High-resolution Detection of Change-Point with Low Coverage Single-cell Sequencing Data

A Dissertation Presented

By

Huan Qi

To

The Graduate School in Partial Fulfillment of the Requirements for the

Degree of

Doctor of Philosophy

In

Applied Mathematics and Statistics

Stony Brook University

December 2015

Stony Brook University

The Graduate School

Huan Qi

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Haipeng Xing – Dissertation Advisor
Associate Professor, Department of Applied Mathematics and Statistics

Wei Zhu - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Song Wu – Committee Member of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Xuefeng Wang – Committee Member of Defense
Assistant Professor, Department of Preventative Medicine

Jiangyong Jia – Outside Member of Defense
Associate Professor, Department of Chemistry

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

High-resolution Detection of Change-Point with Low Coverage Single-cell Sequencing Data

by

Huan Qi

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2015

Advances in next-generation sequencing technologies are revolutionizing our ability to detect copy number variations (CNVs). Single-cell sequencing technology allows for the genome wide copy number analysis within a single nucleus which is isolated form mixed population of cells. It can avoid the disadvantage of genomic differences in complex mixtures of cells. Many statistical methods and tools have been developed for CNVs detection using high-throughput sequencing data, but most methods are not designed for low-coverage sequencing data. In this article, we present a new Bayesian based change-point Model which has never been used for CNVs detection before and propose two similarity scores to discover DNA CNVs with low-coverage single-cell sequencing data and compare with other popular methods.

Table of Contents

Chapter 1

DNA Sequencing Technology and Copy Number Detection

In the late 1860s, Swiss physiological chemist Friedrich Miescher first discovered the existence of DNA. In the next decades, other scientists revealed more details about the structures and functions of DNA based on series of research. In 1953, a ground-breaking conclusion was announced by American biologist James Watson and English physicist Francis Crick that the structure of DNA molecule is double helix [1].

DNA is made up of nucleotides. Each nucleotide is composed by three parts: a sugar molecule, a phosphate molecule and a nitrogenous base. The canonical structure of DNA has four kinds of nitrogenous bases: Thymine (T), Adenine (A), Cytosine (C) and Guanine (G). DNA molecules pair up with each other to form very stable and strong units called base pairs. DNA molecules are held together by hydrogen bonds formed between these four kinds of nitrogenous based. Adenine on one DNA strand binds to Thymine on the complementary DNA strand and Cytosine binds to Guanine. These combinations make the DNA double helix structure is extremely stable. Moreover, DNA double helix has the reform (renature) ability which is the basis of molecular hybridization. But the DNA molecules hydrogen bonds still can break due to environment changes, such as heat or chemicals, DNA molecules are able to reform when the environment becomes favorable again.

As the basis and foundation for all biological research, the determination of the precise sequence of nucleotides in a molecule of DNA is very crucial. There are batches of methods and technologies that are used to determine the order of four nitrogenous bases – Thymine, Adenine, Cytosine and Guanine. DNA sequencing is one of the most important such technologies in bioscience today. As one of the key technologies in genomic research, DNA sequencing not only accelerates basic biological research, but also provides new opportunities in medical and clinical applications.

In the rest of this section, we present an overview of the development on the methods and technologies of DNA sequencing. The methods reviewed are: (i) Early-age methods: Sanger's method and the Maxam & Gillert method; (ii) Shortgun sequencing method; (iii) Next-generation sequencing method; and (iv) Single-cell sequencing method. For each method, we introduce the method's basic idea, its simplified experimental process, real word applications, advantages and disadvantages.

1.1 Basic DNA Sequencing Methods

1.1.1 Sanger's Method and Other Enzymic Methods

The first DNA sequencing method is called 'plus and minus' which introduced by Frederick Sanger and AR Coulson in 1975 [2] through chemical reaction conducted with polymerases. In their experiment, polymerases were resolved by ionophoreasis on denaturing polyacrylamide gels. Two years later, Sanger, Coulson and their coworker Nicklen made a significant improvement with a new method called the chain-termination method or the dideoxynucleotide method [3]. Their method is based on the catalyzed enzymic reaction and also known as enzymic sequencing. They also proposed to anneal a short oligonucleotide primer to a specific point of template DNA which is used as an initial position for the sequencing process. However, the chain-termination method can only applied on short sequence with a few hundreds nucleotides. Later on, more enzymic sequencing methods had been proposed on top of chain-termination method which can be categorized into two major classes: random approach (shortgun sequencing) and direct approach (primer walking sequencing) [4].

Shotgun sequencing technology is a random process since there is no control of which region is going to be sequenced. People usually use this method for determining the sequence of a very large piece of DNA. For large ones, such as BAC DNA, the DNA is first randomly fragmented into small pieces. Then chain-termination method is used for each resultant piece. After the fragments got sequenced, they then are deduced from the original BAC DNA sequence. Due to the randomness of the fragmentation, this process usually is repeated several times through the target DNA.

Another enzymic method for genomic sequencing is known as primer walking [5, 6]. It is also called as direct sequencing method comparing with the random sequencing. In this method, instead of randomly fragmenting the DNA, the enzymic reaction is applied on each fragment piece by piece. The primary advantage of primer walking is the low redundancy [7]. In the meantime, it also requires the synthesis of each new primer which is time consuming and expensive.

The automated chain-termination method had dominated the industry for almost two decades. It made history and led to many remarkable accomplishments including the completion of human genome project [8]. The human genome project is an international cooperation that sequencing the whole genome of the human.

1.1.2 Maxam & Gilbert and Other Chemical Methods

In 1977, Allan Maxam and Walter Gilbert [9] described a DNA sequencing method based on chemical modification. It's also known as chemical sequencing. In this method, DNA fragments require radioactive labelling at one 5' end. Then, these end-labelled DNA fragments are subjected to random cleavage at specific bases. The DNA sequence then can be easily identified from the one or four parallel sequencing gel lanes.

Either double-stranded DNA or single-stranded DNA can be used in chemical sequencing methods. Before end-labelling processed, DNA fragments are digested with certain restriction enzyme [10]. They can be prepared from an rearranged DNA region [11]. These DNA fragments are then labelled at 5' ends.

After Maxam and Gilbert originally first did end-labelling with phosphate and a nucleotide linked to phosphate in 1977, in 1985, Ornstein and Kashdan found an alternative method by using Dideoxyadenosin 5'-(α-thio) triphosphate and deoxynucleotidyltransferase. Comparing with Maxam and Gilbert's original method, this alternative method has higher stability and higher autoradiograph's resolution. Nevertheless, the sequencing method using 21-mer fluorescein labelled M13 sequencing primer was proposed by Ansorge in 1988 where k-mer refers any substrings of length k from a DNA sequencing read. Due to the stability during the chemical reactions, non-radioactive fluorescein dye had been used in new method [12] which was published in 1990. This paper gives us experimental evidence that fluorescein attached to the 5'-phosphate of an oligonucleotide shown to be stable during the reactions of chemical cleavage procedures [13]. Another non-radioactive labelling method used a biotin marker molecule chemically or enzymically attached to an oligonucleotide was proposed by Richterich in 1989 [14].

During the same period, Polymerase Chain Reaction (PCR) was developed by Kary Mullis in 1983 [15] to amplifying DNA fragments. In Polymerase Chain Reaction, the step of end-labelling is automated. The whole process includes several steps: denaturation, annealing, extension and final elongation. In the same year, Wade introduced an automatic DNA sequencing system. This system is compose of a computer controlled microchemical manipulator for the original Maxam-Gilbert DNA sequencing method [16].

Comparing with Sanger's chain termination reaction methods, the significant advantages of Maxam-Gilbert and other chemical methods are: (i) Instead of using enzymic copies in Sanger's methods, Maxam-Gilbert method uses the original DNA fragment to do sequencing; (ii) No requirements of subcloning and Polymerase Chain Reaction (PCR); (iii) It has lower probability of making secondary structure mistakes or enzymic mistakes [17]; (iv) Protocols of Chemical methods are simple and easy. As a result, it's easy to control [18].

1.2 Next-Generation Sequencing

The above methods are considered as first-generation technologies and the novel methods are known as next-generation sequencing (NGS) technologies. Over the last two decades, the first-generation technologies had dominated the industry. However, in recent years, we can significantly realize a fundamental shift away from the applications of the first-generation technologies. The new sequencing technologies are combinations of template preparation, sequencing, imaging, genome alignment, and assembly methods [20].

The emergence of next-generation sequencing technologies provides us high-throughput and low-cost sequencing platform. It doesn't only change our landscape of genomes scientific approaches in basic research, but also ushers in more opportunities for sequencing in applied and clinical research. Due to its outstanding advantages, the next-generation sequencing platforms have been used in many applications. For example, one application of next-generation sequencing is to help us to better understand the genetic different between health and disease tissue.

There are several commercial instruments of next-generation sequencing, including Roche 454 Life Sciences, Illumina, ABI, Life, Helicos BioSciences and Polonator instrument. They have advanced read generation throughput and have achieved base accuracy dramatically since 2004. Different sequencing platform has different advantages and disadvantages.

The 454 system is the first next-generation sequencing platform that available in industry [21]. It can sequences fragments of DNA up to one billion bases in a single day. As all other methods, a double helix DNA needs to be broken up into several short fragments in this system. Using restriction enzymes, DNA strand can be broken at specific points. Short fragment calls adapters that are attached to the DNA fragments. Then, beads are added to the mixture. Based on the DNA complementary theory, DNA fragments can attach directly to the beads. After this step, we break the bonds of the joining double helix. One double strands DNA becomes two single strands. Then these fragments of DNA can be copied millions times on each bead by polymerase chain reaction (PCR). After filtering any beads that have either failed to attach to any DNA or contain more than one type of DNA fragment beads, we put the remaining beads into wells on a sequencing plate. One well only contains one bead. These wells contain the DNA polymerase and primers for sequencing reaction. Then, nucleotide bases are added into the wells. One type of base (adenine, cytosine, guanine and thymine) is added at a time. The camera will catch the light when each type of base is incorporated into the well. The intensity of the light corresponds to the number of nucleotides of the same type that have been incorporated. To get the sequence of the original piece of DNA, the patterns of light intensity can be plotted on a graph. The fragment libraries in this approach can be created by any methods. It can be a mixture of short and adaptor-flanked fragments. However, a major disadvantage of the 454 technology is homopolymers. It has difficulty distinguishing the number of bases in consecutive identical bases, such as AAA or GGG.

There are two kinds of clonal amplification method. In some platforms, like the 454, the Polonator and SOLiD rely on emulsion PCR to amplify clonal sequencing features [22]. In brief, it is the procedure that an $in\ vitro$ constructed adaptor-flanked shotgun library is PCR amplified in a emulsion. Only one of the PCR primers is attached to the surface of micron-scale beads by its $5'$ end. Both the bead and the template molecule present in a productive emulsion compartments. PCR amplicons are captured to the surface of the bead. For example, the 454 uses PCR amplicons to capture the surface of $28-\mu m$ beads. After breaking the emulsion, amplification products carried by beads can be selectively enriched. Another amplification method is called bridge PCR and also referenced as cluster PCR [23, 24]. The Solexa technology uses this method. It also uses an $in\, vitro$ constructed adaptor-flanked shotgun library. However, the different is, using a flexible linker, both primers are attached on a solid substrate surface by their $5'$ ends. Each clonal cluster contains about $1{,}000$ copies of a single member of the template library.

Comparing with conventional sequencing, the major advantage of next-generation sequencing is its ability to produce millions of sequence reads in parallel. This enormous volume throughput makes the sequencing experiment more efficient and less expensive. This feature expands the application of sequencing technology. The sequencing technology is not only used to determine the order of DNA nitrogenous, but also can be applied in other research areas. For example, instead of using microarrays technologies, it’s more popular to use sequencing based methods in the studies of gene-expression in recent years. Also, in the studies of DNA or RNA copy number variations, array-Comparative Genomic Hybirdization (array-CGH) is being replaced by next-generation sequencing. Moreover, due to the ability of sequencing the whole genome in a short time, the large-scale evolutionary studies of normal or tumor cells can be performed. Also, next-generation sequencing can avoid some cloning bias issues because its sequence reads are produced from DNA fragment libraries. This kind of cloning bias happens during vector-based cloning and Escherichia coli-based amplification step which can affects genome representation.

1.3 Single-Cell Sequencing

As we mentioned in the section above, researchers are able to generate a highly quantitative whole-genome sequencing data with remarkably lower cost and shorter time by using next generation sequencing machines. However, since most of previous technologies require bulk DNA or RNA from over 100,000 cells, it is very expensive and limits to provide the average state of the mixed population of cells. Thereby, technologies which can overcome such drawbacks are crucially important.

To address this set of problems, single-cell sequencing is an appealing technology that allows for copy number analysis of an individual cell isolated from mixed population of cells. In order to better understand and study heterogeneity of tumor, Mike Wigler and James Hicks's team developed an approach in 2011. It's known as single-nucleus sequencing (SNS) [25, 26] that can characterizes single-cell copy number profile for the genome wide. The experimental protocol for SNS includes three discrete and important steps: flow sorting of single nuclei, whole-genome amplification (WGA) of DNA and next-generation sequencing on the Illumina platform.

After flow sorting nuclei from a tissue or a cell line material, nuclei are putted into wells on a 96-well plate format. In SNS, WGA relies on a proprietary amplification method that randomly fragments the whole genome. It generates DNA fragments in 200-1,000 base pairs (bps) using a unique combination of primer extension pre-amplification and degenerate oligonucleotide primers. Generally, one can get around 90% success rates in amplification step. Then only these parts of DNA are selected for DNA library construction. After amplification, adaptor sequences of DNA are removed by sonication. In library construction, DNA is processed using a standard Illumina library preparation protocol. To get high-quality libraries, adaptor-ligated libraries are purified by agarose gel electrophoresis. After purification, the popular DNA fragment amplification method, polymerase chain reaction (PCR) developled by Kary Mullis in 1983 [15], is applied. PCR is very popular in current biological and genomic research. This method was based on thermal cycling, repeating heating and cooling cycle for DNA melting and enzymatic replication. Initially, we should sequence each single cell on a lane of Illumina GAIIx instrument. However, in SNS, we sequence many single cells on a single lane by adopting multiplexing using DNA bar codes.

Singe-cell sequencing is the most advanced and state of the art technology in recent developments on DNA sequencing. It provides the ability of revealing genetic information through only one single cell which avoids the problem with genomic differences in complex mixtures of cells. Since singe-cell sequencing is a relatively new sequencing technology, to the author’s best knowledge, no existing statistical method has been developed for modeling single-cell sequencing data yet. On the other hand, the single-cell sequencing data has relatively low coverage and only achieves around 6% coverage of the whole genome of a single cell. Due to this limitation, most of existing change-point detection algorithms cannot be directly applied to get an accurate result. In this dissertation, we will present a novel two-steps change-point detection algorithm to detect DNA CNVs specifically designed for such low coverage single-cell sequencing data and compare the results with popular methods for detecting DNA CNVs.

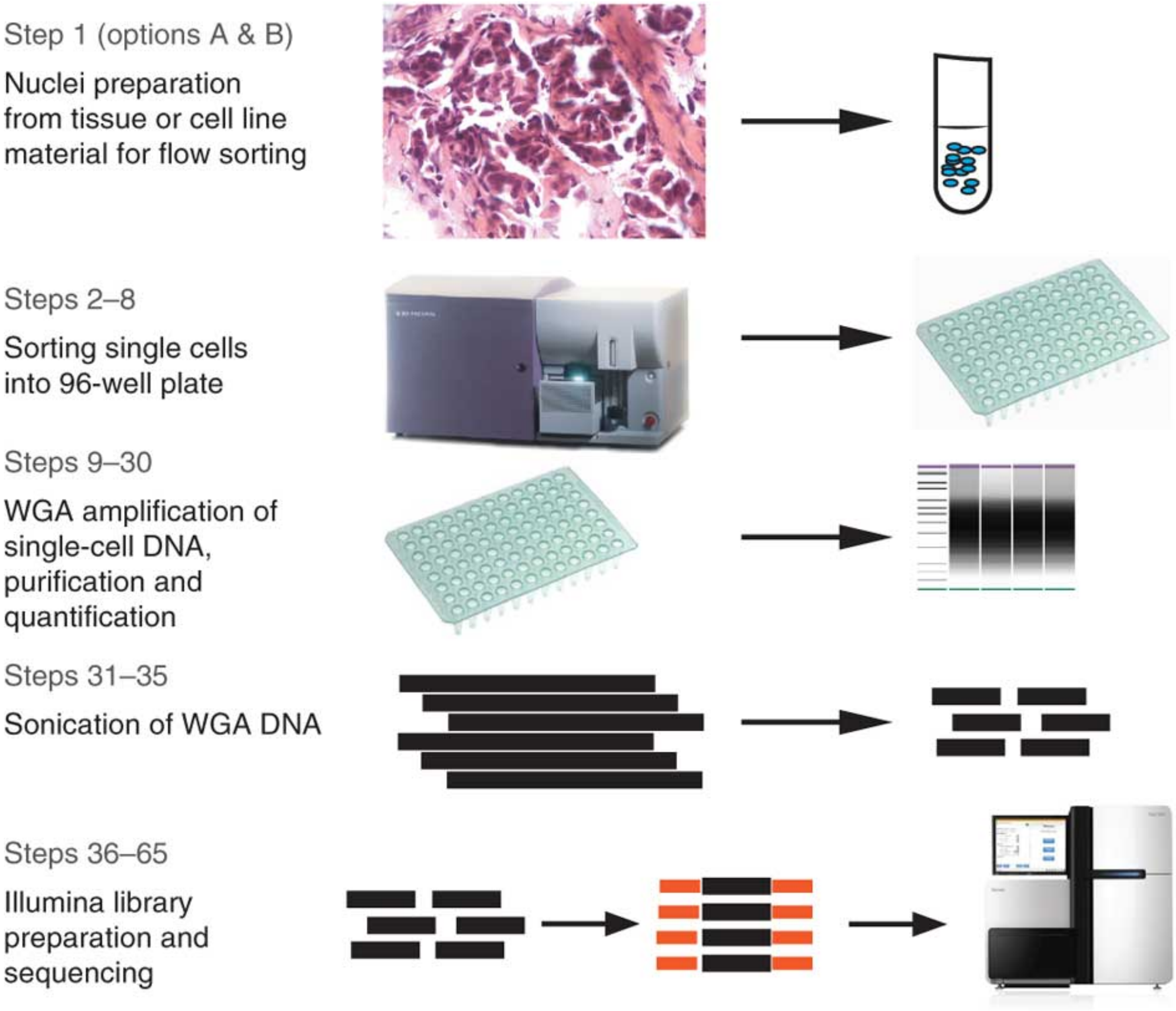

Step 1 (options A & B)
Nuclei preparation from tissue or cell line material for flow sorting
Steps 2–8
Sorting single cells into 96-well plate
Steps 9–30
WGA amplification of single-cell DNA, purification and quantification
Steps 31–35
Sonication of WGA DNA
Steps 36–65
Illumina library preparation and sequencing

Figure1. Schematic of the experimental workflow of SNS [27]

1.4 Copy Number Detection and Analysis

The human genome is comprised of 6 billion nucleotides of DNA. Theoretically, DNA contains two sets of 23 chromosomes with each set inherited from each parent. As a result, each somatic cell almost always presents in two copies of DNA in a certain genomic region. DNA copy number is the number of copies of DNA at a certain region of a genome. However, recent studies have revealed that large segments of DNA can vary in copy number. In 2002, Charles Less found in his experiments that healthy control patients not always had two copies of DNA in the whole genome. As what he found, there were numerous variations in their genetic sequences. Some of these healthy patients had more copies of specific genes than others. Meanwhile, Steven Scherer and Michael Wigler also had made similar findings that indicated large-scale variations in copy number were prevalent in the human genome. While these variations can overlap areas of disease related genes, they also can exist in healthy individuals [28, 29].

This kind of genomic structural variation is named copy number variation (CNV). It is a major category of genetic aberrations [30] that leads to an abnormal number of copies of certain genomic regions in cells. For example, genes that are thought to always possess two copies may somehow present in one, three, or more than three copies. In general, CNVs include deletions, duplications and insertions. These kinds of variations are typically greater than one kilobases (Kb) and less than five megabases (Mb) and the average size is around 250,000 bases. In 2006, after testing 270 individuals, Redon discovered that roughly 12% of the human genome had copy number variants [31]. About 2900 known genes are included in these CNVs. It is known to associate with phenotypes.

Major Influences of copy number variation have been found in disease control and evolution. Among large number of reported CNVs, some do not have negative effect on human health, but some might be associated with complex diseases and phenotypes.

As we mentioned before, copy number variation is very common in healthy individuals, and it is not necessarily have a negative influence. Many researches have revealed the function of copy number variation in healthy individuals. For example, a higher copy number of CCL3L1 has been associated with lower susceptibility to human immunodeficiency virus (HIV) infection [32]. This means that the amplification of CCL3L1 can potently suppress HIV and therefore protect individuals from HIV and save lives. Similarly, some other copy number variations seem to have no function but might just be an evolutionary selection.

Although most CNVs discovered so far are benign variants that will not directly cause disease, approximately half of the copy number variants detected by scientists encompass protein-coding regions. As a result, this kind of copy number variations has been linked to the development and progression of many diseases [33]. Recent studies have shown that copy number variants can directly influence gene dosage through amplifications or deletions, which can result in altered expression or structure of oncogenes and tumor suppressor genes, and potentially cause genetic diseases. For instance, duplications can result in a higher copy number than what is normally expressed by adding an entire gene. A higher copy number usually results in a higher gene dosage. Also gene expression can be influenced by gene dosage as well. In genome-wide association studies, copy number variants have been associated with various diseases. For instance, the EGFR copy number can be higher than normal in non-small cell lung cancer. One of the earliest observed DNA copy number changes is trisomy of chromosome 21 in Down's syndrome. Besides those impacts mentioned above, even some individual copy number variant has no effect to health, but the combination of two or more copy number variants can results in complex diseases.

1.5 Copy Number Variation Detection

Detection of DNA copy number variations can help us to better understand the genomic diversity and pathogenesis of cancer and other diseases. Furthermore, investigations of copy number variations also have been used to guide improved diagnostic and treatment decisions. For example, certain breast cancers are associated with overexpression of the ERBB2 gene. Patients who have ERBB2 gene amplification are more likely to respond to Herceptin treatment. Therefore, measuring the ERBB2 copy number can provide a diagnostic tool for breast cancer and other cancers. In brief, studies on CNVs are very important and high value for clinic applications.

There are many technologies that reveal genome-wide copy number and genetic aberrations at base pair resolution. DNA copy number can be assayed using fluorescence in situ hybridization, microarray-based comparative genomic hybridization (array-CGH) (Pinkel et al., 1998) [34] and next generation sequencing platforms.

1.5.1 Fluorescent In Situ Hybridization (FISH)

Fluorescence in situ hybridization (FISH) is a cytogenetic technique developed by biomedical researchers in the early 1980s [35] that is used to detect and localize the normal and the aberrant of specific DNA sequences on chromosomes. It is the extension of early in situ hybridization (ISH) technique, simply using the high-energy fluorophores instead.

The first in situ hybridization experiment was developed in 1969 by Joseph Gall and Mary Lou Pardue [36]. It is very sensitive and has many variations in the procedure. In early 1960s, the two scientists found that molecular hybridization can be used to identify the position of DNA sequences. Then, in later 1960s, they published a paper on DNA sequences detection that radioactive copies of a ribosomal DNA sequence can be used to identify the complementary DNA sequences.

After Gall and Pardue's work, in 1977, Rudkin and Stollar provided a new type of labels, fluorescent labels, replaced radioactive labels in molecular hybridization probe [37]. Fluorescent-labeled DNA probes are more stable and ease to detect. Today, the most popular in situ hybridization technology, FISH, uses fluorescent probes to detect DNA sequences.

The basic elements of fluorescent in situ hybridization include a labelled DNA probe as the 'magnet' and a target biological sample as the 'needle'. Detecting a DNA sequence basically is the process to find the 'needle' using the 'magnet'. In the first step of Fluorescence in situ hybridization, the DNA probe should be labelled. There are two labeling strategies: indirect labeling and direct labeling. For indirect labelling, we labelled DNA probe with a modified nucleotides that contains a hapten which then can be rendered fluorescent in later procedure. For direct labeling, the DNA probes are labeled with fluorescent nucleotides directly. Before hybridization, both the labeled DNA probe sequences and the target sequences are denatured. Then, mixing the denatured DNA probes and the target DNA sequences allows the annealing of matched DNA sequences. For direct labelled experiment, we can detect the position of DNA sequences directly. For other cases, an additional step is needed to visualize the nonfluorescent hapten before detecting the position. In brief, direct labelling method is faster with fewer steps, whereas the indirect labeling can provides the advantages of signal amplification.

FISH and other in situ hybridization technologies played an important role in human genome history. They led to many landmark accomplishments including the completion of human genome project. Experiment results from FISH and related in situ hybridization methods provided useful data for mapping the positions of genes on chromosomes. These data is collected and provides important genomic information for clinical diagnoses.

1.5.2 Comparative Genomic Hybridization (CGH)

Comparative genomic hybridization was developed to study DNA copy number variations across a whole genome. It is a biological technology that using molecular cytogenetic method to analysis copy number variations between a test sample (for example, tumor tissue) and a reference sample (for example, normal tissue) without the need of culturing cells. It mainly has been used to detect large chromosomal regions deletions or amplifications. With CGH, differentially test and reference genomic DNAs are cohybridized to normal metaphase chromosomes. Fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy number variation.

Conventional comparative genomic hybridization plays an important role in cancer research. It can be used in identifying chromosomal aberrations and has shown efficiency in diagnosing complex abnormalities associated with human genetic disorders. This technology also brings a lot of attentions for researcher in copy number field. Many papers have been published on detecting copy number variations by conventional CGH. There is a standard naming convention to describe the specific location or position of a gene on a chromosome. Usually a name of a location starts with the chromesome number which could be a number, letter ‘X’ or ‘Y’ and followed by the letter, *p* or *q* to represent the arm where the position is located. *p* represents the short arm and stands for *petit* meaning short in French. *q* indicates the long arm and stands for *queue* meaning tail in French. In some cases, there could be addtional number or letter at the end to further represent a particular position on the arm. Some of these copy number variations appear to be common to different kinds of fetal tumors. For example, amplification of chromosomal regions 1*q*, 3*q* and 8*q*, as well as deletions of regions 8*p*, 13*q*, 16*q* and 17*p* are very common in many kinds of tumor, such as breast, prostate and bladder cancer. However, some other CNVs are specific to certain tumor like amplification of 12*p* and X*p*, can only be observed in testicular cancer and more examples, such as 13*q* gain, 9*q* loss in bladder cancer, 14*q* loss in renal cancer and X*p* loss in ovarian cancer. These cancer specific alterations might reflect the unique selection forces operating during cancer development in different organs [38]. Conclusively, conventional CGH makes great contributions to the diagnosis and prognosis of cancer, as well as therapy [39]. Also, it provides more genetic information for study the development of different cancer.

The main limitation of conventional comparative genomic hybridization is the detection resolution. It has been shown in the literature that structural chromosomal aberrations smaller then 5-10Mb cannot be detected using conventional comparative genomic hybridization. In addition, it only can detect chromosomal gains and losses relative to the ploidy level. Any structural chromosomal aberrations without copy number changes, such as mosaicism, cannot be detected by conventional CGH. Furthermore, between individuals, short repetitive DNA sequences are highly variable in chromosomal regions, such as the regions of centromeres and telomeres. However, these chromosomal regions with short repetitive sequences can disrupt CGH analysis [40]. Therefore, we need a novel technology has high-resolution that can identify smaller chromosomal aberrations and can overcomes these limitations.

1.5.3 Array Comparative Genomic Hybridization (array-CGH)

Array-based comparative genomic hybridization, also referred as microarray-based comparative genomic hybridization, is a specific, sensitive, fast and high-throughput technique. In array-CGH, arrays of genomic BAC, P1, cosmid or cDNA clones are used for hybridization instead of metaphase chromosomes in conventional CGH technique. It detects copy number variations at multiple loci simultaneously. Before the emergence of next generation sequencing, array-CGH has been standard technology to detect interesting genomic regions which are associated with copy number variations. It is a molecular cytogenetic technique that can identify DNA copy number aberrations which could cause human genetic diseases, such as deletions, amplifications and breakpoints on a genome wide and high resolution scale.

For each experiment, we prepare two samples, one test sample and one diploid reference sample. Usually, the test sample contains tumor cells and the reference sample contains normal cells. These two samples are labeled with different dyes. The next step is to mix and hybridize two dyed samples into a microarray chip. From the color of the chip, we can calculate the ratio of the fluorescence intensities of the test sample and the reference. Pinkel and Albertson (2005) [33] have reviewed the most recent developments of array-CGH technique and its applications.

To get the chromosome copy numbers from the log intensity measurements, typically we need to segment every chromosome into certain regions first. Then we estimate the real copy number at each location from the array-CGH data. Typically, the number we get is the average copy number at the certain location over all cells in the sample. Due to the heterogeneous of the cell population, there are differences between the real copy number and the estimated number. To avoid this disadvantage of array-CGH data, several statistical algorithms have been published in the recent years. It includes hidden Markov model (HMM, Fridlyand 2004) [41], wavelet approximation (Hsu, 2005) [42], recursive change-point detection (Circular Binary Segmentation, Olshen, 2004) [43] and a Bayes regression approach (Wen, 2006) [44].

The same as conventional CGH, array-CGH is an important technology that provides genetic information for studying genetic diseases and for developing diagnostic and therapeutic targets.

While early techniques such as fluorescence in situ hybridization were only able to locate copy number variations at the whole chromosomal or whole arm level, as a high-throughput technique, array-CGH can detect copy number variations at a level of 5-10 kilobases (1 kilo base pair=1,000 base pairs) of DNA sequences [45]. Comparing with conventional CGH, in array-CGH, the metaphase chromosomes are replaced by short cloned DNA fragments between 100 to 200 kiolbases. The arrays of array-CGH using bacterial artificial chromosome fragments, and contain many regions with some known tumor suppressor genes and oncogenes. Array-CGH data is based on the log-ratios of normalized intensities from test sample and reference, such as solid tumor and normal tissue. For a given gene or region, a negative log-ratio is an indication of a deletion, and a positive log-ratio is an indication of an amplification. If the log-ratio equals zero, the target sample and the reference sample have the same copy number for that given gene or region.

1.5.4 Next-Generation Sequencing

During the last several years, the broadest application of next generation sequencing enables researchers to better understand how genetic differences affect health and disease. As we mentioned in previous sections, traditional methods FISH and array-CGH suffer from low resolution of genomic regions. Comparing with array-CGH technique, recent advances in sequencing technologies enabled that massively parallel sequencing of millions of short sequence reads at remarkably lower costs and shorter time. There are several commercial platforms, including 454 Life Sciences, Roche, Illumine, ABI, and Life Technologies, have advanced read generation throughput and achieved base accuracy dramatically since 2004. Using the ever-increasing output of next generation sequencing machines, researchers were able to generate a highly quantitative whole-genome sequencing data. Moreover, detection of DNA copy number variations from massively parallel sequencing data achieved greater sensitivity as well as greater precision for mapping breakpoints than similar detection based on array-CGH data. In conclusion, the advantages of next generation sequencing technology include lower cost, shorter time, higher coverage and resolution, higher sensitivity of copy number variations detection, and more accurate breaking points location.

However, because most of previous technologies require bulk DNA or RNA from over 100,000 cells, it is very expensive and limits to provid the average state of the mixed population of cells. Genetic heterogeneity is very common in solid tumors, but previous methods are not designed to resolve genomic differences in mixed populations of cells. Solid tumors are complex mixtures of cells including normal cells and multiple clonal subpopulations cancer cells. In a bulk of tumor tissue, there always are some normal cells and such presence can result in inaccurate copy number. Moreover, the characterization of tumor heterogeneity also impacts the purity of tumor tissue, where multiple clonal subpopulations with distinct genomic profiles might be present. Recent studies show that to better understand the evolution of tumor, the study of tumor heterogeneity is necessary. Several papers have described the details of heterogeneous nature of cancer. However, the common sequencing technologies are unable to avoid losing information of tumor cells genetic heterogeneity [26, 46].

Thereby, technologies which can overcome such drawbacks are crucially importance. To address this set of problems, single-cell sequencing is an appealing technology that allows for copy number analysis of an individual cell isolated from mixed populations of cells.

Single-cell genomic methods can help us to investigate tumor population structure and evolution, and enable the discovery of clonal mutations, cryptic cell types or transcriptional features that would be diluted or averaged out in bulk tissue. It leads to greatly improve our essential understanding of evolutionary and metastasize of tumor.

Sequencing data from bulk DNA or RNA from multiple cells provide global information on average states of cell populations. However, with whole-genome amplification and next-generation sequencing, researchers can detect variation in individual cancer cells and dissect tumor evolution. Such cancer genome sequencing will improve oncology by detecting rare tumor cells early, measuring intra-tumor heterogeneity, guiding chemotherapy and controlling drug resistance. The Single-cell sequencing technology explores the latest strategies that influence and aid cancer diagnosis, prognosis and prediction that will lead to individualized cancer therapy.

In order to better understand the tumor evolution and tumor population structure, Michael Wigler and James Hicks lab in Cold Spring Harbor Lab developed a novel approach, single-nucleus sequencing, that we can accurately quantify genomic copy number within an individual nucleus. Single-nucleus sequencing includes three major steps: isolate nuclei by flow-sorting, amplify DNA using whole genome amplification (WGA) and next generation sequencing.

To investigate evolution and population structure of tumor, Nicholas et al. (2011) [26] applied single-nucleus sequencing to analyze 100 single cells of two sets of human breast cancer, one with its matching living metastasis. To cluster 100 single cells profiles, they calculated distances by using pair-wise distances algorithm and built a clustering tree using neighbor joining. The result was shown that one tumor’s 100 profiles have been clustered into four subpopulations: one is flat diploid profiles and other three are advanced tumor subpopulations. These three distinct clonal subpopulations were likely to have originated from a common precursor. In another tumor, the one with matching metastasis, the data indicates that a single clonal expansion formed the primary tumor and seeded the metastasis. The copy number profiles in primary tumors are highly similar to the metastasis, which indicates the metastatic cells are from a main advanced expansion, not from an earlier intermediate subpopulation.

To determining copy number profiles, sequencing data is processed using several different computational and algorithmic tools. Usually, for each nucleus they can get 2 million uniquely mapping reads using SAM tools package and the Bowtie algorithm with defined parameters. These mapping reads offer a low coverage (~6%, mean=5.95%, s.e.m.+-0.229, n=200) of the whole genome of a single cell, sufficient to count copy number from sequence read depth.

Single-cell genomic methods can help us to investigate tumor population structure and evolution, and enable the discovery of clonal mutations, cryptic cell types or transcriptional features that would be diluted or averaged out in bulk tissue. In addition, it provides earlier diagnosis. For example, in the early stage of breast cancer, only small part of cells has pathological changes. If we analysis the bulk tissue, we will lose value information. Using single-cell sequencing, we can diagnose genetic disease through single cell. Earlier diagnosis can be beneficial to the patients as they can undergo early appropriate treatment and prognosis.

1.6 Literature on Copy Number Variants Detection Methods

A cogent way to find cancerigenic genes is to identify genomic regions with recurrent CNVs, amplifications and deletions, in tumor genomes (Beroukhim R, et al., 2007) [47]. An ideal description of a copy number variation should include both accurate positions of breaking points and precise estimation of copy number in each segment. Over the last few years, diverse computational approaches have been developed to detect CNV regions with an unprecedented resolution using next-generation sequencing data. Many methods developed for detecting structural variations can also be used to identify copy number variations. In summary, there are five strategies for CNV detection through sequencing data, including: (1) paired-end mapping (PEM), (2) split read (SR), (3) de novo assembly of a genome (AS), (4) read depth (RD), and (5) combination of the above approaches (CB). In practice, none of these approaches can detect all type of copy number variations. Different methods have their own advantages and limitations. Most PEM-based, SR-based and CB-based approaches are originally designed for structural variations, but can be applied to detect copy number variations. AS-based and RD-based approaches are developed for copy number variations detection.

1.6.1 Paired-End Mapping Approach

Previous computational approaches for detecting DNA copy number variations and structural variations using next generation sequencing data are primarily based on paired-end read mapping (PEM), such as the method provided by Tuzun et al. (2005) and Korbel et al. (2007) [48, 49]. PEM is a large-scale genome sequencing method to identify copy number variations. It mapped DNA sequencing reads onto a reference genome, and compares the length between mapped read pairs to the average insert size of the genomic library [48]. Notably, PEM is only applied for paired-end sequencing reads, not single-end reads. In paired-end sequencing data, distances between every two ends of a read pair has a specific distribution. PEM strategy detects SVs/CNVs from incongruous mapped paired-end whose distances are significantly different from the predefined average insert size. In Korbel et al. (2007) [49], they mapped over 1000 copy number variations, and the number of copy number variations of human is much larger than initially hypothesized.

There are two different approaches have been used in PEM-based tools to identify SVs/CNVs, including the clustering method and the model-based method. The clustering method uses a predetermined distance to recognize discordant reads. If the distance of paired-end reads higher then the expected distance, they will be labeled as discordant mapped reads. Meanwhile, model-based method applies a statistical probability test to identify the uncommon distance between mapped paired reads in comparison to the distance distribution in genome. The popular tool for PEM is so-called BreakDancer which includes two modules, BreakDancerMax and BreakDancerMini [50]. BreakDancerMax is a clustering-based method, while BreakDancerMini is a model-based method to detect smaller insertions and deletions range from 10 to 1000 base pairs. However, in BreakDancer, each read can only be assigned to one cluster. As a result, the reads that can be aligned to multiple genomic regions are dumped, even if they are mapped with high quality. To overcome this limitation, VariationHunter allows reads be assigned into multiple clusters to improve sensitivity [51].

The biggest advantage of PEM is this method can identify DNA copy number variations in a relative small fragment. It can detect deletions within 1 kb size, and locate breakpoints within small regions. The disadvantage of PEM is that some certain kinds of copy number variations are not easy to be identified, such as insertions larger than the average insert size of the genomic library [52] and variants located within complex genomic regions rich in segmental duplications.

1.6.2 Split Read-base Approach

SR-based methods are also only can be applied to pair-end reads sequencing data. The basic idea of SR-based method is for each pair reads, if one read is aligned to the reference genome while the other one fails to map or only partially maps to the reference genome, then those unmapped or partially mapped reads potentially provide accurate breaking points at the single base pair level for SVs/CNVs. Pindel is the first SR-based method tool to identify breaking points of large deletions(1bp-10kbp) and middle size insertions (1bp-20bp) (Ye K et al.,2009) [53]. Alignment with Gap Excision tool can identify breaking points of copy number variations with base pair level using a more strict local alignment algorithm (Abyzov A et al., 2011) [54]. SR-based methods can efficiently identify wide range of SV classes. The disadvantage of this kind of methods is they are limited by the length of reads and is only usable to the unique regions of the reference genome.

1.6.3 Assembly of a Genome

Comparing with PEM and SR methods, the AS-based methods do not need to align NGS reads to the known reference genome before the CNVs detection. Instead, AS-based methods first reconstruct DNA fragments (contigs), from short reads by assembling overlapping reads. Then they detect copy number variations by comparing the assembled contigs to the reference genome. As far, there are not so many AS-based methods being developed. The most recent one is Magnolya, which estimates copy number rely on two or more samples by applied a Poisson mixture model (Nijkamp JF et al., 2012) [55]. AS-based methods require a certain read coverage to detect overlapping fragments, although high coverage will increase the complexity of short read assembly.

1.6.4 Read Depth Methods

Nowadays, mainly due to the advent of high-throughput sequencing technologies, we can get the accumulation of high-coverage NGS data. Therefore, RD-based methods have recently become the most popular method to detect CNVs, since it can identify CNV regions with an unprecedented resolution. Different from the PEM and SR-based methods, the RD-based methods can estimate the exact copy numbers, which other methods can’t because they only use the information of position. Moreover, RD-based methods are able to detect large copy number variations in complex genomic regions, which is difficult for PEM and SR methods (Yoon S et al., 2009) [56].

When we do sequencing, we assume the sequencing process is uniform, which means each genome region should have the same probability being mapped. Another underlying assumption of RD-based methods is that the number of reads mapping to a genome region is proportionate to the copy number of the same region. For example, if a genome region lost copy number, then this region should has a lower intensity than expected (Teo SM et al., 2012)[57] .

Generally, RD-based methods follow three fundamental steps: Data preparation, Data normalization and copy number variation identification. Here, we will discuss every step in details one by one.

1.6.4.1 Data Preparation

The first fundamental step is mapping a set of short reads to the reference genome. Once short reads have been aligned to the reference genome, we need to perform several steps before read count estimation, including duplicated sequences removal, mapping quality filtering and window/bin size estimation. The main goal of removing duplicated reads is to alleviate the effects of Polymerase Chain Reaction (PCR) amplification bias produced during library construction. If multiple reads have the same exact external coordinates, only retain the read which has the highest mapping quality. Here, the samtools package (Li,H. et al., 2009) [58] can provide duplicate removal and other various utilities for operating alignments in the SAM format, such as sorting and merging. After removing duplications, we need to consider how to handle low mapping quality (MQ) sequences. Several aligner tools can provide MQ score for each sequence aligned to the reference genome. Low MQ score represents the sequence falling in repetitive regions of the reference genome or have low base quality. For example, if a read has MQ = 0, it means that there are at least two regions of the genome can perfectly match this read. Conversely, if a read as MQ = 30, it means there are a few or only one region of the genome can match this read. For these reasons, a simple way to solve this problem is removing all sequences with MQ < 30. (Magi et al. 2011 and Yoon et al., 2009) [56, 59]. Then the next step is to estimate the best window size.

As we mentioned before, we assume the read depth at a certain position is expected to be proportional to the copy number at that position. However, this simple concept is complicated by the fact that genomes are not sequenced deeply enough to enable base-pair resolution. Then the binning procedure is necessary. In such methods, the whole genome is bucketed into several non-overlapping bins, and the read depth is calculated according to the number of mapped reads in each predefined bins. We assign each read only once by its start position. Following the assumption, the copy number of any genomic region can be estimated by counting the number of mapped reads aligned to that particular region. A large bin size would provide less precision in locating breaking points of copy number variations, while a small bin size would result in losing information and noise analysis results. To date, only two papers have mentioned a method to do optimal bin size selection, Miller et al. (2011) [60] and Xie et al. (2009) [61]. After we decided bin size, we can calculate read counts in each bin.

1.6.4.2 Data Normalization

The second step focuses on normalization and correction of potential biases in RCs data. Read depths are affected by two main sources of bias, including local DNA GC content and the genomic mappability.

GC content is the percentage of guanine and cytosine bases in a genomic region. The relationship between GC content and read coverage has been studied in several papers, for example, Harismendy et al. (2009) [62], Dohm et al. (2008) [63] and Hillier et al. (2009) [64]. They all have the same conclusion that there is a positive correlation between read coverage and GC content. They tested different sequencing data sets which are generated by different technologies, including the Roche 454, Illumina GA and the LT SOLiD. The read depth of coverage decreases with increasing AT content for all the three platforms. GC-rich sequences, like genic and exotic region, as well as GC-poor regions are often under-represented (Bentley et al.2008) [65] mainly caused by amplification steps in the protocol.

The influence of GC content bias can be effectively cancelled out by comparing the pair of disease and normal samples directly at each region. Local GC content normalization has been mentioned in a few papers. The tool CNAseg (Ivakhno S., 2010) [66]applied an algorithm using locally weighted scatterplot smoothing (LOWESS) regression to adjust GC bias between the two paired samples in each 10Kb region. Chiang et al. (2009) [67] proposed a method to relieve the dependence between local GC content and read coverage by using the ratio of the number of reads in tumor DNA sample and its paired normal sample. Yoon et al. (2009) [56] proposed another method to adjust RCs by using the observed deviation in coverage for a given GC percentage. Each read count is normalized according to the following formula:

$$\overline{RC_\iota} = RC_i \cdot \frac{m}{m_{GC}}$$

Where RC_i read counts of the ith window, m_{GC} is the median RC of all the windows that have the same GC percentage as the ith window, and m is the overall median of all the windows.

Another factor mappability bias, which can introduce a bias in sequencing, is due to the fact the genome contains many repetitive elements and aligning reads to these positions leads to ambiguous mapping (Miller et al., 2011) [68]. In a simple word, the mappable position pattern is not unique for some sequencing reads. An aligning read can be mapped perfectly with more than two regions in the genome. Sequence uniqueness within the genome plays an important role when attempting to map sequence, especially short sequence, like next-generation short sequencing reads. Usually, sequencing reads which can be mapped to multiple regions are often discarded. As a result, genomic regions with high sequence degeneracy (low mappability scores) show lower mapped read coverage than unique regions and create a routine bias.

Mappability bias can be measured by several different tracks, including the Broad alignability track, the Duke uniqueness track and so on. Each track will calculate their own mappability score. The alignability track measures how uniquely k-mer sequences align to a region of the genome (Derrien T, et al 2012) [69] where k-mer refers any substrings of length k from a DNA sequencing read. Usually, k can be set to 36, 40, 50, 75 or 100 nts, For each window of k-mers, a mappability score is computed as S=1/(number of matches found in the genome). Thus, the score range of alignability track is from 0 to 1. S=1 means one match in the genome, S=0.5 means two matches in the genome, and so on. Please note that alignability would allow up to two mismatches [69]. The uniqueness track measures how unique each sequence with particular length is on the positive strand starting at a particular base. Thus, the k-bp track, where k usually can be set to 20 and 36, indicates the uniqueness of all k base sequences with the score being assigned to the first base of the sequence. The uniqueness score also ranges from 0 to 1. A score of 1 indicates a completely unique sequence, and score of 0 represents a sequence that occurs more than 4 times in the genome. It’s obvious that a sequencing read with a higher score is more unique.

At present, there are two methods proposed and used for correcting read count data for sequencing biases due to mappability. Miller et al.(2011) [68] proposed to adjust RCs by multiplying the number of reads in a given window by the inverse of the percent mappability in the same region. Ivakhno et al.(2010) [66] proposed to correct for mappability by using an undecimated discrete wavelet transform (DWT) to smooth read counts in the regions which have low mappability.

1.6.4.3 CNV Regions Identification

Once we corrected the read counts data from GC contend and mappability, the next step is to estimate the read count for each bin in order to identify copy number variations. Notably, after data normalization, the data we obtain from next generation sequencing experiments is very similar to the signal obtain from array-CGH data in mathematical view. Thus, some classical algorithms were originally designed to analysis array-

CGH data can also been applied to next generation sequencing data to detect copy number variations. Generally, RD-based methods can be classified into two categories. One is statistical significance test and the other is segmentation method.

Yoon et al. (2009) [56] developed an advanced method, event-wise testing (EWT), for DNA copy number variation identification based on significance testing that works on intervals of data points. EWT is a computational analysis based on read depth. In this method, read depth is measured by counting the number of mapped reads in 100-bp windows while each read only be assigned once by its start position. After adjustment of GC content, they used the GC-adjusted read depth within 100-bp windows as a quantitative measurement of genome copy number. The basic idea of EWT method is to identify regions of consecutive 100-bp windows with significantly increased or reduced read depth using corrected Z-score. It searches the entire genome for specific classes of small events that meet criteria of statistical significance, and then grouped them into large events. To identify which 100bp bin's read depth is significantly duplication or deletion, the first step is to transfer the GC-adjusted read count of each bin to a Z-score. A Z-score is calculated based on the number of reads mapped in each bin according to a two-tailed normal distribution. Another method based on statistical significance test is provided by Xie and Tammi (2009) [61], named CNV-seq. This method analyzes the ratios between read counts from normal and tumor samples using a sliding window approach. Then they converted each read count into a t-statistic, and infer altered regions by using the distribution of t-statistic.

A lot of statistical models have been proposed as the segmentation algorithm to find the CNVs, such as Circular Binary Segmentation (CBS), Mean Shift-Based (MSB), Shifting Level Model (SLM), Expectation Maximization (EM), and Hidden Markov Model (HMM). The basic idea of segmentation methods is to group the adjacent bins into segments with the same expected copy number. CBS method is the most popular segmentation method, which is originally designed for array-CGH data (Olshen et al., 2005) [43]. The algorithm recursively identifies the breaking points by changing genomic positions until the chromosomes are divided into segments with the same copy numbers that are significantly different from their adjacent regions. R package of circular binary segmentation method is very useful in today's copy number research. Hsu et al. (2005) [42] developed a smoothing algorithm based on wavelets. Chromosomal Aberration Region Miner method (Myers et al. 2005) involves EM algorithm, and can be used to locate copy number variations edges more precisely. Fridlyand et al. (2004) [41] proposed a complex modeling method using Hidden Markov Model to measure copy number, in which underlying copy numbers are the hidden states with certain transition probabilities. Another method named Stochastic Change-Point Model (SCP), which is also based on Hidden Markov Model, is developed by Lai et al. in 2008 [70]. We will introduce this model, Stochastic Change-Point Model (SCP), in Chapter 3.

Chapter 2

Change-point Detection

Detecting abrupt points in time series data is referred as change-point detection which attracts many scientists in statistics. Abrupt point stands for the point in a sequence at which the statistical properties of the sequential observations change. Detecting change-point is important in many research areas, such as finance, biology, genetic, climatology [71], and political. There is a growing need to identify the locations of multiple change-points within time series data.

Time series data vary over time, and conventional statistical models fail to capture temporal variations in regression relationships. Change-point models are very popular to analyze time series data. One of the challenges in change-point analysis is the ability to identify the location of multiple change points within a given time series or sequence data. During the last decade, many algorithms of change-point analysis have been developed to overcome this challenge.

There are several ways to classify change-point detection related problems. Depending on the number of change points, change-point detection can be classified into two groups: Single change-point detection and multiple change-point detection. Depending on the data settings, it can also be classified into fixed sample problem and sequential setting problem. Otherwise, depending on the delay of detection, it can be grouped into another two categories: real-time detection and retrospective detection.

2.1 Change-point Detection for a Sequence of Random Variables

In the procedure of data collection, if data points are in successive order and collected over a time interval, then we call this kind of data sequential data. In this section, we will focus more on the sequential data. We will begin with one point detection problem, and then introduce multiple change-points detection. The methods developed for solving change-point problems include maximum likelihood estimation, Bayesian estimation, isotonic regression, piecewise regression, non-parametric regression and so on. For independent and identically distributed random variables, there are several well-developed theories under maximum likelihood estimation.

Let $\{y_i\}, i = 1, \dots n$ be a sequence of observations of a time series data. For one change-point detection, we assume the change-point occurs at time $\tau_{1:m}, \tau \in \{1, \dots, n-1\}$. The statistical properties of $\{y_1, \dots, y_\tau\}$ and $\{y_{\tau+1}, \dots, y_n\}$ are different in certain way. For the ith segment, it can be denoted using parameter set $\{\theta_i, \phi_i\}$, where θ_i is the set of parameters that represent the distribution of the segment, and ϕ_i is the set of nuisance parameters. To detect the single change-point in a sequential data, we can perform a hypothesis test.

$$H_0 : No\ change - point, m = 0$$

$$H_1 : Having\ a\ single\ chang - point, m = 1$$

Firstly, we should calculate the maximum log-likelihood value for both null and alternative hypotheses. For the null hypothesis, the maximum log-likelihood value is simply $\log p(y_{1:n}|\hat{\theta})$, where $p(\cdot)$ is the probability density function of the distribution of the data and $\hat{\theta}$ is the maximum likelihood estimate of the segment parameter.

Under the alternative hypothesis, we assume there is only one change-point in the sequential data, and the position of the change-point is $\tau, \tau \in \{1, \dots, n-1\}$. Then we will write the maximum log-likelihood function as

$$logL(\tau) = \log p(y_{1:\tau} | \widehat{\theta_1}) + \log p\big(y_{(\tau+1):n} \big| \widehat{\theta_2}\big)$$

The maximum value of log-likelihood under the alternative hypothesis is $\max_\tau logL(\tau)$. It considers all possible change-point positions. Then the test statistic is given as

$$\lambda = 2[\max_\tau logL(\tau) - \log p(y_{1:n} | \hat{\theta})]$$

Then we need to choose a threshold so that if λ is larger than this threshold, we should reject the null hypothesis. If we reject the null hypothesis, we need to find a way to detect the position of change-point. To detect the change-point, we estimate its position as $\hat{\tau}$, the value of τ that maximizes $logL(\tau)$. How to get the appropriate value of the threshold is also an interesting research question. It depends on significant level and other information criteria. Many researchers have published papers on this topic, including Guyon and Yao (1999) [72], Chen and Gupta (1997) [73], Lavielle (2005) [74], and Birge and Massart (2007) [75].

Another similar method was named penalized likelihood approach which is more naturally extend to the multiple change-points detection than the likelihood ratio statistic approach [76]. Consider M_k which corresponds to the model with k change-points, with parameters p_k. Denote the associated parameter vector by $\Theta_k = (\tau_{1:k}, \theta_{1:k+1})$, and the likelihood by $\mathrm{L}(\Theta_k)$. Then the penalized likelihood is written as

$$PL(M_k) = -2 \log max \mathrm{L}(\Theta_k) + p_k \phi(n)$$

$\phi(n)$ is the penalty function which is an non-decreasing function of the data length n. Obviously, the results will depends on the choice of the penalty function $\phi(n)$. Different penalty functions can be considered, for example Akaike's information criterion (AIC) [77], Schwarz information criterion (SIC) and Hannan-Quinn information criterion [78]. The definition of each kind of penalty functions is as following:

$$AIC: \phi(n) = 2$$

$$SIC: \phi(n) = \log n$$

$$Hannan - Quinn: \phi(n) = 2 \log \log n$$

Among these kinds of penalty functions, AIC is the most popular one. It has been shown that AIC asymptotically overestimates the correct number of parameters. Both SIC and Hannan-Quinn criteria asymptotically estimate the correct number of parameters.

For detecting the position of a single change-point, two penalized likelihoods need to be calculated. One assumes the existence of one change-point and another assumes no change-point. This step is similar with the calculation of the likelihood maximization step that described previously. Both of them are comparing the maximum log-likelihood of the two models corresponding to the scenario of one and no change-point. If the log-likelihood of one change-point is greater than certain threshold, a change-point is detected. The differences between penalized likelihood approaches and the likelihood ratio test approaches is how to calculate the threshold.

Beside the maximum likelihood estimations, Bayesians methods are also very popular for detecting the change-point within the sequential data. Before performing the Bayesian analysis, we need to explain some notations. Firstly, we introduce a family of distributions $p(\theta|\psi)$ with the hyperparameters ψ. Then, we have a conditional probability.

$$p(\theta_{1:m+1}|\psi) = \prod_{i=1}^{m+1} p(\theta_1|\psi)$$

Either we known the value of ψ, or the model will be completed through an appropriate hyperprior on ψ. Note that the prior distribution, $p(\theta|\psi)$, can be interpreted as describing the variability of the parameters across segments. If the hyperparameter ψ are known, and we have a segment which contains observation $y_{t_1:t_2}$, for $0 < t_1 < t_2 < n$, then the marginal likelihood of this segment is written as

$$Q(\,t_1, t_2; \psi) = \int p(y_{t_1:t_2}|\theta)p(\theta|\psi)d\theta$$

In Bayesian methods, it is important that all segments marginal likelihood can be calculated. That is, for all $t_1,\ t_2\ and\ \psi$, $Q(\,t_1, t_2; \psi)$ can be calculated.

For the method of Bayesian analysis, we need to specify a prior probability for the case of a change-point, $P(M = 1)$. If there is no change-point, we use $P(M = 0)$ denotes the prior probability, note that $P(M = 0) = 1 - P(M = 1)$.

Firstly, considering the simply case that the hyperparameters ψ are known. In this case, the posterior distribution in terms of marginal likelihoods, $Q(\,t_1, t_2; \psi)$, is very straightforward. It is defined as,

$$P(M = 0|y_{1:n}) \propto P(M = 0)Q(1, n; \psi)$$

$$P(M = 1, \tau|y_{1:n}) \propto P(M = 1)p(\tau)Q(1, \tau; \psi)Q(\tau + 1, n; \psi)$$

for $\tau = 1, \dots, n-1$. Here $p(\tau)$ is the probability the position τ is the change-point.

In this case, the posterior is simple to calculate since the marginal likelihoods $Q(\, t_1, t_2; \psi)$ can be calculated analytically. To calculate the posterior probabilities, we can extend the above expression to get the likelihood ratio of whether there is a change-point.

$$\frac{P(M=1|y_{1:n})}{P(M=0|y_{1:n})} \propto \frac{P(M=1)}{P(M=0)} \left(\frac{\sum_{\tau=1}^{n-1} p(\tau) Q(1,\tau;\psi) Q(\tau+1,n;\psi)}{Q(1,n;\psi)}\right)$$

We call the last term the Bayes Factor. The posterior ratio of probabilities of having a change-point to no change-point is the prior ratio multiplied by the Bayes Factor.

Obviously, the value of ψ decides the posterior distribution. The mis-specification of ψ can have significant effect on the posterior probability of having a change-point [79].

There are two popular methods of choosing ψ, in the absence of prior distribution information. The marginal likelihood of ψ is defined as

$$ML(\psi) = P(M=0) Q(1,n;\psi) + \sum_{\tau=1}^{n-1} P(M=1) p(\tau) Q(1,\tau;\psi) Q(\tau+1,n;\psi)$$

Let define $p(\psi)$ is the prior distribution of ψ. Then the marginal posterior of ψ is proportional to $p(\psi)ML(\psi)$, which can be explored by Markov Chain Monte Carlo (MCMC) [80].

Another method of choosing ψ is the empirical Bayes approach. It basically uses the underlying data to get a point estimate for ψ. We can find the value of ψ which maximized $ML(\psi)$. The disadvantage of this approach is it ignores the effect of uncertainty in the choice of ψ.

2.2 Multiple Change-points Detection

The problem can be extended from detecting one single change-point to multiple change-points. Let us assume we have m change points and the positions of change-points are $\tau_{1:m} = (\tau_1, \dots \tau_m)$. Each position of change-point is an integer between 0 and m. We ask the change-points are orders so that $\tau_i < \tau_j$ if and only if $i < j$. As a result, the whole dataset will split into $m + 1$ segments by m change-points, and each segment has a set of parameters, and adjacent segments have different set of parameters. For the ith segment, it contains data $y_{(\tau_{i-1}+1):\tau_i}$ and it can be denoted using parameter set $\{\theta_i, \phi_i\}$, where θ_i is the set of parameters that represents the distribution of the segment, and ϕ_i is the set of nuisance parameters.

First, we need to test how many segments are needed to describe the data, how many change-points exits and then estimate the value of parameters associated with each segment. It is obvious that the likelihood test statistic can be extended from one change-point question to multiple change-points question simply by summing the likelihood of all m segments. Then the problem changes to estimate the maximum of $ML(\tau_{1:m})$ over all possible combinations of $\tau_1, \dots \tau_m$.

In theory, many ideas of detecting single change-point can be adapted to the detecting of multiple change-points. However, as the number of possible change-point increases quickly, the complex of computation increases exponentially and thus much more challenging. For example, if we have a sequence with 1,000 data points. For single change-point problem, the change-point could be at one of 999 possible positions.

For two change-points problem, the change-points could have 499,500 different combinations.

The problem of testing for a parameter change in statistical models was firstly introduced to public by Page in 1955 [81]. After several years, Quandt developed the likelihood ratio approach to detect change-points. He derived a statistic to test for the change of parameters of general linear regression model [82, 83]. However, Quandt didn't derive the distributions of the likelihood ratio test statistic in small samples or asymptotic approximations.

In 1964, Ghernoff and Zacks derived a test statistic to detect the parameter change for normal distribution using a Bayesian approach [84]. They studied the problem of estimating current mean of sequential variables of independent normal distribution whose means are subjected to change in time. Let $\{y_i\}, i = 1, \dots n$ be independent random variables follow normal distribution, $y_i \sim N(\mu_i, \sigma^2), i = 1, \dots, n$. The null hypothesis

$$H_0: \mu_1 = \cdots = \mu_n = \mu_0, -\infty < \mu_0 < \infty$$

$$H_1: \mu_1 = \cdots = \mu_\tau = \mu_0$$

$$\mu_{\tau+1} = \cdots = \mu_n = \mu_0 + \delta$$

$$\tau \in \{1, \dots, n-1\}, \delta > 0$$

In Chernoff and Zacks's method, they assumed μ_0 is unknown and the test statistic is

$$T_n = \sum_{j=1}^{n-1} p(j) \sum_{i=j+1}^{n} (y_i - \overline{y_n})$$

$$\overline{y_n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where $p(j)$ represents prior probabilities of the change-point τ.

Two years later, depending on Ghernoff and Zacks' results, Kander and Zacks extended it from normal distribution to the exponential family distribution [76] with density functions

$$f(x;\theta) = h(x)\exp\{\psi_1(\theta)U(x) + \psi_2(\theta)\}$$

Where $y_i = U(X_i), i = 1, \dots, n. \{X_j\}_{j=1}^{n}$ is a sequence of independent random variables of exponential densities. $\psi_1(\theta)$ and $\psi_2(\theta)$ have continuous derivatives and $\psi_1'(\theta) > 0$. The null hypothesis is

$$H_0: \theta_1 = \cdots = \theta_n = \theta_0 (\theta_0 is\ known)$$

And the composite alternative

$$H_1: \theta_1 = \cdots = \theta_\tau = \theta_0$$

$$\theta_{\tau+1} = \cdots = \theta_n = \theta_0 + \delta$$

$$\tau\{1, \dots, n-1\}, \delta > 0$$

However, they found the test statistic showed a weak convergence to normal distribution. They suggested using an Edgeworth expansion approximation for the distribution of the test statistic when the sample size is not very large.

In 1969, Gardner derived a new statistic based on Chernoff and Zacks's approach. He solved the problem for normal random variables using two-sided hypotheses with $\delta \neq 0$ in the alternative hypothesis [85]. The new statistic is

$$Q_n = \sum_{j=1}^{n-1} p(j) \sum_{i=j+1}^{n} (y_i - \overline{y_n})^2$$

Under the null hypothesis H_0

$$\frac{6n}{n^2-1}Q_n \sim \sum_{k=1}^{n-1} \lambda_k U_k^2$$

Where $U_1, \dots, U_{n-1}$ are $i.i.d.$ standard normal random variables, and

$$\lambda_k = \frac{6n^2}{\pi^2(n^2-1)k^2}\left[\frac{2n}{k\pi}\cos\left(\frac{k\pi}{2n}\right)\right]^{-2}$$

where $k = 1, \dots, n-1$. One can derive:

$$\frac{6n}{n^2-1}Q_n \overset{d}{\rightarrow} \frac{6}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}U_k^2$$

as $n \rightarrow \infty$. However, Gardner didn't get the asymptotic distribution of the test statistic under the alternative hypothesis.

The first change-point detection using a maximum likelihood based method was proposed by Hinkley in 1970. Hinkley is the first person that derived the asymptotic distribution of the test statistic. He tried to find a change-point in mean within normally distributed observations.

MacNeill also derived the asymptotic distribution of a test statistic. He used methods of weak convergence to approximate the distribution in 1974 [86]. The test statistic

$$T_n = \sum_{j=1}^{n-1} p(j)\left[\sum_{i=j+1}^{n} \frac{(\psi_1'(\theta_0)Y_i + \psi_2'(\theta_0))\sqrt{\psi_1'(\theta_0)}}{\psi_1''(\theta_0)\psi_2'(\theta_0) - \psi_1'(\theta_0)\psi_2''(\theta_0)}\right]^2$$

Where $Y_i = U(X_i), i = 1, \dots, n$. The mean and variance of a random variable X_i are

$$\mu(\theta) = -\frac{\psi_2'(\theta_0)}{\psi_1'(\theta_0)}$$

$$\sigma^2(\theta) = \frac{\psi_1''(\theta_0)\psi_2'(\theta_0) - \psi_1'(\theta_0)\psi_2''(\theta_0)}{(\psi_1'(\theta_0))^3}$$

Then, the test statistic T_n is

$$T_n = \sum_{j=1}^{n-1} p(j) \left[\sum_{i=j+1}^{n} \frac{y_i - \mu(\theta_0)}{\sigma(\theta_0)} \right]^2$$

Scott and Knott (1974) [87] and Sen and Srivastava (1975)[88, 89] performed the binary segmentation search algorithm to identify the change-point. In 1993, Venkatraman explained details of consistency results of the binary segmentation approach in multiple change-points problem under various conditions [90].

Binary segmentation approach is an iterative algorithm can be applied with any single change-point method, in theory, together to detect multiple change-points. The entire process is similar to binary search. It starts with the entire sequence as a whole and try to detect one change-point upon a time. If no change-point is detected, the algorithm stops. Otherwise, the sequence will be split into two sub-sequences by the detected change-point. Then the algorithm repeats the above steps in each sub-sequence to see if the change-point still exists. If a change-point is detected in either sub-sequence, the sub-sequence will be split further into two and repeat the detection part. The procedure keeps splitting the sequence until no further change-point is detected.

In Sen and Srivastava's paper, for a sequence with n positions, let $S_t = X_1 + X_2 + \cdots + X_t, 1 \leq t \leq n$ be the sum of partial data up to position t, where X_t is the observation of position t. When X_t is normally distributed with a known variance, the likelihood ratio statistic for testing the null hypothesis against the alternative at an unknown location t is given by

$$Z_B = max_{1 \le t \le n} |Z_t|$$

Where

$$Z_t = \{\frac{1}{t} + 1/(n-t)\}^{-1/2} \{\frac{S_t}{t} - (S_n - S_t)/(n-t)\}$$

If the statistic Z_B exceeds the threshold C so-called critical value, we reject the null hypothesis of no change-point.

Sen and Srivastava used Monte Carlo simulations to determine the critical value for the hypothesis test. It also can be calculated using the approximation method which was introduced by Siegmund(1986).

The significant advantage of binary segmentation methods is the computation speed and efficiency. The computational complexity of binary segmentation is $O(n)$ which is linear to the length of data set. It has also been shown by venkatraman that binary segmentation approach usually provides consistency results under various conditions [90]. The difficult is how to decide the rejection threshold C. Obviously, the different choice of C will lead to the differences in the number of estimated change-pints. Also, due to the iterative nature of the binary segmentation procedure which only detects one single change-point at a time, it's hard to detect a small change-point that buried in the middle of a large segment.

Another general search method was developed in 1998 by Braun, named the segment neighborhood search, also referred as global segmentation [91, 92]. Unlike binary segmentation algorithm that detects change-point one by one, global segmentation tries to split the entire sequence into multiple segments at once and segment boundaries are change-points. In this algorithm, a measure of data fit $R(\cdot)$ was introduced to measure the goodness of fit for each segment. It was recommended to set $R(\cdot)$ as the negative maximum log-likelihood estimator.

$$R\left(y_{t_1:t_2}\right) = -\log p(y_{t_1:t_2}|\hat{\theta})$$

where $t_1 < t_2$, and t_1, t_2 belong to a single segment. With a maximum number of segments M, the algorithm will detect no more than $M - 1$ change-points.

Segment neighborhood search applies dynamic programming to find the best way to split the entire data set into $m + 1$ segments for $m = 0,1, \dots, M - 1$. They claimed the best segmentation partition must minimize the cost function below,

$$\sum_{i=0}^{m} R(y_{\tau_i:\tau_{i+1}})$$

where the change-points positions are at $\tau_1, \tau_2, \dots, \tau_m$.

Auger and Lawrence [93] further improved the segment neighborhood search and provided better segmentation partition by improving the dynamic programming to maximize the log-likelihood directly. However, the disadvantage of neighborhood search is the high computational cost. The segment neighborhood search requires a computational complexity of $O(n^2)$. With large data sets and long sequences, the increased computational cost is not neglectable. However, compared with the binary segmentation algorithm, such cost increase does result in the improvement of predictive performance in simulation studies [92].

Bayesian–based method has also been proposed to detect multiple change-points. For Bayesian method, it is more like a nature extension from single change-point detection to multiple change-points scenario. Firstly, the method needs a prior for both the number and positions of change-points. There are two different ways to specify the prior. The first one is to specify the prior for the number of change-points first, and then specify the prior for the change-points positions given the number of change-points [94]. Another method is to specify the number of change-points and their positions together indirectly through a prior distribution of the length of each segment. Comparing with the first one, the second one has many advantages [95]. With the second one, the prior does not need to be adapted based on the period of time. Also, it is easier to apply statistical inferences from similar data sets to construct appropriate priors. Plus, the second one also provides computational advantages.

To specify the prior distribution of the length of each segment, first, let's denote the probability mass function $g(\cdot;\psi)$ as the mass function for the length of each segment. In the mass function, we allow unknown parameters which will be the hyperparameters of the model. A survivor function $S(\cdot;\psi)$ will be associated with the mass function.

$$S(t;\psi) = \sum_{i=t}^{\infty} g(i;\psi)$$

Then, if we have multiple change-points at positions $\tau_1, \tau_2, \dots, \tau_m$, the prior probability for this m change-points will be

$$p(m, \tau_{1:m}|\psi) = (\prod_{i=1}^{m} g(\tau_i - \tau_{i-1};\psi))S(\tau_{m+1} - \tau_m;\psi)$$

Usually, geometric distribution is selected as the distribution of the segment length with parameter p. In this case, $g(t;\psi) = p(1-p)^{t-1}, S(t;\psi) = (1-p)^{t-1}$ and $p(m,\tau_{1:m}|\psi) = p^m(1-p)^{n-m-1}$. Note that the binomial distribution is the prior for the number of change-points, and the conditional uniform is the prior for the positions of change-points.

Then we need to calculate the posterior for the number of change-points and their positions. For a fixed value of ψ, we can derive the posterior as follow.

$$p(m,\tau_{1:m}|\psi,y_{1:n}) \propto (\prod_{i=1}^{m} g(\tau_i - \tau_{i-1};\psi))Q(\tau_{i-1}+1,\tau_i;\psi))$$

$$\times S(\tau_{m+1} - \tau_m;\psi)Q(\tau_m + 1, \tau_{m+1};\psi)$$

Where $Q(t_1, t_2;\psi)$ is the segment marginal likelihood.

Then we try to estimate the value of the segment parameters by simulating the posterior distribution given the change-points positions. There are two standard methods to generate samples from the posterior $p(m,\tau_{1:m}|\psi,y_{1:n})$. One is the Markov Chain Monte Carlo method (MCMC) [96], and another is reversible jump Markov Chain Monte Carlo method [94]. However, MCMC is very computationally intensive and usually has difficulties of diagnosing convergence of the MCMC algorithm. Therefore, MCMC algorithm is very time-consuming and requires a very long time to run. Otherwise, the simulation result will be incorrect.

To improve the computationally efficiency on sample generation from the posterior, researchers have proposed various algorithms which can be categorized as: forward filtering and backward filtering. The basic idea was firstly proposed by Yao in 1984 [97]. Barry and Hartigan in 1992

[98] and Liu and Lawrence in 1999 [99] also developed similar algorithms for detecting multiple change-points. In 2008, Fearnhead combined both procedures and proposed a forward-backward algorithm with hidden Markov models [100].

For this algorithm, we denote K_t to be the most recent change-point before time t, thus $K_t \in \{0,1,\dots,t-1\}$. If $K_t = 0$, then there is no change-point before time t. Note, if the most recent change-point before t is at time $t-1$, then $K_t = t-1$. If there is no change-point at time $t-1$, we have $K_t = K_{t-1}$.

For forward step, we need to calculate $P(K_t = i|y_{1:t}, \psi)$ for $i = 0,1,\dots,t-1$based on the following recursions. All recursions are initiated with $P(K_t = 0|y_1) = 1$. For $t = 2,3,\dots,n$, assuming there is no change-point at time $t-1$, therefore, $K_t = K_{t-1}$. Then the survive function and segment marginal likelihood terms correspond to the prior probability. The likelihood of the next new observation given condition $K_{t-1} = i$ respectively is as follow:

$$P(K_t = i|y_{1:t}, \psi) \propto P(K_{t-1} = i|y_{1:t-1}, \psi)(\frac{S(t-i;\psi)}{S(t-i-1;\psi)})(\frac{Q(i+1,t;\psi)}{Q(i+1,t-1;\psi)})$$

for $i = 0,1,\dots,t-2$.

Then assuming it is a change-point at time $t-1$, therefore $K_t = t-1$. Then denote $Q(t,t;\psi)$ as the new likelihood function of observation at time t. Respectively, we can have:

$$P(K_t = t-1|y_{1:t}, \psi) \propto Q(t,t;\psi)\sum_{j=0}^{t-2} P(K_{t-1} = j|y_{1:t-1}, \psi)(\frac{g(t-j-1;\psi)}{S(t-j-1;\psi)})$$

The last sum term is the probability that given observation $y_1, \ldots, y_t$, we have a change-point at time $t-1$. More technical details about the procedure of derivation can be found through Fearnhead and Liu's paper [101]. In this paper, they also showed how to use the output of these recursions to calculate the marginal likelihood $Q(\cdot)$ for ψ.

Then, the next step is the backward step in which we generate samples from the posterior of m and K_t. Firstly, we simulate the last change-point from the distribution of K_n given observation $y_1, \ldots, y_n$. The probability mass function is written as

$$P(K_t = i | y_{1:n}, C_{t+1} = t, \psi) \propto P(K_t = i | y_{1:t}, \psi) \left(\frac{g(t-i;\psi)}{S(t-i;\psi)}\right)$$

For $i = 1,2, \ldots t-1$.

The event $C_{t+1} = t$ only happened when there is a change-point at time t. This mass function is written conditioned on a change-point at time t. The observations after this change-point position are independent of the observations before this change-point. We recursively simulate change-points backwards until we get $C_t = 0$.

2.3 Computation Packages of Change-Point

As the increasing needs to identify the location of change-points within time series data, many change-point detection packages had been developed in R or C++ environment. Some of them only provide single test statistic, like sde[102] , bcp [103]. Some of them are designed for specific research areas, like cumSeg [104, 105], DNAcopy [106]. Other comprehensive R packages are also available, for example, strucchange

[107] can be used to detect the changes in regression models and cpm can be used to detect change-point using parametric and nonparametric methods [108]. Contrary to the most packages that only provide one search algorithm for detecting multiple change-points, some other packages implement a choice of several search algorithms. For example, the R packages ‘changepoint’ [109] allows the user to select from a popular set of change-point algorithms and penalty types.

Chapter 3

A Novel Two-Step Change-point Detection Method

3.1 Data Preparation

3.1.1 Data Transformation

The raw data we get from array-CGH and next-generation sequencing platforms is in a special data format, bam file, which can't be used as input file directly in any algorithm. Therefore, before applying any algorithm, the data need to be pre-processed. Bowtie is a fast and accurate read aligner which can be used to align sequencing reads to the human genome. After using Bowtie, we can get experimental information, such as read ID, strand, chromosome, chromosome start position and read length from bam files. SAM (Sequence Alignment Map) tools provide a utility for transferring data format form bam file to text file. After transformation and binning process, the binned data has 4 columns, “chromosome”, “chromosome start position”, “absolute start position” and “read counts”. This is the input data of segmentation and change-point detection algorithms.

3.1.2 Window Size Estimation

Segmentation model requires binned data. Before using any segmentation algorithm, bin boundaries and the window size needs to be decided to get the binned data. Deciding the correct window size is not trivial. A larger window size would loses more information and causes the model to be insensitive on small changes, while a smaller window size would results in lower signal-to-noise ratio and causes the model to be less accurate and robust. However, there is no standard approach to decide the correct bin size. Researchers usually use a fixed bin size based on their own judgments and domain knowledge. Few literatures has discussed on how to decide the window size. Xie in 2008 [61] proposed a data-driven approach for determining the reasonable window size.

As we mentioned earlier, modern sequencing technology often requires two sets of data, the tumor sample and a normal control sample. Let N and T denote the total number of aligned single-cell sequencing reads from normal and tumor sample respectively.

In Xie's approach, assuming there is no breaking point in a window of length w, the number of reads in the window approximately follows a Poisson distribution with parameter

$$\lambda = \frac{S \cdot w}{L}$$

where S is the total number of sequencing reads in the sample, L is the size of the whole genome, w is the sliding widow size where $w \ll L$. For normal sample, N follows a Poisson distribution with parameter $\lambda_N = \frac{N \cdot w}{L}$. For tumor sample, T follows a Poisson distribution with parameter $\lambda_T = \frac{T \cdot w}{L}$. When the average number of reads per window is greater than 10, we can approximate the Poisson distribution with a Gaussian distribution,

$$Poission(\lambda_N) \approx Gaussian(\mu_N = \lambda_N, \sigma^2{}_N = \lambda_N)$$

$$Poission(\lambda_T) \approx Gaussian(\mu_T = \lambda_T, \sigma^2{}_T = \lambda_T)$$

Then, the predicted window size w is

$$w = \frac{(N \cdot r^2 + T) \cdot L \cdot t^2}{(1-r)^2 \cdot N \cdot T}$$

where $t = \frac{T \cdot z - N}{\sqrt{\sigma^2{}_T \cdot z^2 + \sigma^2{}_N}}, r = z \cdot \frac{T}{N}\ and\ z$ is the actual ratio of two samples' read counts.

3.2 Bayesian Change Point Model

As we discussed in the previous chapter, the change point model is originally designed for detecting change-points and structural break. Bayesian-based change point models are a major group in this domain and have draw lots of attention lately due to the improved performance. Comparing with majority Bayesian Change point models which require Markov Chain Monte Carlo implementation [94] which is computationally expensive, a new Bayesian Change Point model (BCP) developed by Lai et al [110] provided an analytical formula for the posterior means and significantly improved the efficiency by avoiding simulation-based inference through MCMC. In their paper, BCP model has also shown improvement compared with widely used segmentation methods in genomic studies. More recently, Xing et al. applied this method on ChIP-seq data to identify diffuse gene domains and demonstrated the new method outperformed the existing segmentation methods [111]. Detecting CNV is essentially to find the boundaries where the copy number count changes on chromosome. Although there is no literature that applied this methodology on CNV detections, the BCP framework naturally fits into this problem and therefore draws our attention.

3.2.1 Model Specification

As a change point model, BCP can calculate the posterior means by explicit formula and its piecewise constant property will help for calling segmentations. Here we also use bounded Complexity Mixture smoother procedure (BCMIX) to handle the weight calculation for forward and backward filter which is much more time efficient compared with other methods. Under BCP framework, assuming the sequencing process is uniform and the sequence reads are randomly mapped to the genome, the number of reads aligning to a region can be modeled as a Poisson distribution with mean directly proportional to the size of the region.

Below is the BCP model assumptions for CNV detection:

1. Let y_t be the read count in the tth window, where $t = 1, \dots, n$.
2. The observations y_t follow a Poisson distribution with parameter θ_t, where θ_t represent the means of y_t

$$y_t \sim Poisson(\theta_t)$$

3. θ_t are independent and identically distributed Gamma random variables,

$$\theta_t \sim Gamma(\alpha, \beta)$$

We assume the read count within the tth window y_t follows a Poisson distribution with parameter θ_t. and θ_t follows a gamma distribution which is a conjugate prior of Poisson, where $t = 1, \dots, n$. Please note, $\{y_t\}$ are pairwise independent and $\{\theta_t\}$ are piecewise constant. Moverover, given $\{\theta_t\}$, $\{y_t\}$ are independent.

Our goal is to estimate the expected read count θ_t in order to find the position where the expected read count changed. Comparing with common Hidden Markov Models (HMMs) with finite states, we allow infinite states (values) for $\{\theta_t\}$.

Before going into the model details forthis specific application, I will describe the general framework for BCP in exponential families mentioned in Lai and Xing's paper [70]. Considering a multiparameter exponential family of densities

$$f(y|\theta) = \exp\{\theta y - \psi(\theta)\}$$

If the prior density is given as

$$\pi(\theta; a_0, \mu_0) = c(a_0, \mu_0) \exp\{a_0\mu_0\theta - a_0\psi(\theta)\}$$

where $\frac{1}{c(a,\mu)} = \int_\theta \exp\{a\mu\theta - a\psi(\theta)\}\, d\theta$, and θ is the parameter we need to estimate, $\psi(\theta)$ is the function of θ. Then we can have the posterior density of θ given the observation $y_1, \dots, y_m$, $f(\theta|y_1, \dots, y_m)$ is,

$$\pi\left(\theta; a_0 + m, \frac{a_0\mu_0 + \sum_{i=1}^{m} y_i}{a_0 + m}\right)$$

3.2.2 Forward Filter

In our model, we base the estimation on the full sequence by combining both forward filtering and backward filtering. We first derive the forward filter $f(\theta_t|y_1, \dots, y_t)$. For the tth window, we denote $I_t = 1_{\{\theta_t \neq \theta_{t-1}\}}$ as the indicator variable which are independent and identically distributed bernoulli random variable with constant success rate p. That is,

$$I_t = \begin{cases} 1\; when\; \theta_t \neq \theta_{t-1} \\ 0\; when\; \theta_t = \theta_{t-1} \end{cases}$$

$$P(I_t = 1) = p$$

Then we denote K_t as the most recent change point before window t, and p_{it} is the conditional probability if $K_t = \mathrm{i}$.

$$K_t = \max\{s \leq t : I_s = 1\}$$

$$p_{it} = P(K_t = i | y_1, \dots, y_t)$$

Where $\sum_{i=1}^{t} p_{it} = 1$.

Given $K_t = \mathrm{i}$, that is $\theta_{i-1} \neq \theta_i = \theta_t$, then based on the independent and identically distributed assumption on $\{\theta_t\}$, the posterior distribution of θ_t given $K_t = \mathrm{i}$ is independent from $\{y_1, \dots, y_{i-1}\}$ and follows the gamma distribution with parameters $(\alpha_{it}, \beta_{it})$. Since gamma is the conjugate posterior of Poisson, we can get the nice property that $\alpha_{it} = \alpha + \sum_{k=i}^{k=t} y_k\, ; \beta_{it} = ({}^{1}\!/_{\beta} + \mathrm{t} - \mathrm{i} + 1)^{-1}$. Then we can get the posterior distribution of θ_t given $y_1, \dots, y_t$ as

$$f(\theta_t | y_1, \dots, y_t) = \sum_{i=1}^{t} f(\theta_t, K_t = i | y_i, \dots, y_t) = \sum_{i=1}^{t} p_{i,t} \cdot Gamma(\theta_t; \alpha_{it}, \beta_{it})$$

where $p_{i,t} = P(K_t = i | y_1, \dots, y_t)$ and it can be solved recursively.

Here we consider the situation under two different conditions. The first one is when the most recent change-point position is at t, and the second situation is when the most recent change-point position is before t.

When $i = t$, since $f(y_t | y_1, \dots, y_{t-1})$ is a constant, so we have $p_{i,t} = P(K_t = t | y_1, \dots, y_t) = \frac{f(K_t = t, y_t | y_1, \dots, y_{t-1})}{f(y_t | y_1, \dots, y_{t-1})} \propto p \cdot f(y_t | I_t = 1)$.

where

$$f(y_t|I_t = 1) = \int_{t=1}^{n} f(\theta_t) \cdot f(y_t|\theta_t) d\theta_t = \frac{\pi_{0,0}}{\pi_{t,t}}$$

When $i < t$, $p_{i,t} = P(K_{t-1} = i|y_1, \dots, y_t) = \frac{f(K_{t-1}=i, y_t|y_1,\dots,y_{t-1})}{f(y_t|y_1,\dots,y_{t-1})} = (1-p) \cdot p_{i,t-1} \cdot f(y_t|y_i, \dots, y_{t-1}, K_t = i)$.

where

$$f(y_t|y_i, \dots, y_{t-1}, K_t = i) = \int_{t=1}^{n} f(y_t|\theta_t) \cdot f(\theta_t|y_i, \dots, y_{t-1}, K_t = i) d\theta_t = \frac{\pi_{i,t-1}}{\pi_{i,t}}$$

In summary, we have:

$$p_{i,t} = p_{i,t}^* / \sum_{s=i}^{t} p_{s,t}^* \propto p_{i,t}^*$$

$$:= \begin{cases} p \cdot f(y_t|I_t = 1), & if\ i = t \\ (1-p) \cdot p_{i,t-1} \cdot f(y_t|y_i, \dots, y_{t-1}, K_t = i), & if\ i < t \end{cases}$$

$$= \begin{cases} p \cdot \frac{\pi_{0,0}}{\pi_{t,t}}, & if\ i = t \\ (1-p) \cdot p_{i,t-1} \cdot \frac{\pi_{i,t-1}}{\pi_{i,t}}, & if\ i < t \end{cases}$$

where $\pi_{0,0} = \beta^{-\alpha}/\Gamma(\alpha)$, $\pi_{i,j} = \beta_{i,j}^{-\alpha_{i,j}}/\Gamma(\alpha_{i,j})$

3.2.3 Backward Filter

To get the backward filter $f(\theta_t|y_{t+1}, \dots, y_n)$, the calculation and derivation are quite similar with the forward filter. The different is we use a location-reversed procedure with the information from y_{t+1} to y_n in backward filtering. For the tth window, we denote $\widetilde{I_t} = 1_{\{\theta_t \neq \theta_{t+1}\}}$ as the indicator variable which are independent and identically distributed bernoulli random variable with constant success rate p. That is,

$$\widetilde{I_t} = \begin{cases} 1 \; when \; \theta_t \neq \theta_{t+1} \\ 0 \; when \; \theta_t = \theta_{t+1} \end{cases}$$

$$P(I_t = 1) = p$$

Then we denote $\widetilde{K_t}$ as the most recent change point after window t, and q_{jt} is the conditional probability if $\widetilde{K_t} = j$.

$$\widetilde{K_t} = \min\{s \geq t : I_s = 1\}$$

$$\widetilde{p_{jt}} = q_{jt} = P(\widetilde{K_t} = j | y_t, \dots, y_n)$$

Where $\sum_{j=t}^{n} q_{jt} = 1$.

If t is the change point, θ_t follows Gamma(α, β). Otherwise, let's denoted $q_{j,t+1}$ as the probability that j is the first change point after t. Similarly, we could get the posterior distribution of θ_t given $y_{t+1}, \dots, y_n$ as

$$f(\theta_t | y_{t+1}, \dots, y_n) = p \cdot f(\theta_t) + (1-p) \cdot \sum_{j=t+1}^{n} q_{j,t+1} \, f\left(\theta_t \middle| \widetilde{K_{t+1}} = j, y_{t+1}, \dots, y_n\right)$$

$$= pGamma(\alpha, \beta) + (1-p) \cdot \sum_{j=t+1}^{n} q_{j,t+1} \, Gamma\left(\theta_t; \alpha_{j,t+1}, \beta_{j,t+1}\right)$$

For $q_{j,t}$ the formula is almost the same with $p_{i,t}$, expect we use $t+1$ instead of $t-1$.

$$q_{j,t} = q_{j,t}^{*} / \sum_{s=i}^{t} q_{s,t}^{*} \propto q_{j,t}^{*} = \begin{cases} p \cdot \dfrac{\pi_{0,0}}{\pi_{t,t}}, & if \; j = t \\ (1-p) \cdot q_{j,t+1} \cdot \dfrac{\pi_{j,t+1}}{\pi_{j,t}}, & if \; j > t \end{cases}$$

As the above, where $\pi_{0,0} = \beta^{-\alpha}/\Gamma(\alpha)$, $\pi_{i,j} = \beta_{i,j}^{-\alpha_{i,j}}/\Gamma(\alpha_{i,j})$.

If $j = t$

$$q_{jt} = P\left(\widetilde{K_t} = j \middle| y_t, \dots, y_n\right) = P(\widetilde{K_t} = j, \widetilde{I_t} = 1 | y_t, \dots, y_n)$$

For backward filter, if at time t, $I_t = 1$, then the information after t, $y_{t+1}, \dots, y_n$, are useless, that is $P(\widetilde{K_t} = j, \widetilde{I_t} = 1|y_t, \dots, y_n)$ = $P(\widetilde{K_t} = j, \widetilde{I_t} = 1|y_t)$. In addition, since data $y_t, \dots, y_n$ are observed, so we get $f(y_t) = f(y_t|y_t, \dots, y_n)$. Then,

$$\begin{aligned} q_{jt} = P\left(\widetilde{K_t} = j, \widetilde{I_t} = 1 \middle| y_t\right) \propto f\left(\widetilde{I_t} = 1\right) f\left(y_t \middle| \widetilde{I_t} = 1\right) = p \cdot f\left(y_t \middle| \widetilde{I_t} = 1\right) \\ = p \cdot \frac{\pi_{0,0}}{\pi_{t,t}} \end{aligned}$$

If $j < t$

$$q_{jt} = P\left(\widetilde{K_t} = j \middle| y_t, \dots, y_n\right) \propto (1-p) \cdot q_{j,t+1} \cdot f(y_t|\widetilde{K_{t+1}} = j, (y_{t+1,} \dots, y_j))$$

Here we use the similar derivation in forward filter for $f(y_t|\widetilde{K_{t+1}} = j, (y_{t+1,} \dots, y_j)$. Then,

$$q_{jt} \propto (1-p) \cdot q_{j,t+1} \cdot \frac{\pi_{j,t+1}}{\pi_{j,t}}$$

3.2.4 Smoothing

After calculating the posterior distributions of θ_t with forward and backward filter, we can apply Bayes' theorem,

$$f(\theta_t|y_1, \dots, y_n) \propto \frac{f(\theta_t|y_1, \dots, y_t) f(\theta_t|y_{t+1}, \dots, y_n)}{f(\theta_t)}$$

to combine both posteror and yield the posterior distribution of θ_t given $y_1, \dots, y_n$ as

$$f(\theta_t|y_1, \dots, y_n) = \sum_{1 \le i \le t \le j \le n} \delta_{ijt} \cdot Gamma\big(\theta_t; \alpha_{ij}, \beta_{ij}\big)$$

where $\delta_{ijt} = \delta^*_{ijt}/P_t$, $P_t = p + \sum_{1 \le i \le t \le j \le n} \delta^*_{ijt}$

$$\delta_{ijt} \propto \delta^*_{ijt} \coloneqq \begin{cases} p \cdot p_{it}, & if\ i = t \\ (1-p) \cdot p_{it} q_{j,t+1} \cdot \dfrac{\pi_{it}\pi_{t+1,j}}{\pi_{ij}\pi_{00}}, & if\ i < t \end{cases}$$

and $\pi_{0,0} = \beta^{-\alpha}/\Gamma(\alpha)$, $\pi_{i,j} = \beta_{i,j}^{-\alpha_{i,j}}/\Gamma(\alpha_{i,j})$

The above formula gives the estimate of θ_t

$$\mathrm{E}(\theta_t|y_1, \dots, y_n) = \sum_{1 \le i \le t \le j \le n} \delta_{ijt} \cdot \big(\alpha_{ij} \cdot \beta_{ij}\big)$$

We have forward filter, backward filter and prior density function of θ_t

$$f(\theta_t|y_1, \dots, y_t) = \sum_{i=1}^{t} p_{it}\pi(\theta_t; a_0 + t - i + 1, \frac{a_0\mu_0 + \sum_{k=i}^{t} y_k}{a_0 + t - i + 1})$$

$$f(\theta_t|y_{t+1}, \dots, y_n) = p\pi(\theta_t; a_0, \mu_0) +$$

$$(1-p) \sum_{j=t+1}^{n} q_{j,t+1}\pi(\theta_t; a_0 + j - t, \frac{a_0\mu_0 + \sum_{k=t+1}^{j} y_k}{a_0 + j - t})$$

$$f(\theta_t) = \pi(\theta_t; a_0, \mu_0)$$

Based on Bayes' theorem

$$f(\theta_t|y_1, \dots, y_n)$$

$$\propto p \sum_{i=1}^{t} p_{it}\pi\left(\theta_t; a_0 + t - i + 1, \frac{a_0\mu_0 + \sum_{k=i}^{t} y_k}{a_0 + t - i + 1}\right) + (1$$

$$- p) \sum_{i=1}^{t} \sum_{j=t+1}^{n} p_{it}q_{j,t+1} \frac{\pi_{it}\pi_{t+1,j}}{\pi_{00}\pi_{ij}} \pi(\theta_t; a_0 + j - i$$

$$+ 1, \frac{a_0\mu_0 + \sum_{k=i}^{j} y_k}{a_0 + j - i + 1})$$

Since $\sum_{i=1}^{t} P_{it} = 1$,

$$P_t = p + \sum_{1 \le i \le t \le j \le n} \delta_{ijt}^{*} = p \sum_{i=1}^{t} P_{it} + (1 - p) \cdot \sum_{i=1}^{t} \sum_{j=t+1}^{n} p_{it}q_{j,t+1} \cdot \frac{\pi_{it}\pi_{t+1,j}}{\pi_{ij}\pi_{00}}$$

Then, under our specific model assumptions,

$$f(\theta_t | y_1, \dots, y_n) = \sum_{1 \le i \le t \le j \le n} \frac{\delta_{ijt}^{*}}{P_t} \cdot \pi(\theta_t; a_0 + j - i + 1, \frac{a_0\mu_0 + \sum_{k=i}^{j} y_k}{a_0 + j - i + 1})$$

$$= \sum_{1 \le i \le t \le j \le n} \delta_{ijt} \cdot Gamma\big(\theta_t; \alpha_{ij}, \beta_{ij}\big)$$

Since $\mathrm{E}\left(Gamma\big(\theta_t; \alpha_{ij}, \beta_{ij}\big)\right) = \alpha_{ij} \cdot \beta_{ij}$,

$$\mathrm{E}(\theta_t | y_1, \dots, y_n) = \sum_{1 \le i \le t \le j \le n} \delta_{ijt} \cdot \big(\alpha_{ij} \cdot \beta_{ij}\big)$$

3.2.3 Hyperparameters Estimation

The estimate of θ_t, $\mathrm{E}(\theta_t | y_1, \dots, y_n)$, includes the hyperparameters $p, \alpha\ and\ \beta$, which can be calculated using the empirical Bayes approach.

From the definition of $p_{i,t}^{*}$ and the posterior distribution of $\theta_t|y_1, \dots, y_t$, the likelihood function of $p, \alpha\ and\ \beta$ is

$$\prod_{t=1}^{n} f(y_t|y_1, \dots, y_{t-1}) = \prod_{t=1}^{n} (\sum_{i=1}^{t} p_{i,t}^{*})$$

Since $y_1, \dots, y_t$ are exchangeable random variables in our model, we can first estimate α and β using the method of moments. Then putting the estimated hyperparameters α and β into above equation, we can estimate the relative frequency p of change-points by maximizing the log-likelihood function $l(p) = \sum_{t=1}^{n} \log\left(\sum_{i=1}^{t} p_{i,t}^{*}\right)$ which can be conveniently computed with grid search[110].

$$l(p) = \sum_{t=1}^{n} \log\left(\sum_{i=1}^{t} p_{i,t}^{*}\right)$$

$$= \sum_{t=1}^{n} \log\left(p \cdot f(y_t|I_t = 1) + \sum_{i=1}^{t-1} f(y_t|K_{t-1} = i, y_{t-1}) p_{i,t-1}(1-p)\right)$$

$$= \sum_{t=1}^{n} \log\left(p \cdot \frac{\pi_{00}}{\pi_{tt}} + (1-p) \cdot \sum_{i=1}^{t-1} p_{i,t-1} \frac{\pi_{i,t-1}}{\pi_{i,t}}\right)$$

Since $\frac{\pi_{00}}{\pi_{tt}}$ and $\sum_{i=1}^{t-1} p_{i,t-1} \frac{\pi_{i,t-1}}{\pi_{i,t}}$ can be calculated through the prior and posterior distribution, let us denote $\frac{\pi_{00}}{\pi_{tt}} = c_1$ and $\sum_{i=1}^{t-1} p_{i,t-1} \frac{\pi_{i,t-1}}{\pi_{i,t}} = c_2$. Then the likelihood function can be re-writed as,

$$l(p) = \sum_{t=1}^{n} \log(p \cdot c_1 + (1-p) \cdot c_2)$$

Then just take the derivative and set to zero to get:

$$\frac{\partial l(p)}{\partial p} = \sum_{t=1}^{n} \frac{c_1 - c_2}{c_1 p + c_2(1-p)} = \sum_{t=1}^{n} \frac{1}{p + \frac{c_2}{c_1 - c_2}} = 0$$

Numerically, we can use grid search method to find p, which has the form $\{2^j/n : j_0 < j < j_1, j \in \mathcal{Z}, j_0 \leq 0 \leq j_1\}$.

3.3 Circular Binary Segmentation

Circular binary segmentation (CBS) is the most popular and the standard segmentation method in today's copy number research and has been widely used and applied in genetic copy number study. The performance of CBS is usually very good with array-CGH and next generation sequencing data as reported in literature. Therefore, this is also the first algorithm out of three I am going to compare with in the experimental results.

Circular binary segmentation is original designed for array copy number data by Olshen and Venkatraman[43] in 2004. CBS is essentially a modified version over the binary segmentation method initially developed by Sen and Srivastava in 1975 [88, 89]. The underlying algorithm for CBS is still a recursive algorithm, which iteratively identifies all the potential change-points in the data. However, Olshen et al. made several modifications which made CBS the golden-standard algorithm and outperformed the basic binary segmentation method.

The first and the most important modification is the test statistics used in CBS. In majority segmentation based methods, some kind of statistics is calculated to represent the gains or losses of copy number and then if the statistics is above certain threshold, the null hypothesis of no change-point is rejected. In CBS, the statistics is inspired from the likelihood ratio test statistic proposed by Levin and Kline (1985). Let y_t be the array copy number in the t th window, where $t = 1, \dots, n$. The observations y_t has a common distribution function F_t. For each window t:

$$H_0\text{: there is no change} - \text{point}$$

$$H_1\text{: there is exactly one change} - \text{point at window } t$$

Let $S_t = y_1 + y_2 + \cdots + y_t, 1 \leq t \leq n$ be the sum of partial data up to window t. The test statistics of CBS is:

$$Z_C = max_{1 \leq i \leq j \leq n} \left| Z_{ij} \right|$$

where Z_{ij} is a likelihood ratio test statistics for testing if two arcs have different mean:

$$Z_{ij} = \{\frac{1}{j-1} + \frac{1}{n-j+i}\}^{-1/2} \{\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\}$$

Same as the basic binary segmentation approach, if the statistic Z_C exceeds a certain threshold, a change-point is declared. The procedure is applied recursively until there is no any change-point in any segments. Also, this test statistics allows for both single change-point $(j = n)$ and multiple change-points $(j < n)$.

CBS also introduced an additional "undo" step after all the changes are detected to remove the false change-points due to edge effect. It is observed in experiments that CBS tends to detect additional change-point when the actual change-point is close to the start of the end of the sequence. Therefore, after all change points are identified, CBS will re-calculated the Z_{ij} for all change-point pairs if either i is close to 1 or j is close to n to check if Z_{ij} still above the threshold. If not, the change point will be removed.

Another modification is that CBS offers a permutation-based algorithm to construct the reference distribution when the data is not normally distributed. Under the null hypothesis, $\{X_i\}, where\ i = 1, \ldots n$, are identically distributed. Then, let's denote the reference data $\{X_i^*\}$ is a random permutation sample of the original sequence $\{X_i\}$, and corresponding $Z_C^* = max|Z_{ij}^*|$. For a given significant level α, the stopping criteria of the permutation procedure is when the number of $Z_C^* > Z_C$, exceeds $\alpha * (the\ number\ of\ permutations)$ for the first time. For really large datasets, CBS also offers to split the data into K equal-sized overlapping windows and then run the algorithm for each windows separately.

Besides the general modifications above, CBS also have two additional modifications specifically for array DNA copy number data, which CBS is originally designed for, in order to reduce the overall noise in the data. One is automatic outlier smoothing as part of the data preprocessing and another is a 'pruning' procedure to stop earlier in the iterative process.

The nice property of CBS is that CBS gives a natural way to split the entire chromosome into several contiguous segments and uses a permutation reference distribution to round parametric modeling of the data. Although CBS is originally designed for array copy number data, the methodology has became the standard for next generation sequencing data as well. R package 'DNAcopy' is one of the most popular package for applying CBS method to detect change-point.

3.4 Stochastic Change-Point Model (SCP)

Stochastic Change-Point model, which is developed by Lai et al. in 2008 [70], was applied on intensity ratio estimation with array-CGH data and showed promising result. SCP is based on Hidden Markov Model. Compared with CBS, SCP is more similar with BCP regarding to the modeling framework, but with different target function and distribution assumptions.

As we discussed in chapter 2, Hidden Markov model is one of the popular statistical models for change-point modeling. In this paper, they proposed a new model by using the Hidden Markov models to measure intensity ratio between normal and cancel cells assuming underlying copy numbers are the hidden states with certain transition probabilities. SCP model uses the same idea by using Hidden Markov model to disclose the underlying copy numbers for each position spot. Comparing with other models that are based on pseudo-likelihood or Monte Carlo approximations, SCP has an outstanding advantage. Similar with BCP, SCP also gives the explicit formulation of the posterior distributions of the latent variable. As a result, SCP model is a latent variable model with attractive statistical and computational properties.

For each sample data from array-CGH experiments, we can get an ordered sequence of y_t, where t represents the position in the whole genome and y_t is the log ratio of the tumor and normal samples intensities in the position t. $\{y_i\}$ is a sequential number which naturally ordered by the genetic location along the whole genome.

Comparing with BCP, SCP has different target and input data. In BCP, y_t represents the read count in the tth window which is non-negative integer and therefore can be modeled as a Poisson distribution to estimate the parameter θ_t. However, in SCP, y_t is the log ratio of intensitives of the dyes between tumor and normal tissues in array-CGH experiment at each spot. This intensity represents the copy number ratio of tumor and normal tissues. Then, SCP estimates the expected intensity ratio.

Model specification of SCP is as follow. SCP assumes the log ratio of tumor and normal cells at position t, y_t, follows the a normal distribution with mean θ_t and variance σ.

$$y_t = \theta_t + \sigma\epsilon_t, \epsilon_t \sim N(0,1)$$

where θ_t is an unknown step function of t and ϵ_t is independent standard normal random variables.

The prior distribution of θ_t is modeled as a three-state reversible Markov Chain with one baseline state zero and two $i.i.d$ non-zero states which follows normal distribution $N(\mu, v)$. In this setting, at any time t, if θ_t is in the baseline state, then at time $t+1$, it can choose to stay in the baseline state or jump to any non-zero states with equal probability. Then if θ_t is in one of the non-zero state, it can choose to stay in the same state, jump to another non-zero state or move back to the baseline state. Let's denote:

- $p/2$ is the probability from the baseline state to non-zero state
- a is the probability of staying in any of the non-zero state
- b is the probability of jumping from a non-zero state to another non-zero state
- c is the probability of jumpy from a non-zero state back to baseline state.

Then we can construct the transition probability matrix as follow

$$P = \begin{pmatrix} 1-p & \frac{1}{2}p & \frac{1}{2}p \\ c & a & b \\ c & b & a \end{pmatrix}$$

Now you can see, SCP has a different model specification and distribution assumptions due to the nature of the underlying data. However, after the model specification, SCP follows the exact same procedure and technics for the estimation part as BCP, which includes the forward and backward filtering, smoothparameters are derived through the forward filtering and backward filtering. Smoothing based on bayes' theorem and hyperparameter estimation. For each position t, let K_t be the most recent change-point position before or equal to t, p_t is the probability that θ_t is in baseline state and $q_{i,t}$ is the probability that θ_t is in the non-zero state. Then we have:

$$K_t = \max\{s \leq t: \theta_s = \cdots = \theta_t, \theta_{s-1} \neq \theta_s\}$$

$$p_t = P\left(\theta_{K_t} = 0 \middle| y_1, \ldots, y_t\right) = P(\theta_t = 0 | y_1, \ldots, y_t)$$

$$q_{i,t} = P\left(\theta_{K_t} \neq 0, K_t = i \middle| y_1, \ldots, y_t\right)$$

for $1 \leq i \leq t$. Then one can easily derive the forward filter, the posterior distribution of θ_t given $y_1, \ldots, y_t$ass a mixture of normal distributions:

$$\theta_t | y_1, \ldots, y_t \sim p_t \delta_0 + \sum_{i=1}^{t} q_{i,t} N(\mu_{i,t}, v_{i,t})$$

where,

$$\mu_{i,t} = (\frac{\mu}{v} + \sum_{k=i}^{t} \frac{y_k}{\sigma^2}) v_{i,t}$$

$$v_{i,t} = (\frac{1}{v} + \frac{t-i+1}{\sigma^2})^{-1}$$

Let $\phi_{\mu,v}$ denotes the density function of the normal distribution $N(\mu, v)$.

$$\phi_{\mu,v}(y) = (2\pi v)^{-1/2} \exp\{-\frac{1}{2}(y-\mu)^2/v\}$$

Then applying the same recursion technic used in section 3.2.2, we can easily get the formula of p_t and $q_{i,t}$ as follow:

$$p_t \propto {p_t}^* \coloneqq (1-p)p_{t-1} + cq_{t-1}$$

$$q_{i,t} \propto {q_{i,t}}^* \coloneqq \begin{cases} \frac{(pp_{t-1} + bq_{t-1})\psi}{\psi_{t,t}}, i = t \\ \frac{aq_{i,t-1}\psi_{i,t-1}}{\psi_{i,t}}, i < t \end{cases}$$

where $q_t = \sum_{i=1}^{t} q_{i,t} = 1 - p_t$, $\psi = \phi_{\mu,v}(0)$ and $\psi_{i,j} = \phi_{\mu_{i,j},v_{i,j}}(0)$ for $i \le j$. Then,

$$p_t = {p_t}^*/[{p_t}^* + \sum_{i=1}^{t} {q_{i,t}}^*]$$

$$q_{i,t} = {q_{i,t}}^*/[{p_t}^* + \sum_{i=1}^{t} {q_{i,t}}^*]$$

After the calculation of p_t and $q_{i,t}$, we can get the estimation of θ_t given $y_1, \dots, y_t$ as:

$$E(\theta_t | y_1, \dots, y_t) = \sum_{i=1}^{t} q_{i,t}\mu_{i,t}$$

As we mentioned before in BCP method, we can reverse the position and use the data after position t to get the backward filter estimation. The calculation method is very similar with the forward filter estimation shown above. So first is to derive the posterior distribution of θ_{t+1} given $y_{t+1}, \dots, y_n$:

$$\theta_{t+1} | y_{t+1}, \dots, y_n \sim \tilde{p}_{t+1}\delta_0 + \sum_{j=i+1}^{n} \tilde{q}_{j,t+1} N(\mu_{t+1,j}, v_{t+1,j})$$

Then we can calculate $\tilde{p}_s$ and $\tilde{q}_{j,s}$ by the recursion technics as we did for p_t and $p_{i,t}$ to get:

$$\tilde{p}_s \propto \tilde{p}_s{}^* \coloneqq (1-p)\tilde{p}_{s+1} + c\tilde{q}_{s+1}$$

$$\tilde{q}_{j,s} \propto \tilde{q}_{j,s}{}^* \coloneqq \begin{cases} \dfrac{(p\tilde{p}_{s+1} + b\tilde{q}_{s+1})\psi}{\psi_{s,s}}, j = s \\ \dfrac{a\tilde{q}_{j,s}\psi_{s+1,j}}{\psi_{s,j}}, j > s \end{cases}$$

where $\tilde{q}_{s+1} = \sum_{j=s+1}^{n} \tilde{q}_{j,s+1} = 1 - \tilde{p}_{s+1}$. Then, the backward estimation of θ_t given data $y_{t+1}, \dots, y_n$

$$\theta_t | y_{t+1}, \dots, y_n \sim [(1-p)\tilde{p}_{t+1} + c\tilde{q}_{t+1}]\delta_0 + (p\tilde{p}_{t+1} + b\tilde{q}_{t+1})N(\mu, v) + a\sum_{j=t+1}^{n} \tilde{q}_{j,t+1} N(\mu_{t+1,j}, v_{t+1,j})$$

Based on the Bayes' theorem, we can combine the forward filter and the backward filter together to derive the posterior distribution of θ_t given data $y_1, \dots, y_n$, which is a mixture of normal distribution:

$$\theta_t | y_1, \dots, y_n \sim a_t\delta_0 + \sum_{1 \le i \le \mathrm{t} \le \mathrm{j} \le \mathrm{n}} \beta_{ijt} N(\mu_{ij}, v_{ij})$$

Where

$$a_t = a_t^*/A_t$$

$$\beta_{ijt} = \beta_{ijt}^*/A_t$$

$$A_t = a_t^* + \sum_{1\le i\le t\le j\le n} \beta_{ijt}^*$$

$$a_t^* = p_t[(1-p)\tilde{p}_{t+1} + c\tilde{q}_{t+1}]/c$$

$$\beta_{ijt}^* = \begin{cases} \dfrac{q_{i,t}(p\tilde{p}_{t+1} + b\tilde{q}_{t+1})}{p}, i \le t = j \\ \dfrac{aq_{i,t}\tilde{q}_{j,t+1}\psi_{i,t}\psi_{t+1,j}}{p\psi\psi_{i,j}}, i \le t < j \end{cases}$$

and the final expectation of θ_t given $y_1, \dots, y_n$ is:

$$E(\theta_t|y_1, \dots, y_n) = \sum_{1\le i\le t\le j\le n} \beta_{ijt}\mu_{ij}$$

At the end, the figure below is the comparison among SCP, CBS and HMM from Lai's paper [70]. SCP is more sensitive then the other two methods on small change-points. In the simulation section, we used R package 'cnv' to apply SCP method on our data and compared the result.

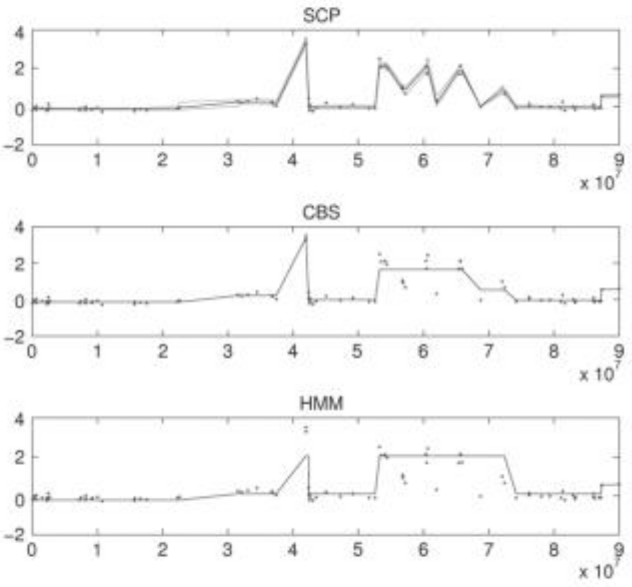


Figure 2. Array-CGH profile for chromosome 17 in cell line BT474. The lines are the signal levels estimated using SCP (top plot), HMM (middle plot), and CBS (bottom plot) [70].

3.5 Cumulative Sum Control Chart

The last method we introduce here is Cumulative Sum (CUSUM) control chart. It's a very traditional method for data visualization. In recent years, some novel developments based on CUSUM's theory are published with different stopping rule or delay of the time.

Cumulative Sum control chart is a change-point analysis technique which is developed by Page E.S. in 1954 [112]. It can be used to monitor the change-point in a sequential data. Let us define the sequential data as $\{y_t\}$, $t = 1,2, \dots, n$. y_i represents the value at position i. To build CUSUM charts, we need to calculate and plot a sequential cumulative sum based on $\{y_t\}$. Let $S_0, S_1, \ldots, S_n$ represent the cumulative sums at each position. One thing needs to be noticed is that n data points will have $n+1$ (0 through n) cumulative sums. To calculate the cumulative sums, we set $S_0 = 0$ first. Then, the cumulative sums are calculated as

$$S_i = S_{i-1} + (X_i - \bar{X}), \text{ where } i = 1, \ldots, n, \bar{X} = \frac{X_1, X_2, \ldots, X_n}{n}$$

Here, S_i is not the traditional cumulative sum of the values. Instead, it actually is the cumulative sum of divergence from the sequential average. It's easy to know that S_i begins at zero, $S_0 = 0$, and also ends at zero. $S_n = S_0 + (X_1 - \bar{X}) + \cdots + (X_n - \bar{X}) = 0$. Therefore, an upward trend of the CUSUM control chart indicates this period values are above the average and a downward trend means this period values are below the average.

As I mentioned previously, CUSUM is a very traditional, yet efficient method. In 2013, Mei introduced a novel change-point detection method based on CUSUM approach. One advantage of this method is that the distribution change does not affect all data streams. Moreover, the stopping rule has higher computational efficiency. However, when the signal is small, the delay of the detection is significant.

In the simulation section, I used R package 'qcc' to apply CUSUM method to detect change-point.

3.6 Similarity Score

All four methods discussed in the previous sections use the binned data. That is, we only get several candidate windows with potential change points instead of the actual change-points positions. We assumed the read depth at a certain position is expected to be proportional to the copy number at that position. However, this simple concept is complicated by the fact that genomes are not sequenced deeply enough to enable base-pair resolution. However, many mutation events encompassing or partially deleting those genes would be much smaller than the resolution of a segmentation profile for this type of data. In this case, segmentation method will fail to detect such small events. Moreover, segmentation approaches are limited to the bin sizes and would not be able to accurately detect the exact location of the events. Since BCP is still a segmentation based method, the same limitations hold for BCP as well. To overcome the limitation of binning, we tried to identify a more accurate position of change point though Distribution-based similarity score: Jensen-Shannon Divergence (JSD) and Knowledge-based similarity score: Different Score (DS). In theory, the similarity score will work well with any segmentation methods.

Within each segment region identified by a segmentation method as change-point, a similarity score is calculated for every mappable position to measure the similarity of both sides and then used to assess the likelihood of being a CNV point. Here we denoted x as the middle position between two adjacent reads, x_{L} as left boundary position, x_R as right boundary position. To determine window's left and right boundaries, we used the following rules: for a tumor middle position x, we define local windows that include exactly w consecutive reads in its matched normal sample to the left and to the right of position x. w is a custom defined integer number. In other words, we can denote:

$$n(x_L, x) = n(x, x_R) = w$$

3.6.1 Jensen-Shannon Divergence

For the distribution-based similarity score, we assumed the existence of single-cell sequencing read at every mappable position follows Bernoulli distribution with success probability P. After determining the left and right local window's boundaries, x_L and x_R, we calculated the number of single-cell sequencing reads in tumor samples in each local window. Let's assume the left local window has t_L tumor reads, and the right local window has t_R tumor reads. Then, we assumed every position in left region $x_L x$ follows $Bernoulli(P_L)$ and every position in right side xx_R follows $Bernoulli(P_R)$.

$$P_L = t_L/w \text{ and } P_R = t_R/w.$$

For normal cell with no CNV, we have $P_L = P_R$. Therefore, detecting CNV becomes to find the position x where $P_L \neq P_R$. To solve this problem, we used Jensen-Shannon divergence [113] to measure the distribution difference. It can measure the similarity between two probability distributions accurately and efficiently. Jensen-Shannon divergence is based on the Kullback-Leibler divergence with some useful improvements.

Kullback-Leibler divergence is a very popular statistics to measure the distance between two distributions. Here is the definition of Kullback-Leibler divergence on two discrete distributions P and Q,

$$D_{KL}(P||Q) = H(P,Q) - H(P) = -\sum_i P(i)\, log\, Q(i) + \sum_i P(i)\, log\, P(i)$$

$$= \sum_{i} log\left(\frac{P(i)}{Q(i)}\right)P(i)$$

where $\sum_i P(i) = 1, \sum_i Q(i) = 1$

Kullback-Leibler divergence is strictly positive, but the problem is that it is a directional and non-symmetric measurement, which means the Kullback-Leibler divergence distance of P from Q is not the same as the Kullback-Leibler divergence distance of Q from P.

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

To avoid these disadvantages, Jensen-Shannon divergence is a smoothed version of Kullback-Leibler divergence. It has a lot of good properties. It is symmetric , non-negative and monotonic. That is the reason we choose to use Jensen-Shannon divergence as one of the scores to locate the change point.

$$JSD\ (P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

where $\mathrm{M} = \frac{1}{2}(P+Q)$

In our case, we have two Bernoulli distributions with parameter P_L and P_R. The Jensen-Shannon divergence is given as below:

$$\begin{aligned} JSD\ (P||Q) &= \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \\ &= \sum_{i=1}^{2} \log\left(\frac{P(i)}{M(i)}\right)P(i) + \sum_{i=1}^{2} \log\left(\frac{Q(i)}{M(i)}\right)Q(i) \end{aligned}$$

$$= \log\left(\frac{P_L}{\frac{P_L + P_R}{2}}\right) \cdot P_L + \log\left(\frac{1 - P_L}{\frac{P_L + P_R}{2}}\right) \cdot (1 - P_L)$$

$$+ \log\left(\frac{P_R}{\frac{P_L + P_R}{2}}\right) \cdot P_R + \log\left(\frac{1 - P_R}{\frac{P_L + P_R}{2}}\right) \cdot (1 - P_R)$$

Jensen-Shannon divergence is symmetric, non-negative and monotonic. When it equals 0, it means two distributions are the same. The higher the Jensen-Shannon divergence is, the more different the two distributions are.

3.6.2 Difference Score

The second similarity score we used to detect and locate CNVs from single-cell sequence data is Difference Score. Biologically, if there is no CNV, the read count in the tumor cell should be comparable with the read count in the same region of a normal cell. Otherwise, there is likely CNV.

Let n_x and t_x denote the total number of aligned single-cell sequencing reads from normal and tumor sample respectively in window x. The tumor-normal copy number ratio of this window, R, is defined as:

$$R(x_L, x_R) = \frac{t_x}{n_x}$$

where $n_x > 0$.

Then we set up a log-ratio difference statistic D. In our algorithm, we used this local difference of log ratios statistics to identify significant copy number changes. We calculated the difference in log ratio between the right window and the left window.

$$D(x) = log(\ R(x_L, x)) - log(\ R(x, x_R))$$

Where

$$R(x_L\ , x) = \frac{t_L}{n_L} and\ R(x, x_R) = \frac{t_R}{n_R}$$

Similar with Jensen-Shannon divergence, Difference-Score measures the difference and reflects the likelihood of being a breaking point. However, one advantage is that Difference-Score could be either positive or negative. A higher absolute value of Difference-Score indicates a greater likelihood of being a breaking point. Positive score means the position’s right side has higher copy number than its left side and vice versa.

After calculating two different scores for every point in the potential regions, to decide the significance of the existence of CNV, we approximated the p-value of every point by permutation method.

Chapter 4

Simulation Study

In this section, we present the numerical results on simulation data where the ground truth is known and comparisons among our novel two-step method and the other two popular segmentation method CBS [43] and SCP [70]. CBS is the most popular segmentation method in genetic study. SCP is a similar algorithm with BCP. Moreover, we also applied CUSUM as a change-point detection method. Before showing the simulation results, we will introduce the simulation mechanism and how the data is generated. Firstly, there are segmentation results from the BCP model, CBS model, SCP model. Then, we add change-point detection results from CUSUM model and followe by the improvement on the accuracy of change-point identification added by similar scores. Except the accuracy of segmentation results and change-point position, we also compare the computational efficiency of all algorithms.

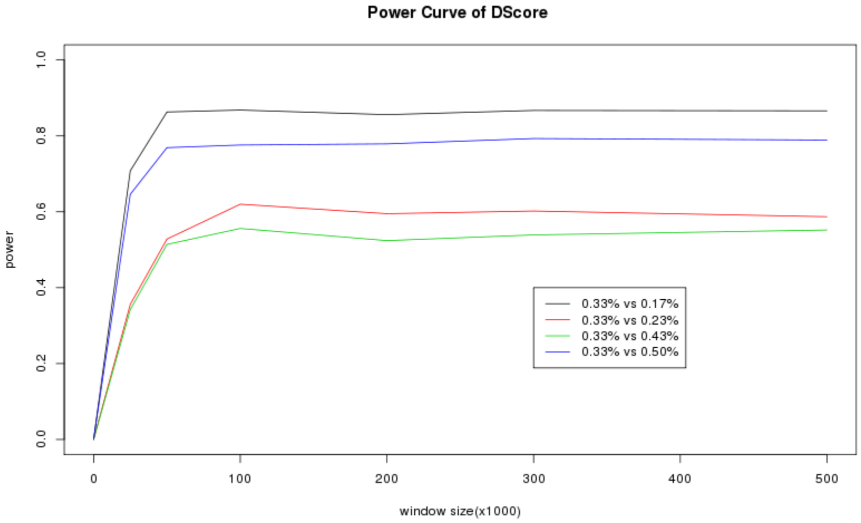
4.1 Experiment Setting

To study the performance of the new algorithm, we set up three simulation studies: no change-point simulation, one change-point simulation and multiple change-points simulation. Each simulation study covers several different scenarios with different reads coverage and different amplitude. For each scenario, 1000 samples of size $s = 5{,}000{,}000$ were generated to evaluate the performance of CBS, SCP, CUSUM and our new algorithm BCP with two similarity score methods.

As suggested by Xie's method, we selected 50K as the window size for segmenting data. Also, we will use the same window size to calculate the similarity scores. We also performed power analysis on JS and DScore to verify the window size is reasonable.

On average, there are around 10 million reads on a full chromosome with length $3 * 10^9$ bps which yields a probability of $\frac{10^7}{3*10^9} = 0.0033$. In the power analysis, we set $P_L = 0.0033$ and simulated four scenarios with different effect size (magnitude of the copy number change) of 50%, 80%, 120% and 150%. In each scenario, 1 million samples are drawn from each distribution. Then both Jensen-Shannon divergence and Difference-Score are calculated with varying local window size.

Figure 3 shows the estimated power with different window size. For both scores, power increased rapidly at the beginning and didn't gain much once the window size reached 50K. This is in line with Xie's suggestion.



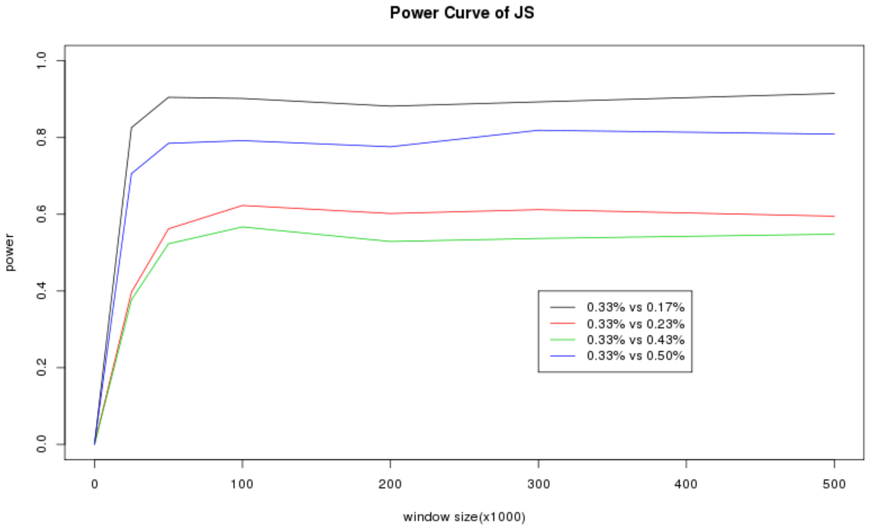


Figure 3. The power test of window size for different read coverage settings

For comparing the performance among different segmentation algorithms, we use the Kullback-Leibler (KL) divergence and the mean Euclidean Error (EE) to measure the accuracy of segmentation esitimation. The Kullback-Leibler divergence can measure the difference between two distributions. In our study, we need to measure the distance between two Poisson ditributions. The Kullback–Leibler divergence of Pois($\hat{\lambda}$) from Pois(λ) is giving by

$$D_{KL}(\lambda||\hat{\lambda}) = \hat{\lambda} - \lambda + \lambda * log\frac{\lambda}{\hat{\lambda}}$$

The Euclidean Error is defined as

$$EE(\lambda,\hat{\lambda}) = \sqrt{\sum_{i}^{1}(\lambda_i - \widehat{\lambda_i})^2}$$

Besides comparing the estimation accuracy, we also compared their capability of identifying the breaking point. In the results tables, we showed the count of false detected change-points and the count of true detected change-points.

Except the comparison of estimation of segmentation, we think the efficiency of the calculation is also important. To compare the algorithm efficiency, we recorded the system runtime of all simulation samples. Here, we present the average computational time of 1000 samples for each algorithm setting. Therefore, the results are based on $1000 * n$ samples, where n is the number of scenario for each simulation study.

4.2 No Change-point Simulation

The first simulation study includes six synthetic chromosome scenarios with different reads coverage setting (0.002, 0.004, 0.008, 0.016, 0.032 and 0.064). For each scenario, 1000 independent samples of size $s = 5{,}000{,}000$ were generated to compare the performance of CBS, SCP, CUSUM and BCP. For each sample, there is no change-point exists. The whole data was randomly drawn from Bernoulli distribution with parameter P. The settings of P are listed in the table 1.

P
5million reads/ the length of genome ≈ 0.002
10million reads/ the length of genome ≈ 0.004
20million reads/ the length of genome ≈ 0.008

40million reads/ the length of genome ≈ 0.016
80million reads/ the length of genome ≈ 0.032
160million reads/ the length of genome ≈ 0.064

Table 1. No Change-point Simulation Settings

Firstly, we try to compare the segmentation results of BCP model, CBS model as well as SCP model when there is no change-point exists in data. Since the true P is known in the simulation, we first compare the accuracy of the empirical estimate of expected read counts. Two measurements results, the Kullback-Leibler (KL) divergence and the mean Euclidean Error (EE), are shown in table 2.

		KL	EE
P=0.002	BCP	0.0099 (0.0051)	1.1643 (0.3206)
	CBS	0.0073 (0.0048)	0.8863 (0.3162)
	SCP	0.0053 (0.0032)	0.7286 (0.2938)
P=0.004	BCP	0.0145	2.0464

		(0.0063)	(0.5107)
	CBS	0.0067 (0.0048)	1.1927 (0.4170)
	SCP	0.0052 (0.0028)	1.1046 (0.3743)
P=0.008	BCP	0.0248 (0.0084)	4.0089 (0.7747)
	CBS	0.0063 (0.0056)	1.6804 (0.5602)
	SCP	0.0050 (0.0028)	1.5837 (0.5000)
P=0.016	BCP	0.0443 (0.0117)	7.9141 (1.1286)
	CBS	0.0061 (0.0049)	2.3299 (0.7544)
	SCP	0.0049 (0.0019)	2.2072 (0.5916)
P=0.032	BCP	0.0835 (0.0163)	15.8090 (1.6001)

	CBS	0.0056 (0.0041)	3.1994 (1.0379)
	SCP	0.0047 (0.0019)	3.0022 (1.0032)
P=0.064	BCP	0.1637 (0.0219)	31.8133 (2.1765)
	CBS	0.0053 (0.0042)	4.4075 (1.4345)
	SCP	0.0040 (0.0013)	4.2705 (1.2748)

Table 2. Comparison between BCP, CBS and SCP on simulation data

To compare their capabilities of identifying the change-point, we calculated the number of points which had been falsely detected by BCP, CBS, SCP and CUSUM (Table 3). In addition, we also measured the average running time from 5000 samples between among these four different methods (Table 4).

	Method	False Changepoint (1000 samples)
P=0.002	BCP	0
	CBS	7
	SCP	175

	CUSUM	584
P=0.004	BCP	0
	CBS	18
	SCP	199
	CUSUM	550
P=0.008	BCP	0
	CBS	19
	SCP	214
	CUSUM	590
P=0.016	BCP	0
	CBS	18
	SCP	250
	CUSUM	604
P=0.032	BCP	0
	CBS	13
	SCP	261
	CUSUM	622
P=0.064	BCP	0
	CBS	10

	SCP	320
	CUSUM	650

Table 3. Number of false detected change-points by BCP, CBS, SCP and CUSUM on simulation data

	Setting	Runtime	Average
BCP	P=0.002	0.018s	0.017s
	P=0.004	0.018s	
	P=0.008	0.016s	
	P=0.016	0.016s	
	P=0.032	0.019s	
	P=0.064	0.015s	
CBS	P=0.002	0.230s	0.255s
	P=0.004	0.252s	
	P=0.008	0.235s	
	P=0.016	0.267s	
	P=0.032	0.274s	
	P=0.064	0.273s	

SCP	P=0.002	0.020s	0.018s
	P=0.004	0.020s	
	P=0.008	0.020s	
	P=0.016	0.019s	
	P=0.032	0.016s	
	P=0.064	0.014s	
CUSUM	P=0.002	0.020s	0.018s
	P=0.004	0.020s	
	P=0.008	0.020s	
	P=0.016	0.019s	
	P=0.032	0.016s	
	P=0.064	0.014s	

Table 4. Running time of BCP, CBS, SCP and CUSUM on simulation data

Table 2 gives the Monte Carlo estimates of $n^{-1}\sum_{t=1}^{n} D_{KL}(\lambda_t||\hat{\lambda}_t)$ and $n^{-1}\sum_{t=1}^{n} EE(\lambda_t||\hat{\lambda}_t)$ and their standard errors (in parentheses) for each simulation scenario. The result demonstrates that the estimates of the true copy number from BCP model are more accurate than from other three models. Moreover, BCP didn't detect any change-point comparing that CBS detected even more than 10 change-points in 1000 samples, SCP detected 200 times number of change-point and CUSUM detected 600 times.

4.3 Single Change-point Simulation

The second simulation study covered ten synthetic chromosome scenarios with different reads coverage (0.002 and 0.004) and different changing amplitudes (x1.5, x2, x4, x10 and x20). For each scenario, 1000 independent samples of size $s = 5{,}000{,}000$ were generated to evaluate the performance of four methods, BCP, CBS, SCP and CUSUM. Here, the simulation focuses on the case of one change-point. But it can be easily generalized to case of multiple change-points. For each sample, the change point was set as the middle point. Then the first half were randomly drawn from Bernoulli distribution with parameter P_1 and the second half were randomly generated from Bernoulli distribution with parameter P_2. The settings of P_1 and P_2 are listed in the table 5.

P_1	P_2
10million reads/ the length of genome ≈ 0.004	0.006,0.008,0.016,0.04,0.08
5million reads/ the length of genome ≈ 0.002	0.003,0.004,0.008,0.02,0.04

Table 5. One Change-point Simulation Settings

After the data was generated, we first ran the step 1, BCP model as well as the other three popular methods, CBS, SCP and CUSUM model, to evaluate the segmentation results. The comparison contains two parts. The first part is comparing the accuracy of the empirical estimate of expected read counts. We use the Kullback-Leibler (KL) divergence and the mean Euclidean Error (EE) to measure the estimate's accuracy. In biological research , the more important thing is the method's ability to detect and identify all the CNVs. Therefore, in the second part, we also compared their capability of identifying the breaking point. True Positive Rate (TPR) can be used to measure the percentage of change points detected at the correct location by the model in all 1000 samples (Table 8). TPR is defined as:

$$TPR = \frac{\#\ of\ detected\ change\ points\ at\ correct\ position}{total\ \#\ of\ change\ points}$$

		KL	EE
P_1=0.002,P_2=0.003	BCP	0.5561 (0.2266)	8.0290 (1.6822)
	CBS	0.8751 (0.2686)	9.3793 (1.3337)
	SCP	0.0053 (0.0032)	0.7286 (0.2938)
P_1=0.002,P_2=0.004	BCP	1.9343	15.9341

		(0.5027)	(2.0137)
	CBS	2.4848 (0.3643)	17.2997 (0.3695)
	SCP	0.0052 (0.0028)	1.1046 (0.3743)
P_1=0.002,P_2=0.008	BCP	9.6499 (0.7399)	50.5418 (1.1198)
	CBS	10.3157 (0.2121)	50.4357 (0.2469)
	SCP	0.0050 (0.0028)	1.5837 (0.5000)
P_1=0.002,P_2=0.02	BCP	36.3709 (0.3340)	150.6906 (0.2408)
	CBS	38.6002 (0.3652)	150.4173 (0.2365)
	SCP	0.0049 (0.0019)	2.2072 (0.5916)
P_1=0.002,P_2=0.04	BCP	82.4686 (0.5054)	317.3518 (0.2384)

	CBS	87.1415 (0.5482)	317.0760 (0.2065)
	SCP	0.0047 (0.0019)	3.0022 (1.0032)

Table 6. Comparison among BCP, CBS and SCP on simulation data of P_1=0.002

		KL	EE
P_1=0.004,P_2=0.006	BCP	1.1474 (0.3567)	16.4897 (2.4729)
	CBS	1.6008 (0.2176)	17.6707 (0.9347)
	SCP	0.0040 (0.0013)	4.2705 (1.2748)
P_1=0.004,P_2=0.008	BCP	3.9643 (0.8396)	33.0266 (3.1815)
	CBS	4.7621 (0.5991)	33.9787 (1.0323)
	SCP	0.0040	4.2705

		(0.0013)	(1.2748)
P_1=0.004,P_2=0.016	BCP	19.1872 (0.6595)	101.3197 (0.3621)
	CBS	20.7010 (0.2804)	100.6187 (0.3919)
	SCP	0.0040 (0.0013)	4.2705 (1.2748)
P_1=0.004,P_2=0.04	BCP	72.7272 (0.4755)	301.3126 (0.3673)
	CBS	77.2537 (0.5180)	300.5813 (0.2779)
	SCP	0.0040 (0.0013)	4.2705 (1.2748)
P_1=0.004,P_2=0.08	BCP	164.8824 (0.6932)	634.6556 (0.3736)
	CBS	174.3066 (0.7503)	633.9297 (0.3512)
	SCP	0.0040 (0.0013)	4.2705 (1.2748)

Table 7. Comparison among BCP, CBS and SCP on simulation data of P_1=0.004

		TPR
P_1=0.002,P_2=0.003	BCP	97.00%
	CBS	48.95%
	SCP	45.50%
	CUSUM	40.35%
P_1=0.002,P_2=0.004	BCP	83.60%
	CBS	50.00%
	SCP	50.00%
	CUSUM	44.70%
P_1=0.002,P_2=0.008	BCP	51.85%
	CBS	50.00%
	SCP	50.00%
	CUSUM	50.00%
P_1=0.002,P_2=0.02	BCP	50.00%
	CBS	50.00%
	SCP	50.00%
	CUSUM	50.00%

P_1=0.002,P_2=0.04	BCP	50.00%
	CBS	50.00%
	SCP	50.00%
	CUSUM	50.00%
P_1=0.004,P_2=0.006	BCP	92.40%
	CBS	49.90%
	SCP	43.00%
	CUSUM	38.40%
P_1=0.004,P_2=0.008	BCP	72.20%
	CBS	50.15%
	SCP	50.00%
	CUSUM	48.70%
P_1=0.004,P_2=0.016	BCP	51.85%
	CBS	50.00%
	SCP	50.00%
	CUSUM	50.00%
P_1=0.004,P_2=0.04	BCP	50.00%
	CBS	50.00%
	SCP	50.00%

	CUSUM	50.00%
P_1=0.004,P_2=0.08	BCP	50.00%
	CBS	50.00%
	SCP	50.00%
	CUSUM	50.00%

Table 8. Comparison of True Positive Rate among BCP, CBS, SCP and CUSUM on simulation data

		Runtime	Average
BCP	P_1=0.002	0.024s	0.024s
	P_1=0.004	0.023s	
CBS	P_1=0.002	2.326s	2.326s
	P_1=0.004	2.326s	
SCP	P_1=0.002	0.018s	0.018s
	P_1=0.004	0.017s	
CUSUM	P_1=0.002	0.028s	0.029s
	P_1=0.004	0.031s	

Table 9. Runtime among BCP, CBS, SCP and CUSUM on simulation data

As in no change-point simulation study, table 6 and 7 give the Monte Carlo estimates of $n^{-1}\sum_{t=1}^{n} D_{KL}(\lambda_t||\hat{\lambda}_t)$ and $n^{-1}\sum_{t=1}^{n} EE(\lambda_t||\hat{\lambda}_t)$ and their standard errors (in parentheses) for each simulation scenario. BCP model seems more sensitive than other methods in small changes detection. For cases with small amplitude changes (X1.5 and X2), BCP model has smaller D_{KL} and EE than other models. It demonstrates that the estimate of the true copy number from BCP model is more accurate than other models. For large amplitude situations, there is no significant different among these three methods in terms of the estimation accuracy. In table 8, the results of TPR suggest that BCP does a better job in identifying the correct segments for change-points with small amplitude changes. But with large amplitude changes, all methods only can reveal around half of the change-points. It becomes the copy number of segment which contains the real change-point will more close to the amplified segment side.

Furthemore, no matter in genetic study, financial study or other area's study, the computational efficiency is an important evaluation parameter of an algorithm. Since the large number of high-throughput data has been created every day, only the method which has high computational efficiency can be applied in real product. In table 9, we also evaluated the efficiency of all algorithms. The average runtime is based on 10000 simulations for each algorithm. The BCP model with BCMIX implementation outperforms CBS 100 times in computational time without loss of accuracy. Comparing with SCP, BCP has the same level of computational time, but with much fewer false positive detected change-points.

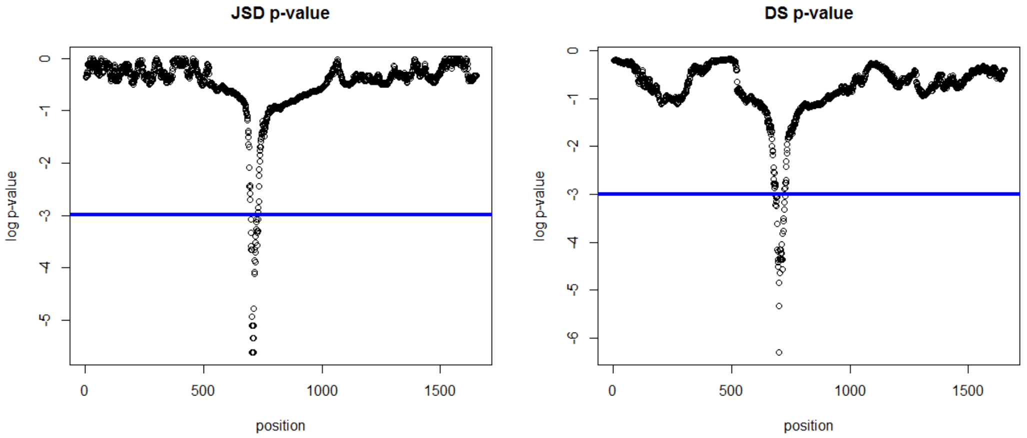


Figure 3. p-value of JSD and DS

After comparing the performance among segmentation methods, we also compare different methods' ability of detecting change-point. Except three segmentation methods, two similarity scores and CUSUM will also be mentioned in change-point detection comparison. We calculated both JSD and DS scores for all positions within the adjacent segments where the segment boundary had been detected as a change-point by BCP model in the previous step. The empirical p-value is calculated by using permutation approach. Then, p-value is used to identify the change-point location at a confidence level of 0.05.

To evaluate the performance among new two-step method and other three popular methods, we calculated the absolute error between the detected change-point and the synthetic change-point to measure the accuracy. In Table 10, we present the absolute error of the results from CBS, SCP, CUSUM and our complete two-step approach, BCP + JSD and BCP + DS. For CBS and SCP, the segment boundary is used as the detected change-point. If the algorithm gives more than one change-point, we choose the one which is the most nearest one to the real change-point position.

As shown in Table 10, our new two-step method has higher accuracy in detecting the change point location compared to segmentation methods which is limited by the segmentation bin boundary positions. Also, the absolute error of both JSD and DS scores are reduced as the coverage of reads. In Figure 4, both scores tends to yield smaller error with High coverage setting $(P_1 = 0.004)$ compared with low coverage setting $(P_1 = 0.002)$. Compared between distribution-based JSD score and knowledge-based DS score, BCP + JSD outperformed BCP + DS with smaller absolute error in the simulation data. One possible explanation is that in simulation setting, there is no comparative normal cell information for DS score. It may cause DS score to be worse. Also, the randomly generated sample may favor the distribution-based score.

One may argue that such comparison may not be fair given CBS, SCP and CUSUM are segmentation-based method but ours is not. We are aware of this. However, to our best knowledge, there is no CNV approach proposed on the raw position level in the literature yet. CBS is the most commonly used and widely adopted approach in genetic copy number analysis area. SCP is the most recent method which applied the similar idea with BCP. CUSUM is the most famous and widely used approach in change-point study. We felt it is worth of pointing out the improvement on the accuracy by taking the extra step to analyze on the raw position level.

	BCP+JSD	BCP+DS	CBS	SCP	CUSUM
P_1=0.002,P_2=0.003	11855 (600)	11977 (600)	26099 (243)	25099 (243)	25099 (243)

P_1=0.002,P_2=0.004	3228 (146)	4013 (208)	24999 (0)	24999 (0)	24999 (0)
P_1=0.002,P_2=0.008	554 (24)	1088 (53)	24999 (0)	24999 (0)	24999 (0)
P_1=0.002,P_2=0.02	133 (5)	634 (33)	24999 (0)	24999 (0)	24999 (0)
P_1=0.002,P_2=0.04	52 (2)	513 (28)	24999 (0)	24999 (0)	24999 (0)
P_1=0.004,P_2=0.006	5501 (270)	6047 (311)	25099 (71)	24999 (51)	24999 (83)
P_1=0.004,P_2=0.008	1473 (61)	1830 (82)	24999 (0)	24999 (0)	24999 (0)
P_1=0.004,P_2=0.016	273 (12)	571 (30)	24999 (0)	24999 (0)	24999 (0)
P_1=0.004,P_2=0.04	61 (2)	343 (19)	24999 (0)	24999 (0)	24999 (0)
P_1=0.004,P_2=0.08	24 (0)	293 (16)	24999 (0)	24999 (0)	24999 (0)

Table 10. Absolute Distance Error between the detected change point and the true change point position

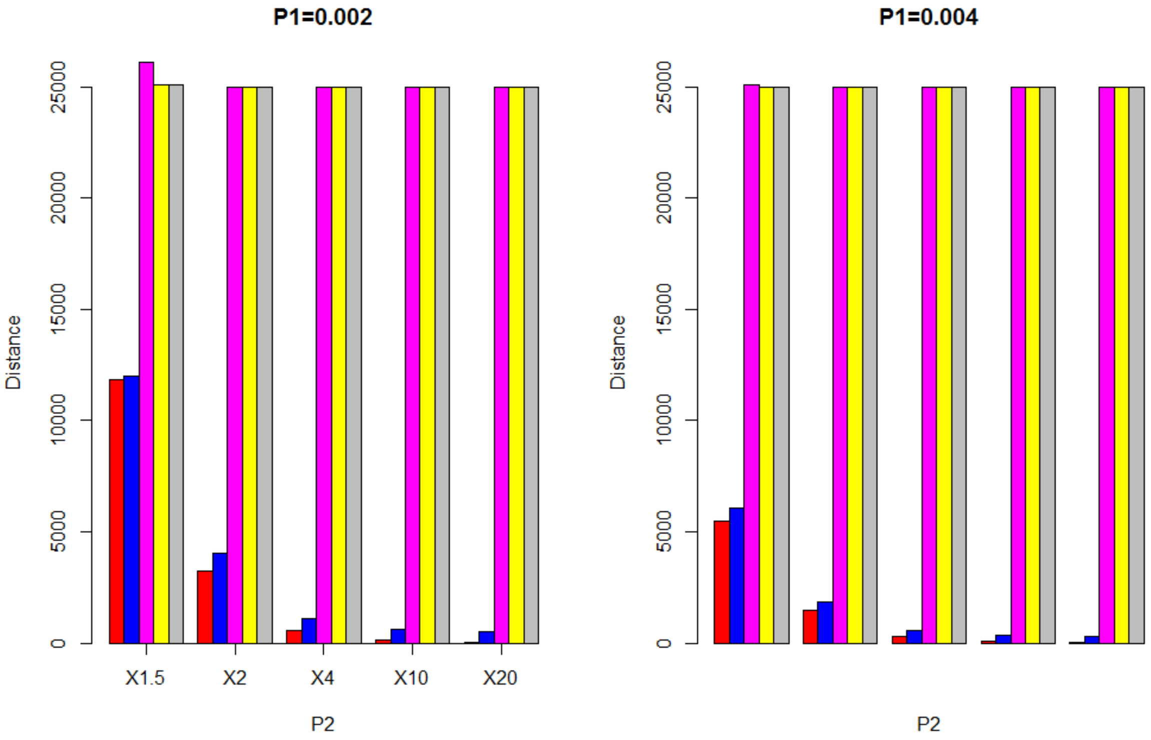


Figure 4. Absolute Distance Error of One Change-point Simulation (Red-BCP+JSD, Blue-BCP+DS, Pink-CBS, Yellow-SCP, Grey-CUSUM)

4.4 Multiple Change-points Simulation

After no change-point simulation study and single change-point simulation study, the last simulation study includes ten synthetic scenarios with different reads coverage and different change amplitudes (x1.5, x2, x4, x10 and x20). As the other simulation study, each scenario has 1000 independent samples. Each sample includes $s = 5{,}000{,}000$ positions. In this simulation study, we tried to simulate cases with two change-points. But it can be easily generalized to cases with multiple change-points. For each sample setting, here are two different distances between two change-points, $500{,}000$ and $750{,}000$. For length $500{,}000$, the positions of change-point are $2{,}250{,}000$ and $2{,}750{,}000$. For length $750{,}000$, the positions of change-point are $2{,}125{,}000$ and $2{,}875{,}000$. Then the first third and the last third were randomly drawn from Bernoulli distribution with parameter P_1 and the second third were randomly generated from Bernoulli distribution with parameter P_2. The settings of P_1 and P_2 are the same with one change-point simulation study (Table 5).

Following the same step with the previous simulation study, BCP, CBS as well as SCP have been compared with their segmentation results. Since we know the true values of $P_1\ and\ P_2$ in the simulation, the true read counts in each segment (window) has been calculated. Then, Kullback-Leibler (KL) divergence and the mean Euclidean Error (EE) is used to measure the accuracy between calculated results and true read counts. True Positive Rate (TPR) is the percentage of change points detected at the correct location by the model in all 1000 samples. We use it to compare the ability of identifying change-points among new two-step mothed, CBS, SCP and CUSUM.

Comparing with one change-point simulation, the results of KL, EE, TPR and runtime of each method doesn't change a lot. So I will not list these result tables here for two change-point simulations.

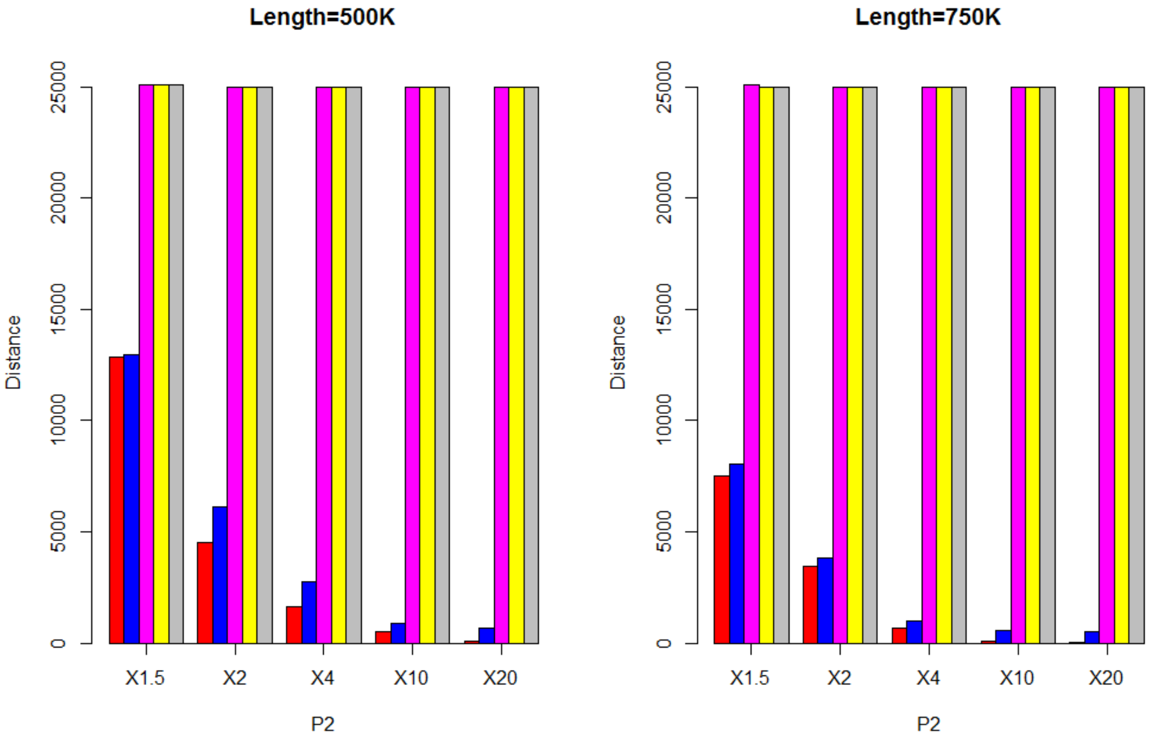


Figure 5. Absolute Distance Error of Two Change-points Simulation (Red-BCP+JSD, Blue-BCP+DS, Pink-CBS, Yellow-SCP, Grey-CUSUM)

As shown in Figure 5, comparing with other three methods, our new two-step methods have higher accuracy in detecting the change-point location. The reason is traditional method is limited by the segmentation bin boundary positions. In figure 5, the absolute error of both JSD and DS scores reduced with the increase of the amplification. The reason is obvious that the larger amplification or deletion is easier to be detected. Also, both scores tend to yield smaller error with larger distance between two change-points $(length = 750K)$ compared with smaller distance $(length = 500K)$. That's why the short change-point segment in copy number data is not easy being discovered.

Chapter 5

Conclusion

In this dissertation, we developed a novel two-step approach for detecting CNVs from single-cell sequencing data. One key advantage of our method is that it can accurately locate change-point positions beyond the locations of bin boundaries. Based on simulation studies, our numerical results show, comparing with CBS, SCP and CUSUM, BCP gives a more accurate estimation of copy number and shows more sensitive with small variations. The BCMIX implementation makes BCP significantly efficient than CBS in computation time. Comparing with SCP and CUSUM, CBS has much lower false positive rate and higher true positive rate especially for detecting small amplitudes. Furthermore, to overcome the limitation of bin boundaries, we proposed two similarity scores, distribution-based JSD score and knowledge based DS score, to find out the accurate position of breaking point. Based on the results from the first step, two similarity scores were calculated for certain areas. Simulation results show the new method is very robust in detecting the true breaking point positions. Another conclusion we got from simulation studies is the accuracy of both scores would increase with the coverage of reads. It would be very interesting to apply this new two-step algorithm on real chromosome data.

[1] Watson JD, Crick FH. The structure of DNA. Cold Spring Harbor symposia on quantitative biology 1953;18:123-31.

[2] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology 1975;94:441-8.

[3] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 1977;74:5463-7.

[4] Griffin HG, Griffin AM. DNA sequencing. Recent innovations and future trends. Applied biochemistry and biotechnology 1993;38:147-59.

[5] Studier FW. A strategy for high-volume sequencing of cosmid DNAs: random and directed priming with a library of oligonucleotides. Proceedings of the National Academy of Sciences of the United States of America 1989;86:6917-21.

[6] Martin-Gallardo A, McCombie WR, Gocayne JD, FitzGerald MG, Wallace S, Lee BM, et al. Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. Nature genetics 1992;1:34-9.

[7] Voss H, Wiemann S, Grothues D, Sensen C, Zimmermann J, Schwager C, et al. Automated low-redundancy large-scale DNA sequencing by primer walking. BioTechniques 1993;15:714-21.

[8] International consortium completes human genome project. Pharmacogenomics 2003;4:241.

[9] Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America 1977;74:560-4.

[10] Maxam AM, Gilbert W. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods in enzymology 1980;65:499-560.

[11] Maxam AM. Sequencing the DNA of recombinant chromosomes. Federation proceedings 1980;39:2830-6.

[12] Ansorge W, Rosenthal A, Sproat B, Schwager C, Stegemann J, Voss H. Non-radioactive automated sequencing of oligonucleotides by chemical degradation. Nucleic acids research 1988;16:2203-6.

[13] Rosenthal A, Sproat B, Voss H, Stegemann J, Schwager C, Erfle H, et al. Automated sequencing of fluorescently labelled DNA by chemical degradation. DNA sequence : the journal of DNA sequencing and mapping 1990;1:63-71.

[14] Richterich P. Non-radioactive chemical sequencing of biotin labelled DNA. Nucleic acids research 1989;17:2181-6.

[15] Bartlett JM, Stirling D. A short history of the polymerase chain reaction. Methods in molecular biology 2003;226:3-6.

[16] Wada A, Yamamoto M, Soeda E. Automatic DNA sequencer: computer-programmed microchemical manipulator for the Maxam-Gilbert sequencing method. The Review of scientific instruments 1983;54:1569-72.

[17] Boland EJ, Pillai A, Odom MW, Jagadeeswaran P. Automation of the Maxam-Gilbert chemical sequencing reactions. BioTechniques 1994;16:1088-92, 94-5.

[18] Negri R, Costanzo G, Di Mauro E. A single-reaction method for DNA sequence determination. Analytical biochemistry 1991;197:389-95.

[19] Morgan MJ, Wallace SE. The international human genome project: An overview. Cross-Cultural Biotechnology 2004:15-23.

[20] Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics 2010;11:31-46.

[21] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.

[22] Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences of the United States of America 2003;100:8817-22.

[23] Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, et al. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic acids research 2000;28.

[24] Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic acids research 2006;34.

[25] Baslan T, Hicks J. Single cell sequencing approaches for complex biological systems. Current opinion in genetics & development 2014;26:59-65.

[26] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90-4.

[27] Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, et al. 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. Genome research 2004;14:179-87.

[28] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nature genetics 2004;36:949-51.

[29] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science 2004;305:525-8.

[30] Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome research 2006;16:949-61.

[31] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54.

[32] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434-40.

[33] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nature genetics 2005;37 Suppl:S11-7.

[34] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature genetics 1998;20:207-11.

[35] Langer-Safer PR, Levine M, Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences of the United States of America 1982;79:4381-5.

[36] Gall JG, Pardue ML. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proceedings of the National Academy of Sciences of the United States of America 1969;63:378-83.

[37] Rudkin GT, Stollar BD. High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence. Nature 1977;265:472-3.

[38] Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP. Genome screening by comparative genomic hybridization. Trends in Genetics 1997;13:405-9.

[39] Weiss MM, Kuipers EJ, Meuwissen SGM, van Diest PJ, Meijer GA. Comparative genomic hybridisation as a supportive tool in diagnostic pathology. J Clin Pathol 2003;56:522-7.

[40] Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clinical genetics 2004;66:488-95.

[41] Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 2004;90:132-53.

[42] Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, et al. Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 2005;6:211-26.

[43] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557-72.

[44] Wen CC, Wu YJ, Huang YH, Chen WC, Liu SC, Jiang SS, et al. A Bayes regression approach to array-CGH data. Statistical applications in genetics and molecular biology 2006;5:Article3.

[45] Ren H, Francis W, Boys A, Chueh AC, Wong N, La P, et al. BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints. Human mutation 2005;25:476-82.

[46] Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, et al. Genome-wide copy number analysis of single cells. Nature protocols 2012;7:1024-41.

[47] Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America 2007;104:20007-12.

[48] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nature genetics 2005;37:727-32.

[49] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420-6.

[50] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 2009;6:677-81.

[51] Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome research 2009;19:1270-8.
[52] Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature methods 2009;6:S13-20.
[53] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865-71.
[54] Abyzov A, Gerstein M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics 2011;27:595-603.
[55] Nijkamp JF, van den Broek MA, Geertman JM, Reinders MJ, Daran JM, de Ridder D. De novo detection of copy number variation by co-assembly. Bioinformatics 2012;28:3195-202.
[56] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research 2009;19:1586-92.
[57] Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 2012;28:2711-8.
[58] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.
[59] Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for DNA copy number variants detection. Bioinformatics 2012;28:470-8.
[60] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature 2011;470:59-65.
[61] Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC bioinformatics 2009;10:80.
[62] Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biology 2009;10:R32.
[63] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research 2008;36:e105.
[64] Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in C. elegans. Nature methods 2008;5:183-8.
[65] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53-9.
[66] Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavare S. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics 2010;26:3051-8.

[67] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 2009;6:99-103.

[68] Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PloS one 2011;6:e16327.

[69] Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, et al. Fast computation and applications of genome mappability. PloS one 2012;7:e30377.

[70] Lai TL, Xing H, Zhang N. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. Biostatistics 2008;9:290-307.

[71] Reeves J, Chen J, Wang XLL, Lund R, Lu QQ. A review and comparison of changepoint detection techniques for climate data. J Appl Meteorol Clim 2007;46:900-15.

[72] Guyon X, Yao JF. On the underfitting and overfitting sets of models chosen by order selection criteria. Journal of Multivariate Analysis 1999;70:221-49.

[73] Chen J, Gupta AK. Testing and locating variance changepoints with application to stock prices. J Am Stat Assoc 1997;92:739-47.

[74] Lavielle M. Using penalized contrasts for the change-point problem. Signal Process 2005;85:1501-10.

[75] Birge L, Massart P. Minimal penalties for Gaussian model selection. Probab Theory Rel 2007;138:33-73.

[76] Kander Z, Zacks S. Test Procedures for Possible Changes in Parameters of Statistical Distributions Occurring at Unknown Time Points. Annals of Mathematical Statistics 1966;37:1196-&.

[77] Akaike H. Citation Classic - a New Look at the Statistical-Model Identification. Cc/Eng Tech Appl Sci 1981:22-.

[78] Hannan EJ, Quinn BG. Determination of the Order of an Autoregression. J Roy Stat Soc B Met 1979;41:190-5.

[79] Bartlett MS. A comment on D.V.Lindleys statistical paradox. Biometrika 1957;44:533-4.

[80] Wijsman EM, Yu D. Joint oligogenic segregation and linkage analysis using bayesian Markov chain Monte Carlo methods. Molecular biotechnology 2004;28:205-26.

[81] Page ES. A Test for a Change in a Parameter Occurring at an Unknown Point. Biometrika 1955;42:523-7.

[82] Quandt RE. The Estimation of the Parameters of a Linear-Regression System Obeying 2 Separate Regimes. J Am Stat Assoc 1958;53:873-80.

[83] Quandt RE. Tests of the Hypothesis That a Linear-Regression System Obeys 2 Separate Regimes. J Am Stat Assoc 1960;55:324-30.

[84] Chernoff H, Zacks S. Estimating Current Mean of Normal-Distribution Which Is Subjected to Changes in Time. Annals of Mathematical Statistics 1964;35:999-&.

[85] Gardner LA. On Detecting Changes in Mean of Normal Variates. Annals of Mathematical Statistics 1969;40:116-&.

[86] Macneill IB. Tests for Change of Parameter at Unknown Times and Distributions of Some Related Functionals on Brownian-Motion. Ann Stat 1974;2:950-62.
[87] Scott AJ, Knott M. Cluster-Analysis Method for Grouping Means in Analysis of Variance. Biometrics 1974;30:507-12.
[88] Sen A, Srivastava MS. Tests for Detecting Change in Mean When Variance Is Unknown. Ann I Stat Math 1975;27:479-86.
[89] Sen A, Srivastava MS. Tests for Detecting Change in Mean. Ann Stat 1975;3:98-108.
[90] Vostrikova LI. Detection of the Disorder in Multidimensional Random-Processes. Dokl Akad Nauk Sssr+ 1981;259:270-4.
[91] Braun JV, Muller HG. Statistical methods for DNA sequence segmentation. Stat Sci 1998;13:142-62.
[92] Braun JV, Braun RK, Muller HG. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. Biometrika 2000;87:301-14.
[93] Auger IE, Lawrence CE. Algorithms for the optimal identification of segment neighborhoods. Bulletin of mathematical biology 1989;51:39-54.
[94] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995;82:711-32.
[95] Fearnhead P. Exact and efficient Bayesian inference for multiple changepoint problems. Stat Comput 2006;16:203-13.
[96] Lavielle M, Lebarbier E. An application of MCMC methods for the multiple change-points problem. Signal Process 2001;81:39-53.
[97] Yao YC. Estimation of a Noisy Discrete-Time Step Function - Bayes and Empirical Bayes Approaches. Ann Stat 1984;12:1434-47.
[98] Barry D, Hartigan JA. Product Partition Models for Change Point Problems. Ann Stat 1992;20:260-79.
[99] Liu JS, Lawrence CE. Bayesian inference on biopolymer models. Bioinformatics 1999;15:38-52.
[100] Fearnhead P. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. Stat Comput 2008;18:151-71.
[101] Fearnhead P, Liu Z. On-line inference for multiple changepoint problems. J Roy Stat Soc B 2007;69:589-605.
[102] SM I. sde: Simulation and Inference for Stochastic Differential Equations. 2009.
[103] Erdman C EJ. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. Journal of Statistical Software 2007;23:1-13.
[104] Muggeo VM, Adelfio G. Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics 2011;27:161-6.
[105] VMR M. cumSeg: Change Point Detection in Genomic Sequences. 2012.
[106] Seshan VE OA. DNAcopy: DNA Copy Number Data Analysis. 2008.
[107] Zeileis A LF, Hornik K, Kleiber C. strucchange: An R Package for Testing for Structural Change in Linear Regression Models. Journal of Statistical Software 2002;7:1-38.

[108] GJ R. cpm: Sequential Parametric and Nonparametric Change Detection. 2013.

[109] Killick R, Eckley IA. changepoint: An R Package for Changepoint Analysis. Journal of Statistical Software 2014;58:1-19.

[110] Lai TL, Xing H. A simple bayesian approach to multiple change-points. Statistica Sinica 2011;21:539-69.

[111] Xing H, Mo Y, Liao W, Zhang MQ. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS computational biology 2012;8:e1002613.

[112] Page ES. Continuous Inspection Schemes. Biometrika 1954;41:100-&.

[113] KUllback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics 1951;22:79-86.