):262-272.

Sun, W. and Wei, Z. Multiple testing for pattern identification, with applications to microarray time course experiments. *Journal of the American Statistical Association* 2011;106(493):73-78.

Teer, J.K., *et al.* VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics* 2011.

Teichert, I., *et al.* Combining laser microdissection and RNA-seq to chart the transcriptional landscape of fungal development. *BMC Genomics* 2012;13:511.

Tian, B., *et al.* A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 2005;33(1):201-212.

Trapnell, C., *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;31(1):46-53.

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

Ulitsky, I., *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res* 2012;22(10):2054-2066.

Vallania, F.L., *et al.* High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res* 2010;20(12):1711-1718.

Wang, E.T., *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;456(7221):470-476.

Wang, K., *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am J Hum Genet* 2010;86(5):730-742.

Wang, K., Li, M. and Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38(16):e164.

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

Wang, Z.Y., *et al.* LDH-A silencing suppresses breast cancer tumorigenicity through induction of oxidative stress mediated mitochondrial pathway apoptosis. *Breast Cancer Res Treat* 2012;131(3):791-800.

Wei, Z., *et al.* Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* 2009;25(21):2802-2808.

Wei, Z., *et al.* SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* 2011;39(19):e132.

Williams, V.S., Jones, L.V. and Tukey, J.W. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* 1999;24(1):42-69.

Worsley, K.J. The Power of Likelihood Ratio and Cumulative Sum Tests for a Change in a Binomial Probability. *Biometrika* 1983;70(2):455-464.

Worsley, K.J. Confidence Regions and Tests for a Change-Point in a Sequence of Exponential Family Random Variables. *Biometrika* 1986;73(1):91-104.

Wu, M.C., *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* 2011;89(1):82-93.

Zhang, H., *et al.* PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res* 2005;33(Database issue):D116-120.

Zhang, Z., *et al.* Dysregulation of synaptogenesis genes antecedes motor neuron pathology in spinal muscular atrophy. *Proceedings of the National Academy of Sciences of the United States of America* 2013;110(48):19348-19353.