STATISTICAL MODELS AND SNP DETECTION METHODS FOR FLASH SEQUENCING

by

Qinglin Pei

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2013

Date of final oral examination: 09/26/2013

The dissertation is approved by the following members of the Final Oral Committee:

Michael A. Newton, Professor, Statistics

David C. Schwartz, Professor, Chemistry and Genetics

Karl Broman, Professor, Biostatistics

Cecile Ane, Associate Professor, Statistics and Botany

Christina Kendziorski, Professor, Biostatistics

ACKNOWLEDGMENTS

Many thanks to my advisor Professor Michael A. Newton and our collaborator Professor David C. Schwartz

DISCARD THIS PAGE

TABLE OF CONTENTS

Page

Page

Appendix

Page

DISCARD THIS PAGE

LIST OF TABLES

Table Page

DISCARD THIS PAGE

LIST OF FIGURES

Figure Page

Figure Page

Appendix
Figure Page

Appendix
Figure Page

Appendix
Figure Page

Appendix
Figure Page

Appendix
Figure Page

ABSTRACT

Flash sequencing is a new technology for whole genome analysis. In the absence of any cycles of biochemistry, it queries large numbers of single DNA molecules for sequence properties near sites in a dense set of physical map locations. These locations are cognate sites of two nicking enzymes: one that reveals a medium-resolution map ($\approx 8kb$) at full digestion, and a second that reveals a random sample of a high-resolution map ($\approx 16b$) at partial digestion. Nick translation and sequence-specific fluorochrome labeling in a neighborhood of each nick site provide partial sequence information. The medium-resolution nick site structure of the first nicking enzyme provides information for the molecule alignment and is used to detect structural variations. The related algorithm and statistical analytical methods have been well studied by previous research [6]. The high-resolution nick site map of the second nicking enzyme provides information which is missed by the data generated by the first nicking enzyme. Our focus is on the observations coming from this high-resolution map. From a statistical perspective the high-resolution map yields a large number of bivariate measurements: each one records the position of a nick site and a sequence-content optical signature in that site's neighborhood. We propose a statistical model for these bivariate data in which uniform discrete mixing over nick sites describes partial digestion and in which location and signature errors are independent Gaussians. We develop a numerical strategy suited to testing the *single-nucleotide polymorphism* (SNP) content of a test genome. The algorithm identifies *regions of interest* (ROI) per potential SNP by efficiently localizing a subset of the bivariate sample space where the sampling model differs highly between the reference genome and the test genome, which is generated by applying this SNP to the reference genome. This approach leads to much

higher detection rates than the classical test statistics which involve very many degrees of freedom while failing to exclude the effects caused by nearby SNPs. A sequential SNP selection algorithm is proposed to identify the full SNP genotype. The methodology is tested by performance evaluation on sets of synthetic data which are generated based on the empirical error distributions of bi-variate observations of the second nicking enzyme.

In the first chapter, we introduce the *flash sequencing system*(FS). First, a basic philosophical question is addressed: what are the potential advantages of FS compared with other next-generation sequencing methods? Secondly, the mechanism and experimental protocols of FS are illustrated, together with a description of the data format of the observed values, which areb bivariate measurements of the nick sites generated by the second nicking enzyme.

In the second chapter, we propose a model for the FS data. The empirical error distributions of the bivariate measurements are built based on two pilot experiments. Since FS is a novel genome sequencing approach which is still under development, there are little experimental data available besides the two pilot experiments at this moment. For this reason we use the empirical error distributions of the bivariate observations to generate sets of synthetic data for the test genome. In this thesis, we use the human build 36 as the reference genome; while the test genome is generated by applying the 3.3 million James Watson SNPs [5] to the human build 36. With these sets of synthetic data, we aim to explore the feasible statistical methods that can reveal the mutations of the test genome based on FS observations.

In the third chapter, a statistical algorithm is proposed to detect the mutations of the test genome. The methodology is applied to sets of synthetic data. For each set of synthetic data, the detection rate is defined to be the percentage of discovered variations among those on the test genome. The detection rates are highly related to the number of molecules measured in experiment. This relationship will be the major issue discussed in this chapter. To calculate the detection rates, we start from the simplest situation when there is only SNP in the target range of the test genome.

Then we consider the more general situations when multiple SNPs exist in the same range. When multiple SNPs happen in a very short region, their interactions need to be considered. Meanwhile, a competing mechanism is employed to select the subset of SNPs with highest likelihood from a given SNP database.

The last chapter lists some possible improvements to FS. We will discuss the effect of more precise measurements and incomplete databases of SNPs.

Chapter 1

Genomic Background and Overview of Thesis

1.1 Introduction of Flash Sequencing System

The analysis of genomes and their variation is central to much of biomedical science. A sampled genome is fully described by the complete ordered base sequence of its constituent chromosomes; however, the cost and difficulty of obtaining such data with current technologies continue to be prohibitive [11]. A reliable reduced-information representation of the genome would be useful in many applications, just as an MP3 recording is a lossy but effective representation of a true audio signal[1]. For example, we might bin the genome and report the relative frequency of each base in each bin. Variation among genomes in these frequency distributions reflects some degree of variation of the whole genome, and would be valuable in many biomedical applications. Furthermore, reduced-information data often complement data produced by direct sequencing.

With sequencing information, we can reveal cancer related gene polymorphism or mutations, which are of significant importance in many research areas. Numerous studies were conducted to analyze these variations and explore possible solutions, such as gene therapy. For example, germline mutations have been discovered to be related with several female malignancies, including breast, ovarian and endometrial cancers [24]. And gene therapy starts to make a contribution to the treatment of debilitating, highly penetrant genetic diseases that have proved intractable to other regimens[27].

In the past 40 years, DNA sequencing technology has been developed at a rapid speed and has a significant effect on the acceleration of biological research and discovery. From the early $1970s$, when the first DNA sequence was obtained, tremendous efforts were conducted to sequence DNA molecules easier and faster. In 1977, two important papers [14] [15] by Fred Sanger used the dideoxynucleotide triplhosphates (ddNTPs) as DNA chain terminators. This first generation Sanger method dominated the DNA sequencing field for nearly 30 years. However, it did not satisfy the need for a faster and more economical sequencing. The *Next Generation Sequencing* (NGS) methods emerged in 2005 with the publication of the sequencing-by-synthesis technology developed by 454 life science [23]. NGS, as a new generation of the non-Sanger-based sequencing technology, is more powerful and cost effective. For example, it will cost less than $\$2000$ for a human genome at depth $13,000X$ by Illumina MiSeq system[2].

NGS greatly benefits research activities in genomic research, such as the 1000 Genome Project. Launched in 2008, this international project has gathered the most detailed database of human genome variations [9]. Its goal is to find the majority of all genomic variations in existence among people around the world. The publicly available database now contains genome data from 1700 people. The 1000 Genome project generates many promising biomedical research opportunities.

All these exciting databases and research activities are based on genome sequencing methods which can detect all kinds of genomic variations. Figure 1.1 is reproduced from [29] with the recent read lengths of several well known NGS systems[22]. It shows the possible variations including SNPs, indels,inversions, translocations, copy number of variations (CNVs), segment duplications, etc. The existing NGS platforms have great difficulties to detect the structural variations in the range from $1kb$ to $3Mb$ because their read lengths are usually less than $3kb$. There are several problems with short read lengths:

1. The alignment of the test genome to the reference genome is a significant computational challenge.

2. The shorter the reads are, the less information we can obtain for the alignment and the chance of mismatching will be bigger.

3. A large portion of NGS short reads cannot be uniquely aligned to the reference genome when the reads are too short and the reference is too complex.

4. The repeats in reference genome also reduce the chance of unique alignment.

Thus, a sequencing system capable of resolving structural details is in high demand.

Optical mapping (OM) is a single-molecule non-sequencing approach to genomic analysis. It measures a reduced-information version of genome and has a number of advantages deriving from its handling of long genomic molecules[6]. In OM, recognition sites of a restriction endonuclease enzyme define a whole genome restriction map for the genome on test. A map is assembled computationally from millions of DNA molecules that are sheared from the test genome, isolated and anchored to a surface, chemically processed, and individually image-analyzed. Genomic variations become evident when the map is compared to the *in silico* restriction map of a reference genome. OM provides an unparalleled view of structural genomic variation (e.g., insertions, deletions, translocations), and has been critical to whole-genome assembly efforts where repeat regions confuse sequence assembly. OM has been used in many applications, such as construction of whole-genome restriction maps of bacteria[31], parasites[3] and fungi[16]. Technological advances in the Schwartz lab have extended OM to the analysis of larger eukaryotic genomes, including human[19] and mouse[12]. Although advances in surface chemistry, microfluidics, instrumentation and algorithms make OM viable for the analysis of complex eukaryotic genomes, there are technical challenges that limit its resolution and throughput.

A new system, *next generation optical mapping system* (NGOM), was proposed by replacing the restriction enzymes in OM with nicking enzymes, which only cut one strand of a double-stranded DNA molecule at specific nick sites. The intact nicked DNA molecules become decorated with labeled nucleotides that replace DNA locally in a neighborhood of the nick site through the

process of nick translation. The nick site map is used to align molecules and detect structural variations of the test genome from the reference genome. The key reason for using nicking enzymes instead of restriction enzymes is the labeling of nick sites prior to anchoring to the surface, thus minimizing surface chemistry as in OM, and enabling more precise measurements. Also there is very little manipulation of the genomic DNA (e.g. no cycles of biochemistry). Much like OM with restriction maps, this NGOM can also be used to detect structural variations. Meanwhile, it shares same intrinsic limitations as OM that there is no information gathered between the nick sites.

Flash sequencing system (FS), under development in Professor Schwartz's lab, aims to overcome the limitations of OM or NGOM. Like NGOM, FS uses a nicking enzyme to cut a single strand of DNA rather than cleave both strands. FS obtains genomic information from many individually-analyzed DNA molecules sheared from a test genome. Like OM, FS also provides a genome-wide physical map composed by nick sites generated by a medium resolution nicking enzyme, which is referred to as the *mapping enzyme*. Meanwhile FS uses a second nicking enzyme that cuts a single strand of DNA with much higher distance density, which we call the *sequencing enzyme*. The sequencing enzyme provides the missing information between the nick sites of the mapping enzyme.

Since the mapping enzyme data of FS is similar to the restriction maps of OM, FS inherits the advantages of OM in detecting structural variations. The algorithm and analytical methods developed for OM can be applied directly to the mapping enzyme data of FS. Our focus in this thesis is on the additional data generated by the sequencing enzyme. Next we compare FS with NGS, OM and NGOM to illustrate the potential advantages of FS.

The first key advance shared by OM, NGOM and FS is that they have a much greater ability to detect structural mutations compared with existing commoditized NGS methods, such as Illumina's Genome Analyzer, Life Technologies' SOLiD system, Roche's 454 GS FLX Helicos' Heliscope Sequencer, Pacific BioSciences PacBio platform, and Life Technologies' Gen-3 system.

Table 1.1: The comparison of NGS, OM, NGOM, and FS

	NGS	OM	NGOM	FS
Long read length	N	Y	Y	Y
Ability to detect structural variations	Weak	Strong	Strong	Strong
Enzyme Type		Restriction	Nicking	Nicking
Information Within Interval		N	N	Y

Despite the existing success of these NGS methods, they carry common limitations in detecting structural variations because of the short read lengths. A comparison of several NGS platforms is shown in Figure 1.1. The read lengths of existing commoditized NGS systems are usually less than 3Kb. As stated in [29], structural variations typically affect a sequence with a length ranging from $1Kb$ to $3Mb$, and approximately 5% of the human genome are defined as structurally variant in the normal population, involving more than 800 independent genes. The short read lengths of NGS limit the ability of these platforms to detect structural variations. On the other hand, OM, NGOM and FS have much more powerful ability to detect the structural variations because their effective read lengths can be several Mbs.

In contrast to OM, which uses surfaces to isolate and image the individual molecules, FS and NGOM use nicking enzymes, which keeps one strand connected. An important advance shared by FS and NGOM is that biochemistry happens essentially in one tube with DNA in solution rather than on a surface. The labeled molecules are then presented for measurements.

In contrast to NGOM, which only has one nicking enzyme in the system, FS uses two distinct nicking enzymes. The mapping enzyme operates at a medium resolution at full digestion; the corresponding nick site map of a molecule is used to align it to the reference genome. The very high resolution of the sequencing enzyme prohibits optical measurement if nicking and labeling

are processed at full digestion. Instead, each molecule is labeled at a random subset of the potential nick sites when the digestion rate is limited to some partial value. By compiling data from a large number of single-molecule profiles, there emerges a distribution reflecting information in the high-resolution map. In this sense, FS not only collects information from mapping enzyme data as in NGOM, but also gathers data from the regions between the nick sites of the mapping enzyme.

In summary, FS has the potential to be a cost effective genome analysis method that is able to detect genomic variations and produces reduced representations of the gnome. Although FS does not seek to obtain the whole base sequence, the mapping enzyme data and additional data generated by the sequencing enzyme reveal massive information about the test genome. Statistical models can be used to extract important mutation information of the test genome. Any deviation of the test genome from the reference genome can alter the underlying distributions of FS data.

The form of polymorphism and mutations can be SNP (Single-nucleotide polymorphism), indels (insertion and deletion), or structural variations, such as translocation, frame shift, and copy number variation. FS has capability to detect all sorts of variations, among which SNPs are the most difficult to detect by FS. We will focus on SNPs in the subsequent analysis.

1.2 Further Details of Flash Sequencing

FS is a new technology for whole genome analysis. In the absence of any cycles of biochemistry, it queries large numbers of single DNA molecules for sequence properties near sites in a dense set of physical map locations. These locations are cognate sites of two nicking enzymes. The mapping enzyme reveals a medium-resolution nick site structure ($\approx 8kb$) at full digestion. The sequencing enzyme reveals a random sample of a high-resolution map ($\approx 16b$) at partial digestion. Nick translation and sequence-specific fluorochrome labeling in a neighborhood of each nick site provide some degree of the sequence information. From a statistical perspective, the sequencing enzyme of FS yields a large number of bivariate measurements: each one records the

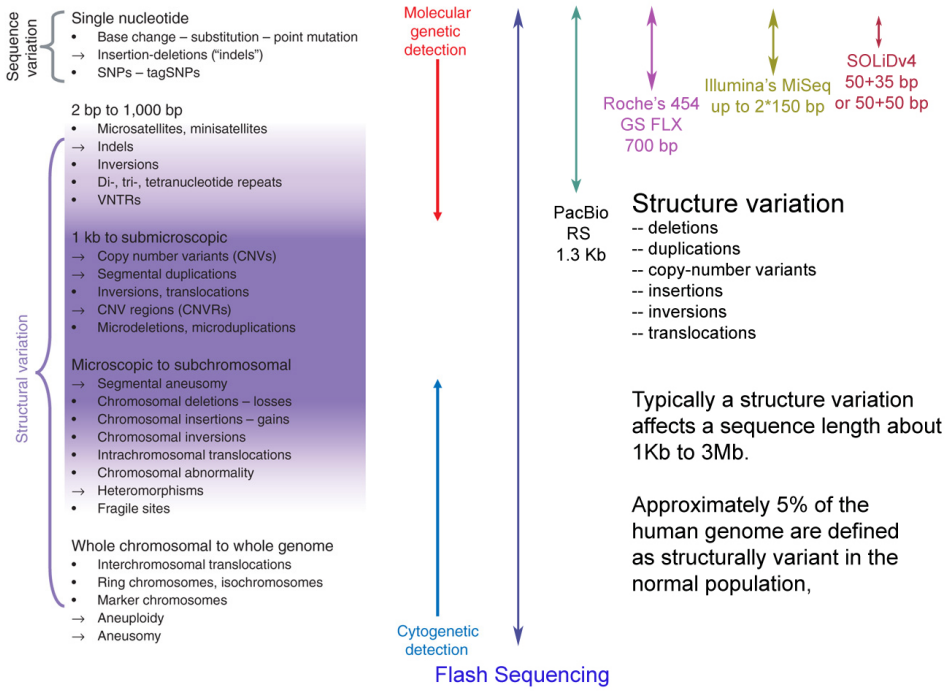

Figure 1.1: The read lengths of existing NGS methods are usually less than $3Kb$, while the structural variations range from $1Kb$ to $3Mb$[29] and are very difficult to detect by these existing NGS methods because of their short read lengths[22]. On the other hand, the Flash Sequencing System (FS) can provide sufficient long read length that covers the range of structural variations.

peak value of the sequence-content optical intensity and the approximate location of a nick site. These bivariate observations are called *signature* and *position* pairs. In the following we discuss the measurement procedure in more detail.

1.2.1 Mapping Enzyme

The mapping enzyme generates a medium resolution physical map of the genome on test. This is used to align the test genome to the reference genome. It recognizes a specific nucleotide sequence and afterwards cuts one strand of the DNA duplex. An example used in Professor Schwartz's lab is Nt.SapI, which recognizes the nucleotide pattern "GCTCTTCN||", where "N" denotes any one of the four nucleotides. The regions between two recognition sites of the mapping enzyme are called *intervals*. When Nt.SapI is used as the mapping enzyme, the average length of intervals is approximately 8000 bases in the human genome.

Visualizing and recording the nicked sites involve fluorescence microscopy and a process called *nick translation*, during which each site revealed by the nicking enzyme has certain nucleotides replaced by fluorochrome labeled analogs. Various labeling schemes are possible; in each case DNA polymerase incorporates nucleotides from the nick site up to a sequence-specific termination site.

On a sufficiently long molecule, the nick sites of the mapping enzyme characterize that molecule's physical position in the reference genome. Each molecule will be aligned to an in silico map of the reference genome as shown in Figure 1.3. If the test genome carries mutations which break the nick site structure of the mapping enzyme, the alignment algorithm will map the corresponding molecules to the most similar region of the reference genome. The difference in the mapping enzyme data reveals structural variations in the test genome. The nick site structure of the mapping enzyme generated by FS is similar to the restriction maps generated by OM. The alignment algorithm and analytical method to detect structural variation are studied extensively for OM and thus

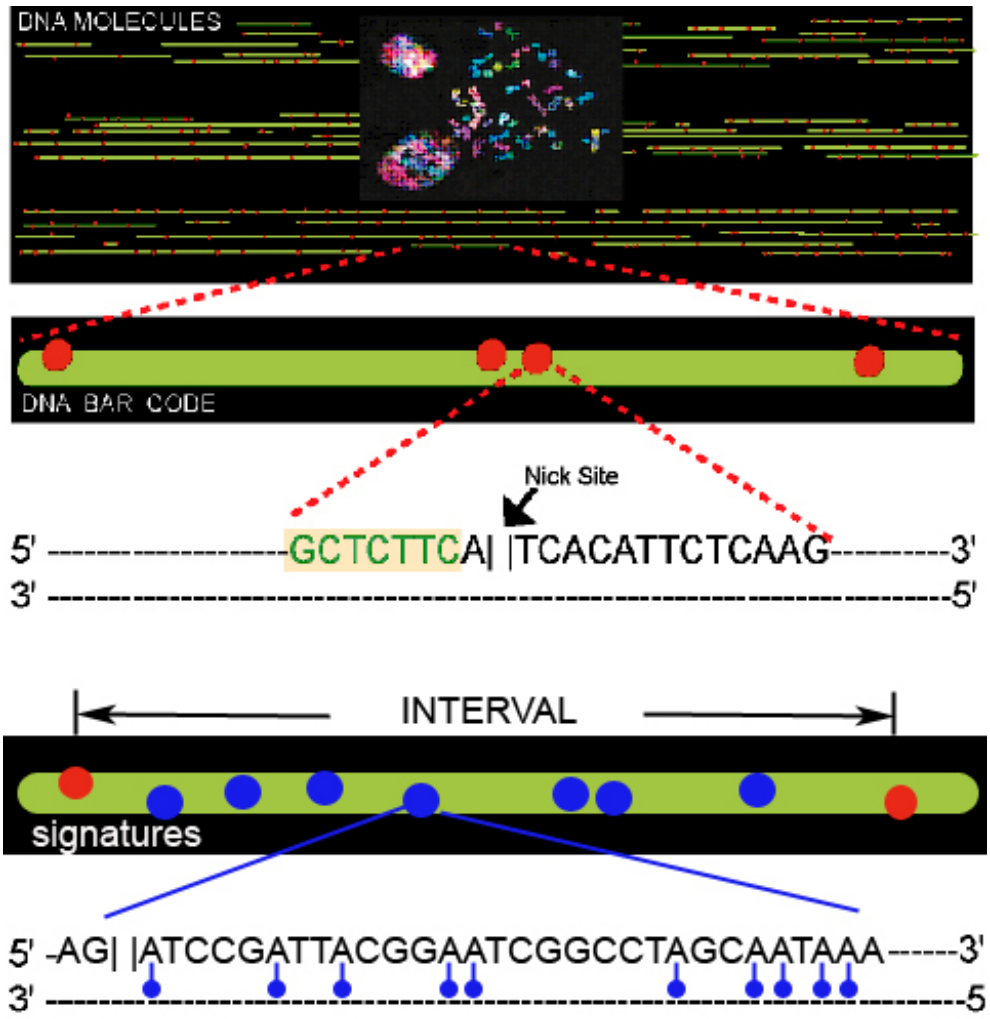

Figure 1.2: Top panel represents DNA molecules in solution and shows some aligned physical maps. The mapping enzyme cut at pattern GCTCTTCN||. The interval between two nick sites has an average length of 8000 bases in the human genome. Lower panel shows how the sequencing enzyme is used to obtain the genomic information within the intervals of the mapping enzyme.

are not the focus of the current work.

1.2.2 Sequencing enzyme

The mapping enzyme by itself does not generate high-resolution information within the intervals. To overcome this limitation, the FS employs a second nicking enzyme. Initial experiments use Nt.CviQII, which recognizes nucleotide patterns "R‖AG" (R=A or G). Such sites occur every 16 bases in the human reference on the average. An average-sized interval from the mapping enzyme contains about $8000/16 = 500$ cognate sites of the sequencing enzyme. The distance between these sites is too small to be measured by a standard microscope. Thus, a *partial digestion* mechanism is deployed. The effect is that a relatively small random subset of the potential nick sites are actually cut and revealed for nick translation of the sequencing enzyme on any given molecule. In Figure 1.2, the red dots represent the nick sites of the mapping enzyme, while the blue dots between the red dots represent the partial digestion of the sequencing enzyme nick sites. Each partially digested molecule reveals a potentially different random subset of the 500 nick sites. After alignment, the signature and position information are gathered by measuring these blue dots in each interval from large amount of molecules. Figure 1.4 indicates a sample of aligned molecules (aligned by the nick site structure of the mapping enzyme) and their labeled nick sites (blue). The top line in Figure 1.4 indicates the positions of the nick sites at full digestion, while the remaining lines are partially digested molecules from the test genome. Each copy of the molecules which are aligned to cover a certain interval will generate a few randomly selected nick sites of the sequencing enzyme inside of this interval. The signature and position pairs of the labeled nick sites are measured and form the observed data set.

The locus defined by nick translation of the sequencing enzyme (from nick site to termination site) is called the *neighborhood* of the nick site. The number of labeled nucleotides in the neighborhood is the underlying true value of the signature measurements of this nick site and the distance from the left end of the interval to the middle of the neighborhood is the true value of

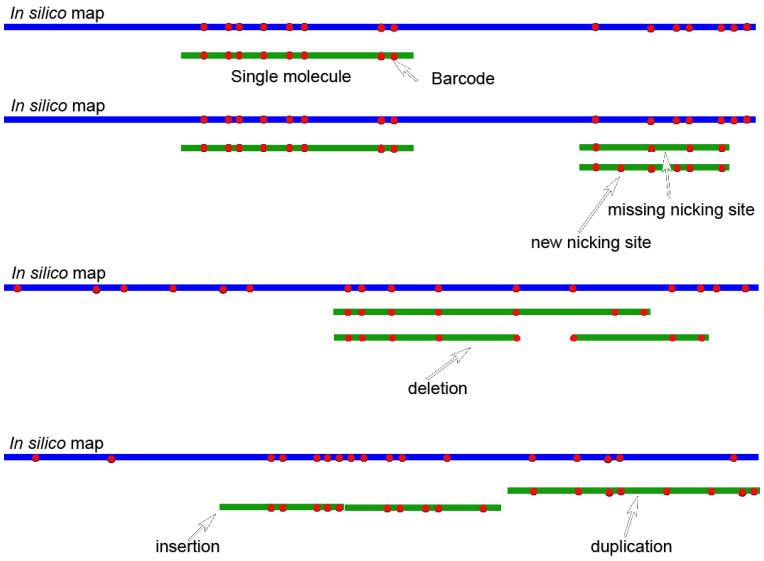

Figure 1.3: Alignment of molecules from the test genome to reference genome by the nick site structure of the mapping enzyme. Some structural variations can be detected by the mismatches during the alignment. This figure shows how the deletion, insertion and duplication can be detected by simply comparing the nick site structure of the mapping enzyme between the test genome and reference genome.

position measurements.

The nick translation procedure of sequencing enzyme is shown in Figure 1.2, in which the red dots denote the nick sites of the mapping enzyme while blue dots denote the nick sites of the sequencing enzyme. For the sequencing enzyme, the nucleotide **A** is labeled blue (others are not labeled). The polymerase terminates the first time it sees three successive **A**'s. (Steric interactions disable the polymerase reaction when three labeled nucleotides require successive incorporation.) In this paper, the nick translation procedures of the mapping enzyme and the sequencing enzyme are similar.

In an experiment deploying FS on a test genome, we have four options of labeling: A, C, G or T. Different options create different neighborhoods and possibly different sets of underlying true values, which will determine the underlying distribution of the observations. In this thesis, our study includes all data sets generated from the four labeling options to increase the detection power.

In summary, FS results in single molecules carrying fluorescently labeled nick sites that: (1) enable genomic alignment and structural variation detection by the nick site structure of the mapping enzyme, and (2) reveal local sequence content by partial digestion of the sequencing enzyme. For each fluorescently labeled nick site the signature and position pair is measured.We will focus on the data generated by the sequencing enzyme

1.3 Overview of Statistical Techniques Developed

Our goal is to develop a model-based testing approach to detect changes between a test genome and the reference genome. Modeling is at the level of bivariate data on measurement events of the sequencing enzyme nick sites. Two major challenges emerge when statistical tests are conducted based on the joint distribution of the bivariate observations to detect variations in the test genome.

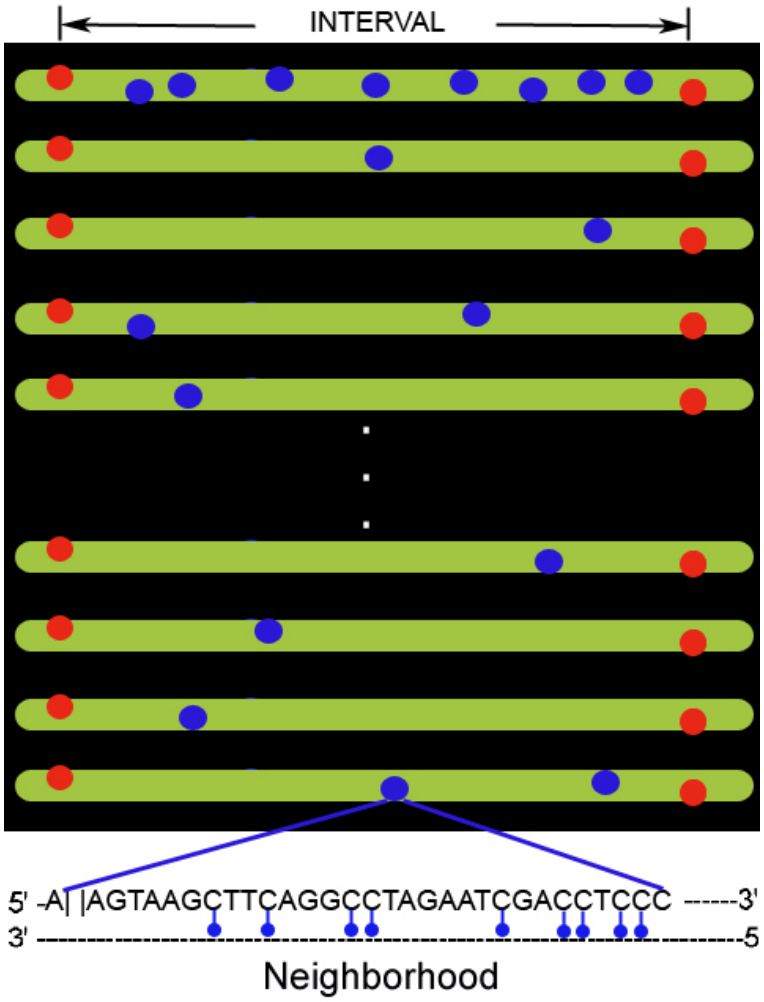

Figure 1.4: Partial digestion: The top line measures an interval under full digestion of the sequencing enzyme. The fluorescently labeled sites in this line will be too close for the microscope to detect. Thus partial digestion is proposed to allow the molecules only react with the sequencing enzyme at certain digestion rate. The lines beneath the first line give examples of the molecules under partial digestion. They are aligned by the nick site map generated by the mapping enzyme (red dots). So for each copy of molecule, we will only digest a small subset of the nick sites generated by the sequencing enzyme in one interval.

First, the measurements of individual nick sites carry relatively large measurement errors while the SNP effects on FS data are very subtle. Classical goodness of fit tests have great difficulty in detecting the SNPs [Section 3.2]. To increase the detection rate, we limit the alternative hypotheses to the known variations from an existing SNP database by using a Regions of Interest (ROI) approach, in which the test statistics are calculated on observations from some small regions of the joint distribution.

The second challenge is that there can be numerous possible variation candidates in one interval, but only a few of them may be likely in the test genome. To increase the chance of identifying correct variations, we not only detect whether there is strong evidence that a variation can happen, but also calculate the likelihood of each variation and always select the most likely ones among those with strong evidence of existence. The entire procedure is recursively executed and at each run one variation is picked until there is no SNP in the database meets the selection criterion.

The details of modeling and detecting methodology will be discussed in Chapter 2 and Chapter 3

1.4 Outline of the Thesis

In Chapter 2 we introduce a statistical model for the sequencing enzyme FS data. We focus on the measurements within one single interval from the mapping enzyme. To reduce the computation complexity, we discretize the two-dimension joint distribution panel of the bivariate observations. After the discretization, the joint distribution is represented by a probability matrix. Under the null hypothesis that the test genome is identical to the reference genome, the test genome has the same probability matrix as that of the reference genome.

The proposed methodology is developed in Chapter 3. From FS data on numerous molecules, the method aims to detect any SNPs that may be present in the test genome. To increase testing

power, we restrict the alternative hypotheses to a SNP database and calculate the test statistics with observations from some small regions of the discretized joint distribution panel, which are referred as *Regions of Interest* (ROI)[21]. Meanwhile, because there are multiple candidates in the database we use *Analysis of Competing Hypotheses*(ACH) [17] to sequentially select the SNPs with the highest likelihood.

To check whether our methodology efficiently selects SNPs on the test genome, we generate many sets of synthetic FS data and repeatedly apply our method. The reference genome is Human build 36 and the test genome is this reference genome with 3.3 million James-Watson SNPs applied. With the synthetic data sets and a database with 15 millon possible SNPs, we aim to detect the James-Watson SNPs. For each synthetic data set, the percentage of detected James Watson SNPs among all James Watson SNPs is calculated and called as *detection rate*.

The detection rate depends on the average number of observations per nick site of the sequencing enzyme, which is defined as *coverage*. We address the following questions : (1) With certain coverage, what percentage of these James-Watson SNPs can be detected when the false positive rate is under a certain threshold? (2) How does coverage affect the detection rate? (3) What is the reason behind the false negatives? All of these questions will be answered in chapter 3.

In chapter 4, we discuss further possible extensions of FS. We investigate how much the detection rate will be improved with more precise measurements and the situation when the SNP database does not include all the James-Watson SNPs.

Chapter 2

Modeling Data From Flash Sequencing

2.1 Overview

There are two nicking enzymes in FS: the medium-resolution mapping enzyme and the high-resolution sequencing enzyme. The nick-site structure of the mapping enzyme enables the alignment of the test genome molecules to the reference genome. Variations in the test genome may cause change of the nick-site structure of the mapping enzyme. However, we concern ourselves here only with variations associated with sequencing enzyme, since those of the mapping enzyme can be analyzed with the existing technologies.

The sequencing enzyme Nt.CviQII recognizes nicking sites at an average distance of 16 bases in human. This short distance not only allows us to obtain information between the mapping enzyme nick sites but also raises the issue of how to distinguish nearby nick sites because they are too close for microscopes to distinguish. To solve this issue, FS partially digests the nick sites of the sequencing enzyme. That is, only a small proportion of nick sites are allowed to react with the sequencing enzyme on any given molecule. Experimentally this is controlled by chemical methods, such as dilution of the sequencing enzyme. In this chapter we model the data from the partial digestion of the sequencing enzyme.

In Section 2.2, the basic measurements generated the sequencing enzyme are described. We discuss the measurement error of signatures based on one pilot experiment in Section 2.3. Section

2.4 gives more details about the error distribution of the position measurements. The joint distribution of signature and position is described in Section 2.5. In Section 2.6, we discretize the joint distribution to reduce computational complexity and facilitate later analysis.

2.2 Basic Measurements

In nick translation of a digested nick site, the polymerase replaces the bases on test molecule with fluorescently labeled bases in the solution until certain ending patterns, such as *AAA* (Figure 1.4). The spanned region from the nick site to the ending pattern are called the *neighborhood*. Under a microscope, this neighborhood with florescent labeled bases looks like one colored shape. Its peak value of intensity is measured and named as *signature*, which we denote by S. The *position* is denoted by X and is defined to be the distance from the left end of the interval to the center of the colored shape. Our basic measurement is the bivariate observation (S, X) of a single nicking and labeling event.

The expected signature value is determined by the number of bases available to be labeled in the neighborhood of the corresponding nick site. A pilot experiment in the Schwartz's lab collected 368 measures in the case of a single available base. In this experiment the single nucleotide (dUTP: 2-Deoxyuridine, 5-Triphosphate) labeled with alexa fluor dyes were imaged under microscope using high resolution digital camera and red laser illumination. The digital image was analyzed by PeakFinder (a custom software used for fluorescent dot detection in the microscopic digital image). The peak fluorescent intensities of all the excited single fluorophores were recorded. A model of signature error is developed based on this experiment in Section 2.3.

In a given mapping enzyme interval, we obtain the position measurements in two ways: (1) measuring the length from the left end of the interval to the peak intensity position (2) subtracting the length from the peak intensity position to the right end of the interval from the entire length of the interval. The weighted average of these two length measurements gives us a better estimation

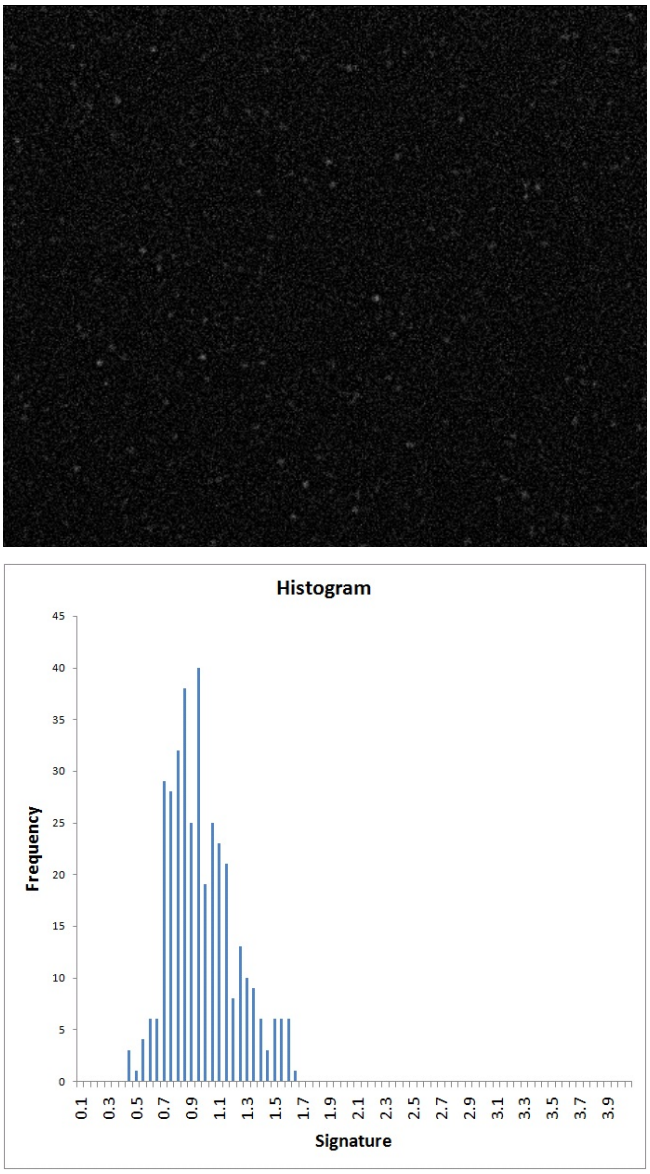

Figure 2.1: Empirical error distribution of signatures S. The top figure shows many fluorescently labeled nick sites on a square slide from a pilot experiment when the true value μ equals to 1. The peak intensity values of these florescent labeled sites are recorded. In total there are 368 observations, which are summarized in the histogram at the bottom. The histogram indicates that the signature measurements are approximately normally distributed with mean value 0.95 and standard deviation 0.24 .

of the position than either one alone. Under the assumption that these two length measurements follow independent normal distributions, the weighted average of them will also follow a normal distribution. A pilot experiment in Schwartz lab obtained 4000 measurements of lengths, whose true values are known. In this experiment the geomic DNA molecules of Mesoplasma florum genome was nanocoded using alexa fluor labeled dUTP from the BspQI nicking sites by E. coli polymerase, and imaged under microscope using two-color imaging system developed in house. The lengths of each fluorescently labeled fragments were measured. With these data we develop a model of position error in Section 2.4.

2.3 Error of Signature

As described in Section 2.2, the basic measurement event produces a signature measurement S and a position measurement X for one labeled nick site on one molecule. While multiple events can (and often do) occur on single molecules, thus inducing statistical dependence, we ignore that feature and focus on the marginal probability distribution of (S, X). (It is as if we restrict attention to a single randomly sampled measurement event on each molecule that yields at least one event.) Naturally, S and X are queries of some nick site and its neighborhood, but this identity is not known precisely, even after measurement.

We take the null hypothesis that the test genome has exactly the same base sequence as the reference genome in the mapping enzyme interval. Thus under null hypothesis we know the exact base order information, from which we can determine the number of nick sites in this interval, which is denoted by J. For the j^{th} nick site ($j = 1, \ldots, J$), we calculate the true number of fluorescently labeled bases and the center position of the neighborhood, which are denoted by μ_j and ν_j respectively. Let $\theta = \{(\mu_j, \nu_j) : j = 1, 2, \ldots, J\}$, where:

$$\begin{aligned} \mu_j &= \text{the number of fluorescently labeled bases within the } j^{th} \text{ neighborhood} \\ \nu_j &= \text{center position of the } j^{th} \text{ neighborhood, i.e., the number of bases from} \\ &\quad \text{the left end of the interval to the center of the neighborhood .} \end{aligned}$$

Let Z denote the nick site being measured by (S, X), with values in $\{1, 2, \ldots, J\}$. Suppose that partial digestion is uniform then:

$$Z \sim \text{Uniform}\{1, 2, \ldots, J\}.$$

Following prior work in fluorescent microscopy [4] and [30] gives

$$S|Z = j \quad \sim \quad \text{Normal}\left(c_1\mu_j, \mu_j\sigma_1^2\right),$$

which allows heteroscedastic errors.

In our pilot experiment, we obtain observations $t_1 = \{t_{1,1}, t_{1,2}, t_{1,3} \ldots t_{1,r_1}\}$ as shown in Figure 2.1. We obtain the estimated error distribution derived from a Bayesian analysis where uncertainty in c_1 and σ^2 has been integrated away [APPENDIX B]:

$$S|Z = j, t_1 \quad \sim \quad \text{Normal}\left(\widehat{c_1}\mu_j, \mu_j\widehat{\sigma_{11}}^2 + \mu_j^2\widehat{\sigma_{12}}^2\right),$$

where it is estimated that $\widehat{c_1} = 0.95$, $\widehat{\sigma_{11}} = 0.061$ and $\widehat{\sigma_{12}} = 9.90e - 5$.
This distribution is denoted by $f(s|Z = j)$.

2.4 Error of Position

Each labeled nick site is associated with a position to be estimated from two length measurements: One is from the peak fluorescent intensity to the left end of the interval; while the other is to the right end of the interval. For the convenience of calculation, the length measurements are transformed to counts of bases. We consider a simple model in which the two measurements X_1 and X_2 are independent. Suppose the $\mathcal{L}$ is the length of the interval, the weighted average of X_1 and $\mathcal{L} - X_2$ is denoted by X and considered as the optimal measurement of position.

The first step in modeling the error of position is to find out how the error distribution of X_1 and X_2 depend on ν_j, where ν_j is the number of the bases from the left end of the interval to the center of the neighborhood. There is no previous work describing the modeling of position. So we build

our model based on around 4000 length measurements from a pilot experiment. The measurements are denoted by $t_2 = \{t_{2,1}, t_{2,2}, \ldots, t_{2,r_2}\}$. In this experiment the true values ν_j for all measurements are known. This pilot experiment suggests that the error distribution of X is approximately normal. Figure 2.2 demonstrates the relationship between the mean/variance and the true values. It shows that the mean value of observations is approximately equal to ν_j while the corresponding variance value has a linear relationship with $\nu_j + \nu_j^2$. When we consider $\theta = \{(\mu_j, \nu_j) : \ j = 1, 2, \ldots, J\}$ and Z as parameters, then

$$X_1|Z = j \quad \sim \quad \text{Normal}\left(\nu_j, (\nu_j^2 + \nu_j)\sigma_2^2\right).$$

Assume $\mathcal{L}$ is the length of this interval, then

$$\mathcal{L} - X_2|Z = j \quad \sim \quad \text{Normal}\left(\nu_j, ((\mathcal{L} - \nu_j)^2 + (\mathcal{L} - \nu_j)) * \sigma_2^2\right).$$

Let $\lambda_1 = \frac{((\mathcal{L}-\nu_j)^2+(\mathcal{L}-\nu_j))\sigma_2^2}{(\nu_j^2+\nu_j)\sigma_2^2+((\mathcal{L}-\nu_j)^2+(\mathcal{L}-\nu_j))\sigma_2^2}$ and $\lambda_2 = 1 - \lambda_1$. Then under the assumption that X_1 and X_2 are independent the weighted average $\lambda_1 * X_1 + \lambda_2 * (\mathcal{L} - X_2)$ should be the optimal measurement of position. However when we collect data we don't know λ_1 because ν_j cannot be identified. So in this study we use the average $X = 0.5 * X_1 + 0.5 * (\mathcal{L} - X_2)$. The estimated error distribution of the average measurement follows:

$$X|Z = j, t_2 \quad \sim \quad \text{Normal}\left(\nu_j, 0.25 * (\nu_j^2 + \nu_j) * \widehat{\sigma_2}^2 + 0.25 * ((\mathcal{L} - \nu_j)^2 + (\mathcal{L} - \nu_j)) * \widehat{\sigma_2}^2\right).$$

where it is estimated that $\widehat{\sigma_2}^2 = 0.0095$ with unit $\frac{base^2}{base^2+base}$. [APPENDIX B]. The density of this distribution is denoted by $g(x|Z = j)$.

2.5 Joint Distribution of (S, X) for one interval

In the absence of other information, is is natural to assume that S and X are conditionally independent given $Z = j$

$$p(s, x|Z = j, \theta) = f(s|Z = j)g(x|Z = j)$$

where $f(S|Z = j)$ and $g(X|Z = j)$ are error distributions of the signature and position as described in the previous sections. When averaged over Z, the joint distribution for (S, X) is a

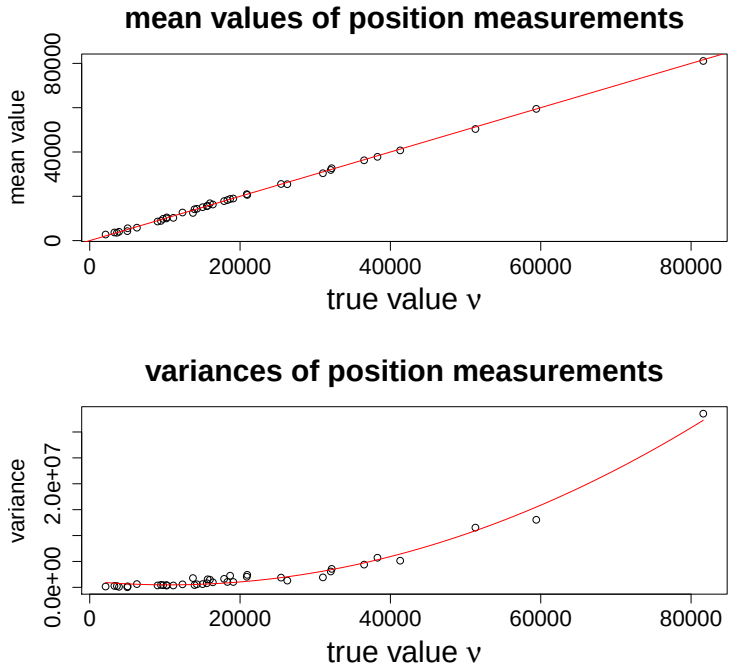

Figure 2.2: Pilot experiment of position (X). We have 39 distinct true values of ν_j with nearly 4000 measurements in this experiment. The top panel shows the relationship between the mean values of the length measurements and the underlying true values. The X axis gives the underlying true values and the Y axis gives the mean values of observed values. The redline represents y=x. Similarly, the relationship between the variance and the true values is shown in the bottom figure.The X axis shows the 39 underlying true values while the Y axis indicates the variances of corresponding observations. The red line is the fitted line of model $var \sim \nu + \nu^2$, where ν is the true value of the positions. The adjusted R-square for this model is 0.98, which indicates that the model fits observed data very well.

mixture of bivariate normals:

$$p(s,x|\theta) = \frac{1}{J}\sum_{j=1}^{J} f(s|Z=j)\, g(x|Z=j) \qquad [2.1]$$

Figure 2.3 shows the joint density (2.1) associated with one interval on human chromosome 1, reference genome.

For computational reasons we consider both the model $p(s,x|\theta)$ and the data D in one interval as summarized on a two-dimensional grid panel, rather than a continuous planar region. As shown in Figure 2.4 when the variances of signature or position go up, the grid cells will be accordingly larger to catch the changes of the joint distributions. The entire panel is divided into M rows for the position measurements and K columns for the signature measurements. We use a common $K = 88$ for all intervals, though M depends on interval length.

2.6 Data Structure

We envision that genome-wide FS data will be collected in four sets of assays for each test genome. Each assay will be associated with one of the four labeled bases (A,C,G or T). The mapping enzyme will result in same interval structure for all the four sets of assays. However within each interval, the sequencing enzyme will produce different information of the test genome since the labeled base is different. To ensure the maximum detection rate, the test statistics are calculated based on FS data from all four sets of assays.

To distinguish the data sets obtained from different fluorescent labels on the same test genome, we append a superscript to indicate the labeled base. D^B denotes the observed data, where B is the base type being labeled $\{A, C, G, T\}$. With the grid cells, we convert D^B to $O^B = \{o^B_{m,k} : m = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K\}$, where $o^B_{m,n}$ is the frequency of the observed values fall into the grid cell located in m^{th} row and k^{th} column of a given mapping enzyme interval when base B

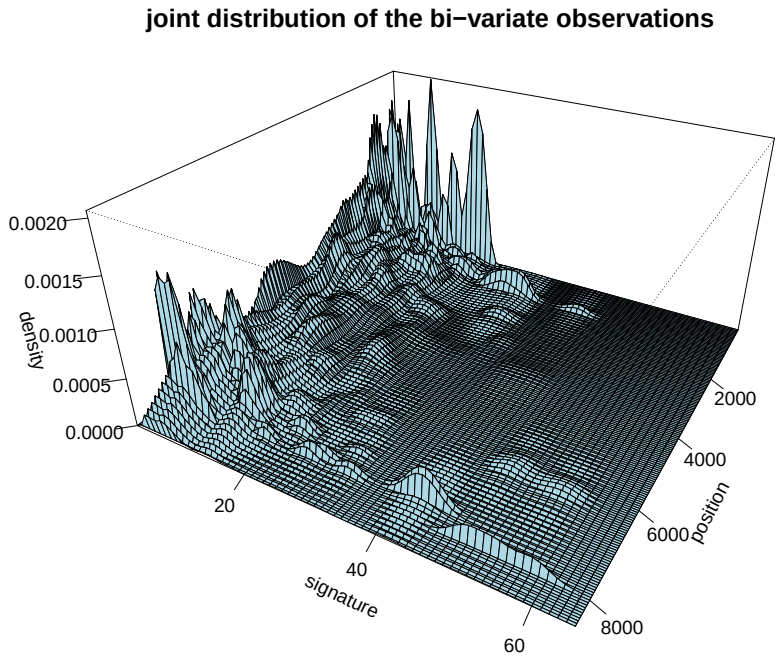

Figure 2.3: Joint distribution of signature S and position X for one interval of Human build 36 chromosome 1, and labeling **A**'s

is labeled.

A central concept in this work is that the distribution of FS data depends on the underlying genome, and thus, genomic variation is detectable by comparing the empirical distribution of the data against expectations under various specifications of the genome. One approach to detect variation from a reference genome is to apply a goodness-of-fit test to the gridded data. Expected grid cell counts are computable from the human reference genome. The interval of the reference genome corresponds to a probability matrix $q^{0,B} = \{q^{0,B}_{m,k} : m = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K\}$, which may be altered by the mutations in the test genome. Goodness-of-fit of the reference genome is tested most simply via the likelihood ratio test or Pearson's chi-square test. With $o^B_{\cdot,\cdot} = \sum_{m,k} o^B_{m,k}$, the likelihood ratio test statistics $2\sum_{m,k} o^B_{m,k} \log \frac{o^B_{m,k}}{q^{0,B}_{m,k} o^B_{\cdot,\cdot}}$ and the Pearson's chi-square test statistics $\sum_{m,k} \frac{(o^B_{m,k} - q^{0,B}_{m,k} * o^B_{\cdot,\cdot})^2}{q^{0,B}_{m,k} o^B_{\cdot,\cdot}}$ A major problem with traditional goodness of fit tests is that the detection power is very conservative because the difference of the frequencies is only carried by a few of the grid cells. The goodness-of-fit test seeks evidence of *any* departures from reference. However certain kinds of departures may be more probable, depending on the application. A test focused on detecting specific departures from the reference genome is bound to have improved sensitivity when they occur. Methodology that develops this concept further is presented in the next chapter.

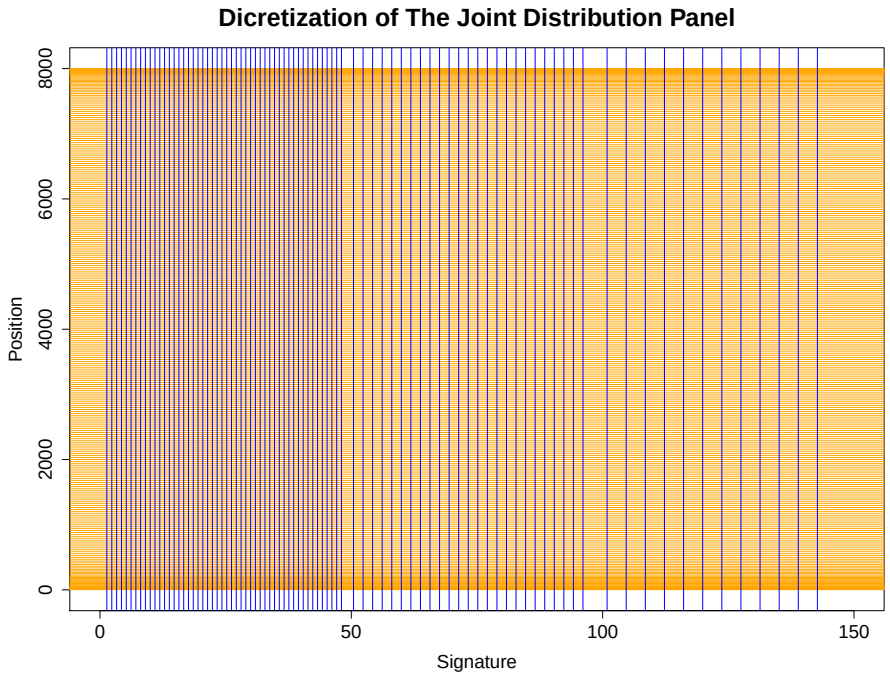

Figure 2.4: Division of the panel. The entire panel is divided into many grid cells. The column spacing increases with increasing signature values because of the increment of the measurement variances. The row spacing peaks at the middle of the range of position. The width of spacing depends on variances of signature and position measurements.

Chapter 3

Identifying SNPs from Flash Sequencing Data

3.1 Overview

A typical FS experiment will collect signature/position data from millions of test genome molecules. In Chapter 2, we explained how the error distribution of signature and position were guided by previous research and two pilot experiments. By discretizing the (S,X) joint distribution panel, the joint distribution of (S,X) in one mapping enzyme interval is multinomial, with probabilities determined from the mixture of bivariate normals. In this chapter we develop an inference method from this model that allows us to detect SNPs in a test genome using FS data on this genome. Our analysis combines results from separate mapping enzyme intervals, and so we focus on the methodology for a single interval to develop the methodology.

Within a mapping enzyme interval, genomic sequence variations can be of many types, including SNPs, indels and translocations. For FS, the SNPs are perhaps the most subtle variations because each SNP affects a single base and its impact on FS could be quite small. If FS can detect SNPs, it is expected that other variations can also be detected since they can alter multiple bases, which will reflected by larger deviations from the reference genome in FS data. Therefore, we focus on inference for SNPs in this chapter.

SNPs can alter the distribution of FS data in many ways. In Section 3.2, we group the SNPs into 8 types based on their effect on FS data. The most common type of SNP is demonstrated by

an example, while more examples of other types are given in APPENDIX C.

To illustrate our proposed methodology, we generated many synthetic data sets. In Section 3.3, the procedure generating these synthetic data sets is introduced. In Section 3.4, two classical goodness of fit test statistics are discussed. We use an example to show how conservative these tests can be, which indicates the need of a much more powerful test statistics for FS data. To increase the detection power, in Section 3.5 we first limit the alternative hypothesis to the existence of a given SNP. Then we propose an ROI (region of interest) test statistic, which is calculated from some small regions of the (S, X) distribution panel. These regions carry the majority information of deviations caused by this SNP. We use the same data of the example in Section 3.4 to show that ROI test statistics greatly improve the testing power. In Section 3.5 the details of ROI test statistics is demonstrated under the simplest situation when there is only one SNP candidate in the target interval of the test genome. To compare the performance of ROI and classical test statistics, we introduce another concept *detected probability*. This concept describe the probability of one certain variation to be detected among many synthetic data sets. For each variation, the detected probability is calculated by dividing the number of synthetic data sets in which this variation is detected by the total number of synthetic data sets.

In more general situation, there are multiple SNPs happening on one interval of the test genome. Much more need to be considered unter this situation. First, SNPs can be so close that they appear in same neighborhoods. These SNPs are considered to have *interactions* with each other. The effect of these interactions on FS data can be very complex since we have 8 types of SNPs. To pick up these SNPs correctly we need to know how the interactions affect the FS data. To further increase the testing power, we include second order interactions (only two SNPs are in same neighborhoods) in our SNP database. The probability of two SNPs having interactions is very low due to the fact that the average distance between SNPs is every 1200 to 1500 DNA bases [10] while the average length of the neighborhoods of the sequencing enzyme is 16 bases with sequencing enzyme Nt.CviQII. Three or higher order interactions are still possible but occur with negligible

frequencies. Secondly, nearby SNPs tend to have similar effect on the FS data. To infer which SNPs in the SNP database occur in the test genome, we use *Analysis of Competing Hypotheses (ACH)* besides the ROI test statistics. In section 3.6, we propose a step-wise competing mechanism, which allows us to pick SNPs sequentially with ACH.

In Section 3.7 we generate synthetic data sets of the whole test genome to access the sensitivity and specificity of our method described in Section 3.6. The synthetic data sets are generated based on human build 36 and 3.3 million James-Watson SNPs. We aim to select the James-Watson SNPs out of the NCBI 15 million SNP database.

For each synthetic data set, the percentage of detected James-Watson SNPs is called *detection rate*, which depends on the value of *coverage*. Coverage is defined to be the average number of observations per $\{\mu_j, \nu_j\}$ pair. The coverage value directly determine the number of whole genome copies, each of which is sheared into many molecules. The expected number of whole genome copies can be calculated as following:

$$\text{Number of whole genome copies needed} = \text{coverage}/\text{digestion rate}$$

Since the coverage value is highly related with the detection rate and cost of experiments, we are interested in following questions:

1. Under different coverage, what is the detection rate when the SNP database includes all of the James-Watson SNPs?

2. What are the major reasons for the false negatives?

In this chapter we will discuss the situation when the SNP database is complete, that is, it contains all of the James-Watson SNPs. In Chapter 4, there is further discussion about a SNP database containing only a portion of the James-Watson SNPs.

3.2 SNP effect on distribution of FS data

The fundamental idea of FS detection is that SNPs alter the parameters governing the joint distribution of (S,X). Recall we denote these parameters, in one interval, as $\theta = \{(\mu_j, \nu_j) : \ j = 1, 2, \ldots, J\}$, where μ_j is the number of fluorescently labeled bases within the j^{th} neighborhood in the interval and ν_j is the distance from the left end of the interval to the center of the j^{th} neighborhood. Based on how SNPs alter the θ vector, they are classified into eight types:

1. Those that are covered by some neighborhoods and increase or decrease the corresponding μ_j values by 1;

2. Those that ablate the termination sites (AAA, CCC, GGG, or TTT) of the neighborhoods. Since the polymerase continues the synthesis of bases until the next termination sites, it is possible that the related μ_j values are increased more than 1;

3. Those that create new termination sites of the neighborhoods. In contrast to the previous type, it is possible that the related μ_j values are decreased by more than 1;

4. Those that ablate the cognate sites (AAG or GAG) of the sequencing enzyme. The corresponding $\{\mu_j, \nu_j\}$ pairs generated from these cognate sites are removed from the θ vector;

5. Those that create new cognate sites of the sequencing enzyme and thus generating new $\{\mu_j, \nu_j\}$ pairs, which will be added into the θ vector;

6. Those that are covered by the some neighborhoods while not altering the the corresponding μ_j values because the SNPs do not change the number of bases with fluorochrome attached. This type of SNPs have no effect on the θ vector;

7. Those that are covered by none of the neighborhoods and thus having no effect on the θ vector;

8. Those that alter the mapping enzyme interval structure by creating or removing the nick site pattern (GCTCTTC). We propose that these effects are accommodated in any preprocessing with the mapping enzyme.

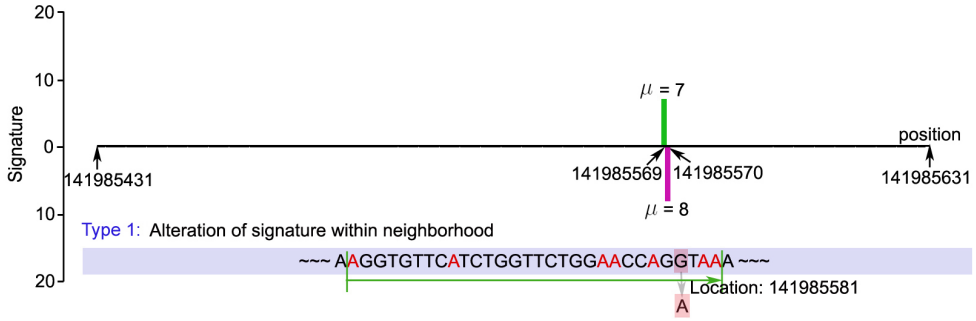

Figure 3.1: Type 1 SNP alters the μ_j value by 1. The X axis denotes the position in first chromosome. The y axis denotes the μ_j value which is the number of labeled bases. The green bar is the original value of μ_j and the purple bar is the value of μ_j being altered by this SNP Here the SNP change base G to A and alter the μ_j value from 7 to 8.

Figure 3.1 shows an example of type (1) which is the most common (around 39.60%) situation for SNP. Other types are shown in APPENDIX C. Note that the SNP type depends on the nucleotide being labeled. In our example, the base A is fluorescently labeled. We get three more data sets by labeling bases C, G and T respectively. The SNP's type may change across the four data sets. For example, one SNP can be type 1 for the data set derived by labeling base A, but it can also be type 6 for the data set derived by labeling base C. So, for most of the SNPs, they are a mixture of different types, with the only exception being type 8. Type 8 SNPs cannot be any other types no matter which type of nucleotide is labeled.

3.3 Synthetic Data Set Generation

To generate synthetic data set and to test our analysis method stated in previous sections, we need the following information

1. a reference genome, which will be used as a baseline for comparison. We chose human build 36 as the reference genome.

2. a test genome, which will be used to generate synthetic data sets.

With the complete base order of the test genome, we are able to identify the mapping enzyme intervals and calculate the θ vector for each interval, from which we can calculate the corresponding frequency matrices. Given a value of coverage, we can simulate vectors of observed counts for the test genome based on the multinomially distributed frequency matrices.

In following sections, synthetic data sets are generated to facilitate the illustration of statistical methodologies. In Section 3.4 we discuss the classical tests under the simplest situation when the test genome is obtained by applying one single SNP ss87153243 to the reference genome. The classical tests are extremely conservative, so the ROI statistics is proposed in Section 3.5 and tested with the same data as in Section 3.4. We discuss the interactions of SNPs in Section 3.6 where there are multiple SNPs in the test genome within the target interval. In section 3.7, we apply the ROI and ACH methodology to the entire genome, where the test genome is obtained by applying 3.3 million James-Watson SNPs to the reference genome. The ROI/ACH approach is applied to genome wide synthetic data sets in Section 3.8.

3.4 Low Power of Classical Tests

In this section, two goodness of fit test statistics are discussed. To illustrate how conservative they can be, an example is given in Section 3.4.2.

3.4.1 Classical Tests: Goodness of Fit

When the test genome is compared with the reference genome within one interval, the null hypothesis is that both genomes have exact same base sequence in that interval. Assume that the (S, X) panel is discretized to M rows for position measurement X and N columns for signature measurement S, we have four data sets by labeling each base (A,C,G and T). For each data set, we have empirical frequencies of the observations falling into each grid cell of the (S, X) panel. Denote the observed frequencies by $O^B = \{o^B_{1,1}, o^B_{1,2}, \ldots, o^B_{m,n}, \ldots, o^B_{M,N}\}$, where $m = 1, \ldots, M$, $n = 1, \ldots, N$ and base $B \in \{A, C, G, T\}$. Under the null hypothesis, the corresponding expected frequencies are $E^B = \{e^B_{1,1}, e^B_{1,2}, \ldots, e^B_{m,n}, \ldots, e^B_{M,N}\}$. Then the classical goodness of fit statistics

are log-likelihood ratio test and Pearson's Chi-squared test based on data set O^B:

$$2\sum_{m,n} o^B_{m,n} \log\left(\frac{o^B_{m,n}}{e^B_{m,n}}\right)$$

or

$$\sum_{m,n} \frac{(o^B_{m,n} - e^B_{m,n})^2}{e^B_{m,n}}$$

When variations happen to the test genome, only a few of the $M * N$ grid cells carry different underlying probabilities from those of the reference sequences. Thus the difference probability matrix is a sparse matrix, which results in conservative behavior of the classical goodness-of-fit tests[25].

3.4.2 Example of The Classical Tests

To illustrate how conservative the above two test statistics can be, we use the 9^{th} interval of the first chromosome of human build 36 as an example, and assume that the only difference between the test genome and the reference genome in this interval is SNP ss87153243. This SNP changes G to A at position 3655 within the interval. It alters multiple $\{\mu_j, \nu_j\}$ pairs when base A, C, G or T is labeled. The detail is in APPENDIX A. Based on the error distributions in chapter 2, when base A is labeled the difference contour plots between the joint distribution of the reference genome and that of the test genome are shown in Figure 3.2. It is observed that the SNP ss87153243 causes a very subtle change of the (S, X) joint distribution. To quantitatively demonstrate the limitation of the classical goodness-of-fit tests, we generated 5000 synthetic data sets with coverage 200 under both null and alternative hypothesis. The values of the classical goodness-of-fit test statistics are shown in Figure 3.3. The 99% empirical quantile of the log-likelihood ratio test statistics under null hypothesis is 9964.689. There are 4.24% of the log-likelihood ratio test statistics under the alternative hypothesis exceeding this threshold. For the Person's Chi-squared test statistics, the histogram is plotted for the log value of these statistics. The 99% empirical threshold is 10.357 and about 1.3% of the test statistics under alternative hypothesis exceed this threshold. The similar results can be obtained when base C, G or T is labeled. Therefore these classical tests are too conservative for FS data. To increase the testing power, we construct a new Regions of Interest

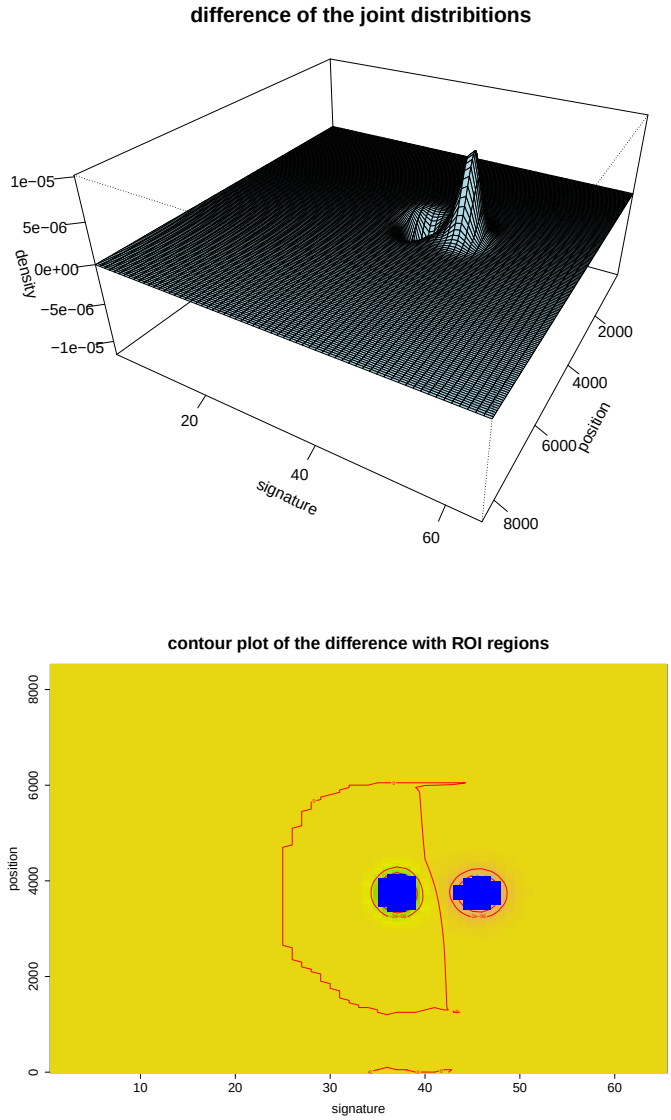

Figure 3.2: The contour plot of the 9^{th} interval for the reference genome is shown in Figure 2.3. We generated a test genome by applying SNP ss87153243 to the reference genome. The contour plot of the underlying joint distribution of the 9^{th} interval is altered and the 3-D difference contour plot is shown in top panel. The 2-D version of the difference is shown in the bottom panel, where the blue area denotes the regions of interest (ROI) identified by our algorithm.

(ROI)[26] test statistics by limiting the alternative to a given database of variations and by only

using data from a small proportion of the grid cells on the discretized panel.

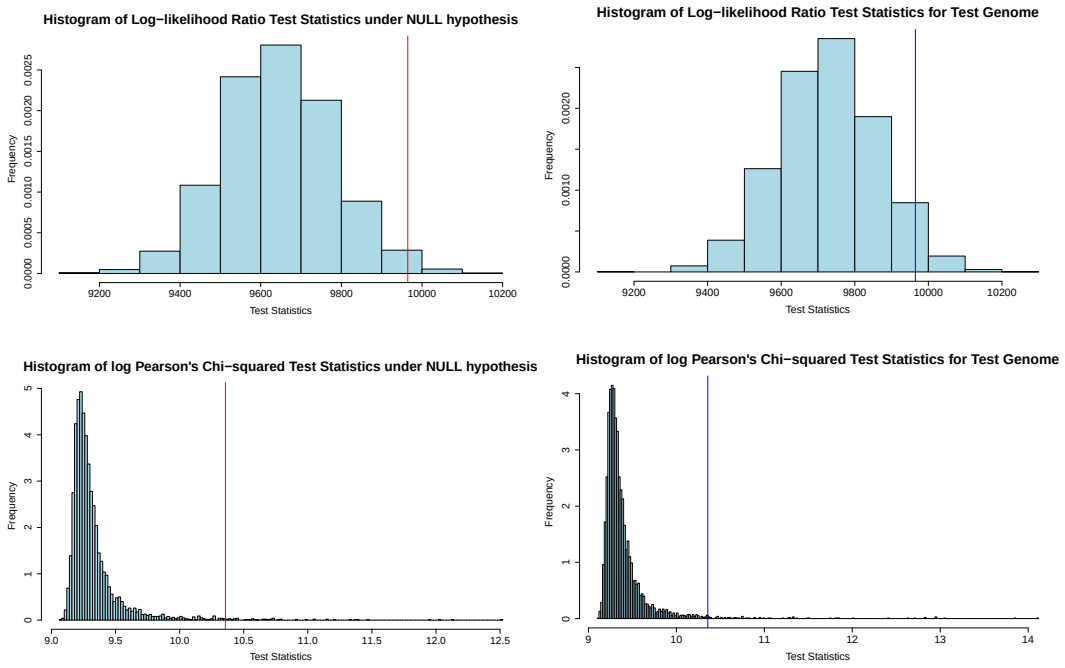

Figure 3.3: The histogram of two classical test statistics. Each plot is histogram of test statistics from 5000 synthetic data sets with coverage 200. The left two plots are calculated under the null hypothesis, while the right two plots are calculated under the the alternative hypothesis (test genome). The empirical 99% quantile (red color lines) for the left two plots are calculated and copied to the right two plots (blue color lines). It is shown in the top right plot that with log-likelihood ratio test variations on the test genome can be detected among 4.24% of the 5000 synthetic data sets. The Person's Chi-squared test statistics are transformed to log scales because the values are too big. It is shown in the bottom right plot that with this type of statistics, variations on the test genome can be detected among 1.3% of the 5000 synthetic data sets.

3.5 ROI test statistics under simplest situation

We start from the simplest situation where there is only one SNP in the target interval to demonstrate the ROI test statistics.

3.5.1 ROI test statistics

Regions of Interest (ROI) is an approach which selects subsets of a data set for a particular purpose. It has been used in various applied fields, including whole-brain voxel-wise analysis [13], medical imaging [8],computer imaging and video processing [20]. It is also known for improving the power of statistical test [26][21].

To increase the testing power, we first limit the alternative to the NCBI dbSNP, which is a database containing 15 million SNPs. For this particular 9^{th} interval of the first chromosome, there are 20 SNPs candidates in the NCBI dbSNP database. Instead of taking the general alternative hypothesis that the test genome is different than the reference genome, we expect to improve the testing power if we restrict the alternatives to these 20 SNPs. Let $k \in \{1, \ldots, \mathcal{K}\}$ denote the index of hypotheses and $\mathcal{K}$ is the total number of SNP candidates in the target interval of the mapping enzyme. Here $\mathcal{K}$ is 20. Then:
H_0 : The test genome is identical to the reference genome
H_{ak} : The k^{th} SNP happens to the test genome. Meanwhile, there is no other SNPs happen on the test genome

k will also be used to index the expected counts or test statistics under the k^{th} alternative hypothesis, while $k = 0$ denotes the situation when the null hypothesis is true. Each of the alternative hypotheses will be tested against the null hypothesis.

Because we know the proportions under the null and the alternative hypotheses, we can calculate the simple vs. simple log-likelihood ratio test. Let $Q^{k,B} = \{q^{k,B}_{1,1}, \ldots, q^{k,B}_{m,n}, \ldots, q^{k,B}_{M,N}\}$

($m = 1, \ldots, M$ and $n = 1, \ldots, N$) denote the density frequencies of observations under k^{th} alternative hypothesis when base B is labeled. When $k = 0$, the proportion vector $Q^{0,B}$ denotes the density frequencies under the null hypothesis. $O^B = \{o^B_{1,1}, \ldots, o^B_{M,N}\}$ is the vector of observed frequencies. Then the log-likelihood ratio test statistics will be :

$$\log\left(\frac{L^{k,B}}{L^{0,B}}\right) = \sum_{m,n} o^B_{m,n} \log\left(\frac{q^{k,B}_{m,n}}{q^{0,B}_{m,n}}\right)$$

From Figure 3.2 we see that most of the probabilities of $q^{k,B}$ are very close to $q^{0,B}$. The majority of the difference is carried by a few small regions, which are called *regions of interest (ROI)* . In our study, the ROI is needed for two reasons:

1. The exact distribution of the simple vs. simple log-likelihood ratio test is difficult to obtain because of the dependencies among variables. With ROI, only the information of a few of the grid cells will be used in the calculation, of which the frequencies can be treated as independent because their sum is far less than 1. Later in this section, we propose an approximate distribution of the ROI test statistics given a large number of observation. The approximation is supported by our simulation study.

2. By using ROI, when we calculate the test statistic of one SNP the effect of nearby SNPs is minimized. In the next section we will discuss some rare cases where two or more SNPs are so close that they have interactions. We propose to process these cases differently, so that the final procedure obtains higher testing power.

Suppose that out of the $M * N$ grid cells, H of them are considered be included in ROI. Then the (S,X) distribution panel can collapse into $H + 1$ regions, where the $H + 1^{th}$ region includes any grid cells out side of the H selected grid cells. Then the test statistics can be written as:

$$\log\left(\frac{L^{k,B}}{L^{0,B}}\right) = C + \sum_{h=1}^{H} d^B_h o^B_h$$

where $C = o^B_{\cdot\cdot} \log\left(\frac{1-sum^H_{h=1} q^{k,B}_h}{1-sum^H_{h=1} q^{0,B}_h}\right)$ and $d^B_h = \log\left(\frac{q^{k,B}_h}{q^{0,B}_h}\right) - \log\left(\frac{1-\sum^H_{h=1} q^{k,B}_h}{1-sum^H_{h=1} q^{0,B}_h}\right)$. Here $o^B_{\cdot\cdot}$ is the total count of observations. When $o^B_{\cdot\cdot}$ is given, C is a constant and o^B_h follows binomial distributions with

parameter $q_h^{k,B}$ under the null hypothesis. Since $\sum_{h=1}^{H} q_h^{k,B} \ll 1$, o_h^B can be considered independent with each other. When $o_{\cdot\cdot}^B$ is large enough, o_h^B will follow approximately normal distribution. Under the null hypothesis we have:

$$O_h^B | o_{\cdot\cdot}^B \sim N\left(o_{\cdot\cdot}^B q_h^{0,B}, o_{\cdot\cdot}^B q_h^{0,B}(1 - q_h^{0,B})\right)$$

and the test statistics $W^B = \sum_h d_h^B o_h$ will follow the distribution:

$$W^B | o_{\cdot\cdot}^B \sim N\left(\sum_{h=1}^{H} o_{\cdot\cdot}^B d_h^B q_h^{0,B}, \sum_{h=1}^{H} o_{\cdot\cdot}^B (d_h^B)^2 q_h^{0,B}(1 - q_h^{0,B})\right) \quad [3.1]$$

An example is given in the next section to show how ROI works and the above distribution is tested by a simulation study with 5000 synthetic data sets under null or alternative hypothesis. The ROI test statistic is calculated for each of the synthetic data set.

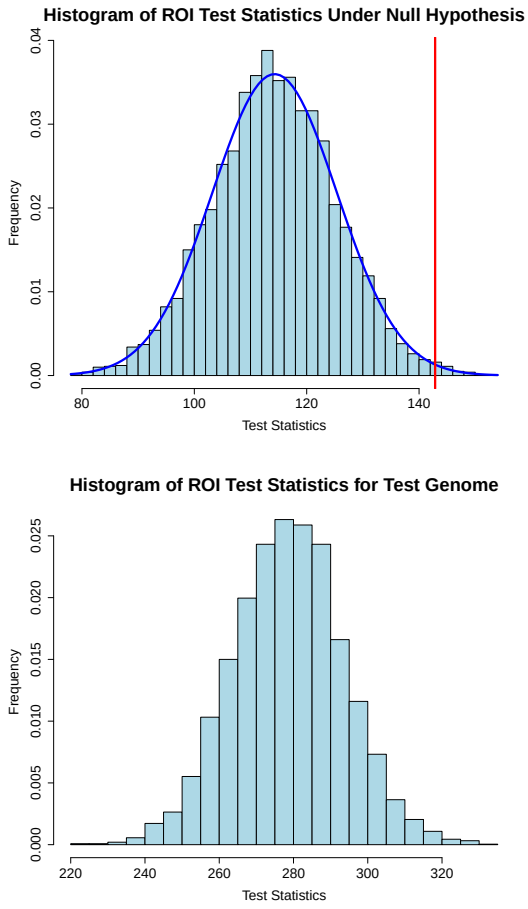

Figure 3.4: The top plot shows the ROI test statistics under the null hypothesis for the 9^{th} interval of the first chromosome. The ROI regions were obtained based on SNP ss87153243. The histogram is obtained from 5000 synthetic data sets with coverage 200 when base A is labeled. Each synthetic data set consists of four data sets with base A,C,G,T labeled respectively. The blue curve denotes the normal distribution obtained from the Formula 3.1. The red line is the threshold of significant findings when the significant level is set to 0.01. The bottom plot shows the ROI test statistics for the test genome, which is obtained by applying SNP ss87153243 to the reference genome.

3.5.2 Example of ROI Test Statistics

By labeling different bases, we have four data sets, which result in four ROI-based test statistics W^A, W^C, W^G and W^T. We propose to calculate the p-value by using a general test statistics $W = W^A + W^C + W^G + W^T$. Same as in Section 3.4, we focus on the 9^{th} interval of human build 3.6. A simulation study of the ROI test statistics is demonstrated by Figure 3.4, which shows the null distribution of W (top panel) and the alternative distribution (lower panel) in the case of a single SNP ss87153243 in the test genome. Both distributions are based on 5000 synthetic data sets with a coverage 200. In the top panel the blue curve is the normal approximation of the general test statistics $W = W^A + W^C + W^G + W^T$ under null hypothesis. This normal approximation is calculated by Formula 3.1. The red line indicates the threshold of the test statistics when the significance level is set to 0.01. It is shown in the bottom panel that all 5000 ROI test statistics exceed the null threshold, thus showing a much higher power in detecting this single SNP than the classical tests described in Section 3.4.

The bottom plot of Figure 3.4 shows that all the ROI test statistics of the 5000 synthetic data sets exceed the 99% threshold under null hypothesis. Thus the detected probability of SNP ss87153243 is 100%. To make a comparison we calculate the ROI p-values for all the 20 SNP candidates (including ss87153243) in this interval in the NCBI database(APPENDIX D). Their p-values from the 5000 synthetic data sets for the test sequence (alternative hypothesis) are shown in the top panel of Figure 3.5, which shows that the other 19 SNPs have slim chance to be selected if we use ROI p-value as the criterion. In general situation, there will be multiple SNPs on the test genome. To select the SNPs explaining the observed data sets best, we use the log-likelihood values as the second criterion. For each of the 20 SNP, the log-likelihood values of alternative hypothesis and null hypothesis are calculated and their differences are shown in the bottom panel of Figure 3.5.

A SNP will be selected if :

1. It has a significantly small p-value generated from the ROI test statistics $W = W^A + W^C + W^G + W^T$. This indicates in the ROI the observation counts deviate significantly from the expected counts under the null hypothesis significantly.

2. The likelihood of the alternative hypothesis (existence of this SNP on test genome) is higher than the null hypothesis.

From Figure 3.5, SNP ss87153243 fits the above two criteria for all the 5000 synthetic data sets under alternative hypothesis, while none of other SNPs fit the two criterion for any of these synthetic data set. This means the detected probability of ss87153243 is 100% and the detected probability of the other 19 SNPs is 0%.

One interesting phenomenon shown in Figure 3.5 is that the SNPs rs28534012, rs28464214, rs13328655 and rs12401368 have likelihood differences 0. This is because these four SNPs happen to be in a gap where no neighborhoods will cover. Thus the frequency matrices of FS data for these SNPs are identical to the one under the null hypothesis. FS cannot detect this kind of SNPs due to the fact that they do not cause any change of the underlying joint distribution of FS data.

3.6 Interactions Between SNPs

On average, the distance between the SNPs on human genomes is approximately 1200 to 1500 bases [10]. The average length of neighborhoods is about 16 bases for the sequencing enzyme. For this reason, usually two SNPs are far away enough from each other and they are not in any one of the neighborhoods simultaneously. In this situation the following relationship holds:

$$Q_0^B - Q_{1,2}^B = Q_0^B - Q_1^B + Q_0^B - Q_2^B$$

where Q_0^B is the proportion vector of the discretized (S,X) panel under the null hypothesis, Q_1^B and Q_2^B are proportion vectors when either SNP1 or SNP2 happens, while $Q_{1,2}^B$ is the situation when both SNP1 and SNP2 happen. This suggests that SNPs can be selected sequentially. We

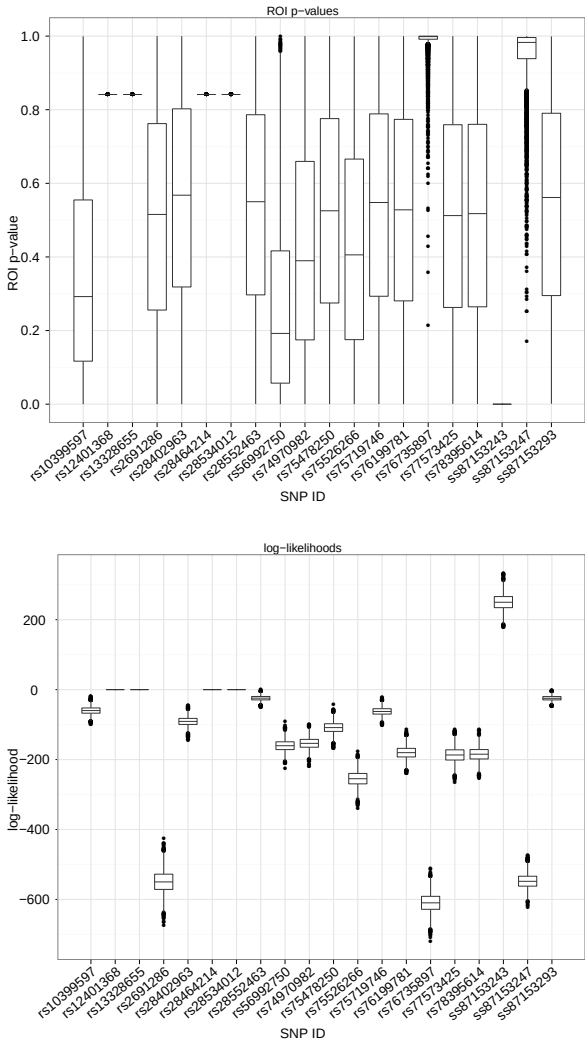

Figure 3.5: This figure shows the distribution of the ROI test statistic and the log-likelihood differences of the 20 candidates in the NCBI dbSNP database in the 9^{th} interval. The results are obtained from 5000 synthetic data sets of the test genome with coverage equal to 200 measures/per sequencing enzyme nick site. The top panel shows the quartiles and ranges of the ROI test statistics for each SNP. The bottom panel gives the difference of the log-likelihoods of the alternative hypothesis (existence of the SNP on test genome) and the null hypothesis (the test genome is identical to the reference genome)

propose to use the SNP-specific ROI algorithm to pick these SNPs sequentially and the effect of their co-existence can be obtained by adding the effect of the occurrence of each individual SNP.

On the other hand, two SNPs can be close enough that they appear in some of the neighborhoods simultaneously, the $\{\mu_j, \nu_j\}$ pairs will be affected by both of the SNPs. In this situation, we consider these two SNPs have second order interaction. In Section 3.1 we listed 8 types of SNPs. Their interactions can be much more complex. Figure 3.6 gives a simple example of when two SNPs happen in same neighborhood and both change one base from G to A. In this example the original number of labeled base A is 7, which is altered to 9 by the 2 SNPs. The complexity can escalate significantly when some SNPs alter the nick sites (AAG or GAG) of the sequencing enzyme or the ending patterns (AAA,CCC, GGG or TTT).

Figure 3.6: Interactions of two SNPs: The number of labeled bases (A) is changed from 7 to 9.

In some rare situations, more than two SNPs can have interactions. We consider these types of interactions as three or higher order interactions, while the second order interactions refer to only two SNPs which have interactions with each other. Since the percentage of second order interactions is not negligible, we need to consider them during analysis. On the other hand the third or higher order interactions are rare enough to be neglected in order to avoid the associated extremely high cost of computational complexity. Each second order interaction will be considered as a new entry in the SNP database and effectively treated as a distinct potential SNP. If an interaction is picked by the algorithm, both of its SNPs will be labeled as selected. This means a SNP with second order interactions usually have two approaches to select: (1) It is selected because it has a significantly small p-value and has a higher likelihood than that under the null hypothesis, (2) Its interactions with other SNPs are selected because of a significantly small p-value and a higher likelihood than that under the null hypothesis. Thus, the p-values for this SNP and its interactions

need to be adjusted by Bonferroni rules. The p-value of a SNP is multiplied by the candidate entries in the database which include this SNP. Since one interaction is compost of two SNPs, to ensure the false positive rate is well controlled, the p-value of the interaction is multiplied by the larger multiplier of the individual SNPs.

The SNP database size is enlarged by considering the second order interactions. In our ongoing example, the 9^{th} interval has 20 SNPs. There are 4 second order interactions among these SNPs. Thus, we have 24 entries in the SNP database for this interval. In next section we discuss how to detect SNPs on test genome when second order interactions are under consideration.

3.7 General Situation: Multiple SNPs in one interval

In Section 3.1 we discussed the situation when there is only one SNP happens in the target interval of the test genome. A more general situation involves multiple SNPs. In this section we introduce an algorithm which uses the idea of *Analysis of Competing Hypotheses (ACH)* to detect SNPs on the test genome.

3.7.1 Sequential Approach with ACH

The ROI method enables us to test individual variations from the SNP database. Then we need a protocol to select an optimal subset of variations. For this task , *Analysis of Competing Hypotheses* (ACH) is used here to select SNP sequentially. ACH has been studied for many years. As a step forward analysis methodology in intelligence analysis, ACH provides an unbiased approach to evaluating multiple competing hypotheses based on observed data. It was first developed by Richards Heuer in the $1970s$[17]. ACH is a step forward methodology for evaluating multiple competing hypotheses. The ACH procedure involves identification of hypotheses and calculation of evidence. Here we set the alternative hypotheses to the NCBI dbSNP database and use ROI test statistics as evidence. Meanwhile we use log-likelihood as competing mechanism among the alternative hypotheses. Each time we only pick up one SNP with the highest likelihood value, which is used to update the reference genome. Then the ROI test statistics for all other SNPs are

updated according to the updated reference genome. The ACH procedure is recursively processed as following:

1. Calculate the ROI test statistic and p-value for each variation in the SNP database. Meanwhile, the difference of log-likelihoods under this alternative hypothesis and the null hypothesis is calculated.

2. Adjust the p-value by Bonferroni correction for the second order interactions or the SNPs included in some second order interactions. The p-values of each SNP will be multiplied by the number of candidate entries including this SNP. The p-values of each interaction will be multiplied by the larger multiplier of its SNPs

3. Select candidates if the adjusted p-values are below a threshold ($\alpha = 0.01$). If no candidate satisfies the requirement then exit.

4. Select the candidate with maximum likelihood from those candidates in step (3). If this maximum likelihood is lower than the likelihood value of the null hypothesis, then exit.

5. Update the reference genome interval by applying the candidate from (4). Mark this variation as selected so that it won't be processed again in the later iterations.

6. If all SNPs in the SNP database are marked as selected then quit. Otherwise recalculate the ROI p-value and the log-likelihood for each entry in the SNP database based on the updated version of the reference genome and go to step (2).

3.7.2 An Example of ROI/ACH

To illustrate ROI/ACH algorithm clearly, we give a detailed example in this section. In this example, the reference genome is Human build 36 and we will focus on the 9^{th} interval of the first chromosome. There are 24 candidates (20 SNPs and 4 second order interactions) in the NCBI dbSNP database. Among the 20 SNPs, 5 SNPs on the test genome: ss87153243, rs7497098, rs56992750, rs28552463, rs13328655. The SNPs rs7497098 and rs56992750 have a second order

interaction. With coverage 200, we generate a synthetic data set and apply the ROI/ACH methodology described in last section. The details are shown in Table 3.1. The selection is done with 4 iterations. p1, p2, p3, p4 are p-values for each iteration, and L1, L2, L3, L4 are the differences of the log-likelihoods under alternative and null hypotheses. Both the SNPs and interactions are listed in this table.

Table 3.1 shows how the SNPs are selected by ACH. In the first round, ss87153243 is selected first because its p-value is less than the significance level 0.01 and it carries the maximum likelihood, which is higher than that under the null hypothesis. In the second round, interaction2 is selected by the same procedure. Since interaction2 is the interaction of SNP rs74970982 and rs56992750, both of them are selected. In the third run SNP rs28552463 is selected. Our final SNP set is { ss87153243, rs74970982, rs56992750, rs28552463 }, while the true SNP set is { ss87153243, rs74970982, rs56992750, rs28552463, rs13328655}. The final selection set includes all the element in the true set with only one exception: SNP rs13328655. This SNP is not detectable because, as we discussed previously, it happens in a gap where there are no neighborhoods.

The example in Table 3.1 shows details of how we process one synthetic data set. To access the detected probabilities of SNPs, we generated 5000 synthetic data sets of this interval for the test genome with coverage 200 and applied our methodology. The result shows that the detected probabilities of { ss87153243, rs74970982, rs56992750, rs28552463 } are 100% while the rest SNPs have 0% detected probabilities.

Two more things need to be mentioned about Table 3.1. First, the p-values for interactions and their single SNPs have been adjusted by Bonferroni correction. Secondly, in the first run the SNP "rs75478250" carries a significantly small p-value and a higher likelihood than that under the null hypothesis. If there is no competing mechanism, this SNP will be falsely selected. With the competing mechanism applied, SNP ss87153243 is selected first because it has maximum likelihood. After applying ss87153243 to the reference genome and updating the p-values and

Table 3.1: Competing hypothesis algorithm for the 9^{th} interval of the first chromosome, "x" denotes selection.

SNP ID	p1	L1	p2	L2	p3	L3	p4	L4
rs78395614	1.00	-216.22	1.00	-216.22	1.00	-216.22	1.00	-216.22
rs76199781	0.98	-160.32	0.98	-160.32	0.98	-160.32	0.98	-160.32
interaction1	1.00	-303.88	1.00	-303.88	1.00	-303.88	1.00	-303.88
rs74970982	0.00	320.14	0.00	320.14	x	x	x	x
rs56992750	0.00	346.45	0.00	346.45	x	x	x	x
interaction2	0.00	484.88	0.00	484.88	x	x	x	x
rs75526266	0.36	-322.01	0.36	-322.01	0.36	-322.01	0.36	-322.01
rs75719746	1.00	-75.37	1.00	-75.37	1.00	-75.37	1.00	-75.37
interaction3	0.32	-442.36	0.32	-442.36	0.32	-442.36	0.32	-442.36
ss87153243	0.00	368.70	x	x	x	x	x	x
rs76735897	1.00	-884.42	1.00	-884.42	1.00	-884.42	1.00	-884.42
rs77573425	1.00	-221.79	1.00	-221.79	1.00	-221.79	1.00	-221.79
interaction4	1.00	-777.07	1.00	-777.07	1.00	-777.07	1.00	-777.07
rs10399597	0.21	-40.25	0.21	-40.25	0.21	-40.25	0.21	-40.25
rs28402963	0.85	-132.51	0.85	-132.51	0.85	-132.51	0.85	-132.51
ss87153247	1.00	-692.83	1.00	-692.83	1.00	-692.83	1.00	-692.83
rs75478250	0.00	54.65	0.17	-43.69	0.17	-43.69	0.17	-43.69
rs2691286	0.77	-504.15	0.77	-504.15	0.77	-504.15	0.77	-504.15
ss87153293	1.00	-110.02	1.00	-110.02	1.00	-110.02	1.00	-110.02
rs28552463	0.00	34.64	0.00	34.64	0.00	34.64	x	x
rs28534012	0.84	0.00	0.84	0.00	0.84	0.00	0.84	0.00
rs28464214	0.84	0.00	0.84	0.00	0.84	0.00	0.84	0.00
rs13328655	0.84	0.00	0.84	0.00	0.84	0.00	0.84	0.00
rs12401368	0.84	0.00	0.84	0.00	0.84	0.00	0.84	0.00

likelihood values, SNP rs75478250 does not have a significant p-value any longer and its likelihood value is smaller than that under the null hypothesis. For those SNPs which are close to each other, the competing mechanism will pick the one with maximum likelihood and apply it to the reference genome and therefore the other nearby SNPs will be prevented from being selected.

3.8 Verification of ROI and ACH

3.8.1 Overview

To access the sensitivity and specificity of ACH/ROI methodology, we extend our analysis method to the genome wide synthetic data set according to following information:

1. a reference genome, which will be used as a baseline for comparison. We chose human build 36 as the reference genome.

2. a test genome, which will be used to generate synthetic data sets. As the second fully sequenced human genome, James Watson sequence [5] has 3.3 million SNPs as compared with the human build 36. The test genome is obtained by applying these 3.3 million SNPs to the reference genome.

3. a SNP database. We get much higher testing power by limiting the alternative hypothesis to an existing knowledge database. We downloaded a SNP database from the NCBI dbSNP website. The database contains 15 million SNPs including the 3.3 million James Watson SNPs.

Among all James-Watson SNPs, a small proportion (0.085%) alter the mapping enzyme interval structure and therefore can be caught by comparing the structure with that of the reference genome. These SNPs can be detected with existing methods used in OM[28]. To simplify the analysis procedure, we assume that these SNPs have been discovered and are applied to the reference genome to generate an updated version, which has the same mapping enzyme interval structure and is used as a baseline for the later analysis. The test genome is obtained by applying all James-Watson SNPs to the reference genome. Based on the proportion vectors of the test genome we

generate synthetic data sets, which are used to detect James-Watson SNPs.

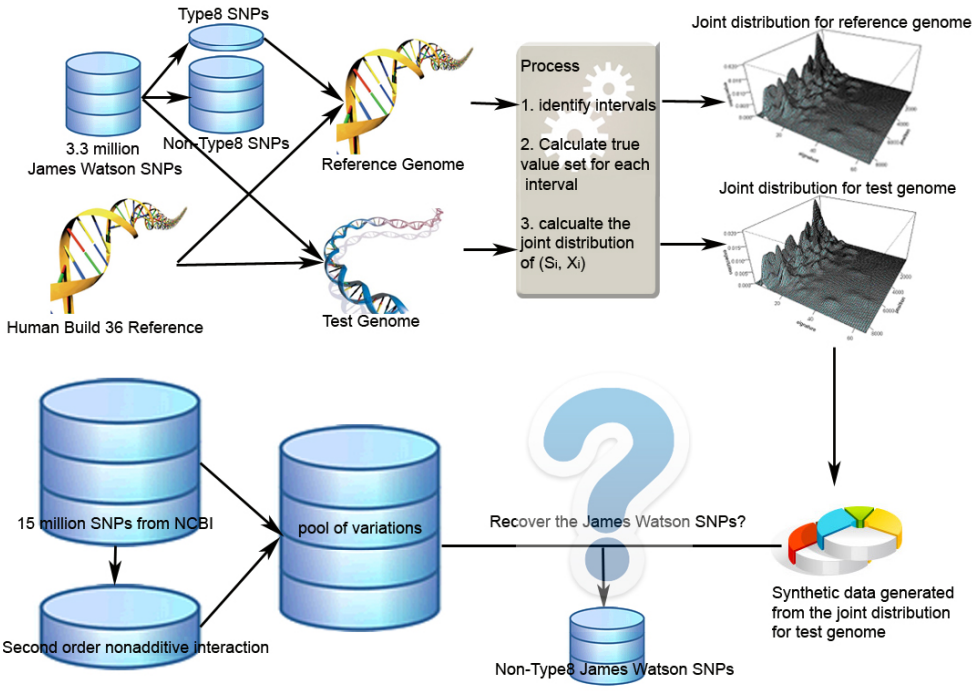

Figure 3.7: Synthetic data set generation procedure. The 3.3 million James-Watson SNPs are divided into two subsets. One contains SNPs that can alter the nick site structure of the mapping enzyme. According to the SNP classification rules, these SNPs are type 8 SNPs. They are applied to the human build 36 and generate an updated version of the reference genome. All other types of SNPs are belong to the second subsets and are applied to human build 36 to generate the test genome. The synthetic data sets are generated based on the frequency matrices measured on the test genome.

The entire procedure for generating genome wide synthetic data sets is described in Figure 3.7. In the next section, we apply the proposed ROI/ACH method to the synthetic data sets and reveal the James-Watson SNPs from the 15 million NCBI dbSNP database.

3.8.2 Detection Rate with Complete Database

As stated before, coverage is directly related with the detection rate and cost. Higher coverage means more data are observed. Hence the detection rate is expected to be higher while the experiment cost is also increased. It is of our major interest to discover how the coverage affects the detection rate, so that we can find a balanced point between detection rate and cost.

To investigate this topic, we identify 5 different coverages: 100, 200, 300, 400 and 500 (measures per nick site). For each of the coverage, three genome wide synthetic data sets were generated and processed using the ROI/ACH algorithm. Then three detection rates are calculated under each coverage, from which the mean and variance values can be calculated. Figure 3.8 shows how the mean value of detection rates will be affected by the coverage values. Meanwhile, the standard deviation of the detection rates are also labeled. Besides of the detection rates, the mean value and standard deviation of false positive rates are also shown in this figure with the right axis as Y axis. To provide an intuitive impression of how many copies of whole genome are needed, we also put the numbers of whole genome copies at the extra X axis according to with corresponding digestion rates.

It is shown in Figure 3.8 that the detection rates and the coverage values have strong positive correlations. When the coverage is above 400 measures per nick site of sequencing enzyme, the increment of detection rate starts to saturate. With higher coverage, the cost and time of the entire procedure increase accordingly. Thus, we need to balance the cost and accuracy when determining the coverage value. Meanwhile, the genome wide false positive rate is always smaller than the significance level of 0.01 and decreases while the coverage values increases.

3.8.3 Reasons of Non-discovered James-Watson SNPs

Among all the James-Watson SNPs, 98.36% altered some $\{\mu_j, \nu_j\}$ pairs and, as a result, are detectable with large enough FS coverage. In the last section we generated three synthetic data sets and the average detection rate for these detectable SNPs is 96.42% when the coverage is equal

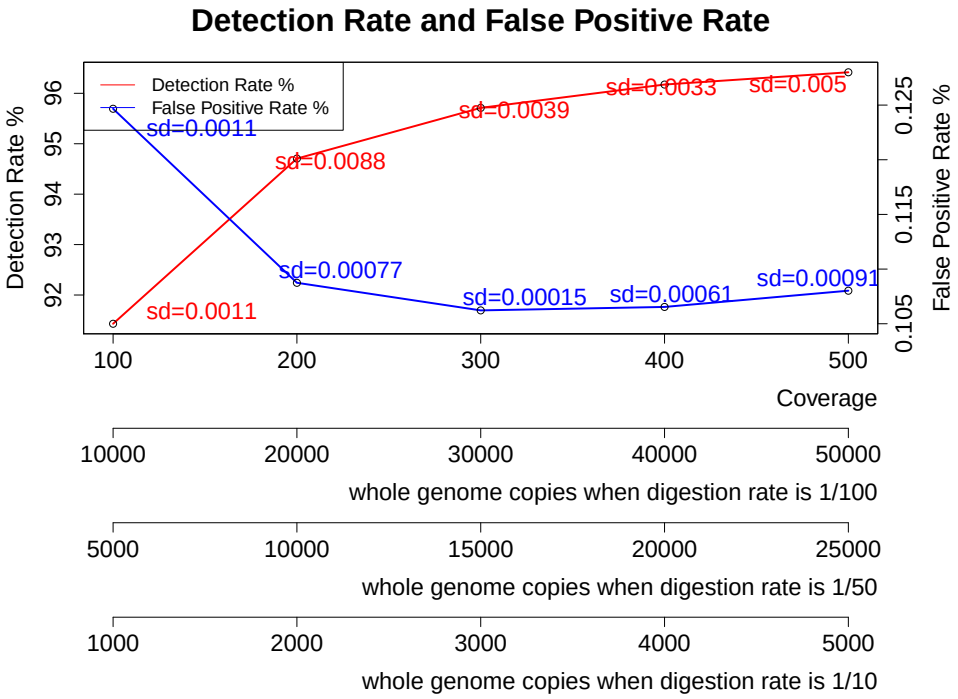

Figure 3.8: The detection rate and false positive rate vs. coverage. The red curve is the detection rate, which increases while the coverage increases. The blue curve is the false positive rate with the y axis on the right side of the plot, and it is controlled under the significance level of 0.01. We generate genome wide synthetic data sets and apply the ROI and ACH method. The entire procedure is repeated for three times. Each time we get a detection rate and a false positive rate. The points in plot are mean values of these rates and the corresponding standard deviations are expressed as text beside the points.

to 500. The remaining 3.58% of them are not detected. It is of interest to investigate the reasons for this. There are several important factors to consider with: (1) Since larger signature and position values will result in larger variance, we are expecting that the detection rates will be affected directly by these two factors. (2) We used ACH to pick SNPs from the SNP database. More candidates indicate a higher chance of false negatives. Therefore, the number of candidates in the database will have an effect on the detection rate.

The first factor we need to check is the minimum distance of a SNP, which is defined to be the shorter distance from the left end or right end of the interval to this SNP. Here we will look into the relationship between the detection rates and the minimum distances of the SNPs. The minimum distances range from 1 to 86470 in this analysis. It is complicated to draw the detection line against each point of the minimum distances. So, we group the minimum distances by the following divisions: $\{1-100, 101-200, ..., 9900-10000, >10000\}$.

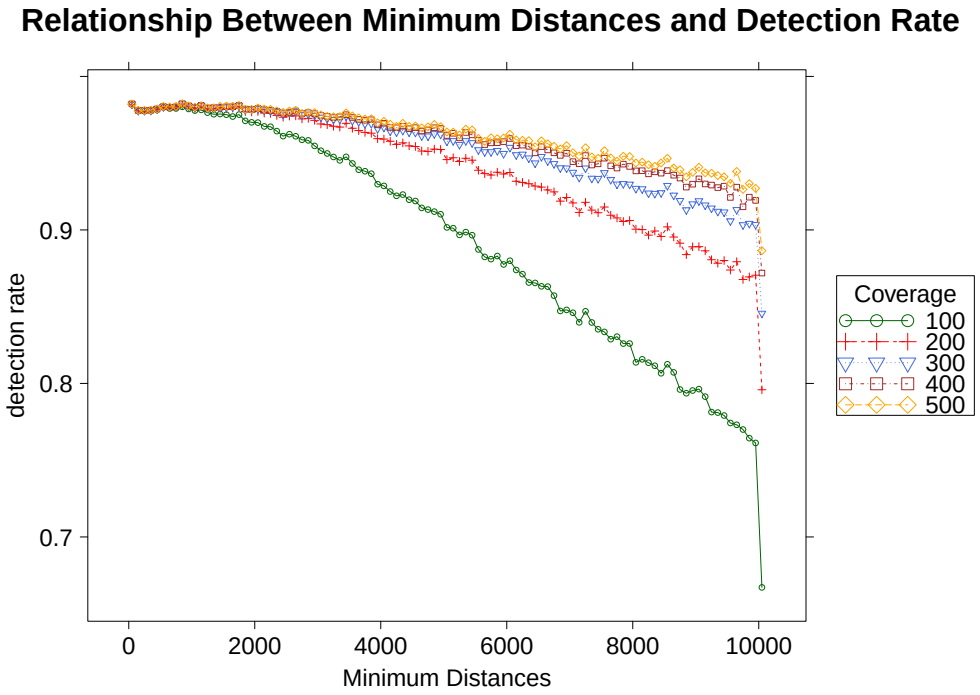

Figure 3.9: The relationship between the detection rate and the minimum distances of the SNP under different coverage.

The relationship between detection rates and the positions of the SNPs is shown in Figure 3.9, which gives us the following information :

1. When the distance value increases, the detection rate will decrease
2. The increment of coverage will benefit the detection rates, especially for those SNPs with larger position values
3. The increment of the detection rate brought by increasing the coverage will start to saturate when the coverage value is larger than 300.

The second factor of interest is the value of signatures, which are determined by the number of fluorescently labeled bases in the neighborhood μ_j. Different from position measurements that are gathered around the SNP positions, there may be several μ_j values altered by a SNP. The signatures from smaller μ_j values are more precisely measured and, thus, whether SNPs alter μ_j values tends to be easier to detect. Figure 3.10 shows the relationship between detection rates and the minimum of μ_j values altered by the SNPs.

Figure 3.10 suggests that a smaller minimum value of μ_j will result in a higher detection rate provided that the coverage is high enough. When the coverage is small there is a local minimum detection rate of around 20. This is caused by measurement error of both the position and the signature. When we limit the position to a more narrow range, as shown in the bottom plot of Figure 3.10, the negative effect of the increment of μ_j values on the detection rate is very clear. The local minimum around 20 will be eliminated when the coverage is large enough, which provides more accurate measurements of both positions and signatures.

The last factor we want to examine is the number of candidates in the SNP database. More candidates in one interval will decrease the chance of the true SNPs being discovered. The type-I error is still under control. The increment of the total number of candidates indicates a higher chance for the James-Watson SNPs being misrepresented by some other non-James-Watson SNPs.

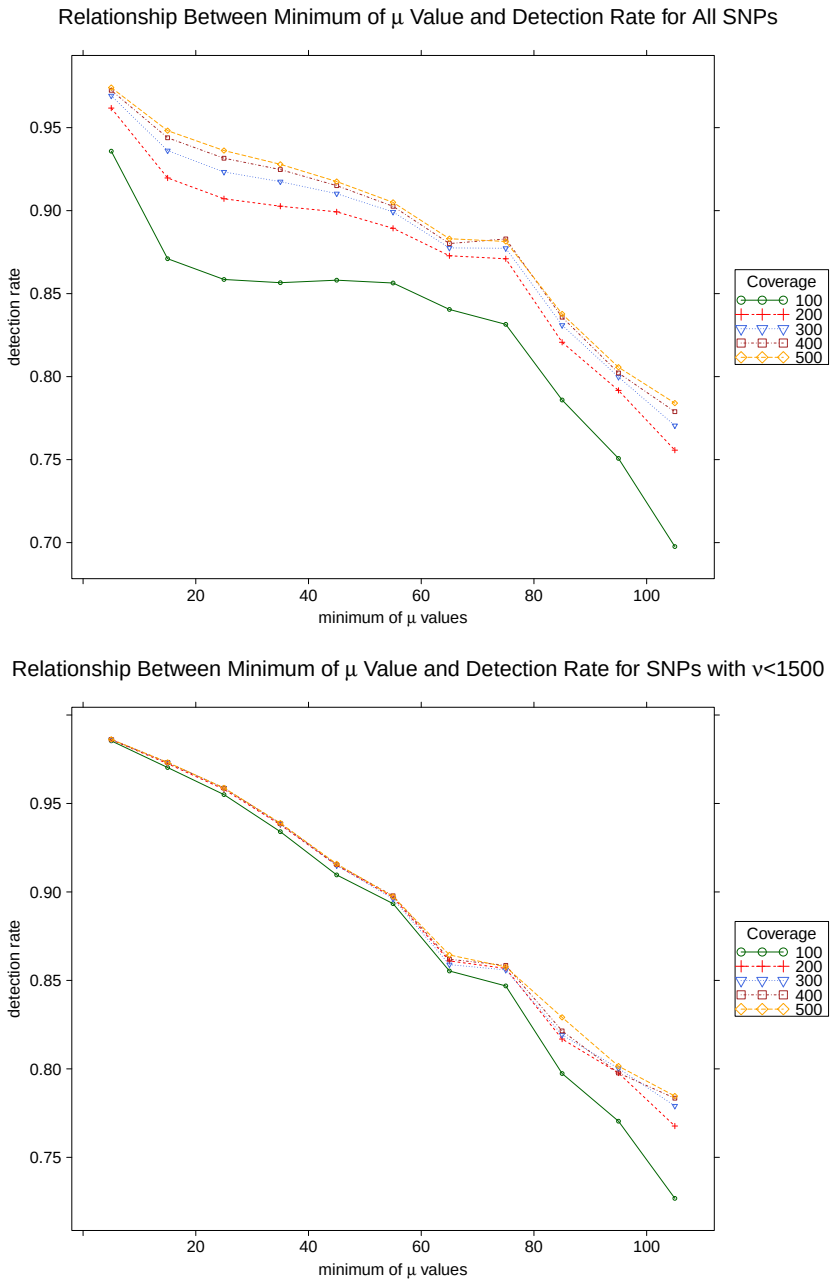

Figure 3.10: The relationship between detection rates and the minimum of μ_j values. The top one is based on the entire population of James-Watson SNPs while the bottom one is based on those James-Watson SNPs with positions from 0 to 1500.

Figure 3.11 shows how the number of candidates in an interval affects the detection rates of SNPs in the same interval.

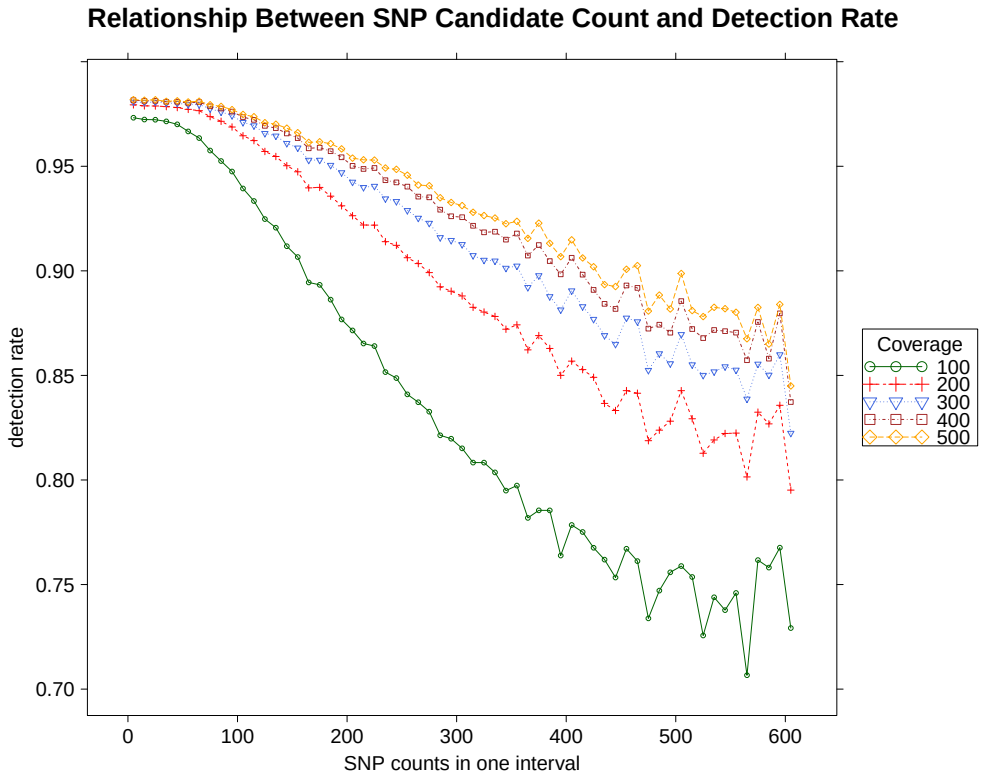

Figure 3.11: The relationship between the detection rate and the number of SNP candidates in the interval.

As anticipated, the three factors (position of SNP, the minimum μ_j value altered by the SNP and the number of SNP candidates in the interval) all have an effect on the detection rate. By fitting a logistic regression model as following, their effects are represented by the negative coefficients:

$$\log\left(\frac{d_r}{1-d_r}\right) = 3.75 - (0.00012) * X_s - (0.025) * S_m - (0.0014) * Count$$

where d_r is the detection rate, X_s is the position value of the SNPs, S_m is the minimum of the μ_j values and $Count$ is the candidate number in the same interval. This model shows that all the three factors have negative effect on the detection rates. In the next chapter we will discuss the situation where we have a more precise measurement, such as nanopore technology, which is expected to result in a higher detection rate.

3.9 Result with More Precise Measurements

In this section we will first discuss how the detection rate can be improved by more precise measurements. New technologies such as nanopore sequencing can greatly improve the measurement precision. In this section, we reduces the measurement errors of signature and position to $1/10$ of those used in Section 3.8.1. The detection rates with these improvement of measurements are calculated and compared with the original ones. We also included a list of possible alternative mapping enzyme for the more precise measurement. The current mapping enzyme Nt.SapI is selected because the average interval length of the intervals (8000 base) is about the precision our current measurement method can detect.

Among the many technologies under development to refine the process of measuring the positions and intensity of fluorescently labeled sites, the nanopore sequencing technology is a promising approach. It was developed in 1995 [7][18] to detect the DNA sequencing order by measuring the electric current when a DNA molecule passes through a nanopore. A nanopore has an internal diameter of 1 nanometer and is usually made from certain porous transmembrane cellular proteins, silicon holes with the ion-beam sculpting method or graphene. As shown in Figure 4.1, a copy of molecule is passing through a nanopore and the electric current level will be changed upon certain events, such as fluorescently labeled sites on the molecules.

As of today, the nanopore sequencing method is still under development and is not ready for commercial application. We expect that it can be used in a FS system to measure the position and signature pairs, which will greatly reduce the measurement error. To check how a more precise measurement method can affect the detection rate, we reduce the standard errors of position and signature to $1/10$ of the ones described in Chapter 2 respectively. Figure 4.2 shows the comparison of detection rates under three situations : (1) original measurement error; (2) when the signature error deviation is reduced to $1/10$ of the original one; and (3)when position error deviation is reduced to $1/10$ of the original one. The detection rates are averages from 3 genome wide synthetic

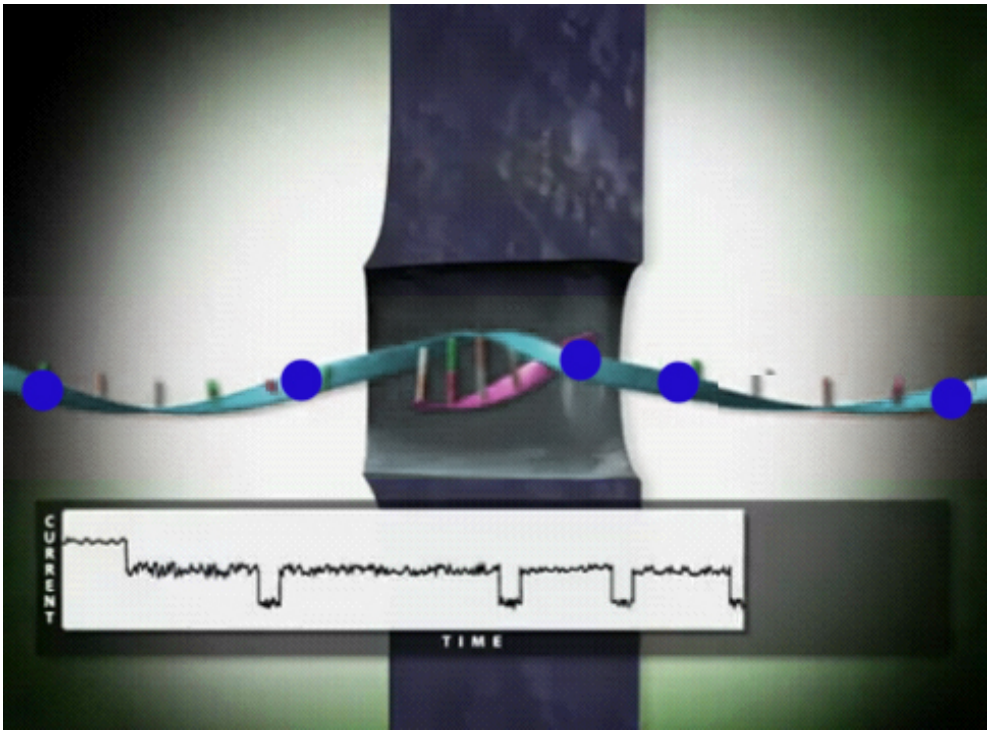

Figure 3.12: Nanopore Sequencing Technology. Measuring the electric current level when the molecules pass through the pore with 1 nanometer diameter.

data sets with coverage from 100 to 500. Figure 4.2 shows that the reduction of position error results in a significant performance improvement of detection rates, especially when the coverage is low. Meanwhile the reduction of signature error brings relatively small improve of the detection rate.

New measurement technologies like nanopore sequencing can detect nick sites with higher dense. Thus it will be possible to choose other mapping enzyme to replace the current Nt.SapI with recognition pattern "GCTCTTCN||" and an average interval length of around 8000. Table 4.1 lists other possible choice of mapping enzyme together with their expected average length of intervals. It is expected that with shorter interval the alignment of molecules and the position measurements of signatures will be more precise.

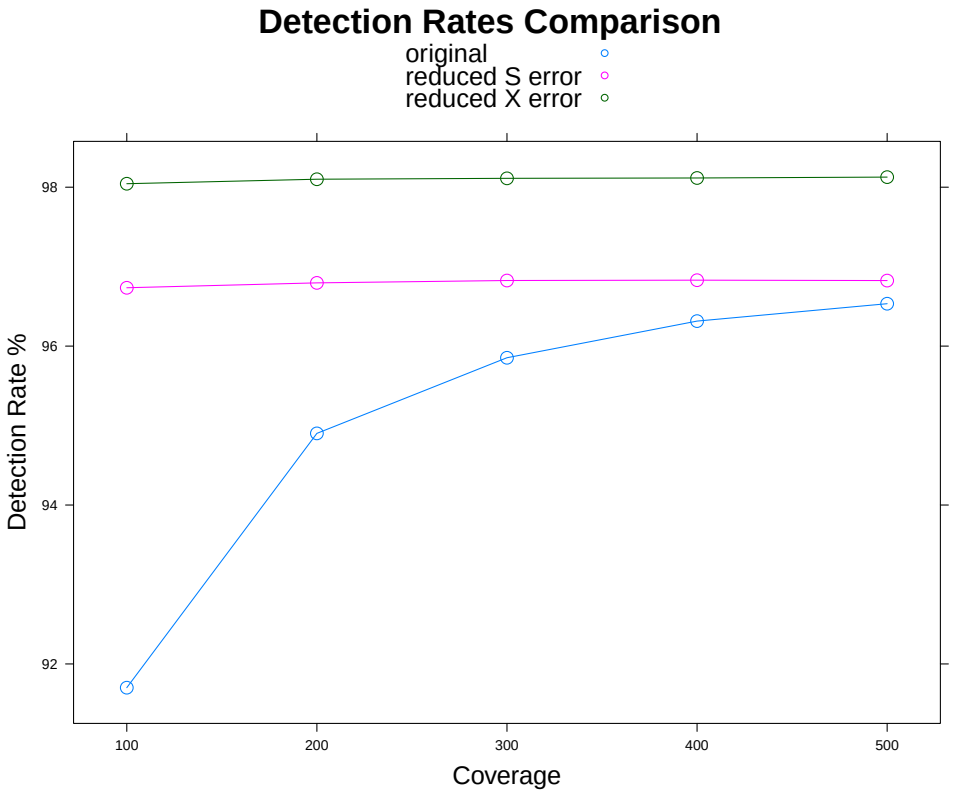

Figure 3.13: Detection rate comparison. The blue color dots are detection rates with error distribution described in Chapter 2. The pink color represents the situations when the signature error deviation is reduced to $1/10$ of the blue color case. The green color gives detection rates when the position error deviation is reduced to $1/10$ of the blue color case. Each detection rate is an average from 3 genome wide synthetic data sets.

Table 3.2: Available nicking enzymes for mapping purposes. Here N denotes any of the four bases.

Enzyme	Recognition Sequence	Suppliers	Average Length of Interval
Nt.A1wI	GGATCNNNN‖	New England Biolabs	512
Nb.BsmI	GAATG‖C	New England Biolabs	2048
Nb.BsmAI	GTCTCN‖	New England Biolabs	512
Nb.Bpu10I	CCTNA‖GC	Thermo Scientific Fermentas	2048
Nt.Bst9I	GAGTCNNNN‖	SibEnzyme Ltd.	512
Nt.BtsI	GCAGTG‖	New England Biolabs	2048
Nb.Mva1269I	GAATG‖C	Thermo Scientific Fermentas	2048

3.10 Detected Probabilities with Incomplete Database

In Section 3.8, our analysis is based on a complete SNP database, which means the database contains all the variations on test genome. In Section 3.10, we discuss the situation when the database is NOT complete. We randomly remove 10%, 50% and 90% SNPs from the complete NCBI dbSNP database. We will focus on the first contig of the first chromosome. Simulation studies are conducted to check how the detected probabilities of James-Watson SNPs are affected. Meanwhile we are interested in the false positive SNPs. It is expected that some non-James-Watson SNPs tend to be selected with higher chance than others because they can explain the changes caused by nearby James-Watson SNPs which are not included in the SNP database. In this section, we will focus on the first contig of the first chromosome. The purpose is to find answers to the following questions:
(1) How are the detected probabilities of the James-Watson SNPs affected?
(2) Are there any patterns of the falsely selected SNPs?

The FS data profile the test genome by generating binary indicators of variations in a database. With a smaller size database, the profile will be more rough. In the sense of carrying reduced information, an analogy is given in Figure 4.3 between the JPEG file formatting and the FS profiling. From the left to the right, the JPEG file size is reduced significantly but the major information is not lost. Similarly, the complete 3G genome sequence is profiled to be binary output of FS data. When the database is not completed, the binary output will be reduced but still carry significant genome information, such as structural variations.

To investigate how FS performs in a situation when the database is not completed, we use the first contig of the first chromosome as an example. In this contig there are 33 intervals and 146 SNP candidates, among which 20 are James-Watson SNPs. Besides of SNPs, there are 39 second order interactions. Thus the complete database will have $146 + 39$=185 variation candidates. We randomly remove the SNPs from the complete database to generate four sets of SNP databases.

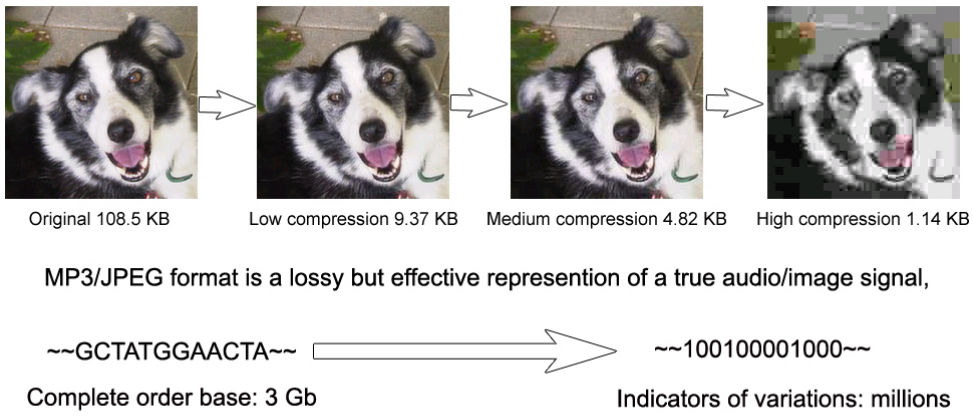

Figure 3.14: Reduced information. The above shows how a image of dog is compressed to a much smaller size by reducing detail information. The lower one shows how the complete order of a genome sequence is represented by an indicator of variations provided by FS.

Whenever a SNP is removed, all its related second order interactions are also removed.

Database	Content
$D0$	Complete database
$D9$	90% of the Complete database
$D5$	50% of the Complete database
$D1$	10% of the Complete database

where $D1 \subset D5 \subset D9 \subset D0$.

For each SNP database, with coverage 500 we generate 500 synthetic data sets for the test genome which is obtained by applying James-Watson SNPs to the reference genome. Then ROI/ACH algorithm is applied to each data set. The purpose is to discover detected probabilities of existing James-Watson SNPs and find a pattern of those falsely selected SNPs. It is expected that some non James-Watson SNPs will be selected with hither probabilities because they explain the change caused by nearby James-Watson SNPs which are not included in the incomplete database.

Table 4.2 shows the detected probabilities for each of the James-Watson SNPs given a certain database. This table indicates that :

1. With a complete SNP database we got the 100% detected probability for 19 of the SNPs. One of the SNP "ss87153195" is in the gap of the intervals and is not detectable by FS. Thus its detected probability is 0%.

2. When we reduce the database to 90% of the complete database, only one James-Watson SNP is removed and its detected probability becomes 0%. The detected probabilities of other James-Watson SNPs remain 100%

3. When the database is reduced to 50% of the complete database, the situation is similar to that of the 90% incomplete database except one SNP: ss87153452. We will investigate later to explain why this SNP has a very low detected probability.

Table 3.3: Detected probability with complete or incomplete SNP database. We listed four situations here: Complete database $D0$, 90% of the complete database $D9$, 50% of the complete database $D5$ and 10% of the complete database $D10$. For each of the SNP databases there are 500 synthetic data sets with coverage 500. The "probability" column gives the detected probability of each SNP among these 500 synthetic data sets given a certain database. The "Included" column indicates whether the SNP is included in the database or not.

SNP ID	Complete DB	90% DB		50% DB		10% DB	
	Probability	Included	Probability	Included	Probability	Included	Probability
ss87152653	100%	1	100%	0	0%	0	0%
ss87153174	100%	1	100%	0	0%	0	0%
ss87153179	100%	0	0%	0	0%	0	0%
ss87153195	0%	1	0%	0	0%	0	0%
ss87153200	100%	1	100%	0	0%	0	0%
ss87153243	100%	1	100%	0	0%	0	0%
ss87153247	100%	1	100%	0	0%	0	0%
ss87153293	100%	1	100%	1	100%	1	100%
ss87153308	100%	1	100%	1	100%	0	0%
ss87153313	100%	1	100%	0	0%	0	0%
ss87153318	100%	1	100%	0	0%	0	0%
ss87153346	100%	1	100%	1	100%	0	0%
ss87153350	100%	1	100%	1	100%	0	0%
ss87153371	100%	1	100%	0	0%	0	0%
ss87153376	100%	1	100%	0	0%	0	0%
ss87153386	100%	1	100%	1	100%	0	0%
ss87153391	100%	1	100%	0	0%	0	0%
ss87153447	100%	1	100%	0	0%	0	0%
ss87153452	100%	1	100%	1	0.4%	0	0%
ss87153462	100%	1	100%	0	0%	0	0%

4. When the database is reduced to 10%, the included James-Watson SNPs still have a 100% detected probabilities.

As stated above, the SNP ss87153452 has unexpected low detection probability for the 50% database. It is found that this SNP has second order interaction with other SNP ss87153447. Their effect on FS data is described in Table 4.3. When both SNPs are included in the database, their second order interaction will also be included and the ROI algorithms will locate the regions carrying significant changes precisely. When only SNP ss87153452 is included in the database, the ROI is identified by the alternative hypothesis that only this SNP is on the test genome, which is shown in the top panel of Figure 4.4. But the test genome actually have both ss87153452 and ss87153447 and its contour plot in this region is shown in the bottom panel of Figure 4.4. We can see that the ROI didn't correctly select the regions due to the fact of incomplete information offered by the database. In the top panel, the maximum absolute difference in the selected regions is $1.02e-05$ while in the bottom panel the maximum absolute difference in the selected regions is $2.42e-08$. This shifting of ROI brings great difficulty to detect the SNP ss87153452 even though it is included in the database.

The second major question we are interested in is what will happen to those 126 non James-Watson SNPs when the database is incomplete. Based on the analysis of the synthetic data set, it is found that the majority of non James-Watson SNPs have 0% detected probabilities. There are four exceptions when the database is 90% of the complete database: "rs58108140" "rs10218492" "rs75609629" "rs76402894". Their detected probabilities are 0.8%, 0.4%, 0.2% and 16.2% respectively. The SNP "rs76402894" has such a high detected probability because it is close to the missing James-Watson SNP ss87153179 and explains part of the difference caused by this James-Watson SNP.

In the above discussion the synthetic data sets are generated based on the test genome. To check the false positive rates under null hypothesis, we generates 500 synthetic data sets for the reference genome. The analysis results show that all the SNPs including the James-Watson SNPs

Table 3.4: This table shows the μ_j and ν_j values changed by SNP "ss87153447" and "ss87153452" for the data set with C base labeled. SNP "ss87153447 "changes base T to base C in the 18^{th} interval at position 4343 while SNP "ss87153452" changes base G to A in the same interval at position 4349. SNP "ss87153452" does not have effect on other three data sets. When only this SNP is in the SNP database, it is very hard to detect it because of shifting of ROI.

SNP ID	Original μ_j	New μ_j	Original ν_j	New ν_j
ss87153447	16	3	4320	4348
ss87153447	14	1	4317	4345
ss87153447	25	18	4318	4322
ss87153452	16	15	4320	4320

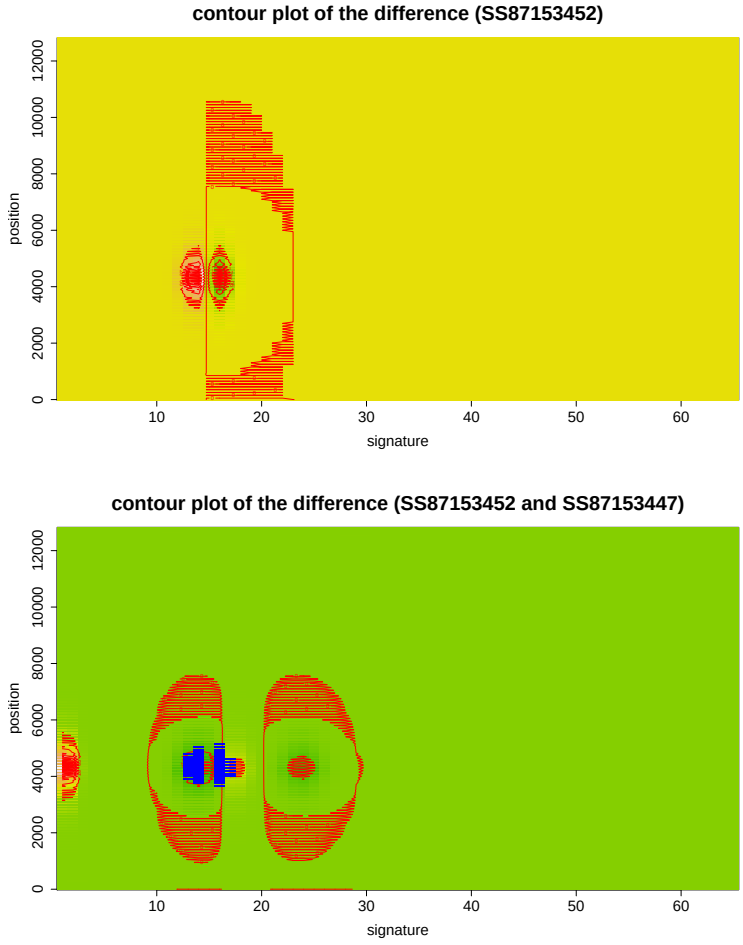

Figure 3.15: The top plot shows the contour plot of difference between proportions under null and proportions when SNP ss87153452 is applied to test genome; while the bottom one shows the situation when both SNP ss87153452 and ss87153447 are applied. Color green and yellow denotes the regions with close to 0 values. The redlines indicate the regions with major differences caused by the SNPs.

and non James-Watson SNPs have detected probability 0 or very close to 0. Thus, the high falsely detected probability will only happen when the test genome is different from the reference genome.

Different SNP database will provide us with different levels of information. But the information carried by the FS data is always there. In this sense, when the variation database is updated, the

analysis can also be updated based on the same FS data. And we only need to redo the analysis on those intervals with new variations in the updated database.

Chapter 4

Summary

4.1 Major Issues

As a novel sequencing method, FS has advantages in catching structural variations. However, the detection of variations like SNPs is difficult due to the fact that the effect of SNPs on FS data is very subtle and the current measurement methods produce relatively large measurement errors. SNPs can be grouped into 8 types based on their effect on FS data, which are described in Section 3.2. Among the 8 types, the first type is the most common type (about 39.60%). The difference between the original distribution of (S, X) and the distribution after applying this SNP is calculated. The contour plot of this difference is shown in Figure 3.2. The maximum of the difference is lower than 1e-5. This subtle difference cannot be detected by classical methods which generate very conservative p-values (Section 3.4). We propose ROI statistics to solve this issue.

The second difficulty of analyzing FS data is how to analyze the interactions among SNPs. Usually two SNPs can be far away from each other so that they won't happen in one neighborhood. However there are still non-negligible amount of SNPs which are close enough so that they can alter one (μ_j, ν_j) pair simultaneously. The interaction can be very complex since there are 8 types of SNPs. To get higher detection rate, we include the second order interactions (only two SNPs happen in same interval simultaneously) in the database, while the third or higher order interactions are ignored because they are very rare situations and considering them will result in extremely high computational cost.

The third challenge during the analysis is the competition among SNPs. We calculate ROI test statistic for each SNP within one interval. The p-values of some SNPs will be lower than the threshold. But not all of these SNPs are on the test genome. The SNPs close to a true SNPs have higher chance to carry small p-values because their effect on FS data is similar to that of the true SNP. To solve this issue we use log-likelihood as the competition criteria and propose an ACH algorithm to select SNPs sequentially.

4.2 Analysis Approaches

Since each interval is equivalent in analysis, we focus on one interval to discuss the analysis methods and modeling. The methodology will then be applied to each interval when we analyze the whole genome data.

Before the analysis, we simulated sets of synthetic data because there is little experimental data available. Based on two pilot experiments, we modeled the empirical error distribution of S and X. (Section 2.3 and 2.4). The joint distribution of (S, X) was proposed in Section 2.5 based on the assumption that S and X are conditionally independent given $Z = j$.

With the synthetic data sets, we can proceed to the analysis phase. Since the classical goodness of fit test statistics are too conservative for FS data, the ROI test statistics is proposed. To avoid of computational complexity and facilitate the ROI test statistics, the joint distribution panel of (S, X) is discretized into many grid cells. The first step of calculating ROI test statistics is to locate those grid cells which carry the major difference. Then the ROI test statistic is given by calculating the log-likelihood ratio of the counts in the selected regions. The test statistic is approximately normal distributed given large amount of observations. The approximation is supported by a simulation study (Figure 3.4). An example of ROI test statistics is given in Section 3.5.2. In this example we calculated the ROI p-values for 5000 synthetic data sets and all the p-values are significant small. (Figure 3.4). On the other hand, the classical goodness of fit test statistics are very conservative. There are 4.24% of the log-likelihood ratio test p-values for the 5000 synthetic

data sets are significantly small and only 1.3% of the Person's Chi-squared test p-values are significantly small. This simulation study shows that ROI can improve the testing power dramatically.

For each of the SNP we calculate the ROI test statistic. When there are multiple SNPs, we need to consider the interactions and competition among SNPs. Two SNPs can be very close so that they can alter one pair of (μ_j, ν_j) simultaneously. In this situation, we consider these two SNPs have a second order interaction. The behavior of the interactions cannot be derived directly from the behavior of the individual SNPs. We propose to consider the second order interactions as entries in the database. Since second order interaction are not negligible, they are included in the database. We omit the three or higher order interactions , which are very rare, to avoid of extremely high computational complexity. One interaction example is shown in Figure 3.6.

When there are usually multiple SNPs in one interval, it is observed that not only the ROI p-values of the true SNPs but also those of the SNPs around these true SNPs will have great chance to be significantly small. To avoid of high false positive rate, we use Analysis of Competing Hypothesis (ACH) to select SNPs sequentially with log-likelihood as competing criterion. The ACH algorithm is described in Section 3.7.1 followed by a detailed example in Section 3.7.2.

To verify the ROI/ACH methodology, we generated three sets of whole genome FS data. The detection rates and false positive rates are shown in Figure 3.8. It is shown that when coverage (equal to number of measurements per (μ_j, ν_j) pair) is 200, the detection rate is about 95%. The false positive rate is always controlled under 1%. We also discussed the potential factors of the detection rates: 1. The minimum distance of a SNP, which is defined to be the shorter distance from the left end or right end of the interval to this SNP; 2. The minimum of μ_j value altered by the SNP; 3. The count of SNP candidates in the interval. The relationship between the detection rates and the above three factors are shown in Figure 3.9, Figure 3.10 and Figure 3.11. The result suggests that a better measurement of signature and position will benefit the detection rate greatly. Meanwhile if the interval length is shorter (by change the current mapping enzyme) then there will

be fewer candidates in one interval , which will result in a better detection rate.

We also did some simulation study to check how better measurement methods of signature and position can improve the detection rate. The result is shown in Figure 3.12, which indicates an urgent need of some better measurements such as nanopore technology.

At the end, we extend our discussion to the situation when the SNP database is not completed. This means that not all SNPs on the test genome are included in the database. We remove 10%, 50% or 90% of the SNPs from the complete database randomly. We then generate 500 synthetic data sets with the coverage 500. The detected probabilities of some typical SNPs given different database are shown in table 3.3.

4.3 Weakness and Limitations

First of all, this study only considers the data from the sequencing enzyme. We didn't consider the mapping enzyme data, which carry extra measurement or alignment errors. So it is expected that the detection rate of a real experimental data should be lower than those in Figure 3.8.

Secondly, we didn't consider alleles. Our reference genome contains only one strand of DNA and there is no allele information inside of the NCBI dSNP database. The analysis methods can be extended to consider alleles when we have knowledge of the allele probabilities, by which we can calculate the joint distribution of (S, X).

The majority of the analysis of this thesis is based on a complete database. In reality, we usually cannot have a database containing all the SNPs on the test genome. This could introduce very low detected probabilities of some SNPs. As shown in Figure 3.15, the regions of interest are not correctly selected because the SNP database is not completed. Both SNP ss87153452 and ss81753447 are on the test genome and they have a second order interaction. The real difference contour plot should be the bottom one, while in the database we only have SNP ss87153452 and

the regions of interest are selected based on the top plot. The miss-election of the ROIs result in extremely low detected probability of ss87153452 (0.4% out of 500 synthetic data sets with coverage 500).

LIST OF REFERENCES

[1] Information technology coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s part 3: Audio. *ISO/IEC*, pages 11172–3, 1993.

[2] An introduction to illumina next-generation sequencing technology for microbiologiests. *www.illumina.com*, pages 1–9, 2012.

[3] Lim A., Dimalanta E.T., Potamousis K.D., Yen G., Apodoca J., Tao C., Lin J., Qi R., Skiadas J., Ramanathan A., Perna N.T., Plunkett G. 3rd, Burland V., Mau B., Hackett J., Blattner F.R., Anantharaman T.S., Mishra B., and Schwartz D.C. Shotgun optical maps of the whole escherichia coli o157:h7 genome. *Genome research*, 11.9:1584–93, 2001.

[4] Ramanathan A., Huff E.J., Lamers C.C., Potamousis K.D., Forrest D.K., and Schwartz D.C. An integrative approach for the optical sequencing of single dna molecules. *Anal Biochem.*, 330(2):227–241, 2004.

[5] David A.W., Maithreyan S., Michael E., Yufeng S., Lei C., Amy M., Wen H., Yi-Ju C., Vinod M., Thomas G.R., Xavier G., Karrie T., Faheem N., Cynthia L.T., Gerard P.I., James R.L., Craig C., Xing zhi S., Yue L., Ye Y., Lynne N., Xiang Q., Donna M.M., Marcel M., George M.W., Richard A.G., and Jonathan M.R. The complete genome of an individual by massively parallel dna sequencing. *Proc Natl Acad Sci USA*, 452:872–876, 2008.

[6] Teague B., Waterman M.S., Goldstein S., Potamousis K., Zhou S., Reslewic S., Sarkar D., Valouev A., Churas C., Kidd J. M., Kohn S., Runnheim R., Lamers C., Forrest D., Newton M.A., Eichler E.E., Kent-First M., Surti U., Livny M., and Schwartza D.C. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107:10848–10853, 2010.

[7] George C., David W.D., Daniel B., Richard B., and John K. Us patent 5,795,782 (filed march 1995) characterization of individual polymer molecules based on monomer-interface interaction. 1998.

[8] Kang C. and Speller R. The effect of region of interest selection on dual energy x-ray absorptiometry emasurements of the calcaneus in 55 post-menopausal women. *The british Journal of Radiology*, 72:864–871, 1999.

[9] The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature*, 467:1061–1073, 2010.

[10] John C.V., Mark D.A., Eugene W.M., Peter W.L., Richard J.M., Granger G.S., and Hamilton O.S. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[11] Schwartz D.C. and Waterman M.S. New generations: sequencing machines and their computational challenges. *Journal of Computer Science and Technology*, 25(1):3–9, 2010.

[12] Church D.M., Goodstadt L., Hillier L.W., Zody M.C., Goldstein S., She X., Bult C.J., Agarwala R., Cherry J.L., DiCuccio M., Hlavina W., Kapustin Y., Meric P., Maglott D., Birtle Z., Marques A.C., Graves T., Zhou S., Teague B., Potamousis K., Churas C., Place M., Herschleb J., Runnheim R., Forrest D., Amos-Landgraf J., Schwartz D.C., Cheng Z., Lindblad-Toh K., Eichler E.E., and Ponting C.P. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, 7.5:e1000112, 2009.

[13] Tor D.W., Matthew C.K., Steven C.L., and John J. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, 26:99–113, 2005.

[14] Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes J.C., Hutchison C.A., Slocombe P.M., and Smith M. Nucleotide sequence of bacteriophage phix174 dna. *Nature Biotechnology*, 265:687–695, 1977.

[15] Sanger F., Nicklen S., and Coulson A.R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74:5463–5467, 1977.

[16] Lin J., Qi R., Aston C., Jing J., Anantharaman T.S., Mishra B., White O., Daly M., Minton K.W., Venter J.C., and Schwartz D.C. Whole-genome shotgun optical mapping of deinococcus radiodurans. *Science (New York, N.Y.)*, 285.5433:155862, 1999.

[17] Richards J.H.Jr. Chapter 8: Analysis of competing hypotheses. *Psychology of Intelligence Analysis*.

[18] John J.K., Eric B., Daniel B., and David W.D. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA*, 24:137703, 1996.

[19] Kidd J.M., Cooper G.M., Donahue W.F., Hayden H.S., Sampas N., Graves T., Hansen N., Teague B., Alkan C., Antonacci F., Haugen E., Zerr T., Yamada N.A., Tsang P., Newman T.L., Tzn E., Cheng Z., Ebling H.M., Tusneem N., David R., Gillett W., Phelps K.A., Weaver M., Saranga D., Brand A., Tao W., Gustafson E., McKernan K., Chen L., Malig M., Smith J.D., Korn J.M., McCarroll S.A., Altshuler D.A., Peiffer D.A., Dorschner M., Stamatoyannopoulos J., Schwartz D.C., Nickerson D.A., Mullikin J.C., Wilson R.K., Bruhn L., Olson M.V., Kaul R., Smith D.R., and Eichler E.E. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:5664, 2008.

[20] Gunhee K. and Antonio T. Unsupervised detection of regions of interest using iterative link analysis. *Annual Conference on Neural Information Processing Systems*, 2009.

[21] Worsley K.J., Marrette S., Neelin P., Vandal A.C., Friston K.J., and Evans A.C. A unified staistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996.

[22] Lin L., Yinhu L., Siliang L., Ni H., Yimin H., Ray P., Danni L., Lihua L., and Maggie L. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 251364, 2012.

[23] Margulies M., Egholm M., and Altman W.E. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[24] Kim M.H., Timothy R.R., and Arnold J.L. Germline mutations and polymorphisms in the origins of cancers in women. *Journal of Oncology*, 2010:11 pages, 2010.

[25] Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, 21:3789–3801, 2002.

[26] Christopher R.G., Nicole A.L., and Thomas N. Thresholding of statistical maps in functional neuroimaging using the flase discovery rate. *NeuroImage*, 15:870–878, 2002.

[27] Cormac S. Gene therapy finds its niche. *Nature Biotechnology*, 29:121–128, 2011.

[28] D.C. Schwartz. Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262.5130:1104, 1993.

[29] Scherer S.W., Lee C., Birney E., Altshuler D.M., Eichler E.E., Carter N.P., Hurles M.E., and Feuk L. Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, 39:s7–15, 2007.

[30] Schmidt T.H., Schutz G.J., Baumgartner W., Gruber H.J., and Schindler H. Imaging of single molecule diffusion. *Pro. Natl. Acad. Sci.*, 93:2926–2929, 1996.

[31] Lai Z., Jing J., Aston C., Clarke V., Apodaca J., Dimalanta E.T., Carucci D.J., Gardner M.J., Mishra B., Anantharaman T.S., Paxia S., Hoffman S.L., Craig Venter J., Huff E.J., and Schwartz D.C. A shotgun optical map of the entire plasmodium falciparum genome. *Nature genetics*, 23.3:309–13, 1999.

Appendix A: Example of a SNP altering the joint distribution

The number of signatures altered by a SNP is hard to predict. The range can be from 1 to dozens. The majority of the SNPs will change the signature by 1 while some of them create or remove nick sites of the sequencing enzyme or ending patterns. In following example the SNP "ss87153243" happens on the 9^{th} interval of the first chromosome. It changes base G to A at position 3655 of this interval and alters three pairs of $\{\mu_j, \nu_j\}$ values as in table A.1. Figure 3.2 shows how the joint distribution is affected by this SNP.

Table A.1: True values altered by the SNP "ss87153243" when base A is labeled

Org. ν_j	Org. μ_j	New ν_j	New μ_j
3737	46	3737	47
3740	43	3740	44
3752	40	3752	41

Table A.2: True values altered by the SNP ”ss87153243” when base C is labeled

Org. ν_j	Org. μ_j	New ν_j	New μ_j
3621	18	3621	17
3623	20	3623	19
3629	21	3629	20
3632	22	3632	21

Table A.3: True values altered by the SNP ”ss87153243” when base G is labeled

Org. ν_j	Org. μ_j	New ν_j	New μ_j
3669	12	3669	11
3657	18	3657	17
3654	20	3654	19

Table A.4: True values altered by the SNP ”ss87153243” when base T is labeled

Org. ν_j	Org. μ_j	New ν_j	New μ_j
3632	12	3632	13
3640	15	3640	16
3653	18	3653	19
3656	19	3656	20
3705	42	3705	43
3709	43	3709	44
3712	44	3712	45
3719	46	3719	47

Appendix B: Error Distribution of (S, X)

This appendix gives details of how $f(S|Z = j, t_1)$ and $g(X|z = j, t_2)$ are derived.

B.1 Error Distribution of S

According to [4] and [30]:

$$S|Z = j \quad \sim \quad \text{Normal}\left(c_1\mu_j, \mu_j\sigma_1^2\right)$$

where c_1 and σ_1^2 are unknown parameters and μ_j is as described in section 2.2:

$$\mu_j \quad = \quad \text{The number of fluorescent labeled bases within the } j^{th} \text{ neighborhood}$$

Suppose the prior of (c_1, σ_1^2) follows normal scaled inverse gamma distribution with parameter $(1, 1, 1, 1)$ and the p.d.f. is

$$(\frac{1}{\sigma_1^2})^{2.5} exp(\frac{2 + (c_1 - 1)^2}{2\sigma_1^2})$$

Our first pilot experiment generates observations $t_1 = \{t_{1,1}, t_{1,2}, \ldots, t_{1,r_1}\}$ with mean value $\overline{t_1}$. These observations were obtained when only one base was fluorescently labeled. Thus they are samples from distribution $Normal(c_1, \sigma_1^2)$. The posterior distribution of (c_1, σ_1^2) is still norm al scaled inverse gamma distribution with parameter $(\gamma, \eta, \alpha, \beta)$, where:

$$\gamma = \frac{1 + r_1\overline{t_1}}{1 + r_1}$$
$$\eta = 1 + r_1$$
$$\alpha = 1 + r_1/2$$
$$\beta = 1 + \frac{1}{2}\sum_i (t_{1,i} - \overline{t_1})^2 + \frac{r_1(\overline{t_1} - 1)^2}{(1 + r_1)^2}$$

Then we have

$$f(S|Z = j, t_1) \propto \int_{c_1,\sigma_1^2} f_1(S|Z = j, c_1, \sigma_1^2) f_2((c_1, \sigma_1^2)|t_1) d(c_1\sigma_1^2)$$
$$\propto \int_{c_1,\sigma_1^2} (1/\sigma_1^2)^{\alpha+1.5} (1/\sigma_1^2)^{0.5} exp(-\frac{(s - c_1\mu_j)^2}{2\mu_j\sigma_1^2}) exp(-\frac{2\beta + \eta(c_1 - \gamma)^2}{2\sigma_1^2}) d(c_1\sigma_1^2)$$

$$\propto \int_{c_1,\sigma_1^2} (1/\sigma_1^2)^{\alpha+2} exp(-\frac{c_1^2(\mu_j+\eta) - 2c_1(s+\eta\gamma) + 2\beta + s^2/\mu_j + \eta\gamma^2}{2\sigma_1^2}) d(c_1\sigma_1^2)$$

$$\propto \int_{c_1,\sigma_1^2} \frac{(\mu_j+\eta)(c_1 - \frac{s+\eta\gamma}{\mu_j+\eta})^2 + 2\beta + s^2/\mu_j + \eta\gamma^2 - \frac{(s+\eta\gamma)^2}{\mu_j+\eta}}{2\sigma_1^2} d(c_1\sigma_1^2)$$

The part inside of the integration is a kernel of another normal scaled inverse gamma distribution with parameter:$(\frac{s+\eta\gamma}{\mu_j+\eta}, \mu_j+\eta, \alpha+0.5, \beta+\frac{\eta(s-\mu_j\gamma)^2}{2\mu_j(\mu_j+\eta)})$.

Thus:

$$f(S|Z=j,t_1) \propto \frac{1}{(\beta+\frac{\eta(s-\mu_j\gamma)^2}{2\mu_j(\mu_j+\eta)})^{\alpha+0.5}}$$

So, we have $\sqrt{\frac{\alpha\eta(s-\mu_j\gamma)^2}{\beta\mu_j(\mu_j+\eta)}} \sim t(2\alpha)$

Since $\alpha = 1+r_1/2$, when r_1 is large enough, we can approximately get: $\sqrt{\frac{\alpha\eta(s-\mu_j\gamma)^2}{\beta\mu_j(\mu_j+\eta)}} \sim Normal(0,1)$, which means:

$$f(S|Z=j,t_1) \quad \sim \quad \text{Normal}\left(\frac{\mu_j(1+r_1\overline{t_1})}{1+r_1}, \frac{\mu_j(\mu_j+1+r_1)((2+\sum_i(t_{1,i}-\overline{t_1})^2)(1+r_1)+r_1(\overline{t_1}-1)^2)}{(r_1+1)^2(r_1+2)}\right)$$

B.2 Error Distribution of Position X

The second pilot experiment suggests that

$$X|Z = j \sim Normal(\nu_j, (\nu_j + \nu_j^2)\sigma_2^2)$$

where σ_2^2 is unknown parameter and ν_j is as described in section 2.2

$$\nu_j \quad = \quad \text{Center position of the } j^{th} \text{ neighborhood .}$$

Suppose the prior of σ_2^2 has inverse gamma distribution:

$$\sigma_2^2 \sim \Gamma^{-1}(\alpha_0, \beta_0)$$

Based on the second pilot experiment, α_0 is set to 2.83 and β_0 is set to 0.011.
The second pilot experiment generates data $t_2 = \{t_{2,1}, t_{2,2}, \ldots, t_{2,r_2}\}$ with known true values $e_2 = \{e_{2,1}, e_{2,2}, \ldots, e_{2,r_2}\}$. Then the posterior distribution of σ_2^2 given t_2 is still inverse gamma distribution with parameter $(2.83 + r_2/2, 0.011 + \sum_i \frac{(t_{2,i}-e_{2,i})^2}{2(e_{2,i}+e_{2,i}^2)})$

Then

$$g(X|Z = j, t_2) \propto \int_{\sigma_2^2} g_1(X|Z = j, \sigma_2^2) g_2(\sigma_2^2) d(\sigma_2^2)$$

$$\propto \int_{\sigma_2^2} exp(-\frac{\beta + \frac{(x-\nu_j)^2}{2(\nu_j+\nu_j^2)}}{\sigma_2^2}) \sigma_2^{-\alpha-1-0.5} d(\sigma_2^2)$$

The part inside the integration is kernel of a inverse gamma distribution with parameter $(\alpha + 0.5, \beta + \frac{(x-\nu_j)^2}{2(\nu_j+\nu_j^2)})$
Thus

$$g(X|Z = j, t_2) \propto \frac{1}{(1 + \frac{(X-\nu_j)^2}{2\beta(\nu_j+\nu_j^2)})^{\alpha+0.5}}$$

so, $\sqrt{\frac{(x-\nu_j)^2\alpha}{\beta(\nu_j+\nu_j^2)}} \sim t(2\alpha)$
Since $\alpha = 2.83 + r_2/2$, when r_2 is large enough, we have

$$\sqrt{\frac{(x-\nu_j)^2\alpha}{\beta(\nu_j+\nu_j^2)}} \sim Normal(0, 1)$$

Thus,

$$X|Z=j,t_2 \quad \sim \quad \text{Normal}\left(\nu_j, \frac{(\nu_j+\nu_j^2)(0.011+\sum_i \frac{(t_{2,i}-e_{2,i})^2}{2(e_{2,i}+e_{2,i}^2)})}{2.83+r_2/2}\right).$$

Appendix C: SNP Types

This appendix shows examples of all the SNP types. Suppose

μ = The number of fluorescent labeled bases within the target neighborhood

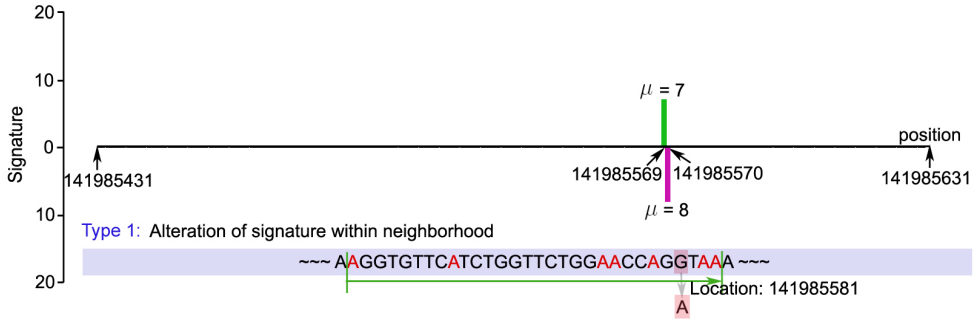

Figure C.1: Type 1 SNP alters the μ value by 1. The x axis denotes the position in first chromosome. The y axis denotes the μ value which is the number of labeled bases. The green bar is the original value of μ and the purple bar is the value of μ being altered by this SNP Here the SNP change base G to A and alter the μ value from 7 to 8.

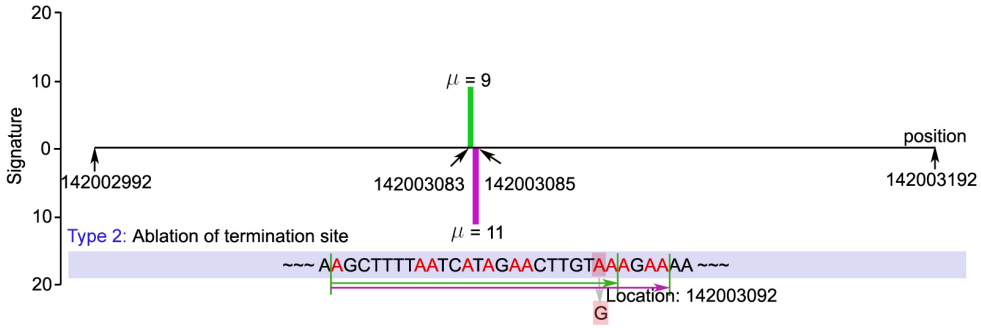

Figure C.2: Type 2 SNP ablates the termination sites. Since the polymerase will continue the synthesis of bases until the next termination sites, the related μ values can be increased by 1 or more. This example shown in this figure ablates ending pattern "AAA" and alter the μ from 9 to 11.

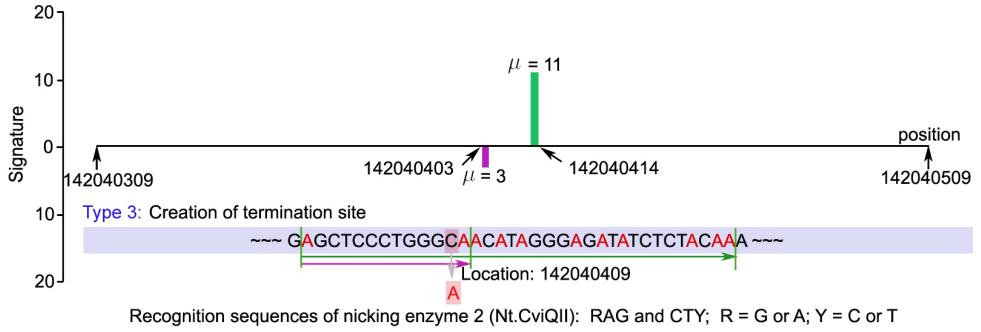

Figure C.3: Type 3 creating a new termination sites. In contrast to the previous type, the related μ values can be decreased by 1 or more.This example shown in this figure creates a new ending pattern "AAA" and alter the μ from 11 to 3.

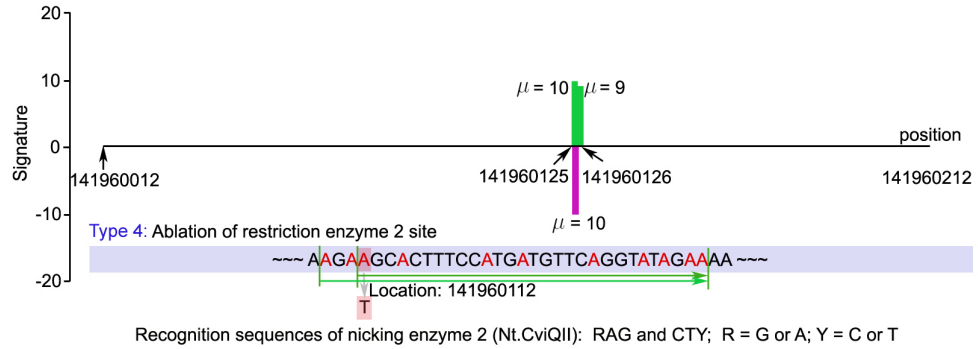

Figure C.4: Type 4 ablating the cognate sites of the second enzyme. The μ values generated from these cognate sites are removed; This example shown in this figure ablates a nick site "AAG" and removes one of the μ values ($\mu = 9$) in this local region. The other μ value is changed from 11 to 10.

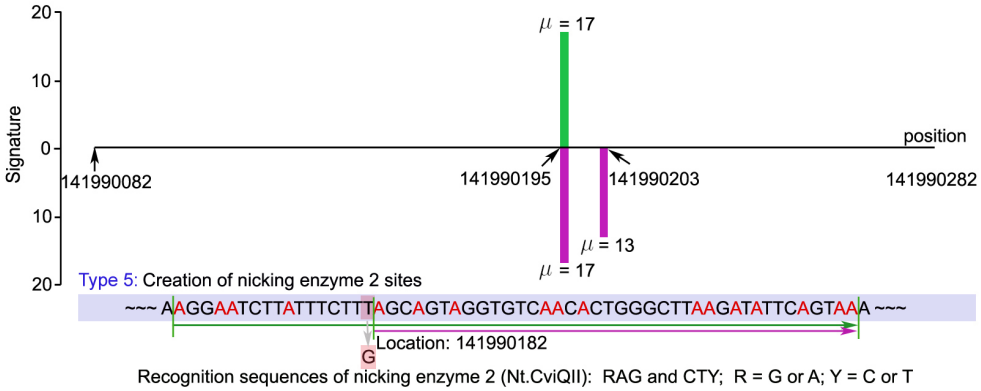

Figure C.5: Type 5 SNP creating new cognate sites of the second enzyme and thus generating new μ values. This example shown in this figure creates a nick site "GAG" and add one new μ value ($= 13$) to this interval

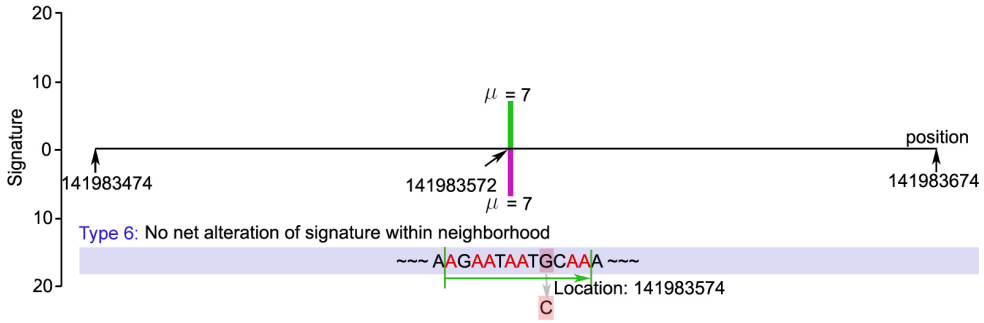

Figure C.6: Type 6 SNP being covered by the some neighborhoods while not altering the μ values because the SNPs don't change the bases with fluorochrome attached. The SNP shown in this figure is covered by a neighborhood, however it changes G to C, when the labeled base is A, it has no effect on the μ value.

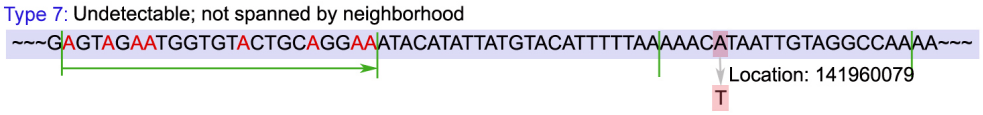

Figure C.7: Type 7 being covered by none of the neighborhoods and thus having no effect.

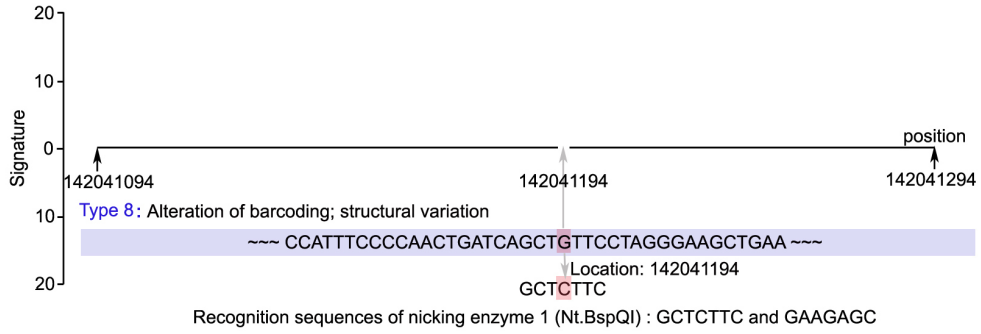

Figure C.8: Type 8 SNP altering the mapping enzyme nick site by creating or removing the nicking patterns. This type of SNPs are considered to be known by applying OM analysis method.

Appendix D: SNP candidates in the 9^{th} interval

Table D.1: SNP candidates in the 9^{th} interval

candidate index	SNP ID	Location	From	To
1	rs78395614	2453	G	A
2	rs76199781	2526	A	G
3	rs74970982	3177	A	G
4	rs56992750	3197	T	A
5	rs75526266	3215	G	C
6	rs75719746	3234	G	A
7	ss87153243	3655	G	A
8	rs76735897	3722	A	G
9	rs77573425	3724	G	C
10	rs10399597	3892	G	A
11	rs28402963	3938	T	C
12	ss87153247	4006	A	G
13	rs75478250	5003	T	C
14	rs2691286	7743	C	G
15	ss87153293	7897	A	T
16	rs28552463	7911	T	A
17	rs28534012	8049	T	A
18	rs28464214	8069	T	A
19	rs13328655	8192	T	A
20	rs12401368	8242	T	A

The above table shows the 20 SNP candidates in the 9^{th} interval. If we consider second order interactions, there are four more candidates:

Table D.2: second order interactions in the 9^{th} interval

candidate index	Interactions	SNP1	SNP2
21	interaction1	rs78395614	rs76199781
22	interaction2	rs74970982	rs56992750
23	interaction3	rs75526266	rs75719746
24	interaction4	rs76735897	rs77573425