BAYESIAN CHANGE POINT ANALYSIS of COPY NUMBER VARIANTS USING HUMAN NEXT GENERATION SEQUENCING DATA

A DISSERTATION IN
Mathematics
and
Molecular Biology and Biolochemistry

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
Jianfeng Meng

MS, University of Colorda Health Science Center, 2007
MS, Indiana University, 2001
MPH, Sichuan University HuaXi School of Public Health, 1991
MD, Sichuan University HaiXi Health Science Center, 1988

Kansas City, Missouri
2014

BAYESIAN CHANGE POINT ANALYSIS of COPY NUMBER VARIANTS USING HUMAN NEXT GENERATION SEQUENCING DATA

Jianfeng Meng, Candidate for the Doctor of Philosophy Degree

University of Missouri-Kansas City, 2014

ABSTRACT

Read count analysis is the principal strategy implemented in detection of copy number variants using human next generation sequencing (NGS) data. Read count data from NGS has been demonstrated to follow non homogeneous Poisson distributions. The current change point analysis methods for detection of copy number variants are based on normal distribution assumption and used ordinary normal approximation in their algorithms. To improve sensitivity and reduce false positive rate for detection of copy number variants, we developed three models: one Bayesian Anscombe normal approximation model for single genome, one Bayesian Poisson model for single genome, and a Bayesian Anscome normal approximation model for paired genome. The Bayesian statistics have been optimized for detection of change points and copy numbers at single and multiple change points through Monte Carlo simulations. Three R packages based on these models have been built up to simulate Poisson distribution data, estimate and display copy number variants in table and graphics. The high sensitivity and specificity of these models have been demonstrated in simulated read count data with known Poisson distribution and in human NGS read count data as well in comparison to other popular packages.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of Graduate Studies have examined a dissertation titled "Bayesian Change Point Analysis of Copy Number Variants Using Human Next Generation Sequencing Data", presented by Jianfeng Meng, candidate for the DOCTOR OF PHILOSOPHY degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Jie Chen, Ph.D., Committee Chair

Department of Mathematics and Statistics

Shui Qing Ye, Ph.D.,

Department of Biomedical & Health Informatics

Gerald J Wyckoff, P.h.D,

Division of Molecular Biology and Biochemistry

Majid Bani-Yaghoub, P.h.D,

Department of Mathematics and Statistics

Xiao-qiang Yu, P.h.D,

Division of Cell Biology and Biophysics

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ACKNOWLEDGEMENTS

I am very grateful to my mentor Dr. Jie Chen for her continuous support, patience and excellent guidance through my PhD study. The comments and advices from my supervisory committee members are really valuable for the preparation of this dissertation and for my career development. I want to thank all of these who have provide helps during my study. The support from my lovely family are always strong and consistent. Without their support, I may not continue the efforts towards this degree. Thanks to the people who made this dissertation possible, especially put together a nice LATEX template for me to use.

Chapter 1

BACKGROUND

Genetic variations refer to the differences in the sequence of genetic materials (DNA) in human genome among individuals. Genetic variations take many forms ranging from large chromosomal anomalies (segmental aneuploidy) to single nucleotide variant (Abecasis et al, 2012). Among them, structural variants (SV) involve quantitative (e.g. copy number variants, indels) or positional (translocations) or orientation (inversion) alterations of DNA sequence with multiple or segmental nucleotides. Copy number variation refers to copy number differences in a segment of DNA among individuals. Copy number variation is evidenced from either failure to detect a certain DNA segment (deletion) or identification of multiple copies of a DNA sequence (insertion) per haploid genome in comparison with reference human genome sequence. A copy number variant (CNV) was usually defined as the duplication, insertion, or deletion of DNA segment larger than 1 kb. With the advent of next-generating sequencing and new generation arrays, several studies used a minimal length of 500bp to define the size of DNA segment for copy number variation (Valsesia et al, 2013).

CNVs have been found genome wide in humans and other species. CNVs could occur in both normal and disease population. The comparison of CNVs among different human populations or across species has been explored to understand the genetic diversity with the consequence of genetic evolution and disease susceptibility (Conrad

and Hurles, 2007). Some CNVs were likely to be associated with common diseases (Craddock et al, 2010).

1.1 Detection of CNVs

Gross CNVs were initially detected by karyotyping in the early days of cytogenetic, where the number and structure of chromosomes in the sample of cells were examined. In past decades, the evolution of CNV detection techniques from Fluorescent In Situ Hybridization (FISH) in early 1980 to new technology made it possible to be more reliable, high throughput and fine resolution in the detection of CNVs (Langer-Safter et al, 1982). The current widely used techniques include comparative genome hybridization (CGH), single nucleotide polymorphism (SNP) array and next generation sequencing based methods.

CGH is developed based on the assumption that CNVs can be detected by the relative ratio of the test sample DNA amounts to the reference genome DNA. The fluorescence labeled test and reference DNA in ratio 1:1 were simultaneously hybridized to a normal metaphase spread of chromosomes or a DNA microarray (aCGH), and bind competitively at the locus of the origin. If there are more copies of test DNA than the reference DNA, a relatively higher intensity of the test sample color in a specific chromosomal region will be captured through fluorescence microscope and computer software. Conversely, a relative lower intensity of the test sample color in comparison to the reference DNA color will indicate the loss of DNA sequence copies in the test samples (Ren et al, 2005; Urban et al, 2006). The location of copy number will be determined by known DNA sequences embedded on the DNA microarray.

SNP array was originally used to detect single nucleotide polymorphisms (SNPs) and to genotype human DNA in populations. The common procedure for SNP array is that the test and reference sample DNA are hybridized to an DNA array containing hundreds of thousands even millions of unique nucleotide probe sequences with

fluorescent which is designed to bind complementarily to a target DNA subsequence. Various computation algorithms have been implemented in converting the raw signals into inferences about the presence or absence of each of the two alleles in different platforms. Meanwhile, the probe signal intensity level either from summarization of multiple SNP probe intensity or from singleton non-polymorphic probes incorporated with genotype information is converted into single measure of row copy number by comparing to that from a panel of reference samples (LaFramboise, 2009). Since SNP array is relatively cheap to genotype individuals for previously known variants, it is popularly used in population genetic studies to identify the association of SNPs and copy number variations with disease traits (Valsesia et al, 2012). Dozens of statistical methods have been developed to infer chromosomal segments of locally constant copy number from the noisy raw copy number measurements. The algorithms range from mainly Hidden Markov Model (HMM) and circular binary segmentation to mixture models, maximum likelihood, regression, wavelet and genetic models (Lai et al, 2005; Wang et al, 2007; Colella et al, 2007; Chen et al, 2009). However, considerable variations in sensitivity and specificity for identifying CNV boundary and determining the number of CNVs called exist among programs. The CNV length, frequency of CNV and features of CNV e.g. deletions or duplication in addition to SNP array platform with density of selected SNPs may contribute to the accuracy of the detection across multiple programs (Zhang et al, 2011).

DNA sequencing is a process of determining the precise order of nucleotides (A, G, C and T) in a strand of DNA. Early DNA sequencing techniques is based on inferring DNA sequences from differentiating labeled DNA fragments in DNA synthesis process either through radiolabelled DNA and chemical cleavage of specific base (Maxam and Gilbert, 1977) or through fluorescence labeled chain-termination inhibitor of DNA polymerase. The high demand for low cost and fast sequencing has driven the development of high throughput sequencing or next generation sequencing (NGS) in

which parallelized sequencing process produces thousands or millions of sequencing reads simultaneously. Several platforms have been developed since the mid 1990s. Among them, 454 Pyrosequencing by Roche Diagnostic, Applied SOLiD technology and Illumina (Solexa) sequencing are the most popularly used. As an example to illustrate the next generation sequencing technique, Illumina Solexa sequencing system is based on reversible dye terminators technology. In this method, DNA fragment with primers are attached to solid surfaces and amplified by PCR so that DNA clusters of clonal DNA are formed. To determine the sequence, four types of reversible terminator nucleotides are added and non-incorporating nucleotides are washed away. After a camera takes the fluorescent image, the dye and 3 blockers are chemically removed so that the next cycling can begin. The DNA chains are extended one nucleotide at one time and an image is acquired sequentially until a DNA sequence is completed. By combined with massively parallelized sequencing technology, very large arrays of DNA colonies can be sequenced concurrently (Margulies et al, 2005). Up to billions of nucleotide reads per run within 1 to 10 days can be generated using Illumina sequencing system. In deep sequencing, total length of DNA sequences generated is many times larger than the length of the sequence under the study due to the depth of the process. Depth of coverage in massively paralleling sequencing refers to the number of times a nucleotide is read during the sequencing process. More NGS methods are under development. The extension of single read sequencing (also called single end sequencing) to paired end reads or mate pairs (also called paired end sequencing) offers additional information to detect copy number variation and other structural variations such as inversion and translocation (Raphael et al, 2012; Medvedev et al, 2009). The comparison of NGS methods can be found in the excellent reviews by Quail et al (2012) and Liu et al (2012). Since significant reduction in cost has been achieved, the application of NGS has been used to sequence whole genome sequence of thousand human individuals.

Compared to array based techniques, copy number analysis by NGS allows the breakpoints of copy number regions to be determined more precisely because it does not rely on predefined probes (Chiang and McCarroll, 2009). It also allows detection of smaller copy number variations by simply increasing the depth of sequencing (Chiang and McCarroll, 2009). Estimation of integer copy numbers from NGS data is more accurate at high copy counts, since depths of coverage scale linearly with copy number and it does not suffer from hybridization saturation (Alkan et al, 2009). More and new allele specific copy numbers may be estimated from sequencing data, while array-based techniques are restricted to predefined alleles. Allele specific copy numbers are of interest because the functional alleles may be mutants leading to disease development (Stratton et al, 2009; Klambauer et al, 2012). Overall, NGS has demonstrated higher sensitivity, in terms of types and sizes of variants that can be detected.

1.2 Detection of CNVs Using Next Generation Sequencing Data

Strategy The commonly used strategies include read depth analysis, paired end mapping, split read approach, sequence assembly and combination algorithms. Read depth analysis and sequence assembly can be applied on single end sequencing, while paired end mapping and split read approach mainly depend on paired end sequencing (PES) (Alkan et al, 2011; Valsesia et al, 2013; Teo et al, 2012).

Paired end mapping (PEM) requires sequencing both ends (also called paired end reads or mate pairs) of a genomic fragment of known size and then mapping the end sequence pair to a reference sequence (Raphael et al, 2012). Since the distance between paired end reads is expected to fall in fixed range, fragments overlapping structural variant events in a test genome may result in discordant paired end sequences that map to different parts of the reference genome. Discordant paired ends in length or direction indicate respectively possible indels or inversions or locations

(Medvedev et al, 2009). When sequenced ends of the fragment map to the reference at a distance longer than expected, it is indicative of an insertion in the studied genome. Vice versa, when sequenced ends of the fragment map to the reference at a distance shorter than expected, it is indicative of a deletion in the studied genome. As an example, if two ends of a fragment are mapped with a wrong orientation, it could be an indication of an inversion. Several analysis tools based on detection of discordant end-pairs and clustering of end pairs have been developed and used for the estimate of SVs and CNVs (Koboldt et al, 2012). Paired end mapping is more specific when lining back to genome resulting from two paired end reads compared to single end mapping. Precise breakpoints can be determined through paired end mapping. In addition to CNV, other SVs can be measured through it. However, the detection resolution is limited to the distance between pairs, neither large nor very small rearrangements can be detected, with the exception of large deletions (Valsesia et al, 2013).

The split read (SR) approach was used to detect deletion and insertion events when one of paired end reads mapped uniquely onto a reference genome but the other one of paired end reads can't be mapped. It is assumed in the split read approach that the unmapped read occurs because the breakpoints of deletion or insertion are inside the reads in test genome. Therefore the analysis strategy is that the mapped end read is taken as anchor point and the unmapped read can be mapped to reference genome by splitting it into two fragments or three fragments. If two fragments are mapped to the reference genome uniquely, it indicates deletion. If two ends of three fragments are mapped to the reference genome uniquely, it indicates an insertion of middle part of the unmapped end read (Ye et al, 2009). Identification of breakpoints with a pattern growth algorithm was applied in the split read analysis (Ye et al, 2009).

De novo assembly aligns a test genome sequence utilizing high sequencing depth data. Then the sequence is compared to the reference sequence for detection of dele-

tions and insertions. The advantage is that different size of deletions and insertions even smaller than paired end insert size can be detected. However, de novo assembly is very difficult for repeat rich regions and until recently was only possible with high read depth (Iqbal et al, 2012; Simpson and Durbin, 2012).

Read depth also called *read count (RC) analysis* is based on depth of coverage (DOC) in the NGS. It assumed that the sequencing process is uniform across a genome and number of reads mapping to a region is expected to be proportional to number of times the region appear in the DNA sample. A region that has deletion (or duplication) is expected to have less (or more) reads mapping to it (Chiang and McCarroll, 2009; Medvedev et al, 2009).

Read Count Analysis Read count analysis are still the main strategy in analysis of read data for copy number variations from both paired and single end sequencing in literatures. After RC analysis was first used by Campbell (Campbell et al, 2008) and Chiang (Chiang and McCarroll, 2009) to detect CNVs, several RC analysis packages have been developed to look for copy number difference in normal and diseased populations (Yoon et al, 2009; Xie et al, 2009; Ivakno et al, 2010; Kim et al, 2010; Xi et al, 2011). The analysis pipeline implemented in current packages for discovering CNVs is conceptually derived from aCGH data analysis and can be divided into four fundamental steps: read data filtering and sequence alignment; assignment to bin/window; CNV region identification (segmentation); Copy number estimation (Magi et al, 2012). Since the mappability of reads onto the reference genome sequence influences the accuracy of read counts assignment, mappability correction has been often applied. GC content in genome sequence and sequencing error among DNA samples have been demonstrated to influence the accuracy of read generation in next generation sequencing technology. Therefore GC content correction and sample normalization are frequently seen in analysis packages.

Segmentation Segmentation is a critical step for identification of copy number vari-

ation. Taking consideration of NGS read data from single genome, paired genomes or multiple genomes three different strategies have been implemented. Various algorithms have been developed for the segmentation process in each strategy. The packages for single genome rely mostly on read depth data that are sequenced from one single person, and sometimes integrated with read pair and read split information. The packages for paired genomes are based on ratio or comparison between targeted individual sequence data and a reference sequence data. A typical example is the detection of CNVs in cancer patients compared with normal tissue or with a control subject. Population based methods utilize read depth data from multiple samples.

Copy number estimation The copy number calling algorithms are based on segmentation and generate results about gain or loss of copy number or rough copy numbers. The copy number of each event was inferred in RDxplorer (Yoon et al, 2009) by rounding the average normalized read counts in each individual to the nearest integer. The normalized read counts is defined as 2 X (read count)/(mean read count over the genome). The segments identified in ReadDepth by Miller et al (2011) were called gains or losses if their mean value exceeded 1.5 standard deviation from the mean probe value. The CNV calling for duplication or deletion in CNVnator (Abyzov et al, 2009) is based on statistical significance t tests with multiple correction p value in comparison with genomic average RD signal. ERDS by Zhu et al (2012) generates initial copy number inference using continuous paired Hidden Markov Model (HMM) for both deletion and duplication based on expected RD data in nonoverlaping windows and refines the deletion and breakpoints of putative deletions by integrating paired end mapping (PEM) and soft-clipping signatures to filter false positive and confirm weak RD signals.

The copy number change based on read count ratio between test and control samples is considered as gain or loss while segmentation algorithms for significantly changed ratio along the genome position are applied (Chiang and McCarroll, 2009;

Xie et al, 2009; Kim et al, 2010; Ivakno et al, 2010; Xi et al, 2011).

Read count analysis requires that reads from sample DNA are mapped to the reference genome sequence. Since reads from a novel sequence can't be mapped to the reference genome sequence, it is unable to detect a new insertion of DNA sequence through read count analysis. Unlike PEM insertion signatures, RC analysis can't detect other SVs such as inversion or translocation. Furthermore uniquely mapping reads to high repeat rich region is very difficult. These deteriorate the accuracy of RC analysis in amounts and location of CNVs. To overcome the limitation of algorithm only based on DOC, more packages applied combination algorithm with information from DOC, PEM and SR as well as other features of sequence data at population level for more accurate detection of CNVs (Medvedev et al, 2010; Miller et al, 2011; Abyzov et al, 2009; Zhu et al, 2012; Klambauer et al, 2012; Bellos et al, 2009).

Overall, Most of these methods showed less sensitive and high false discovery rate ($\sim$ 20% or more) in read datasets. Those may lead to wrong conclusions for identification of disease variants in genetic association studies. Identifying factors that may impact the accuracy of those methods and developing more sensitive and accurate statistical methods with innovative strategy are crucial to improve CNV detection using next generation sequencing technology. Since the segmentation plays a crucial role in determining the copy number change, an improved algorithm is expected to increase the sensitivity and reduce false positive rate in the detection. The change point analysis is among one of these approaches. In the following section, the detailed algorithm and mathematical deduction with respect to change point analysis are illustrated so that improvements can be specified.

1.3 Change Point Data Analysis

Change Point Analysis The change point problem refers to the identification of changes in a series of events including the number of change points and their

locations. Usually statistical inference about change points has two aspects. The first is to detect if there is any change in the sequence of observed random variables. The second is to estimate the number of changes and their corresponding locations. The detection of CNVs along the human genome is actually a change point issue where read counts change in bins corresponding to copy number change. This has been recognized and applied in CNV detection packages.

Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be a sequence of independent random vectors (variables) with probability distribution functions $F_1, F_2, \cdots, F_n$, respectively. Then in general, the change point problem is to test the following null hypothesis,

$$H_0 : F_1 = F_2 = \cdots = F_n$$

versus the alternative:

$$\mathrm{H}_1 : F_1 = \cdots = F_{k_1} \neq F_{k_1+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots = F_{k_q} \neq F_{k_q+1} \cdots = F_n$$

where $1 < k_1 < k_2 < \cdots < k_q < n$, q is the unknown number of change points and $k_1, k_2, \cdots, k_q$ are the respective unknown positions that have to be estimated. If the distributions $F_1, F_2, \cdots, F_n$ belong to a common parametric family $F(\theta)$, where $\theta \in \mathbf{R^P}$, then the change point problem is to test the null hypothesis about the population parameters $\theta_i, i = 1, \cdots, n$:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n = \theta(unknown) \tag{1.1}$$

versus the alternative:

$$\begin{aligned} \mathrm{H}_1 : \theta_1 = \cdots = \theta_{k_1} \neq \theta_{k_1+1} = \cdots = \theta_{k_2} \neq \theta_{k_2+1} \\ = \cdots = \theta_{k_q} \neq \theta_{k_q+1} \cdots = \theta_n \end{aligned} \tag{1.2}$$

where q and $k_1, k_2, \cdots, k_q$ have to be estimated. These hypotheses together reveal the aspects of change point inference: determining if any change point exists in the process and estimating the number and positions of change point(s).

The frequently used methods for change point inference in the literature are the maximum likelihood ratio test, Bayesian test, nonparametric test, stochastic process, information theoretic approach, and so on. Many of them can be found in a book by Chen and Gupta (2012). Most of the change point studies were concentrated on the case of a single change point in the random sequence. The problem of multiple change points were rarely addressed by many authors. A binary segmentation procedure (BSP) proposed by Vostrikova (1981) has been proved to be consistent and been widely used in detecting multiple change points, and it has the merits of detecting the number of change points and their positions simultaneously and saving a lot of computation time. Briefly in the case we described in equation (1) and (2), several steps are generally involved in the final identification of multiple change points. The first step is to test for no change point versus one change point; that is, test the null hypothesis by equation (1) versus the following alternative.

$$H_1 : \theta_1 = \cdots = \theta_k \neq \theta_{k+1} \cdots = \theta_n \tag{1.3}$$

where k is the location of the single change point at this stage. If H_0 is not rejected, then stop. There is no change point. If H_0 is rejected, then there is a change

point and we go to step 2.

In step 2, the two subsequences before and after the change point found in step 1 are tested separately for a change. This process is repeated until no further subsequences have change points. The collection of change point locations found by above steps is denoted by $\{\hat{k_1}, \hat{k_2}, \cdots, \hat{k_q}\}$, and the estimated total number of change points is then q.

The majority of the models proposed for change point problems in the literatures are the normal models (univariate or multivariate), perhaps due to the fact that the normal model is the most common model in practice. These have been addressed by Chen and Gupta (2012). Meanwhile, the gamma and exponential model, regression model, hazard function model, binomial model, Poisson model as well as well smooth and abrupt model have been reported in the literatures.

Application of Change Point Analysis in CNVs Detection The copy number detection in aCGH experiments is based on the ratio of test sample intensity to the reference sample intensity. Let T_i denote the test sample intensity at locus i on the genome and R_i denotes the the corresponding reference sample intensity. The normalized log base 2 ratio of the sample and reference intensities, $log_2 T_i/R_i$, is considered as a random variable used for the derivation of copy number. Here, $log_2 T_i/R_i = 0$ indicates no DNA copy number change at locus i, $log_2 T_i/R_i < 0$ reveals a deletion at locus i, and $log_2 T_i/R_i > 0$ signifies duplication in the test sample at that locus. This random variable is assumed to follow a Gaussian distribution of mean 0 and constant variance σ^2. Then, deviations from the constant parameters (mean and variance) presented in $log_2 T_i/R_i$ data may indicate a copy number change.

Among the many methods used for a CGH data, Olshen et al (2004) proposed a circular binary segmentation (CBS) method to identify DNA copy number changes in an aCGH database on the mean change point model. This CBS method is mainly the combination of the likelihood ratio based test (Sen and Srivastava, 1975) for testing no

change in the mean against exactly one change in the mean with the BSP (Vestrikova, 1981) for searching multiple change points in the mean, assuming that the variance is unchanged.

To illustrate the CBS algorithm, let X_i denote the normalized $log_2 T_i/R_i$ at the ith locus along the chromosome;then $\{X_i\}$ is considered as a sequence of normal random variables taken from $N(\mu_i, \sigma_i^2)$, respectively, for $i = 1, \cdots, n$. Consider any segment of the sequence of the log ratio intensities X_i to be spliced at the two ends to form a circle; Z_{ij}, the likelihood ratio test statistic given by Sen and Srivastava (1975a) for testing the hypothesis that they are from i+1 to j and its complement have different means, is given by:

$$Z_{ij} = \frac{1}{\{1/(j-i) + 1/(n-j+1)\}^{\frac{1}{2}}} \{\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\} \tag{1.4}$$

with $S_i = X_1 + X_2 + \cdots + X_i, \qquad 1 \leq i < j \leq n.$

The test statistic Z_c of the CBS is based on the modified likelihood-ratio test and is given by:

$$Z_c = \max_{1 \leq i < j \leq n} |Z_{ij}| \tag{1.5}$$

Z_c allows for both a single change (j=n) and the epidemic alternative ($j < n$). A change is claimed if the statistic exceeds an appropriate critical value at a given significant level based on the null distribution. However, the null distribution of the test statistic Z_c is not attainable so far in the literature of change point analysis. Olshen et al (2004) suggested that the critical value needs to be computed using

Monte Carlo simulations or the approximation given by Segmund (1986) for the tail probability when X_is are normal. Once the null hypothesis of no change is rejected, the change points are estimated to be i (and j) such that $Z_c = |Z_{ij}|$ and the procedure is applied recursively to identify all the changes in the whole sequence of the log ratio intensities of a chromosome.

Since the p-value given by the CBS for a specific locus being a change point, however, is obtained by a permutation method and the calculation of such a p-value takes a long computation time when the sequence is long, which is the case for high density array data. The original R package was found to be the slowest computation method by Pichard et al (2005). By modifying the analytic formula, an updated version DNA copy (Venkatraman and Olshen, 2007) has improved the computational speed of CBS. The DNA copy package has been widely used in copy number analysis of aCGH data. Miller et al (2011) has applied CBS in the package ReadDepth for analysis of CNV from NGS read counts which has been considered as one of the best packages until now.

Change Point Analysis in NGS Among single genome based approaches, Yoon et al (2009) applied the event-wise testing (EWT) algorithm to define the region with read count changes. Based on standardized read data counted in 100bp windows over whole genome, the events for increased or reduced read depth data are identified by upper or lower tail probability of approximated normal distribution in EWT. A duplication or deletion segment within consecutive windows is defined if the maximum of the p values is lower than multiple corrected false positive rate (FPR). Miller et al (2011) used normalized read count data in the package ReadDepth and adopted circular binary segment algorithm which are widely used in aCGH data copy number estimation to identify the boundaries of the segment with read count changes. CNVnator by Abyzov et al (2009) is based on combining the established mean shift approach with additional multiple bandwidth partitioning approach. The mean shift

process is an iterative procedure that shifts each data points to the density maxim along the mean shift vector, a gradient of probability density function. Boundaries of genomic segments are identified by finding the consecutive pairs of bins with mean-shift vectors switching direction. Iterative read depth signaling merging, segmentation and boundary identification are through greedy algorithm and significance tests. FreeC by Boeva et al (2011) adopted LASSO based algorithms for segmentation of normalized read count data. All these algorithms are based on normal approximation of read counts data which may not reflect the true discrete data distribution.

In the early algorithms that utilize ratio of read counts between tumor and control samples, Segseq (Chiang and McCarroll, 2009) is a hybrid of local change point analysis with a subsequent merging procedure that joints adjacent chromosomal segments. It partitions the genome into windows of fixed size, estimates the test-control ratios for each window and merges the adjacent bins based on significance p value threshold. CNV-seq by Xie et al (2009) transforms the read count ratio of Gaussian variables to a t variable and calls copy number change based on p-value calculation from greater or less than ratio 1 with optimized window size. rSW-seq (Kim et al, 2010) applies a recursive smith waterman algorithm to sequence reads from tumor and matched control genomes that are combined and sorted in a no decreasing order according to their genomic position and weighted differently for tumor (Wt) and control (Wc). A large local positive (or negative) cumulative sum of weighted values indicates a local copy number gain (or loss). For copy number gains, the algorithm searches for the segment such that the partial cumulative sum is maximized and iterated until no more alteration can be found. In package CNAseg, Ivakno et al (2010) transferred the read count difference between tumor and control samples in non-overlapping windows to hidden states of copy number using k-mean clustering algorithm. Merging segments with the same copy number states is based on Pearson χ^2 test for neighboring read counts between tumor and control in the 2x2 contingency

table. The merging threshold is derived from flow cell splitting approach. BIC-seq developed by Xi et al (2011) applied minimizing Bayesian information criterion (BIC) on likelihood of mapped reads from tumor or control in reference genome positions to define optimized breakpoints for segmentation. BIC-seq level function (B, t, λ) (given a list of bins B, consecutive t bins in a window, and a tuning parameter for smooth penalty) is iteratively evaluated until all windows with negative BIC-diff are exhausted. False positives was identified by BIC-seq from the re-sampled data pulled from a pool of tumor and control reads.

Poisson distribution of Read Counts All of the above approaches are based on the assumption of normal distribution or approximation of normal distribution. Actually it has been demonstrated that read counts follows Poisson distribution. Under the assumption that the reads are randomly and independently sampled from any location of the test genome with equal probability, Bentley et al (2008) and Yoon et al (2009) have reported that RCs by Illumina GA follow a Poisson distribution with a slight over dispersion. The distribution of RC counts in fixed bin size window seems to be more dispersed in data from Illumina and SOLiD than that from Roche. Smaller bin size (e.g. 1000 bp compared to 2000 bp and 5000 bp) seems to have a more narrowly distribution. Except Poisson distribution, negative binomial distribution seems to fit the data more likely than Poisson distribution (Magi et al, 2012).

The over dispersion of RC data distribution can be accounted for by three main reasons: the existence of genomic regions of duplications and deletions (CNVs); the correlation between read coverage and the DNA local GC content; the correlation between the read coverage and the mappability (i.e. the inability to map reads into repetitive regions of the genome (Magi et al, 2012). It is estimated that the fraction of a genome subject to variation of copy number is 3.7$\sim$7.0% (Conrad et al, 2010). It was observed by Magi et al (2012) that the removal of genomic regions with known CNVs reduces the index of dispersion of RC data distribution in either of Illumina,

SOLiD and Roche platforms and also in either of low and high coverage data. These indicate that the existence of genomic regions of duplications and deletions (CNVs) is the major reason for over dispersion of the data distribution especially after correction of GC content and mappability error while the majority of RC data follow a Poisson distribution with global average number of reads as its mean.

It was also noted that in RD data, the variance is lowest for deletion states (zero or one copy) and variance increases proportionally with increasing copy number (Bentley et al, 2008; Yoon et al, 2009). This observation is consistent with the Poisson distribution in which the variance is equal to the average mean value. This differs from the characteristic data from microarray CGH data in which variance in probe ratios is lowest for the normal state (two copies) and probe variance increase from copy number changes in both directions (Yoon et al, 2009).

By reducing the bin size to 100 bp, Yoon et al (2009) used normal distribution to approximate standardized RD data. It has been noted by Olshen et al (2004) in copy number detection using aCGH array that centering log ratio intensities around zero for normalization of data to correct for confounding factors makes the copy numbers unidentifiable without some additional modeling. For example, the normalized data from diploid and triploid samples will appear similar. Miller et al (2011) recommended a model of negative binomial distribution to approximate the over dispersed Poisson distribution. As we know, a mixture of Poisson distributions with mean parameter value λ which is drawn from a gamma distribution with shape parameter α and scale parameter β follows a negative binomial distribution. Miller et al (2011) developed a model of negative binomial distribution with λ as the median value of the Poisson distribution and with γ as a variance parameter for λ from a gamma distribution. Introduction of γ alters the variancemean ratio (VMR) and accounts for the excessive variance from the over dispersion. The resulting root mean square error is three times smaller than that of Poisson.

As illustrated above, read counts from NGS follows overdispersed Poisson distribution in which the copy number variants seem to be the main reason accounting for the overdispersion. Under the assumption of unchanged copy number variants, read counts more likely follow homogeneous Poisson distribution. Therefore, the identification of copy number variants is an issue to find change of Poisson distributions for read counts data along the reference genome sequence.

1.4 Bayesian Change Point Analysis

It has long been recognized that Bayesian methods are well suited to find change points (Smith, 1975; Worsley et al, 1986). Bayesian change point methods are based on Bayesian inference according to Bayes' rule. The posterior probability ($\pi(\theta|\underline{X})$ is derived as a consequence of two antecedants, a prior probability ($\pi_0(\theta)$) and a likelihood function (L$(\theta, \underline{X})$ derived from a probability model for the data to be observed.

$$\pi(\theta|\underline{X}) = \frac{L(\theta, \underline{X}) * \pi_0(\theta)}{m(\underline{X})} \tag{1.6}$$

where m($\underline{X}$) is the marginal likelihood of observed values and is same for all hypotheses and often canceled in the relative posterior odds between two hypotheses. All inferences are based on posterior distribution of θ.

Under null hypothesis (equation (1.1)), the integration of the product of prior distribution and likelihood function will generate the posterior probability of hypothesis H_0: $\pi_0(\underline{X}) = P(\theta \in \ominus_0|\underline{X})$. Under alternative hypothesis, that will generate the posterior probability of $H_1 : \pi_1(\underline{X}) = p(\theta \in \ominus_0^c|X)$. Bayesian factor which is the ratio of integrated posterior probability of θ is preferred to be used for comparison of multiple models by hypothesis testing. With specific k change point in mind, the posterior probability of change point at k is: $\pi_1(k|\underline{X}) = p(\theta \in \ominus_0^c(k)|X)$ is frequently used to determine the existence of change point at k. The posterior odds is the ratio

of posterior probabilities of the hypotheses given observed data.

$$
\begin{aligned}
OR_k &= \frac{\pi_1(k|\underline{X})}{\pi_0(\underline{X})} \\
&= \frac{\int_{\theta \in \ominus_0^c} L_1(\theta, k, \underline{X}) \pi_0(\theta \in \ominus_0^c) \pi_0(k)}{\int_{\theta \in \ominus_0} L_0(\theta, \underline{X}) \pi_0(\theta \in \ominus_0)}
\end{aligned} \tag{1.7}
$$

The posterior odds OR_k is used to in this research for comparison of one change point model at k vs H_0. The threshold level will be derived from empirical distribution of the posterior odds in Monte Carlo simulations.

Bayesian inference is used to update the probability estimate for a hypothesis as additional evidence is acquired. If the evidence does not match up with a hypothesis, one should reject the hypothesis. But if a hypothesis is extremely unlikely a prior, one should also reject it, even if the evidence does not appear to match up. When the prior distribution is uninformative, the Bayesian is just like a likelihood ratio test. Bayesian utilized all information in the inferences including both the observed data and prior belief for the parameters. Wald and Wolfowitz (1950) proved that every Bayesian procedure is admissible. Under some conditions, all admissible procedures are either Bayesian procedures or limits of Bayesian procedures (in various senses) in decision theory.

Bayesian methods are very popular in change point analysis. Bayesian analysis of Poisson data similar in spirit to the present work include Raftery and Akman (1986); Gregory and Loredo (1992); West and Ogden (1997); Scargle (1998). However, no Poisson Bayesian model has ever been used in CNV detection esp using next generation sequencing data as we know. The application of Bayesian approaches in big data as that from NGS is also challenging.

In present work, we developed a Bayesian change point model for single genome and a Bayesian change point model for paired genomes based on Anscombe normal

approximation algorithm, and modified a Bayesian Poisson change point model to identify the change points and estimate copy number variants from NGS read count data in single genome. The statistics have been optimized for the detection of change points and copy numbers at single and multiple change points. Three rapid R packages have been developed to simulate Poisson data, estimate and display copy number variants in table and graphics. The sensitivity and specificity are evaluated in both simulated read count data with known Poisson distribution and in human NGS read count in comparison to other popular packages.

Chapter 2

SINGLE GENOME BAYESIAN APPROACHES IN NGS READ COUNT ANALYSIS

2.1 The Models

A reference human genome sequence (or a specific chromosome) with total g nucleotides is divided into total n evenly spaced intervals with each interval having b base pairs (called bin size). The next generation sequencing will produce total T reads and each read has average l nucleotides called read length. The average read counts per bin is equal to T/n.

Let X_i represents read counts in evenly spaced intervals i with bin size b bps, where $X_i = 0, 1, \ldots$ and $i \in (1, 2, \cdots, n)$ along the studied genome sequence and n=int(g/b). We assume X_i follow independent Poisson distributions: $X_i \sim Pois(\lambda_i)$ with intensity parameter $0 \leq \lambda_i < \infty$. Statistical inference for multiple change point analysis is based on the following hypotheses testing. The nulll hypothesis is given by :

$$H_0 : \ \lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda \tag{2.1}$$

versus alternative hypothesis

$$H_1 : \lambda_1 = \cdots = \lambda_{k_1} \neq \lambda_{k_1+1} = \cdots = \lambda_{k_2} \neq \lambda_{k_2+1} = \cdots = \lambda_{k_q} \neq \lambda_{k_q+1} = \cdots = \lambda_n \quad (2.2)$$

Under null hypothesis, all read counts follow identically independent Poisson distribution (i.e. there is no CNV in the genome segment) with parameter λ over the segment, while under alternative hypothesis, there are q change points at $k_1, k_2, \cdots, k_q$, which separate the studied segment into q+1 regions with possibly q+1 different constant rates of reads. Read counts X_is are assumed to follow identically independent Poisson distribution with parameter λ_1 from 1 to k_1, with parameter λ_2 from $k_1 + 1$ to k_2 etc in the genome segment. The no change point model corresponds to null hypothesis. One change point model and multiple change point model correspond to the alternative hypothesis. These models are detailed subsequently in the following.

2.2 Change Point Model Under Normal Transformation

2.2.1 Rationale

Normal approximation of read counts data has been widely used in CNV data analysis especially using the next generation sequencing data although read counts have been demonstrated to follow Poisson distribution. This is partly because it is easier to deduce the mathematical algorithm and popular properties of a normally distributed variable. However, the normal approximation method used currently is not efficient and has been considered to generate errors in probability analysis (Lesch and Jeske, 2009). We developed the following change point model for single genome

using Anscombe (1948)'s normal approximation to Poisson distribution from Bayesian perspective, which is one of popular variance stabilizing algorithms that converge to normality faster than the ordinary normal approximation in which mean and variance of normal distribution are equal to the Poisson intensity parameter (Anscombe, 1948). The latter is often used in current NGS read count data analysis.

2.2.2 No Change Point Model

Let Y_i represents read counts in b_i, where $Y_i = 0, 1, \cdots$ and $i \in (1, 2, \cdots, n)$ along the reference genome sequence. The Y_i is assumed to independently follow Poisson distributions: $Y_i \sim Pois(\lambda_i)$ with mean parameter $0 \leq \lambda_i < \infty$.

Let $X_i = \sqrt{Y_i + \frac{3}{8}}$. The independence of $Y_i s$ results in the independence of X_i because X_i is a function of only Y_i. $Xi \in (\underline{X} : X_1, X_2, \cdots, X_n)$

According to Anscombe (1948),

$$X_i = \sqrt{Y_i + \frac{3}{8}} \qquad \sim approxiN\left(\sqrt{\lambda_i + \frac{1}{8}}, \frac{1}{4}\right), \qquad if\ \lambda_i\ is\ large \tag{2.3}$$

Let assume $\theta_i = \sqrt{\lambda_i + \frac{1}{8}}$, then

$$X_i \qquad \sim approxiN\left(\theta_i, \frac{1}{4}\right), \qquad if\ \theta_i\ is\ large\ and\ \sqrt{\frac{1}{8}} \leq \theta_i \tag{2.4}$$

Since the final deduction of posterior probability will be the result of integration of λ_i and is independent with λ_i, we will use θ_i in the following deduction as a substitute for simplification purpose. Actually we can deduce $\lambda_1 = \lambda_2$ from $\theta_1 = \theta_2$ or deduce $\theta_1 = \theta_2$ from $\lambda_1 = \lambda_2$ based on $\theta_i = \sqrt{\lambda_i + \frac{1}{8}}$. The null hypothesis and alternative

hypothesis for λ_i will become that for θ_i. The probability density function for X_i is:

$$f(X_i|\theta_i) \propto \sqrt{\frac{2}{\pi}} e^{(-2(X_i-\theta_i)^2)}, \qquad i = 1, \ldots, n \tag{2.5}$$

Under no change point assumption, θ_i are constant for all intervals and $\theta_i = \theta$. The independence between the read counts X_i in intervals leads to the likelihood function

$$\begin{aligned} L_0(\theta|X_1, X_2, \ldots, X_n) &= f(X_1, X_2, \cdots, X_n|\theta) \\ &= \prod_{i=1}^{n} f(X_i|\theta) \\ &= \prod_{i=1}^{n} \sqrt{\frac{2}{\pi}} e^{(-2(X_i-\theta)^2)} \\ &= (\frac{2}{\pi})^{\frac{n}{2}} e^{-2\sum_{i=1}^{n}(X_i-\theta)^2} \end{aligned} \tag{2.6}$$

We assume a normal prior for $\theta : \theta \sim N(\mu, \tau^2)$.

$$\pi_0(\theta) \begin{cases} \propto \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} & \sqrt{\frac{1}{8}} \leq \theta, \\ = 0 & \theta < \sqrt{\frac{1}{8}}. \end{cases} \tag{2.7}$$

Integrating the above likelihood times this prior gives the posterior probability of

no change, denoted as $\pi_0(\underline{X})$:

$$\begin{aligned}
\pi_0(\underline{X}) &\propto \int_\theta L_0(\theta|\underline{X})\pi_0(\theta)\,\mathrm{d}\theta \\
&\propto \int_\theta (\frac{2}{\pi})^{\frac{n}{2}} e^{-2\sum_{i=1}^{n}(X_i-\theta)^2} \\
&\cdot \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \\
&= 2^{(\frac{n-1}{2})}\pi^{-\frac{n+1}{2}}(\tau^2)^{-\frac{1}{2}} \int_\theta \\
&exp(-(2\sum_{i=1}^{n}(X_i-\theta)^2 + \frac{(\theta-\mu)^2}{2\tau^2}))\,\mathrm{d}\theta \\
&= 2^{(\frac{n-1}{2})}\pi^{-\frac{n+1}{2}}(\tau^2)^{-\frac{1}{2}} \int_\theta exp(-\Delta)\,\mathrm{d}\theta
\end{aligned} \tag{2.8}$$

Then,

$$\begin{aligned}
\Delta &= 2\sum_{i=1}^{n}(X_i-\theta)^2 + \frac{(\theta-\mu)^2}{2\tau^2}) \\
&= \frac{4\tau^2\sum_{i=1}^{n}(X_i-\theta)^2 + (\theta-\mu)^2}{2\tau^2} \\
&= \frac{4\tau^2\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\theta)^2 + (\theta-\mu)^2}{2\tau^2} \\
&\cdots \\
&= \frac{4\tau^2\sum_{i=1}^{n}(X_i-\bar{X})^2 + 4n\tau^2\bar{X}^2 + \mu^2 - \frac{(4n\tau^2\bar{X}+\mu)^2}{4n\tau^2+1}}{2\tau^2} \\
&+ \frac{(4n\tau^2+1)(\theta - \frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1})^2}{2\tau^2}
\end{aligned} \tag{2.9}$$

Introducing Δ from equation (2.9) into equation (2.8) leads to the posterior probability of no change point

$$
\begin{aligned}
\pi_0(\underline{X}) &\propto 2^{\frac{n-1}{2}}\pi^{-\frac{n+1}{2}}(\tau^2)^{-\frac{1}{2}} \\
&\quad exp(-(\frac{4\tau^2\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\tau^2})) \\
&\quad \cdot exp(-\frac{4n\tau^2\bar{X}^2+\mu^2-\frac{(4n\tau^2\bar{X}+\mu)^2}{4n\tau^2+1}}{2\tau^2}) \\
&\quad \cdot \int_{\theta} exp(-\frac{(4n\tau^2+1)(\theta-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1})^2}{2\tau^2})\,\mathrm{d}\theta \\
&\quad = 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}}e^{-2\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
&\quad \cdot exp(-\frac{4n\tau^2\bar{X}^2+\mu^2-\frac{(4n\tau^2\bar{X}+\mu)^2}{4n\tau^2+1}}{2\tau^2}) \\
&\quad \cdot (4n\tau^2+1)^{-\frac{1}{2}}\cdot(2\pi\tau^2)^{-\frac{1}{2}}(4n\tau^2+1)^{\frac{1}{2}} \\
&\quad \cdot \int_{\sqrt{\frac{1}{8}}}^{\infty} e^{-\frac{(4n\tau^2+1)(\theta-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1})^2}{2\tau^2}}\,\mathrm{d}\theta \\
&\quad = 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}}e^{-2\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
&\quad \cdot exp(-\frac{4n\tau^2\bar{X}^2+\mu^2-\frac{(4n\tau^2\bar{X}+\mu)^2}{4n\tau^2+1}}{2\tau^2}) \\
&\quad \cdot (4n\tau^2+1)^{-\frac{1}{2}}\cdot\delta
\end{aligned}
\tag{2.10}
$$

Let $z=\frac{\theta-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}}$, then $\theta=z\sqrt{\frac{\tau^2}{4n\tau^2+1}}+\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}$, $\mathrm{d}\theta=\sqrt{\frac{\tau^2}{4n\tau^2+1}}\,\mathrm{d}z$. and

$$\begin{aligned}
\delta &= (2\pi\tau^2)^{-\frac{1}{2}}(4n\tau^2+1)^{\frac{1}{2}} \int_{\sqrt{\frac{1}{8}}}^{\infty} e^{-\frac{(\theta-\frac{4n\tau^2\bar{X}}{4n\tau^2+1})^2}{2\frac{\tau^2}{4n\tau^2+1}}} \,\mathrm{d}\theta \\
&= (2\pi)^{-\frac{1}{2}} \int_{\frac{(\sqrt{\frac{1}{8}}-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1})}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}}}^{\infty} e^{-\frac{z^2}{2}} \,\mathrm{d}z \\
&= (1-\phi(\frac{\sqrt{\frac{1}{8}}-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}})) \qquad (2.11)
\end{aligned}$$

where $\phi(\frac{\sqrt{\frac{1}{8}}-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}})$ is a cumulated standard normal distribution with variance 1 and mean 0 at $\frac{\sqrt{\frac{1}{8}}-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}}$. After substituting the results from equation (2.11) into equation (2.10), we get :

$$\begin{aligned}
\pi_0(\underline{X}) \propto\ & 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}} e^{-2\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
& \cdot exp(-\frac{4n\tau^2\bar{X}^2+\mu^2-\frac{(4n\tau^2\bar{X}+\mu)^2}{4n\tau^2+1}}{2\tau^2})(4n\tau^2+1)^{-\frac{1}{2}} \\
& \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\frac{4n\tau^2\bar{X}+\mu}{4n\tau^2+1}}{\sqrt{\frac{\tau^2}{4n\tau^2+1}}})) \qquad (2.12)
\end{aligned}$$

We assume that μ and τ^2 are constants and they can be estimated by $\bar{X}$ =

$\sum_{i=1}^{n} X_i/n$ and $\tau^2 = \frac{1}{4}$ and substitute them in. Then

$$\begin{aligned}
\pi_0(\underline{X}) &\propto 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}} e^{-2\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
&\cdot exp(-\frac{4n\frac{1}{4}\bar{X}^2 + \bar{X}^2 - \frac{(4n\frac{1}{4}\bar{X}+\bar{X})^2}{4n\frac{1}{4}+1}}{2\frac{1}{4}})(4n\frac{1}{4}+1)^{-\frac{1}{2}} \\
&\cdot (1-\phi(\frac{\sqrt{\frac{1}{8}} - \frac{4n\frac{1}{4}\bar{X}+\bar{X}}{4n\frac{1}{4}+1}}{\sqrt{\frac{\frac{1}{4}}{4n\frac{1}{4}+1}}})) \\
&= 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}} e^{-2\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
&\cdot (n+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}} - \bar{X}}{\sqrt{\frac{1}{4(n+1)}}})) \qquad (2.13)
\end{aligned}$$

This will be used to compare with the following posterior probability of one change point model.

2.2.3 One Change Point Model

We assume that one change point separates the genome sequence into two subsegments because the underlying process changes abruptly there. Denote two segment models with constant Poisson rates λ_1 and λ_2 and change point k, the position at which Poisson rates for read counts switch from λ_1 to λ_2. The genome is partitioned into two intervals $1, 2, \cdots, k$ and $k+1, \cdots, n$ containing k bins at positions less equal than change point k, and n-k bins at positions greater than k, respectively.

Let Y_i denote read count in the ith bin and $Y_i \sim Pois(\lambda_1) for\ i = 1, \cdots, k$ and $Y_i \sim Pois(\lambda_2) for\ i = k+1, \cdots, n$. According to Anscombe (1948),

$$X_i = \sqrt{Y_i + \frac{3}{8}} \sim approx N\left(\sqrt{\lambda_1 + \frac{1}{8}}, \frac{1}{4}\right) \qquad for\ i = 1, \cdots, k \tag{2.14}$$

and

$$X_i = \sqrt{Y_i + \frac{3}{8}} \sim Napprox\left(\sqrt{\lambda_2 + \frac{1}{8}}, \frac{1}{4}\right) \qquad for\ i = k+1, \cdots, n \tag{2.15}$$

Let $\theta_1 = \sqrt{\lambda_1 + \frac{1}{8}}$ and $\theta_2 = \sqrt{\lambda_2 + \frac{1}{8}}$. Then, the probability density function for X_i is

$$f(X_i|\theta_1, k) \simeq \sqrt{\frac{2}{\pi}} e^{(-2(X_i - \theta_1)^2)} \qquad for\ i = 1, \cdots, k \tag{2.16}$$

and

$$f(X_i|\theta_2, n-k) \simeq \sqrt{\frac{2}{\pi}} e^{(-2(X_i - \theta_2)^2)} \qquad for\ i = k+1, \cdots, n \tag{2.17}$$

The likelihood, L, of one change point model is, by independence assumption discussed above, just the product of the probabilities of two segments considered separately.

$$\begin{aligned} L(k, \theta_1, \theta_2|X_1, \cdots, X_n) = f(X_1, \cdots, X_k|\theta_1, k)) \\ \cdot f(X_{k+1}, \cdots, X_n|\theta_2, n-k)) \end{aligned} \tag{2.18}$$

We assume normal priors for $\theta_1 = \sqrt{\lambda_1 + \frac{1}{8}}$ and $\theta_2 = \sqrt{\lambda_2 + \frac{1}{8}}$ similar to that under null hypothesis:

$$\pi_0(\theta_1|k) \propto \frac{1}{\sqrt{2\pi\tau_1^2}} e^{-\frac{(\theta_1 - \mu_1)^2}{2\tau_1^2}}, \qquad for\ \sqrt{\frac{1}{8}} \leq \theta_1 \tag{2.19}$$

and

$$\pi_0(\theta_2|k) \propto \frac{1}{\sqrt{2\pi\tau_2^2}} e^{-\frac{(\theta_2-\mu_2)^2}{2\tau_2^2}}, \qquad for\ \sqrt{\frac{1}{8}} \leq \theta_2 \tag{2.20}$$

otherwise, $\pi_0(\theta_1|k) = \pi_0(\theta_2|k) = 0$. We estimate μ_1 by $\bar{X_1} = \frac{1}{k}\sum_{i=1}^{k} X_i$, $\tau_1^2 = \frac{1}{4}$, μ_2 by $\bar{X_2} = \frac{1}{n-k}\sum_{i=k+1}^{n} X_i$ and $\tau_2 = \frac{1}{4}$.

Change point k is assumed to follow a discrete Uniform prior distribution U(n-1):

$$\pi_0(k) = \frac{1}{n-1}, \qquad k = 1, \cdots, n-1 \tag{2.21}$$

Similarly, then the joint posterior distribution of parameters θ_1, θ_2, k under one change point model H_1 can be derived as:

$$\begin{aligned} \pi_1(\theta_1, \theta_2, k|\underline{X}) &\propto f(X_1, \cdots, X_n|\theta_1, \theta_2, k)\pi_0(\theta_1, \theta_2|k)\pi_0(k) \\ &= f(X_1, \cdots, X_k|\theta_1, k)\pi_0(\theta_1|k) \\ &\cdot f(X_{k+1}, \cdots, X_n|\theta_2, n-k)\pi_0(\theta_2|n-k)\pi_0(k) \end{aligned} \tag{2.22}$$

Thus, the posterior probability of one change point model at position k is:

$$
\begin{aligned}
\pi_1(k) &= \frac{\int_{\theta_2}\int_{\theta_1}\pi_1(\theta_1,\theta_2,k,\underline{X})\,\mathrm{d}\theta_1\,\mathrm{d}\theta_1}{\sum_{k=1}^{n-1}\int_{\theta_2}\int_{\theta_1}\pi_1(\theta_1,\theta_2,k,\underline{X})\,\mathrm{d}\theta_1\,\mathrm{d}\theta_1)\Delta k} \\
&\propto \int_{\theta_2}\int_{\theta_1}\pi_1(\theta_1,\theta_2,k|\underline{X})\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\
&= \int_{\theta_2}\int_{\theta_1} f(X_1,\cdots,X_k|\theta_1,k)\pi_0(\theta_1|k) \\
&\cdot f(X_{k+1},\cdots,X_n|\theta_2,n-k)\pi_0(\theta_2|n-k)\pi_0(k)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\
&= \int_{\theta_1} f(X_1,\cdots,X_k|\theta_1,k)\pi_0(\theta_1|k)\,\mathrm{d}\theta_1 \\
&\cdot \int_{\theta_2} f(X_{k+1},\cdots,X_n|\theta_2,n-k)\cdot\pi_0(\theta_2|n-k)\,\mathrm{d}\theta_2\ \pi_0(k) \qquad (2.23)
\end{aligned}
$$

We derive two terms in the equation (2.23) in the following similarly as in no change point model (see equation (2.13)).

$$
\begin{aligned}
A(k) &\propto \int_{\theta_1} f(X_1,\cdots,X_k|\theta_1,k)\pi_0(\theta_1|k)\,\mathrm{d}\theta_1 \\
&= 2^{\frac{k}{2}}(\pi)^{-\frac{k}{2}}e^{-2\sum_{i=1}^{k}(X_i-\bar{X}_1)^2} \\
&\cdot (k+1)^{-\frac{1}{2}}\cdot\left(1-\phi\left(\frac{\sqrt{\frac{1}{8}}-\bar{X}_1}{\sqrt{\frac{1}{4(k+1)}}}\right)\right) \qquad (2.24)
\end{aligned}
$$

and

$$
\begin{aligned}
B(k) &\propto \int_{\theta_2} (X_{k+1}, \cdots, X_n|\lambda_2, n-k)\pi_0(\theta_2|k)\,\mathrm{d}\theta_2 \\
&= 2^{\frac{n-k}{2}}(\pi)^{-\frac{n-k}{2}} e^{-2\sum_{i=k+1}^{n}(X_i-\bar{X}_2)^2} \\
&\cdot (n-k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_2}{\sqrt{\frac{1}{4(n-k+1)}}}))
\end{aligned} \tag{2.25}
$$

where $\bar{X}_1 = \frac{1}{k}\sum_{i=1}^{k} X_i$ and $\bar{X}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n} X_i$.

The product of A(k) and B(k) generates:

$$
\begin{aligned}
A(k)B(k) &\propto 2^{\frac{k}{2}}(\pi)^{-\frac{k}{2}} e^{-2\sum_{i=1}^{k}(X_i-\bar{X}_1)^2} \\
&\cdot (k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_1}{\sqrt{\frac{1}{4(k+1)}}})) \\
&\cdot 2^{\frac{n-k}{2}}(\pi)^{-\frac{n-k}{2}} e^{-2\sum_{i=k+1}^{n}(X_i-\bar{X}_2)^2} \\
&\cdot (n-k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_2}{\sqrt{\frac{1}{4(n-k+1)}}})) \\
&= 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}} e^{-2\sum_{i=1}^{k}(X_i-\bar{X}_1)^2 - 2\sum_{i=k+1}^{n}(X_i-\bar{X}_2)^2} \\
&\cdot (k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_1}{\sqrt{\frac{1}{4(k+1)}}})) \\
&\cdot (n-k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_2}{\sqrt{\frac{1}{4(n-k+1)}}}))
\end{aligned} \tag{2.26}
$$

Finally, we obtain the posterior probability of one change point at position k:

$$\begin{aligned}\pi_1(k) &\propto A(k)B(k) \\ &\propto 2^{\frac{n}{2}}(\pi)^{-\frac{n}{2}}e^{-2\sum_{i=1}^{k}(X_i-\bar{X}_1)^2-2\sum_{i=k+1}^{n}(X_i-\bar{X}_2)^2} \\ &\cdot (k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_1}{\sqrt{\frac{1}{4(k+1)}}})) \\ &\cdot (n-k+1)^{-\frac{1}{2}} \cdot (1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_2}{\sqrt{\frac{1}{4(n-k+1)}}}))\end{aligned} \tag{2.27}$$

2.2.4 H_0 vs.H_1 and Change Point

We define the posterior probability odds ratio (OR) of one change point at k vs. no change point as:

$$OR_k = \frac{\pi_1(k|\underline{X})}{\pi_0(\underline{X})} \tag{2.28}$$

where $\pi_1(k)$ and $\pi_0(\underline{X})$ can be obtained from equation (2.27) and (2.13). The natural logarithm of OR can be subsequently deduced by the substitution of equation (2.13) and (2.27):

$$\begin{aligned}lnOR_k &= 2(\sum_{i=1}^{n}(X_i-\bar{X})^2 - \sum_{i=1}^{k}(X_i-\bar{X}_1)^2 - \sum_{i=k+1}^{n}(X_i-\bar{X}_2)^2) \\ &+ \frac{1}{2}(log(n+1)-log(k+1)-log(n-k+1)) \\ &+ log(1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_1}{\sqrt{\frac{1}{4(k+1)}}})) + log(1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}_2}{\sqrt{\frac{1}{4(n-k+1)}}})) \\ &- log((1-\phi(\frac{\sqrt{\frac{1}{8}}-\bar{X}}{\sqrt{\frac{1}{4(n+1)}}})))\end{aligned} \tag{2.29}$$

In our method, the maximum posterior probability odds ratio (mlnOR) is compared to a lnOR.level for determination of significance in favor of $H_1 vs. H_0$. Since the distribution of mlnOR is not clear, we chosen a lnOR.level from the empirical cumulative distribution of mlnOR through Monte Caro simulations given known read counts for two copy number genome and known read counts for one copy number change (gain or loss) within a segment. Optimizations have been performed so that an optimized OR.level is chosen for high sensitivity and low false positive rate in single change point simulations rather than arbitrary value e.g. 0.

Once mlnOR is less than or equal to lnOR.level, it is claimed as no significance for one change point model in favor of no change point model. Therefore, no change point in the segment is identified. Once a significant result is found, the change point k is identified through the mlnOR: $\hat{k} = argmax_k mlnOR$.

2.3 Bayesian Approach to Poisson Change Point Model

Scargle (1998) developed an efficient Bayesian analysis algorithm called Bayesian Block to analyze structure in photon counting data in astronomical time series study. Scargle adopted the nonuniform but normalized prior, a special case of the gamma distribution. To provide a more flexible choice for prior distribution according to possible difference in multiple change point analysis, we adapted a more generalized gamma prior for Poisson intensity parameter which is an extension of the work by Scargle (1998). The maximum likelihood estimator of Poisson parameter from read count data is assigned as α for the shape parameter and $\beta = 1$ for scale parameter in Gamma distribution. Furthermore, we evaluated and optimized the statistics and their threshold level for significance test and change point identification and adopted sliding window for multiple change point analysis in our approach *BayGamma*.

2.3.1 No Change Point Model (H_0)

For $X \in (X_i, i = 1, 2, , n)$, where the integer X_i is the number of read counts assigned to the ith bin interval. Taking the rate per bin to be constant λ, the counts in a given bin obey Poisson statistics for this rate:

$$P(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \tag{2.30}$$

Independence of the counts X_i yields the likelihood:

$$\begin{aligned} L(\lambda|X_1, X_2, \cdots, X_n) &= P(X_1, X_2, \cdots, X_n|\lambda) \\ &= \prod_{i=1}^{n} P(X_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}) \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!} \end{aligned} \tag{2.31}$$

The maximum of this probability occurs at the value $\lambda = \frac{\sum_{i=1}^{n} X_i}{n}$. Since the denominator in equation (2.31) has the property that its value for an interval is just the product of its value for two or more subintervals, this factor cancels out in a comparison of null hypothesis with alternative hypotheses of a given model and we omit it.

Gamma distribution is commonly used as a prior distribution in Bayesian inference for the intensity parameter of Poisson distribution. We assume that λ follows a gamma distribution with shape parameter α and scale parameter β.

$$P(\lambda|\alpha,\beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}} & 0 \leq \lambda, \alpha, \beta > 0, \\ 0 & \lambda < 0. \end{cases} \tag{2.32}$$

The expected value of λ can be obtained by $E(\lambda) = \alpha\beta$. The introduction of α and β allows us to adjust the prior distribution based on different λ distributions that correspond to likely different copy number variants in segments separated by change points. We assumed known $\alpha = \sum_{i=1}^{n} X_i/n$ and $\beta = 1$ since we expect $E(\lambda) = \sum_{i=1}^{n} X_i/n$ and $var(\lambda) = \sum_{i=1}^{n} X_i/n$.

Integrating the above likelihood times this prior gives the posterior distribution of no change point given observed data $X_i, i \in (1, \cdots, n)$.

$$\begin{aligned} \pi_0(\underline{X}) &\propto \int_0^{\infty} L(\lambda|X_1, \cdots, X_n)P(\lambda)\,\mathrm{d}\lambda \\ &= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\lambda^{(\sum_{i=1}^{n} X_i+\alpha-1)} \\ &\cdot e^{-(n+\frac{1}{\beta})\lambda}\,\mathrm{d}\lambda \\ &= \frac{\Gamma(\alpha+\sum_{i=1}^{n} X_i)}{\Gamma(\alpha)\beta^{\alpha}(n+\frac{1}{\beta})^{\alpha+\sum_{i=1}^{n} X_i}} \int_0^{\infty} \frac{(n+\frac{1}{\beta})^{(\alpha+\sum_{i=1}^{n} X_i)}}{\Gamma(\alpha+\sum_{i=1}^{n} X_i)} \\ &\cdot \lambda^{\sum_{i=1}^{n} X_i+\alpha-1}e^{-(n+\frac{1}{\beta})\lambda}\,\mathrm{d}\lambda \\ &= \frac{\Gamma(\alpha+\sum_{i=1}^{n} X_i)}{\Gamma(\alpha)\beta^{\alpha}(n+\frac{1}{\beta})^{-(\alpha+\sum_{i=1}^{n} X_i)}} \end{aligned} \tag{2.33}$$

This will be used to compare with the following one change point model in which

a observation interval is broken into two subintervals over which read counts are assumed to follow homogeneous distribution within each subinterval but different between two subintervals.

2.3.2 One Change Point Model (H_1)

The point separating such segment is called a change point because the underlying process changes abruptly there. Denote two segment model with constant Poisson rates λ_1 and λ_2 and change point k, the position at which distributions of read counts switch from λ_1 to λ_2. A segment is partitioned into two intervals $1, 2, \cdots, k$ and $k+1, \cdots, n$ containing k bins at positions less equal than change point k and n-k bins at positions greater than k, respectively.

The probability of one change point two segment model is, by independence assumption discussed above, just the product of probabilities of two segments considered separately.

$$\begin{aligned} P(X_1, \cdots, X_n|\lambda_1, \lambda_2, k) &= P(X_1, \cdots, X_k|\lambda_1, k) \\ &\cdot P(X_{k+1}, \cdots, X_n|\lambda_2, n-k) \end{aligned} \tag{2.34}$$

Similarly, the joint likelihood of parameters λ_1, λ_2, k under alternative hypothesis H_1 can be derived as:

$$\begin{aligned} L(\lambda_1, \lambda_2, k|X_1, \cdots, X_n) &= P(X_1, \cdots, X_n|\lambda_1, \lambda_2, k) \\ &= P(X_1, \cdots, X_k|\lambda_1, k) \\ &\cdot P(X_{k+1}, \cdots, X_n|\lambda_2, n-k) \end{aligned} \tag{2.35}$$

We assume that λ_1 follows a gamma distribution with parameters α_1 and β_1, λ_2 with α_2 and β_2, and k follows a uniform discrete distribution with equal probability at each point between 1 and n-1. We assume $\alpha_1 = \sum_{i=1}^{k} X_i/k,\ \beta_1 = 1$ and $\alpha_2 = \sum_{k+1}^{n} X_i/(n-k),\ \beta_2 = 1$ as constants.

The prior probability for λ_1 is:

$$\pi_0(\lambda_1|k,\alpha_1,\beta_1) = \begin{cases} \frac{1}{\Gamma(\alpha_1)\beta_1^{\alpha_1}}\lambda_1^{\alpha_1-1}e^{-\frac{\lambda_1}{\beta_1}} & 0 \le \lambda_1,\ \alpha_1,\beta_1 > 0, \\ 0 & \lambda_1 < 0. \end{cases} \tag{2.36}$$

The expected value of λ_1 can be obtained by $E(\lambda_1) = \alpha_1\beta_1$. The prior probability for λ_2 is:

$$\pi_0(\lambda_2|(n-k),\alpha_2,\beta_2) = \begin{cases} \frac{1}{\Gamma(\alpha_2)\beta_2^{\alpha_2}}\lambda_2^{\alpha_2-1}e^{-\frac{\lambda_2}{\beta_2}} & 0 \le \lambda_2,\ \alpha_2,\beta_2 > 0, \\ 0 & \lambda_2 < 0. \end{cases} \tag{2.37}$$

The expected value of λ_2 can be obtained by $E(\lambda_2) = \alpha_2\beta_2$, and the prior probability for k is:

$$\pi_0(k) = \frac{1}{n-1} \tag{2.38}$$

Thus, the posterior probability of one change point at position k is:

$$
\begin{aligned}
\pi_1(k) \propto & \int_{\lambda_2} \int_{\lambda_1} L(\lambda_1, \lambda_2, k | X_1, \cdots, X_n) \pi_0(\lambda_1 | k, \alpha_1, \beta_1) \\
& \cdot \pi_0(\lambda_2 | n-k, \alpha_2, \beta_2) \pi_0(k) \, \mathrm{d}\lambda_1 \, \mathrm{d}\lambda_2 \\
= & \int_{\lambda_2} \int_{\lambda_1} P(X_1, \cdots, X_k | \lambda_1, k) \pi_0(\lambda_1 | k, \alpha_1, \beta_1) \\
& \cdot P(X_{k+1}, \cdots, X_n | \lambda_2, n-k) \pi_0(\lambda_2 | n-k, \alpha_2, \beta_2) \pi_0(k) \, \mathrm{d}\lambda_1 \, \mathrm{d}\lambda_2 \\
= & \int_{\lambda_1} P(X_1, \cdots, X_k | \lambda_1, k) \pi_0(\lambda_1 | k, \alpha_1, \beta_1) \, \mathrm{d}\lambda_1 \\
& \cdot \int_{\lambda_2} P(X_{k+1}, \cdots, X_n | \lambda_2, n-k) \pi_0(\lambda_2 | n-k, \alpha_2, \beta_2) \, \mathrm{d}\lambda_2 \pi_0(k)
\end{aligned} \tag{2.39}
$$

We introduce two terms $C_1(k)$ and $C_2(k)$ in the equation (2.39) for each subinterval within which no change point is assumed similar to equation (2.33).

$$
\begin{aligned}
C_1(k) &= \int_{\lambda_1} P(X_1, \cdots, X_k | \lambda_1, k) P(\lambda_1) \, \mathrm{d}\lambda_1 \\
&= \frac{\Gamma(\alpha_1 + \sum_{i=1}^{k} X_i)}{\Gamma(\alpha_1) \beta^{\alpha_1} (k + \frac{1}{\beta_1})^{-(\alpha_1 + \sum_{i=1}^{k} X_i)}}
\end{aligned} \tag{2.40}
$$

and

$$
\begin{aligned}
C_2(k) &= \int_{\lambda_2} (X_{k+1}, \cdots, X_n | H_1(\lambda_2, n-k) P(\lambda_2 | H_1) \, \mathrm{d}\lambda_2 \\
&= \frac{\Gamma(\alpha_2 + \sum_{i=k+1}^{n} X_i)}{\Gamma(\alpha_2) \beta^{\alpha_2} (n - k + \frac{1}{\beta_2})^{-(\alpha_2 + \sum_{i=k+1}^{n} X_i)}}
\end{aligned} \tag{2.41}
$$

Finally, we obtain the posterior probability of one change point at k as:

$$\begin{aligned}\pi_1(k) &\propto C_1(k)C_2(k) \\ &= \frac{\Gamma(\alpha_1 + \sum_{i=1}^{k} X_i)}{\Gamma(\alpha_1)\beta^{\alpha_1}(k + \frac{1}{\beta_1})^{-(\alpha_1+\sum_{i=1}^{k} X_i)}} \\ &\cdot \frac{\Gamma(\alpha_2 + \sum_{i=k+1}^{n} X_i)}{\Gamma(\alpha_2)\beta^{\alpha_2}(n - k + \frac{1}{\beta_2})^{-(\alpha_2+\sum_{i=k+1}^{n} X_i)}}\end{aligned} \tag{2.42}$$

2.3.3 H_0 vs.H_1 and Change Point

We define the posterior probability odds ratio (OR) of the one change point at k vs. no change point model as:

$$OR_k = \frac{\pi_1(k)}{\pi_0(\underline{X})} \tag{2.43}$$

where $\pi_1(k)$ and $\pi_0(\underline{X})$ can be obtained from equation (2.33) and (2.42). The natural logarithm of OR can be subsequently deduced by the substitution of equation (2.33) and (2.42):

$$
\begin{aligned}
lnOR_k &= log(\pi_1(k) - log(\pi_0(\underline{X}) \\
&= log(\frac{\Gamma(\alpha_1 + \sum_{i=1}^{k} X_i)}{\Gamma(\alpha_1)\beta^{\alpha_1}(k + \frac{1}{\beta_1})^{-(\alpha_1 + \sum_{i=1}^{k} X_i)}} \\
&\cdot \frac{\Gamma(\alpha_2 + \sum_{i=k+1}^{n} X_i)}{\Gamma(\alpha_2)\beta^{\alpha_2}(n - k + \frac{1}{\beta_2})^{-(\alpha_2 + \sum_{i=k+1}^{n} X_i)}}) \\
&- log(\frac{\Gamma(\alpha + \sum_{i=1}^{n} X_i)}{\Gamma(\alpha)\beta^{\alpha}(n + \frac{1}{\beta})^{-(\alpha + \sum_{i=1}^{n} X_i)}}) \\
&= log(\Gamma(\alpha_1 + \sum_{i=1}^{k} X_i)) + log(\Gamma(\alpha_2 + \sum_{i=k+1}^{n} X_i)) - log(\Gamma(\alpha + \sum_{i=1}^{n} X_i)) \\
&+ log(\Gamma(\alpha)) - log(\Gamma(\alpha_1)) - log(\Gamma(\alpha_2)) + (\alpha_1 + \sum_{i=1}^{k} X_i) log(k + \frac{1}{\beta_1}) \\
&+ (\alpha_2 + \sum_{i=k+1}^{n} X_i) log(n - k + \frac{1}{\beta_2}) - (\alpha + \sum_{i=1}^{n} X_i) log(n + \frac{1}{\beta}) \qquad (2.44)
\end{aligned}
$$

The maximum posterior probability odds ratio (mlnOR) is used as the test statistics for the determination in favor of $H_1 vs. H_0$. The mlnOR is compared to a threshold lnOR.level inferred from the empirical distribution of mlnOR through Monte Caro simulations under the assumption that there is one copy number change (gain or loss) or no change. The read counts in two copy number genome and one copy number genome can be calculated from NGS data. The lnOR.level is chosen so that high sensitivity and low false positive rate can be achieved in single change point data simulations.

When mlnOR is lower or equal to the lnOR.level, we conclued that no change point is found. When mlnOR is higher than the lnOR.level, the significant result for one change point is concluded. The change point k based on mlnOR is then identified as the change point in one change point model: $\hat{k} = argmax_k mlnOR$.

2.4 Multiple Change Point Decomposition

Binary Segmentation Procedure (BSP) and Circular Binary Segmentation (CBS) are well known multiple change point decomposition procedures as discussed in background. Sliding window algorithm has been widely adapted in CNVs analysis in big data such as NGS data. We adapted a sliding window algorithm for identification of change points and segmentation. A sliding window is defined as a segment with fixed number of bins (e.g. l=100) and moved along data sequence to end of the data set. Test statistics mlnOR is computed based on the data within the window. If mlnOR is greater than OR.level, it is claimed a significant positive result in favor of H_1 model. Then the change point within the window is outputted. If mlnOR is less than or equal to OR.level, no positive result is claimed. No change point but the end location of the data set is outputted.

Under multiple change point assumption, a new window is chosen along the genome sequence with certain distance skip parameter (e.g., sp=10) and the same procedure is followed for identification of the next change point until the last sliding window moved to the end of the dataset. The segment between two adjacent change points or between start point and first change point or between last change point and end of the data set is considered following homogeneous distribution. The average read counts per bin within the segment is calculated as the estimated intensity parameter of expected homogeneous distribution in the segment. Copy number for bins in the segment can be calculated from the average read counts as described in the following.

The sliding window procedure allows optimization of algorithm to significantly detect read counts change sensitively and specifically and consistently based on detection resolution requirement and computation speed. Number of multiple comparisons can be controlled so that inflated type I error can be minimized in multiple

comparisons. The influence of factors associated with sliding window on empirical cumulative distribution of mlnOR in Monte Carlo simulations will be illustrated in our result section.

2.5 Data

2.5.1 One Change Point Simulation Procedure

To simulate total n positive integer value with k values following Poisson distribution with parameter λ_1 in the first segment, and n-k values following Poisson distribution with parameter λ_2, we first generate k random positive integer values with SimplePois.Data function in R with parameter λ_1 representing simulated read counts from diploid genome in the first segment, and then simulate n-k random positive integer values with random Poisson function in R with parameter λ_2 representing simulated read counts from diploid genome in the second segment. Finally, we merge the simulated for both segments together in order and obtain total n simulated data in each simulation representing read counts per bin in NGS sequencing data. Let X_{ij} denote read count, then $X_{ij} \sim Pois(\lambda_1)$ for $i = 1, \cdots, k$, and $X_{ij} \sim Pois(\lambda_2)$ for $i = k+1, \cdots, n$ in the jth simulation where $j = 1, \cdots, m$.

Example: One Change Point Read Count Data Simulation

Control Read Count Data Low depth coverage data (x4) is mostly used for read count analysis due to the low cost. When a genome is separated into bins with bin size of 1kb along the reference genome sequence, read counts of low coverage data from NGS machine average around 50 based on NA19239 NGS dataset. Therefore, we generated k=50 random values with parameter $\lambda_1 = 40$ and n-k=50 random values with parameter $\lambda_2 = 40$ for the second segment. Total 100 random positive integer values follow the Poisson distribution with parameter $\lambda = 40$ with no change point assumed:the read count $X_{ij} \sim Pois(40) for\ i = 1, \cdots, 100$ in each simulation. The simulated data can be found in Figure 2.1(top).

One DNA Copy Gain Data We generated k= 50 random values with parameter $\lambda_1 = 40$ following the above procedure, and simulated 50 random numbers similarly with parameter $\lambda_2 = 60$ in the second segment taking into account of a single copy number increase from the first segment to the second segment. We produced total n=100 integer value representing read counts in 100 bins in each simulation. Read count in each bin $X_{ij} \sim Pois(40) for\ i = 1, \cdots, 50$ and $X_{ij} \sim Pois(60) for\ i = 51, \cdots, 100$. The assumed change point for copy number gain should be at k=50. The simulated data can be found in Figure 2.1 (middle).

One DNA Copy Loss Data Similarly, after we generated k=50 random integers with $\lambda_1 = 40$, we simulated n-k=50 random integers with $\lambda_2 = 20$ in each simulation based on the assumption that one copy of genomic DNA was lost in a diploid genome. Total n=100 integers representing read counts in 100 bins in each simulation. Read counts per bin $X_{ij} \sim Pois(40) for\ i = 1, \cdots, 50$ and $X_{ij} \sim Pois(20) for\ i = 51, \cdots, 100$. The assumed change point for one DNA copy loss should be at k=50. The simulated data can be found in Figure 2.1 (bottom).

2.5.2 Multiple Change Point Simulation

Since multiple change points and segment usually need to be discovered in human NGS data, we simulated read count data with multiple change points and segments so that sensitivity and specificity of the proposed approaches can be evaluated under multiple change point assumption. To simulate six segments, we first generate k_1=50 positive integer value following Poisson distribution with parameter $\lambda_1 = 40$ for the first segment, $k_2 - k_1$=50 values following Poisson distribution with parameter $\lambda_2 = 60$ for the second segment, $k_3 - k_2$=50 values following Poisson distribution with parameter $\lambda_3 = 40$ for the third segment, $k_4 - k_3$=50 values following Poisson distribution with parameter $\lambda_4 = 20$ for the fourth segment, $k_5 - k_4$=50 values following Poisson distribution with parameter $\lambda_5 = 40$ for the fifth segment and n-k_5=50 values

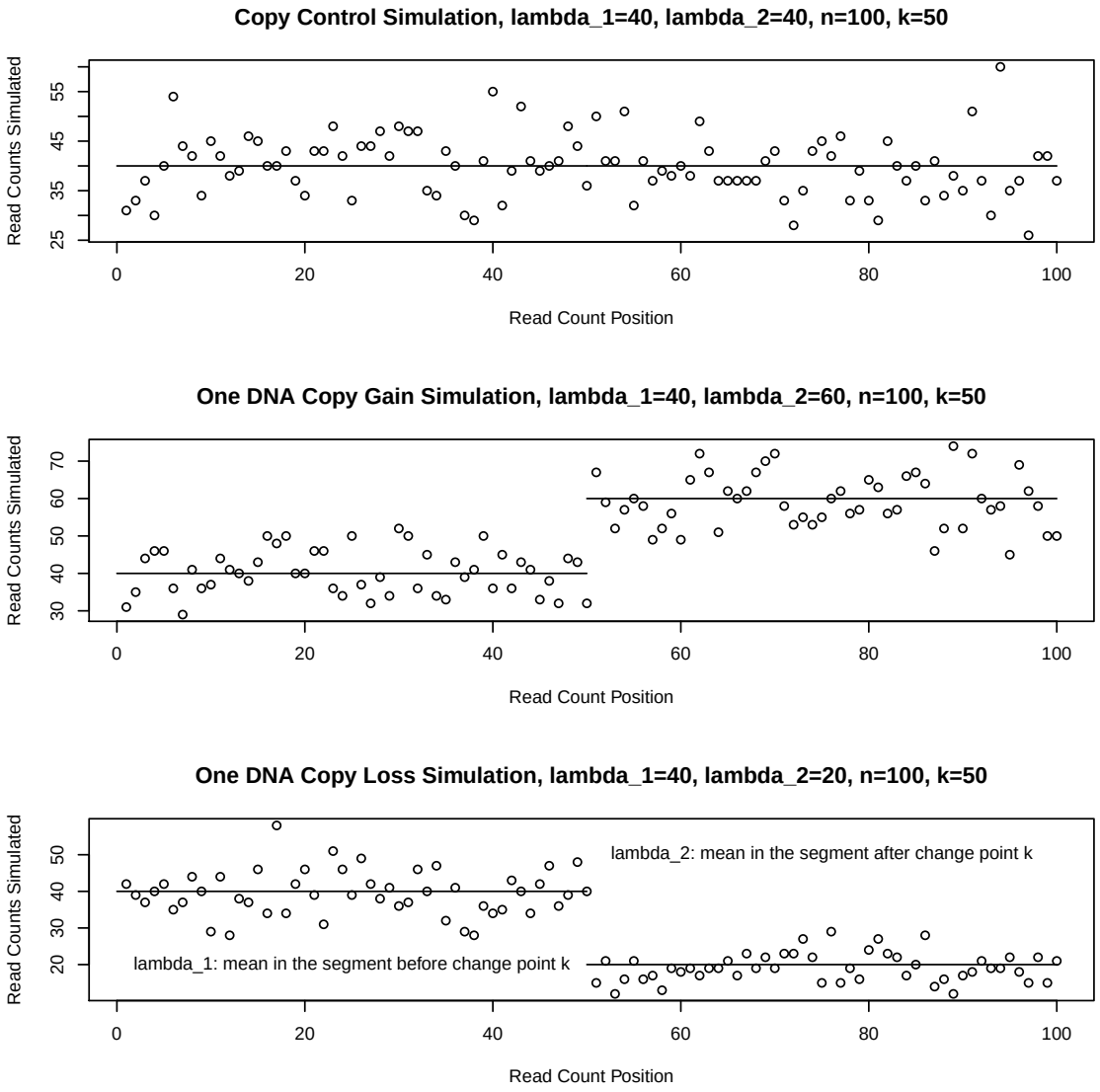

Figure 2.1: One change point read count data simulation with segment length n=100 at assumed changed point k=50. The data are generated by random Poisson function in R with intensity parameter in first segment (λ_1) and in second segment (λ_2).

following Poisson distribution with parameter $\lambda_6 = 40$ for the sixth segment via our R function MultiPois.Data. Finally, we merge the simulated for all of these segments together in order and obtain total n=300 simulated read count in each simulation representing NGS sequencing data. The simulated data can be found in Figure 2.2. The Positive Rate of Change Point Identification of gain will be based on identification of k_1, the positive ration of change point identification of loss will be based on identification of k_3 and the false positive rate of change point identification will be based on identification of any change point in the region from k_4 to n in m=1000 simulations.

2.5.3 Next Generation Sequencing (NGS) Data Sets

The low coverage read count data with bin size 1kb on the chromosome 6 of NA19239 from the 1000Genome Sequencing Project are utilized for identification of change point and estimation of copy numbers.

Chromosome 6 is chosen because it is one of the most varied chromosomes holding important immune related genes e.g. HLA MHC I and II. Low coverage data is chosen because it is mostly used and easier to obtain due to low cost. Figure 2.3 shows read counts of NA 19239 NGS data with low coverage and bin size=1kb.

One copy gain NGS data: Although structural variants have been reported for chromosome 6 of NA19239, they are not considered accurate in terms of location and length of DNA sequence because they have showed variation from different sources. To build up a dataset for evaluation as a positive control, we first simulated a segment with length l=50 read counts which follow Poisson distribution with parameter $\lambda = 25$ representing one DNA copy gain for consecutive 50 bins. Then total m=100 simulated segments were randomly added to the NGS data to denote 100 segments and 2m=200 change points which are considered as known one DNA copy gain segment. As an example, NGS data (top) and simulated one copy gain NGS data (middle) can be

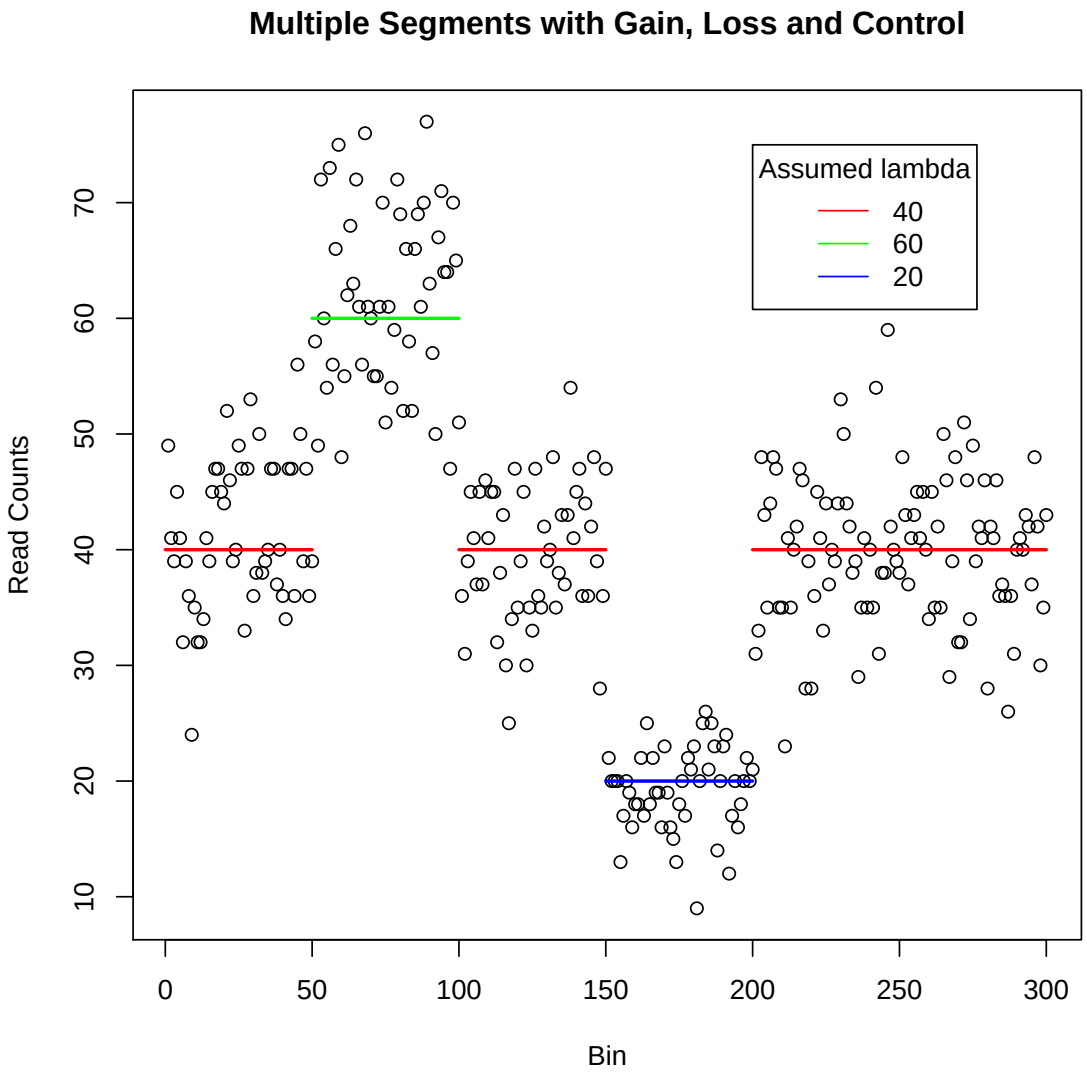

Figure 2.2: Multiple change point data simulation for six segments with change point at $k_1 = 50, k_2 = 100, k_3 = 150, k_4 = 200$ and segment length n=300.

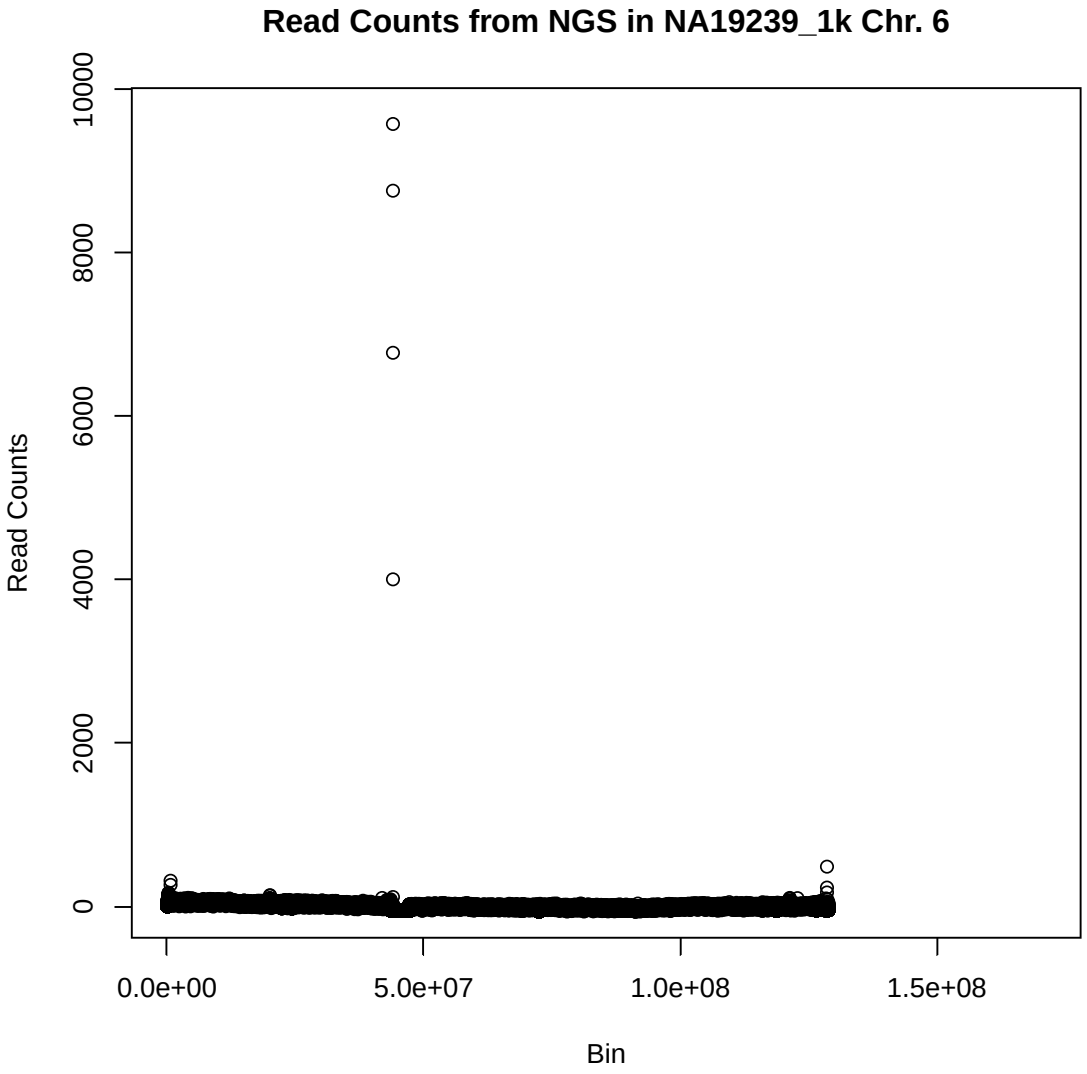

Figure 2.3: Read count data from human NGS of NA 19239 chromosome 6 with 1kb in each bin.

seen in Figure 2.4.

One copy loss NGS data: Similarly we simulated l (.e.g.=50) read counts which follow Poisson distributions with parameter $\lambda(e.g. = 25)$. Then m (e.g. =100) positions were randomly selected and the NGS read data in the following 50 consecutive bins were reduced by the amounts of the simulated data in order representing a segment of l=50 consecutive bins with DNA one copy loss data randomly. Consequently, we have a positive control for DNA copy loss NGS data with 2m (e.g.200) change points and m=100 segments. The NGS with one copy loss data can be seen in Figure 2.4(bottom).

The evaluation indices for identification of these change points and estimation of copy numbers in these regions as described in the following.

2.6 Evaluation

2.6.1 Definition

An assumed Change Point is a known change point that has been confirmed or on which the read count data simulation was based.

Sensitity of DNA copy gain (or loss) detection(PR) is defined as the proportion of assumed DNA copy gains or losses that were detected. These include at least one change point (either gain or loss, or one of multiple change points). The type II error rate is 1- sensitivity.

Flase positive rate of DNA copy gain (or loss) detection (FR) is defined as the proportion of at least one DNA copy gain or loss detected in control region that no change point was assumed. The specificity is calculated as 1-false positive rate (FR).

Estimated Change Point is the change point identified by maximum posterior odds (lnOR) in comparison to a threshold by each approach.

Positive Rate of Change Point Identification (PRCPI), also known as sen-

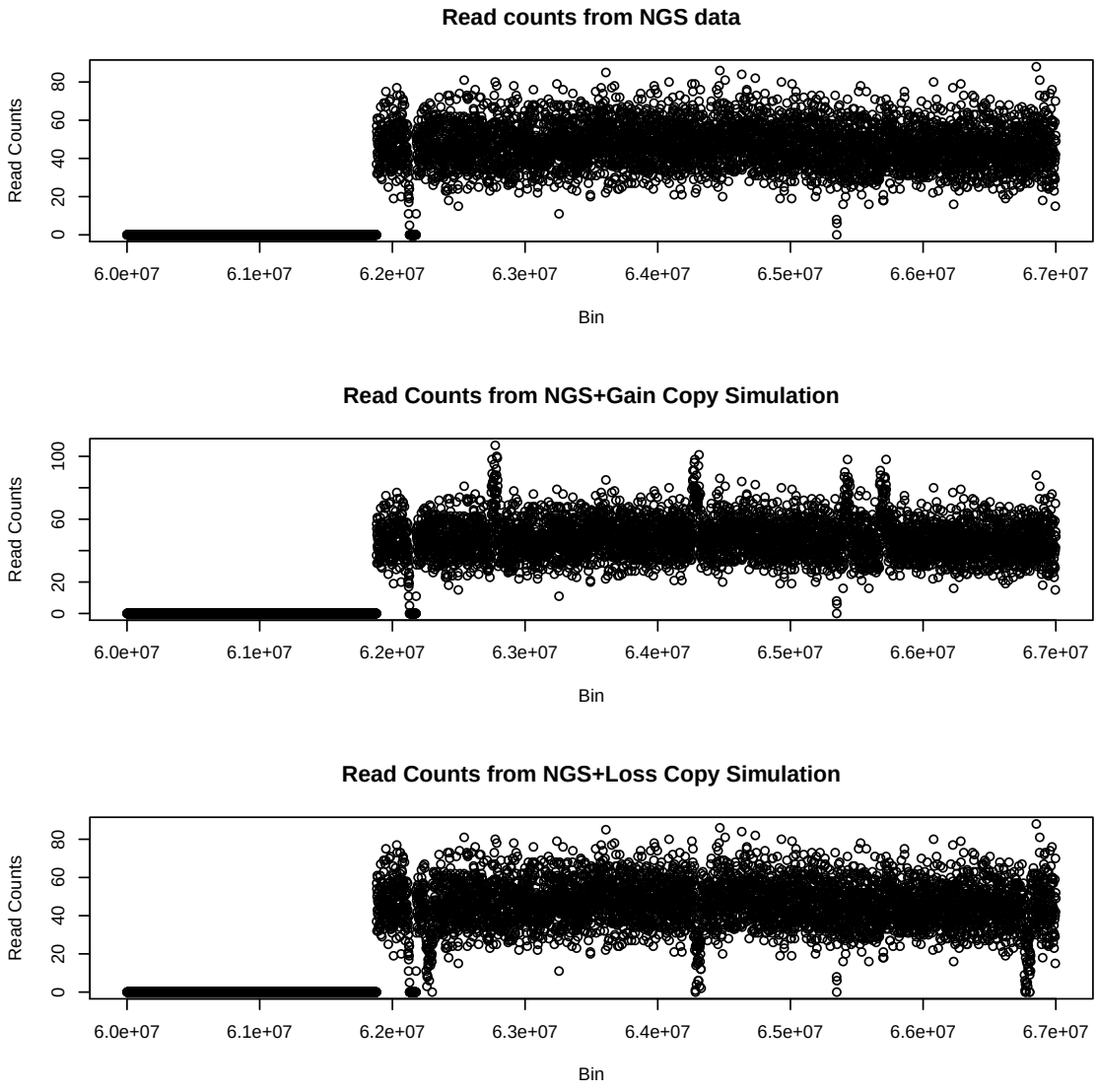

Figure 2.4: Display of read count data from NGS in a segment from bp position 6.0E+07 to 6.7E+07 in reference genome sequence. Random one copy gain data were added (middle) or reduced (bottom) from NGS control (top).

sitivity of specific change point identification, is defined as the proportion of the assumed change points whose locations are identified at certain accuracy range. PRCPI is usually for $\hat{k} = k$, and $PRCPI_1$ for $\hat{k} \in k \pm 1$ and $PRCPI_2$ for $\hat{k} \in k \pm 2$.

False Positive Rate of Change Point Identification (FPRCPI) is defined as the proportion of change points identified around the position where separate simulations for two segments were operated but with no change point assumed. It is similar to PRCRI but applied in control data instead.

False Change Point Rate (FCPR) is defined as the proportion of the identified change points which are not equal to the assumed change points at certain accuracy range.

Read Count Estimate is defined as the estimated value of read counts per bin in segment refined by estimated change points. Read count estimate represents the estimated intensity parameter for the Poisson distribution. Read count estimate can be calculated as the average value of read counts per bin in the segment and is the MLE estimate of Poisson parameter.

Assumed Read Count is defined as the read counts per bin in a segment refined by assumed change points. Assumed read count represents the assumed intensity parameter for the Poisson distribution in simulated or NGS data.

Copy Number Estimate is defined as the ratio of read count estimate with the expected read counts from one copy number sequence.

Assumed Copy Number is defined as the ratio of assumed read count with the expected read counts from one copy number sequence.

2.6.2 Evaluation Indicies for One Change Point and Multiple Change Point Simulation Data

For each data simulation, test statistics (e.g. $mlnOR$) was calculated and used for inference about significance test with rejection of null hypothesis based on thresh-

old level (e.g, $mlnOR > OR.level$). The corresponding estimated change point($\hat{k_j}$) was identified by mlnOR once significance test for copy gain or loss detection was determined for the jth simulation. If no significant change point was found, end location (n) of the data set was reported. Total $\hat{Q_j}$ change points were produced for each simulation. For one change point model, Q is expected to be 2 including k and n in one copy gain or loss data set, but for multiple change point model with four change points, Q is expected to be 5. The estimated read count ($\hat{\lambda}_{jq}$) and the estimated copy number ($\hat{C}_{jq}$) at the q th segment($q \in (1, \cdots, Q)$ before the qth change point were calculated according to the following formulas. Their mean value and mean squared error in m =1,000 data simulations in comparison to the corresponding assumed values were obtained.

Sensitivity and Specificity

Sensitivity of DNA one copy gain (or loss) detection (Positive Rate, PR) based on m times of assumed one copy gain (or loss) data simulations was obtained by:

$$PR = \frac{1}{m} \sum_{j=1}^{m} I(mlnOR_j > OR.level), \tag{2.45}$$

where I is an indicator function with 1 for a true event (>OR.level) and 0 for a false event. The type II error rate is 1- sensitivity.

False Positive Rate of DNA one copy gain (or loss) detection(FPR) is the same as above PR except that the evaluation is based on m times of control data simulations The specificity is 1- FR.

Change Point Estimation

The Positive Rate of Change Point Identification (PRCPI) based on m gain or

loss simulations was calculated as

$$PCPIR = \frac{1}{m}\sum_{j=1}^{m} I(\hat{k}_{jq} = k_q), \tag{2.46}$$

where k_q is the assumed change point in data simulation. For multiple change point detection, $\hat{k}_{jq}$ is the closest estimated change point if $\hat{Q}$ is not equal to Q. $PCPIR_1$ is for $\hat{k}_{jq} \in (k_q \pm 1)$ and $PCPIR_2$ is for $\hat{k}_{jq} \in (k_q \pm 2)$. $I(\hat{k}_{jq} = k_q)$ is an indicator function with 1 for a true event and 0 for a false event.

False Positive Rate of Change Point Identification (FPRCPI) is the same as the above PRCPI except that the evaluation is based on m control read count simulations. The specificity is 1-FPCPIR.

Mean False Change Point Rate (FCPR) was calculated by :

$$FCPR = \frac{1}{m}\sum_{j=1}^{m} \frac{\hat{Q}_j - Q}{\hat{Q}_j} \tag{2.47}$$

where Q is the total number of assumed change points and $\hat{Q}_j$ is the number of estimated change points in the jth multiple change point simulation.

Mean Change Point Estimate of change point k_q was obtained by

$$\bar{k}_q = \frac{1}{m}\sum_{j=1}^{m} \hat{k}_{jq} \tag{2.48}$$

Estimated Mean Squared Error of Change Point Estimate at k_q was obtained by:

$$MSE(k_q) = \frac{1}{m}\sum_{j=1}^{m} (\hat{k}_{jq} - k_q)^2 \tag{2.49}$$

Read Count Estimation

Read Count Estimate in segments before $(\hat{\lambda}_{jq})$ the change point $\hat{k}_{jq}$ in the jth simulation was obtained by:

$$\hat{\lambda}_{jq} = \frac{1}{\hat{k}_{jq} - \hat{k}_{j(q-1)}} \sum_{\hat{k}_{j(q-1)}+1}^{\hat{k}_{jq}} X_{ij} \tag{2.50}$$

where $j = 1, \cdots, m,\ i = 1, \cdots, n$, and $q = 1, \cdots, Q$.

Mean Read Count Estimates: Mean read count estimates before $(\bar{\lambda}_q)$ the estimated change point k_q was obtained by

$$\begin{aligned} \bar{\lambda}_q &= \frac{1}{m} \sum_{j=1}^{m} \hat{\lambda}_{jq} \\ &= \frac{1}{m} \sum_{j=1}^{m} \frac{1}{\hat{k}_{jq} - \hat{k}_{j(q-1)}} \sum_{\hat{k}_{j(q-1)}+1}^{\hat{k}_{jq}} X_{ij} \end{aligned} \tag{2.51}$$

Estimated Mean Squared Error (MSE) of Read Count Estimate for segment before $(\hat{\lambda}_{jq})$ the qth change point was obtained by:

$$MSE(\hat{\lambda}_q) = \frac{1}{m} \sum_{j=1}^{m} (\hat{\lambda}_{jq} - \lambda_q)^2 \tag{2.52}$$

Copy Number Estimation

Copy Number Estimate for the segment before the qth change point was obtained by the following:

$$\hat{C_{jq}} = \frac{2\hat{\lambda}_{jq}}{\lambda_1} \tag{2.53}$$

where 2 is taken with respect to diploid genome, and λ_1 is the assumed read counts in a diploid chromosome.

Mean Copy Number Estimate for the qth segment was obtained by:

$$\bar{C_q} = \frac{1}{m}\sum_{j=1}^{m}\hat{C}_{jq} = \frac{1}{m}\sum_{j=1}^{m}\frac{2\hat{\lambda}_{jq}}{\lambda_1} \tag{2.54}$$

Estimated Mean Squared Error of Copy Number Estimate for the qth segment was obtained by:

$$MSE(\hat{C_q}) = \frac{1}{m}\sum_{j=1}^{m}(\hat{C_{jq}} - C_q)^2 \tag{2.55}$$

where $C_q = \frac{2\lambda_q}{\lambda_1}$

2.6.3 Evaluation Indices for NGS Data

Assumed Change Points: The NGS plus or minus one copy (alse called)gain or loss simulations have Q=2m assumed change points in addition to unknown change points in NA19329 NGS data sets. The change points k_q are randomly located in $k_1, k_{1+l}, \cdots, k_j, k_{j+l}, \cdots, k_m$,$k_{m+l}$. The evaluation indexes are based on identification of these 2m change points and estimation of read counts and copy number in the segments refined by two neighboring change points k_j and k_{j+l}.

Assumed Read Counts Let X_{ij} represent read counts in bin i which is classified in segment $j \in (1, \cdots, m)$. We assume that read counts in segments between two change points are homogeneous and follow Poisson distributions with parameter $\lambda_1, \cdots, \lambda_m$ respectively. Assumed read count in the jth segment is estimated by:

$$\lambda_j = \bar{X_j} = \frac{1}{l}\sum_{i=k_j+1}^{k_j+l} X_{ij}, \tag{2.56}$$

Expected read count (λ) for one copy gain or loss: Since expected read count for one copy gain or loss varies between sequencing platforms probably due to copy number variation, we used trimmed mean of NGS read count data with 10 percent of extreme values (less equal than 5 percent lowest and greater than 5 percent largest) removed. The extreme values are believed to contribute to the dispersion of NGS read counts data due to copy number variation. Then,

$$\lambda = \frac{1}{0.9 * n * 2} \sum_i X_i, \ i \in (i : X_5 < Xi <= X_{95}) \tag{2.57}$$

where X_5 is the 5th quantile value and X_{95} is the 95 quantile value based on empirical distribution of NGS read count data. The number 2 is introduced taking account into diploid chromosomes in human genome. λ is used as the denominator in calculation of estimated copy number in the following.

Assumed copy number was estimated by:

$$C_j = \frac{\lambda_j}{\lambda} \tag{2.58}$$

We expect to generate multiple estimated change points $\hat{Q}$ by the proposed change point algorithms. Then each assumed change point was searched for matching of estimated change points based on either exact match ($\hat{k_q} = k_q$) or with certain range of assumed change points (e.g. $k_q \pm 2$). When no matched change point was found, the closest change point was assigned as the corresponding estimated change point. The following indices are used for evaluation of test efficiency and accuracy.

Change Point Identification

Positive Rate of Change Point Identification (PRCPI, also called sensitivity) was

obtained by

$$PRCPI = \frac{1}{2m}\sum_{q=1}^{2m} I(\hat{k_q} = k_q), \tag{2.59}$$

where 2m is the number of assumed change points and I is an indicator function with 1 for a true event in that estimated change points matched assumed change points and 0 for a false event.

False Change Point Rate (FCPR) was obtained by:

$$FCPR = \frac{1}{\hat{Q}}(\hat{Q} - \sum_{q=1}^{2m} I(\hat{k}_q = k_q) - Q_0) \tag{2.60}$$

where $I(\hat{k_q} = k_q)$ is an indicator function with 1 for true event and 0 for false event, $\hat{Q}$ is total number of change points identified by the proposed algorithm in NGS plus simulation data and Q_0 is that in NGS alone data sets.

Mean Deviation of Change Point Estimate was obtained by

$$\bar{\Delta k} = \frac{1}{2m}\sum_{q=1}^{2m}(\hat{k}_q - k_q) \tag{2.61}$$

Estimated Mean Squared Error of Change Point Estimation was obtained by:

$$MSE(\hat{k}) = \frac{1}{2m}\sum_{q=1}^{2m}(\hat{k_q} - k_q)^2 \tag{2.62}$$

Read Count Estimation

Read Count Estimate was calculated as mean of the read counts in a segment between $\hat{k}_{jq} + 1$ and $\hat{k}_{jq} + l$ obtained by:

$$\hat{\lambda_j} = \frac{1}{l}\sum_{i=\hat{k}_{jq}+1}^{\hat{k}_{jq}+l} X_{ij} \tag{2.63}$$

where $\hat{k}_{jq}$ is the $\hat{k}_q$ classified within the jth segment.

Mean Deviation of Read Count Estimate between estimated read count and assumed read count was obtained by:

$$\bar{\Delta}\lambda = \frac{1}{m}\sum_{j=1}^{m}(\hat{\lambda}_j - \lambda_j) \tag{2.64}$$

Estimated Mean Squared Error (MSE) of Read Count Estimation was obtained by:

$$MSE(\hat{\lambda}_j) = \frac{1}{m}\sum_{j=1}^{m}(\hat{\lambda}_j - \lambda_j)^2 \tag{2.65}$$

Copy Number Estimation

Copy number estimate was obtained by:

$$\hat{C}_j = \frac{\hat{\lambda}_j}{\lambda} \tag{2.66}$$

Mean Deviation of Copy Number Estimate between assumed copy number and estimated copy number was obtained by:

$$\hat{\Delta}C = \frac{1}{m}\sum_{j=1}^{m}(\hat{C}_j - C_j) \tag{2.67}$$

Estimated Mean Squared Error of Copy Number Estimate was obtained by:

$$MSE(\hat{C}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{C}_j - C_j)^2 \tag{2.68}$$

2.7 Results

2.7.1 BayNormal

R programs based on the algorithm of normal approximation change point model called BayNormal have been developed. BayNormal can be used to simulate and detect single change point, multiple change point and human NGS data, graphic display of read count data, estimate of read count and copy numbers and allow us to set OR.level, window size and speed parameter in copy number estimation. The optimization and evaluation results are presented in the following.

2.7.1.1 Optimization of Test Statistics and OR.level

Since the prior distribution in normal approximation bayesian approach is normal, we call it BayNormal as the brief. After running BayNormal algorithms on single change point simulated data with segment length n=100, the log posterior odds ratio(lnOR) values at each k position are showed in Figure 2.5 based on no change point control(top), one copy gain (middle) and one copy loss simulation (bottom) as described in data section. The maximum of log posterior odds ratio (mlnOR) is identified around k=50 for one copy gain or loss data but not for no change point control which is consistent with the assumed change point in data simulation.

Window Size (or Segment Length To find an optimized window size for detection of change points by BayNormal, we simulated m=1000 times of single change point data with segment length n=10 to 200 and with change point located in the middle. The empirical cumulative distribution of statistics mlnOR generated by BayNormal can be seen in Figure 2.6. The empirical cumulative distribution of mlnOR is shifted towards right indicating that mlnOR is increased with increased window size from n=12 to 200 in both one copy gain (middle) and loss (bottom) data simulations but is slightly decreased in copy control data. This can be expected because big sample size

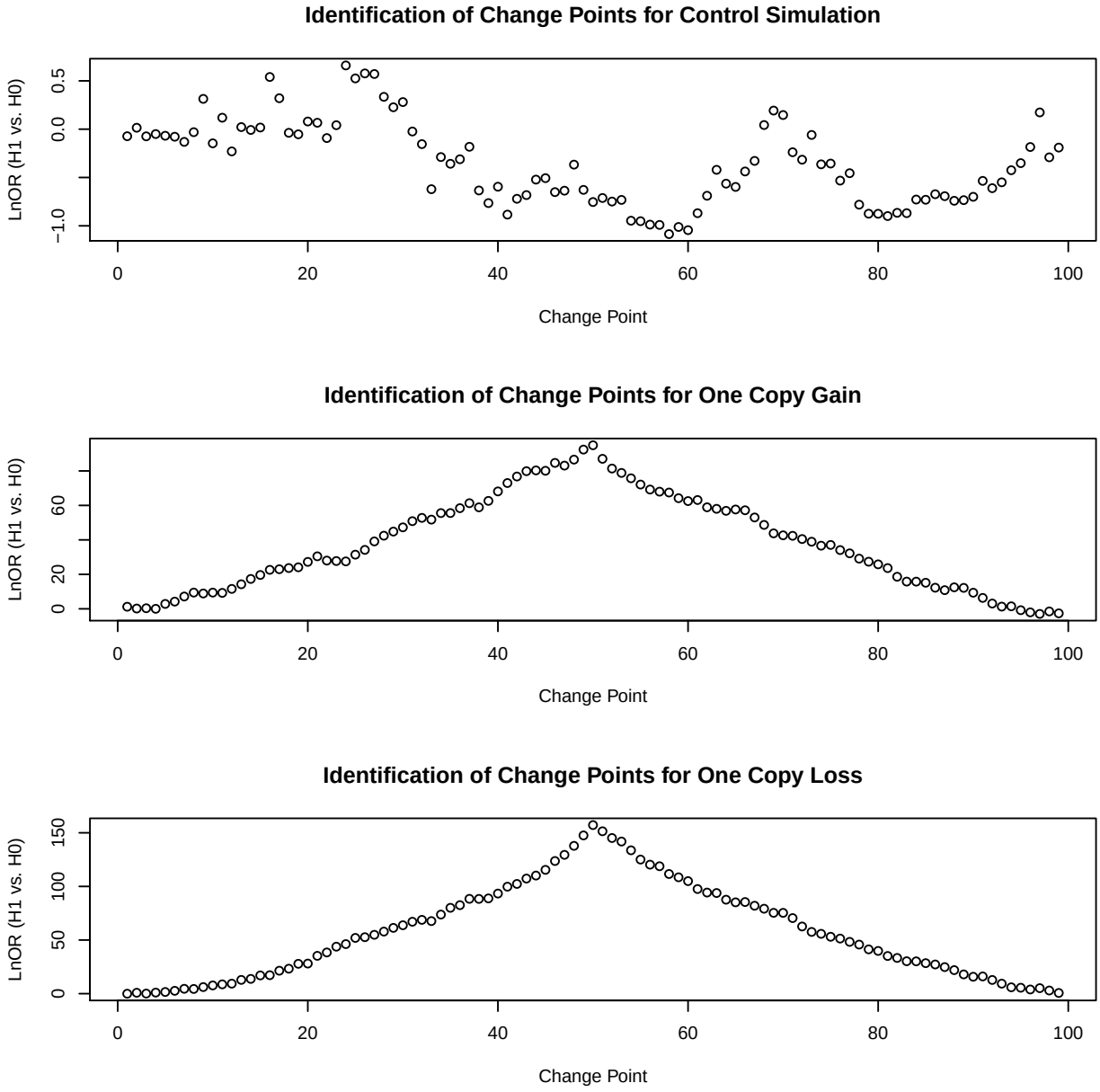

Figure 2.5: Log Posterior Odds ratio (lnOR) generated by BayNormal at each possible change point along the segment based on a control, one copy gain and one copy loss single change point simulation.

will result in a more powerful test and be more specific. The mlnOR for one copy loss data is bigger than that for one copy gain data at the same window size. However, window size 10 leads to significant increase of lnOR in copy control simulations. This indicates that it is easy to obtain false positive result under window size 10 and even fails to significantly detect one copy loss change point. It seems that we can efficiently detect assumed change points with low false positive result in window size greater than 20 if appropriate threshold level (OR.level) is used. Since the calculation and computation speed will be slowed down in a bigger window, we will select window size n=100 for the following change point data analysis.

Location of Change Point To understand the impact of change point location within the window and to explore detection resolution, we simulated m=1000 times of control, one copy gain and one copy loss data with segment length n=100 and change points at k=50, 25, 10 and 5. The empirical distribution of mlnOR showed that mlnORs are increased when change point is located closer to central line in both one copy gain and loss data. mlnORs are not changed significantly in control data simulations (see Figure 2.7). Even at k=5, a threshold level (e.g. mlnOR=5.0) can be identified to make an appropriate inference between no change point and one change point model with high sensitivity and specificity.

OR.level The distributions of mlnORs are unknown and at least do not follow normal distribution based on Shapiro-Wilk normality test Based on the empirical cumulative distribution of mlnORs under no change point assumption in control simulations (see table 2.1), we will make type I error $\leq 0.05, \leq 0.01$, and ≤ 0.00001 type at OR.level=4.0, 5.0 and 8.0 respectively. The selection of OR.level will allow us to detect change points at various degree of resolution with high sensitivity and specificity.

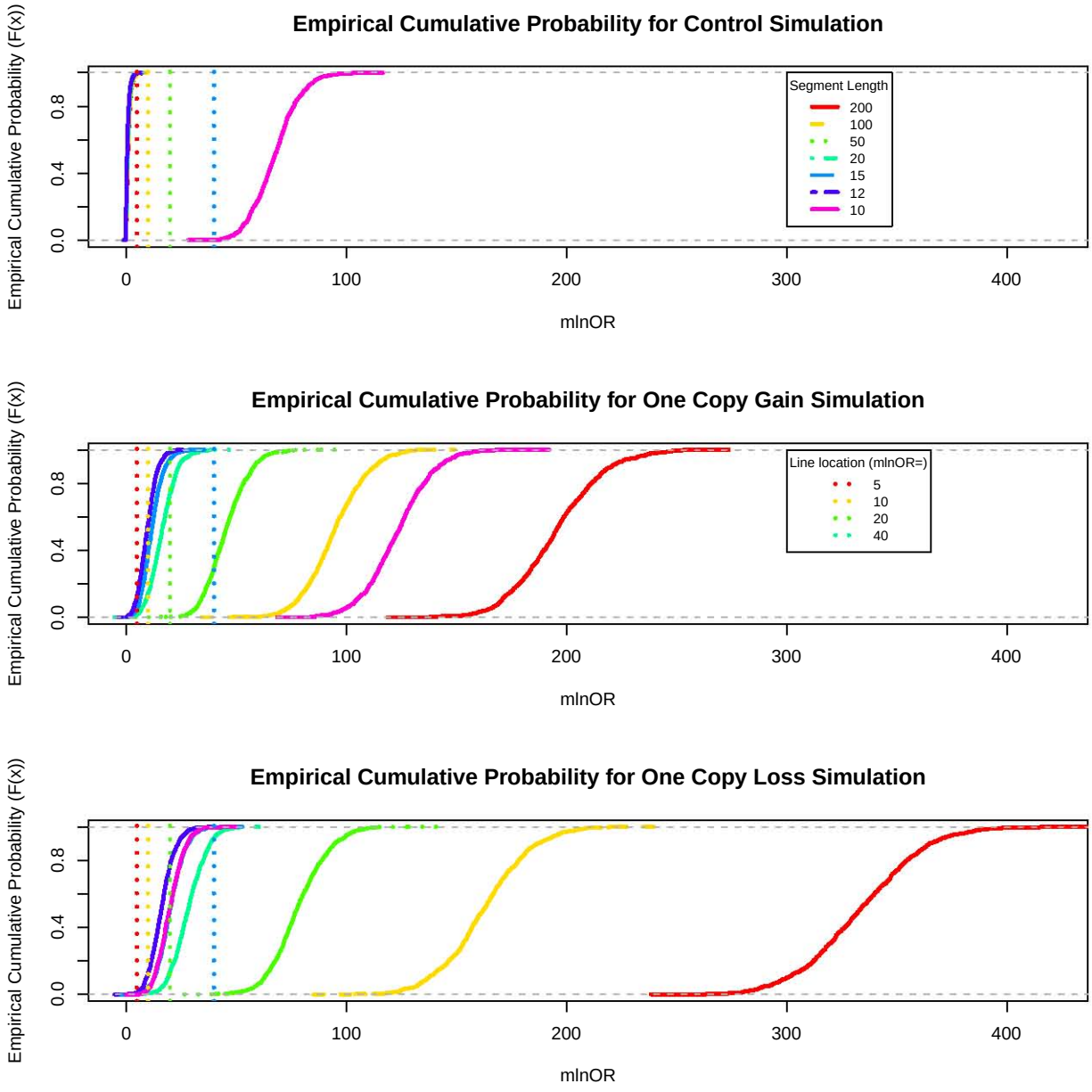

Figure 2.6: The impact of segment length on the empirical distribution of maximum log posterior odds ratio (mlnOR) generated by BayNormal based on m=1000 times of control (top), one copy gain (middle) and one copy loss (bottom) single change point simulations. Segment length n=10, 12, 15, 20, 50, 100, 200 with change point k in the middle

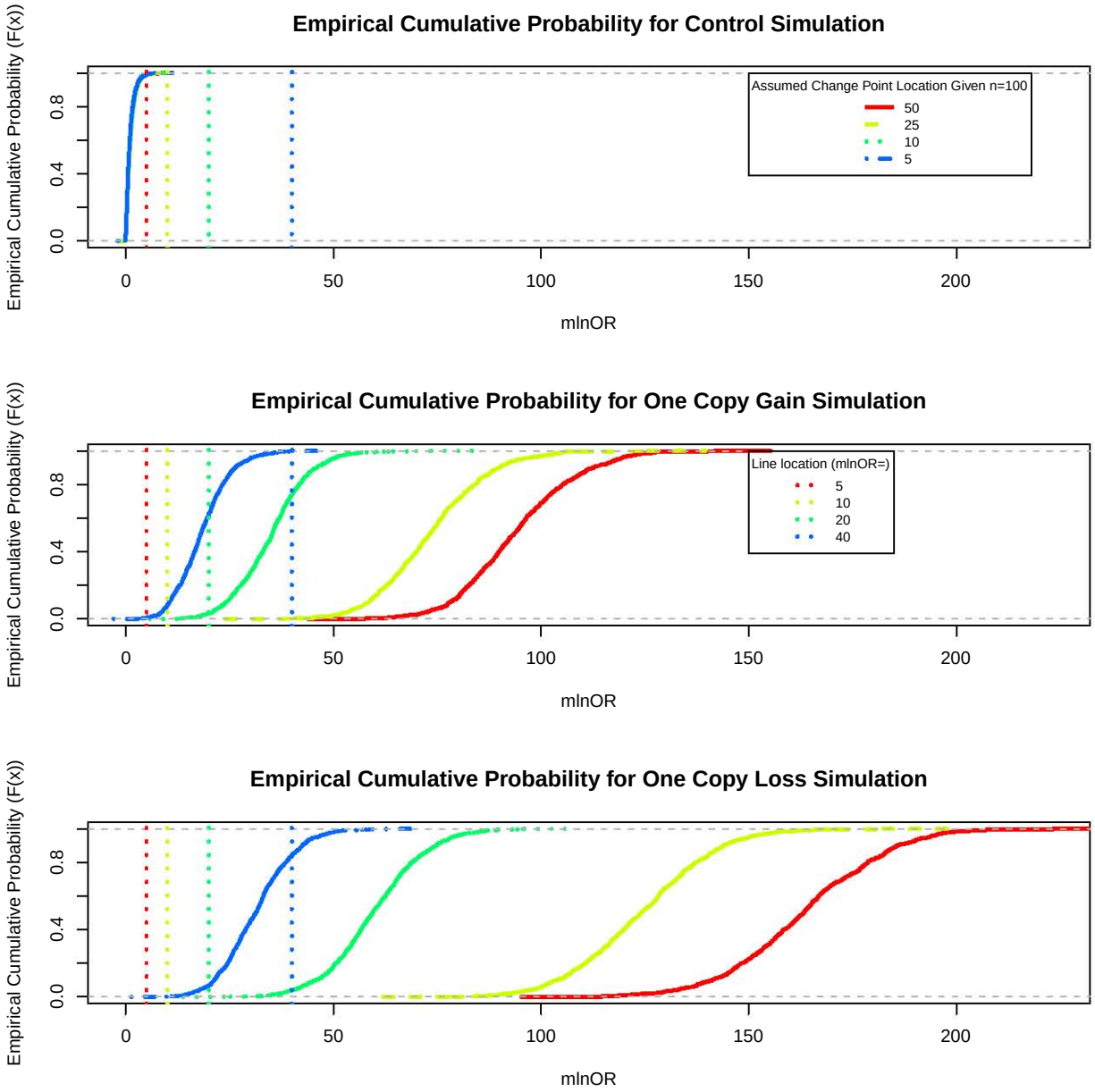

Figure 2.7: Impact of the assumed change point location on the empirical cumulative distribution of mlnOR generated by BayNormal based on m=1000 times of control (top), one copy gain (middle) and one copy loss single change simulations. Segment length n=100 and change points k=5, 10, 25 and 50.

Table 2.1: Empirical Cumulative Distribution of Maximum Log Posterior Odds Ratio (mlnOR) by BayNormal

Prob	**mlnOR**		
	Control	**Gain**	**Loss**
0.00%	-0.105	57.738	110.069
0.01%	-0.104	57.806	110.083
0.05%	-0.099	58.109	110.144
0.06%	-0.098	58.185	110.159
0.08%	-0.096	58.298	110.182
0.10%	-0.093	58.488	110.220
0.50%	-0.061	61.838	117.737
1%	-0.041	64.239	122.794
5%	0.031	71.476	133.934
50%	0.778	94.763	161.449
95%	3.319	119.275	191.869
97.50%	3.904	124.915	197.954
98%	4.064	125.448	201.500
98.50%	4.296	126.754	205.342
99%	4.737	129.867	209.565
99.90%	7.472	144.519	222.749
99.99%	7.511	150.216	229.121
100.00%	7.515	150.785	229.759

2.7.1.2 Evaluation on Single Change Point Data

To evaluate the efficiency of BayNormal, we first simulated no change point control, one copy gain and one copy loss data as described in data section. Then, BayNormal with setting of OR.level =10.0 as mlnOR threshold has been run on m=1000 times of data simulations. Table 2.2 indicates that positive rate for detecting one copy gain and loss can reach 100% but false positive rate in no change control is low to zero. The positive rate of change point identification for change point k which matches within only one bin deviation from the estimated change point $(49 - 51)$ can be more than 95% for one copy gain data and 99.0% for one copy loss data but zero 0% for no change control data showing its high sensitivity in accurate identification of change point location with low false positive rate. The resulting estimates for read count and copy number are showed in table 2.2 also indicating high accuracy and precision.

2.7.1.3 Evaluation on Multiple Change Point Data

Multiple change point data with six segments and four assumed change points were generated according to description in data section. The evaluations of one copy gain , one copy loss and control detection were based on assumed change $k_1 = 50$, $k_3 = 150$, and a region between $k_4 = 200$ and n=300 in m=1000 times of simulations respectively. The results by BayNormal with OR.level=10.0 and sliding window size= 100 in table 2.3. showed that positive rate of change point identification for exact match to the assumed change points can reach to 87.5%, 94.9% for one copy gain and one copy loss respectively. The false positive rate of change point identification with at least one change point identified in the region between $k_4 = 202$ and n=299 is 0.1%. The slight increases in PRCPIs in comparison to single change point detection is probably due to multiple tests on one change point through sliding windows. The false change point rate over the whole segment n=300 is 2.741% on average.

Table 2.2: Evaluation of Single Change Point Detection at OR.level=10.0 by BayNormal

Indices	Gain	Loss	Control
PR	1.000	1.000	0.000
PRCPI	0.833	0.928	0.000
PRCPI1	0.960	0.993	0.000
Mean(k)	49.980	49.991	100.000
MSE(k)	0.400	0.081	0.000
Mean(λ_1)	39.983	39.965	40.000
MSE(λ_1)	0.278	1.243	0.000
Mean(C1)	1.999	1.998	2.000
MSE(C1)	0.000	0.000	0.000
Mean(λ_2)	60.073	20.010	40.000
MSE(λ_2)	5.261	0.106	0.000
Mean(C2)	3.004	1.001	2.000
MSE(C2)	0.000	0.000	0.000

PR: Positive Rate, Sensitivity; PRCPI: positive rate of change point identification; PRCPI1: positive rate of change point identification for k $\in k \pm 1$; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1): mean read count estimate in segment 1; MSE(λ_1): mean squared error of read counts estimate in segment 1; Mean(C1): mean copy number estimate in segment 1 (before the change point k); MSE(C1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C2): mean copy number estimate in segment 2; MSE(C2): mean squared error of copy number estimate in segment 2.

Table 2.3: Evaluation of Multiple Change Point Detection by BayNormal

Indices	Gain	Loss	Control
PRCPI	0.880	0.954	0.009
Mean(k)	49.933	149.949	295.641
MSE(k)	0.199	0.081	216.165
Mean(λ_1)	39.931	40.353	39.791
MSE(λ_1)	0.817	0.955	0.431
Mean(C_1)	1.997	2.018	1.990
MSE(C_1)	0.002	0.002	0.001
Mean(λ_2)	59.635	20.377	39.791
MSE(λ_2)	1.337	0.519	0.431
Mean(C_2)	2.982	1.019	1.990
MSE(C_2)	0.003	0.001	0.001

PRCPI: positive rate of change point identification; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1):mean read count estimate in segment 1; MSE(λ_1): mean squared error of read count estimates in segment 1; Mean(C_1):mean copy number estimate in segment 1 (before the change point k); MSE(C_1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C_2): mean copy number estimate in segment 2; MSE(C_2): mean squared error of copy number estimate in segment 2.

2.7.1.4 Evaluation on Human NGS Data

The simulations of NGS data with extra 100 copy gain or copy loss data and 200 extra assumed change points was conducted according to the description in data section. As an example, the estimated read counts and copy numbers of a segment in NGS plus one copy gain data by BayNoraml can be seen in Figure 2.8. The results by BayNormal showed that positive rate of change point identification for both one copy gain and loss data can reach greater than 90%. The false change point rate for identified extra change points in addition to simulated and unknown existing change points in human NGS data are less than 10% (see table 2.4). These results provide consistent evidences that BayNormal offers a powerful and reliable tool to identify copy number variants in human NGS data.

Table 2.4: Evaluation on Human NGS+Simulation Data by BayNormal

Indices	Gain	Loss
$PRCPI_2$	0.905	0.940
FCPR	0.034	0.078
Mean(Δk)	-0.005	1.770
MSE(Δk)	34.755	324.460
Mean($\Delta\lambda$)	-0.705	-0.307
MSE($\Delta\lambda$)	12.558	0.737
Mean(ΔC)	-0.028	-0.012
MSE (ΔC)	0.020	0.001

$PRCPI_2$: positive rate of change point identification for k$\pm$2; FCPR: False Change Point Rate; Mean(Δk): Mean Deviation of Change Point Estimate; MSE (Δk): mean squared error of change point estimate; Mean($\Delta\lambda$): mean deviation of read count estimate; MSE($\Delta\lambda$): mean squared error of copy number estimate; Mean(ΔC): mean deviation of copy number estimate; MSE(ΔC): mean squared error of copy number estimate;

2.7.2 BayGamma

R programs based on the algorithm of Bayesian approach to Poisson Change Point Model named BayGamma have been established. Similar to BayNormal, BayGamma

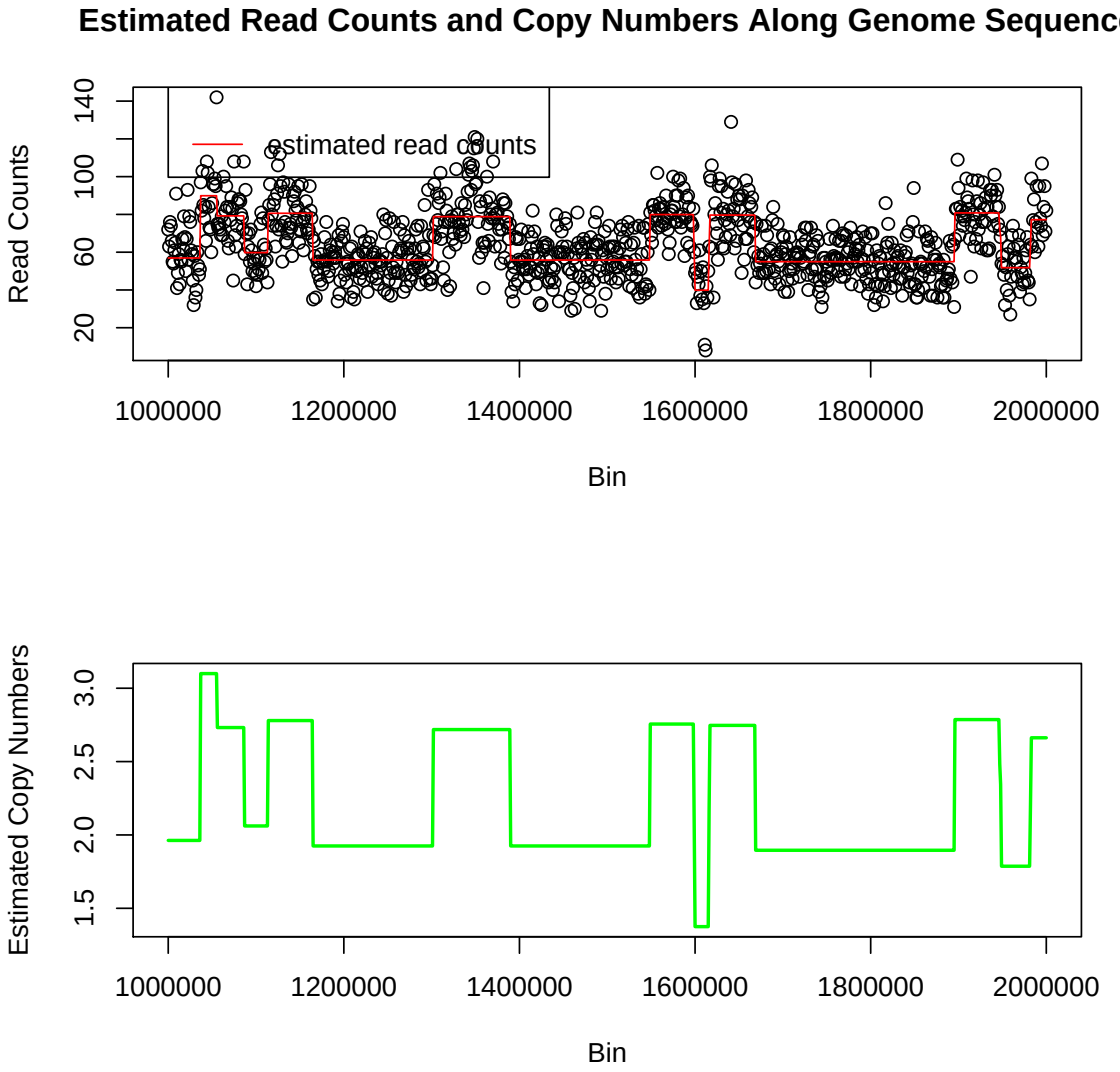

Figure 2.8: Estimated read counts and copy numbers of human NGS read count data by BayNormal.

can be used as alternative options to detect single change point, multiple change point and human NGS data in addition to common functions such as data and copy number report as well as graphic display. The optimization of test statistics threshold, window size in detection of change point and evaluation of copy number estimation in single change point, multiple change point and human NGS read count data are described in the following.

2.7.2.1 Optimization of Test Statistics and OR.level

BayGamma is named because we assume the prior distribution follows Gamma distribution which is distinguished from Scargle's prior assumption. The log posterior odds ratio (lnOR) over potential change points k=1,$\cdots$, n-1 in a segment n=100 on no change control, one copy gain and one copy loss simulations are showed in Figure 2.9. The maximum of log posterior odds is around k=50 for one copy gain and loss data but very low for no change control data.

Window Size To understand the impact of window size on statistics, BayGamma was run for m=1000 times of no change control, one copy gain and one copy loss simulations with window size from n=10 to 200. The empirical cumulative distributions of statistics mlnOR are shown in Figure 2.10. The right shift of the empirical distribution of mlnOR indicating that mlnOR is increased with longer segment length for one copy gain and one copy loss data, but mlnOR is slightly decreased for no change control. An appropriate lnOR threshold (e.g. OR.level=5.0) can be found to distinguish between one copy gain or loss even at n=10 window size. Since computation with larger window size will cause the slowdown of the program, we choose n=100 as sliding window size in the following optimization and change point analysis.

Location of Change Points To understand the impact of location of change points on test statistics and detection resolution, BayGamma was conducted on m=1000

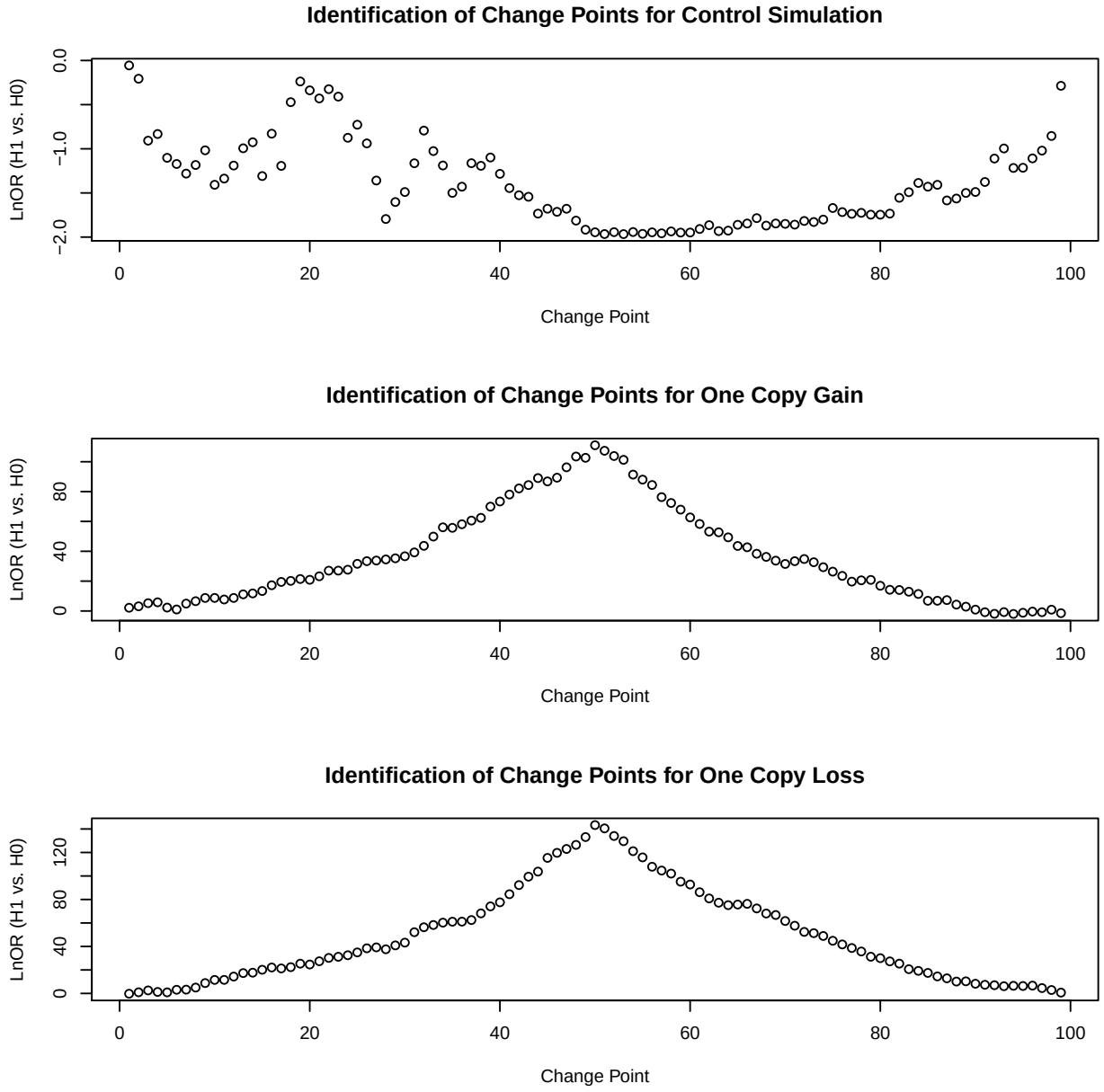

Figure 2.9: Difference in Log Posterior Odds (lnOR) generated by BayGamma at each possible change point along the segment based on a control, one copy gain and one copy loss single change point simulation.

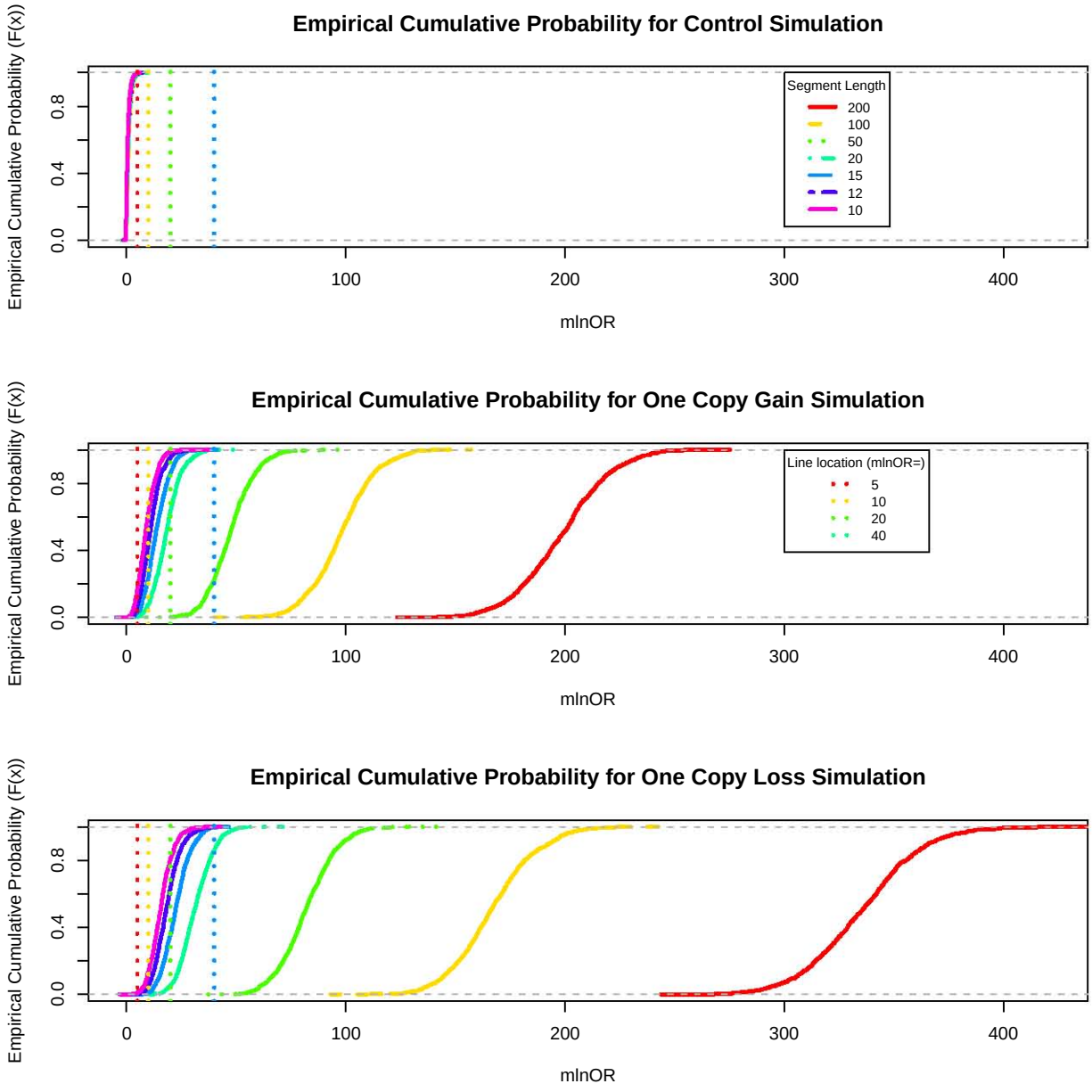

Figure 2.10: The impact of segment length on the empirical distribution of maximum log posterior odds (mlnOR) generated by BayGamma based on m=1000 times of control (top), one copy gain (middle) and one copy loss (bottom) single change point simulations. Segment length n=10, 12, 15, 20, 50, 100, 200 with change point k in the middle.

times of no change control, one copy gain and one copy loss simulations. The empirical cumulative distributions of mlnOR in Figure 2.11 indicate that mlnORs are increased when change point is closer to the central line in window size n=100. We can use appropriate OR.level (e.g., 5.0) to differentiate the one copy gain or loss from control even for a segment with 5 bins with high power and specificity. By setting different value of OR.level, we can increase the resolution of change point identification.

OR.level Most empirical cumulative distributions of mlnOR on control, one copy gain or one copy loss for window size n=100 are not consistent with normal distribution based on Shapiro-Wilk normality test . Given no change hypothesis assumed for control data simulations, type I error 5%, 1%, and 0.001% may be produced if OR.level=4.0, 6.0 or 8.0 are set as mlnOR threshold level for inference between no change point model and one change point model based on empirical distributions of mlnOR (see table 2.5). These indicate that BayGamma provides a powerful and reliable tool to identify a change point with high confidence. With respect to possible type I inflation in multiple comparisons which may occur in sliding window algorithm, OR.level =10 was used in following change point analysis.

2.7.2.2 Evaluation on Single Change Point Data

The evaluation of BayGamma on single change point data was first conducted in m=1000 times of no change control, one copy gain and one copy loss data simulations as described in data section. By setting OR.level=10.0, positive rate for one copy gain and one copy loss can reach 100% while false positive rate for no change control is zero (see table 2.6) indicating high sensitivity and low false positive rate for the inference about existence of one change point. The positive rate of change point identification showed that more than 96% of the assumed change points can be matched to one bin deviation from each estimated change point in one copy gain data

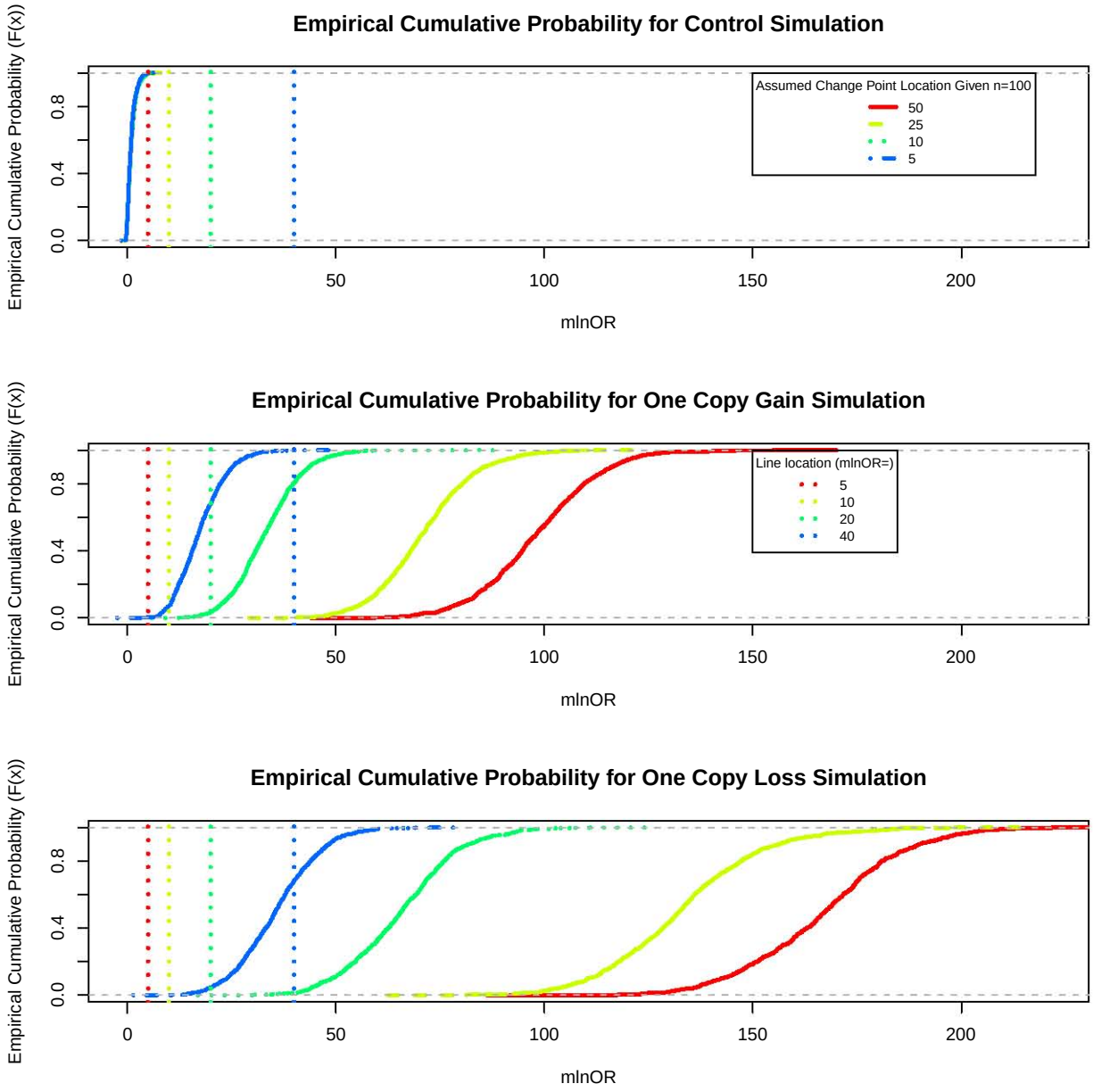

Figure 2.11: The impact of assumed change point location on the empirical cumulative distribution of mlnOR generated by BayGamma based on m=1000 times of control (top), one copy gain (middle) and one copy loss single change simulations. Segment length n=100 and change points k=5, 10, 25 and 50.

Table 2.5: Empirical Cumulative Distribution of Maximum Posterior Odds (mlnOR) by BayGamma

prob	**mlnOR**		
	Control	**Gain**	**Loss**
0.00%	-0.339	62.507	115.336
0.01%	-0.337	62.588	115.388
0.05%	-0.328	62.948	115.617
0.06%	-0.326	63.038	115.674
0.08%	-0.322	63.173	115.760
0.10%	-0.317	63.397	115.903
0.50%	-0.301	67.687	119.719
1%	-0.292	69.528	124.515
5%	-0.169	76.776	136.300
50%	0.633	97.389	166.321
95%	3.224	120.982	195.396
97.50%	4.288	124.625	201.846
98%	4.401	127.310	205.510
98.50%	4.528	129.595	207.336
99%	5.277	131.354	209.148
99.90%	7.086	142.200	215.231
99.99%	7.397	146.965	217.012
100.00%	7.428	147.441	217.190

simulations, and more than 99% in one copy loss data simulations. No change point has ever been produced in 1000 times of control data simulations. The BayGamma also produced accurate and precise estimates of read counts and copy numbers as showed in table 2.6.

Table 2.6: Evaluation of Single Change Point Detection at OR.level=10.0 by BayGamma

Indices	Gain	Loss	Control
PR	1.000	1.000	0.000
PRCPI	0.834	0.932	0.000
PRCPI1	0.965	0.993	0.000
Mean(k)	49.999	50.011	100.000
MSE(k)	0.001	0.121	0.000
Mean(λ_1)	40.052	39.987	40.017
MSE(λ_1)	2.661	0.178	0.290
Mean(C1)	2.003	1.999	2.001
MSE(C1)	0.000	0.000	0.000
Mean(λ_2)	60.038	19.993	40.017
MSE(λ_2)	1.460	0.049	0.290
Mean(C2)	3.002	1.000	2.001
MSE(C2)	0.000	0.000	0.000

PR: Positive Rate, Sensitivity; PRCPI: positive rate of change point identification; PRCPI1: positive rate of change point identification for k $\pm$1; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1):mean read count estimate in segment 1; MSE(λ_1): mean squared error of read counts estimate in segment 1; Mean(C_1):mean copy number estimate in segment 1 (before the change point k); MSE(C_1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C_2): mean copy number estimate in segment 2; MSE(C_2): mean squared error of copy number estimate in segment 2.

2.7.2.3 Evaluation on Multiple Change Point Data

The evaluation of BayGamma on multiple change point data was conducted for m=1000 times of six segment and four assumed multiple change point data simulations as described in data section. The positive rate of change point identification showed that in 86.8%, 96.1% of simulations exact assumed change point at k=50 for one copy

gain and at k=150 for one copy loss were identified respectively (table 2.7). Only 1% of simulations in a region from k=200 to 300 where no change point was assumed was accidently shown to have at least one change point based on OR.level =10.0. The mean change point estimate and MSE of change point estimate showed accurate and precision in identification of change points in one copy gain and loss data simulations although the false estimate of change point spread across the control region. The accuracy and precision can also be evidenced from the read count estimates and copy number estimates. The false change point rate with change point not within k±2 in the whole region is 1.71% on average indicating high precision in change point detection.

Table 2.7: Evaluation of Multiple Change Point Detection By BayGamma

Indices	Gain	Loss	Control
PRCPI	0.874	0.961	0.011
Mean(k)	49.958	149.965	293.305
MSE(k)	0.188	0.053	336.068
Mean(λ_1)	40.016	40.386	39.751
MSE(λ_1)	0.821	0.926	0.447
Mean(C_1)	2.001	2.019	1.988
MSE(C_1)	0.002	0.002	0.001
Mean(λ_2)	59.682	20.365	39.751
MSE(λ_2)	1.326	0.510	0.447
Mean(C_2)	2.984	1.018	1.988
MSE(C_2)	0.003	0.001	0.001

PRCPI: positive rate of change point identification; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1):mean read count estimate in segment 1; MSE(λ_1): mean squared error of read count estimates in segment 1; Mean(C_1):mean copy number estimate in segment 1 (before the change point k); MSE(C1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C_2): mean copy number estimate in segment 2; MSE(C_2): mean squared error of copy number estimate in segment 2.

2.7.2.4 Evaluation on Human NGS Data

The evaluation of copy number variants detection over human NGS data NA19239 plus random 100 extra one copy gain and one copy loss segments and 200 extra change point simulations have been performed. The results in table 2.8 showed that over 90% of the assumed change points for both one copy gain and one copy loss can be captured. The positive rate of change point identification for one copy loss is higher than that for one copy gain simulations. The false change point rate for the estimated change points which does not belong to assumed change points or unknown change points in original human NGS read data is less or equal than 10%. The false change point rate is higher for copy loss than for copy gain. The accuracy and precision of BayGamma on human NGS data can be reflected by the deviation and MSE of change point identification, read count and copy number estimation from the assumed change points.

Table 2.8: Evaluation on Human NGS+Simulation Data by BayGamma

Indices	Gain	Loss
$PRCPI_2$	0.915	0.935
FCPR	-0.004	0.100
Mean(Δk)	0.090	10.295
MSE(k)	48.090	5416.465
Mean($\Delta\lambda$)	-0.458	0.050
MSE($\Delta\lambda$)	8.835	2.546
Mean(ΔC)	-0.018	0.002
MSE (ΔC)	0.014	0.004

PRCPI2: positive rate of change point identification for k$\pm$2; FCPR: False Change Point Rate; Mean(Δk): Mean Deviation of Change Point Estimate; MSE (k): mean squared error of change point estimate; Mean($\Delta\lambda$): mean deviation of read count estimate; MSE($\Delta\lambda$): mean squared error of copy number estimate; Mean(ΔC): mean deviation of copy number estimate; MSE(ΔC):mean squared error of copy number estimate;

2.7.3 Comparisons

2.7.3.1 BayNormal vs. BayGamma

From above evaluation results, BayNormal and BayGamma showed comparative sensitivity and specificity in single change point, multiple change point simulations and human NA19239 NGS read count analysis. One exception is that at n=10 segment length, BayNormal may generate high false positive rate and lower power to detect one copy gain and even failure to detect one copy loss singe change point simulated data. This may be that small sample size may violate the normal assumption we applied in BayNormal to approximate the Poisson distribution data.

2.7.3.2 BayNormal and BayGamma vs. CBS

CBS is a popular change point detection package that is considered to be the most powerful and reliable algorithm in CNV estimation using either aCGH or next generation sequencing data. Normally distributed data is usually assumed for the application. We applied CBS algorithm to detect change point and estimate copy number in single change point and multiple change point simulated data and human NGS read count data. The results showed in table 2.9, table 2.10 and table 2.11 supported its sensitivity and specificity in detecting change points. BayGamma and BayNormal offer a better at least comparative sensitivity and specificity in change point and copy number estimation.

Meanwhile we noticed that the variation in change point estimation is bigger for CBS than BayGamma and BayNormal in singe change point detection (e.g., 15.01 vs. 0.12 for loss) (see table 2.9and table 2.6). To further understand the sources of the variations among three algorithms, we observed that the estimated change points for one copy gain and loss by CBS follow a more skewed distribution than that by BayGamma and BayNormal accounting for the difference in change point estimation (see Figure 2.12, Figure 2.13 and Figure 2.14).

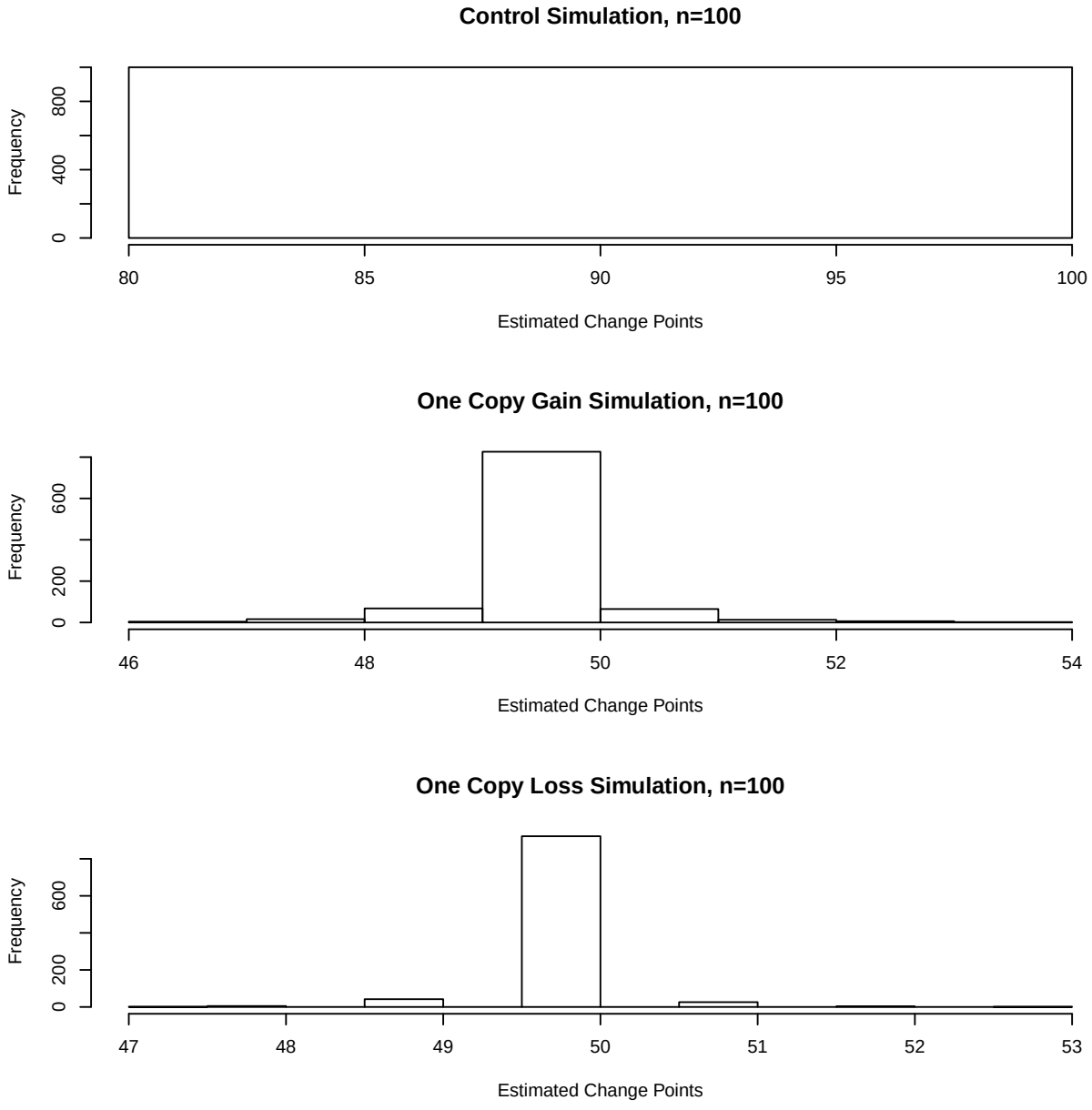

Figure 2.12: Histogram of change point identification by BayNormal in m=1000 times of control, one copy gain and one copy loss single change point simulations.

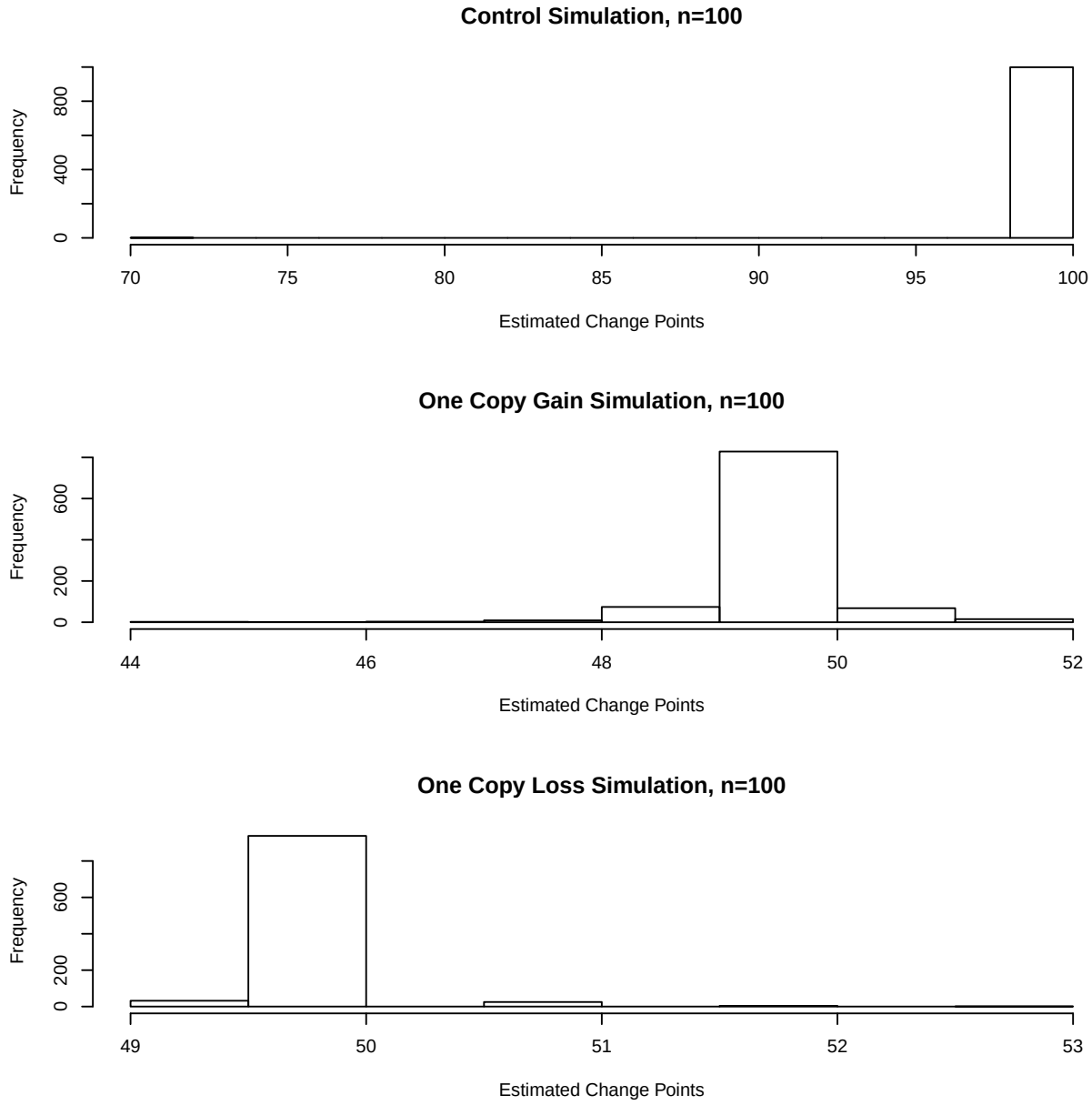

Figure 2.13: Histogram of change point identification by BayGamma in m=1000 times of control, one copy gain and one copy loss single change point simulations.

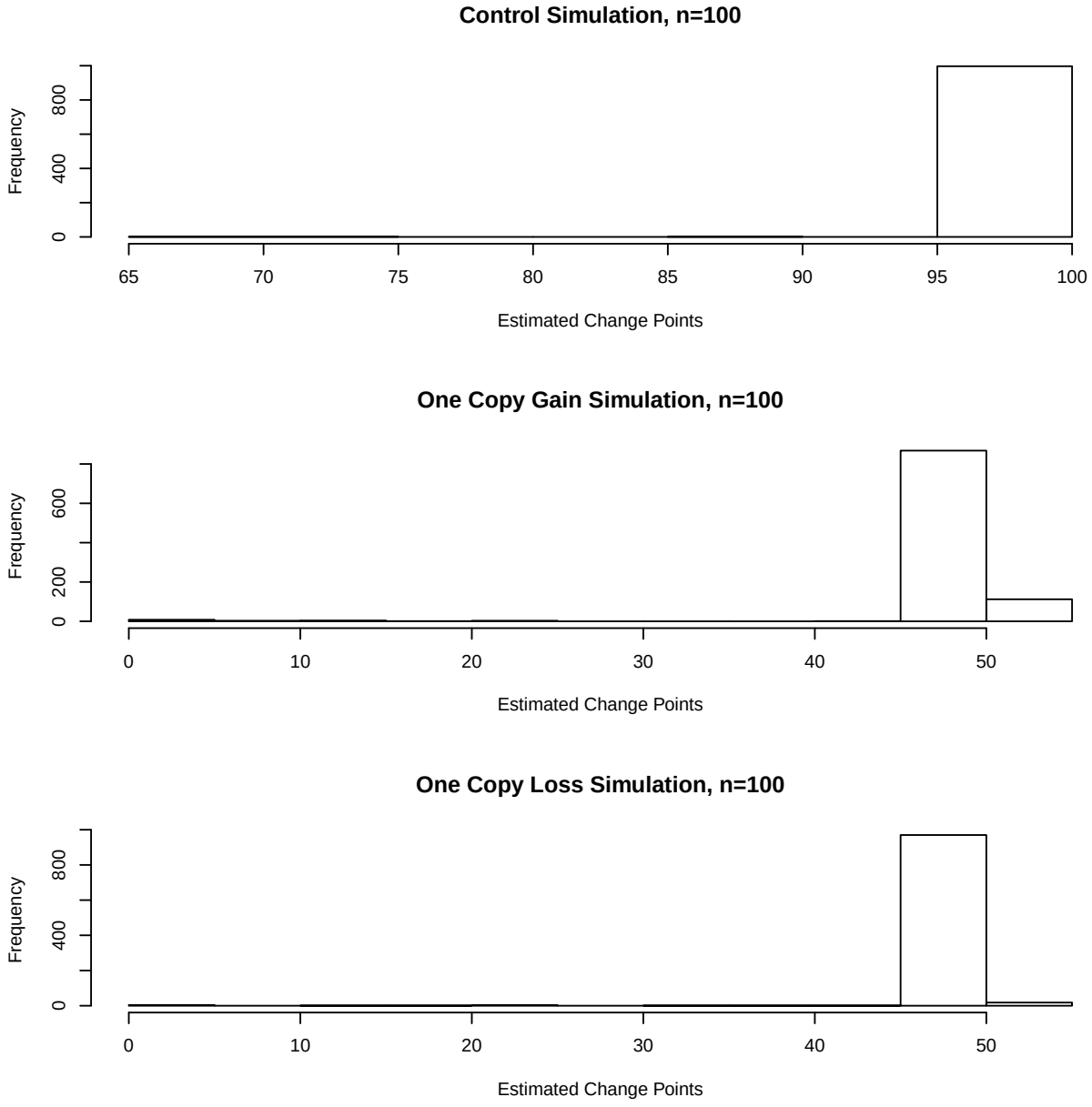

Figure 2.14: Histogram of change point identification by circular binary segmentation (CBS) in m=1000 times of control, one copy gain and one copy loss single change point simulations.

Table 2.9: Evaluation of Single Change Point Detection by CBS

Indices	Gain	Loss	Control
PR	1.000	1.000	0.009
PRCPI	0.822	0.917	0.000
PRCPI1	0.951	0.982	0.009
Mean(k)	49.625	49.543	99.458
MSE(k)	17.867	15.011	39.904
Mean(λ_1)	40.015	39.993	39.994
MSE(λ_1)	0.217	0.044	0.040
Mean(C_1)	2.001	2.000	2.000
MSE(C_1)	0.000	0.000	0.000
Mean(λ_2)	59.881	20.287	NA
MSE(λ_2)	14.169	82.606	NA
Mean(C_2)	2.994	1.014	NA
MSE(C_2)	0.000	0.000	NA

PR: Positive Rate, Sensitivity; PRCPI: positive rate of change point identification; $PRCPI_1$: positive rate of change point identification for k $\pm$1; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1):mean read count estimate in segment 1; MSE(λ_1): mean squared error of read counts estimate in segment 1; Mean(C_1):mean copy number estimate in segment 1 (before the change point k); MSE(C_1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C_2): mean copy number estimate in segment 2; MSE(C_2): mean squared error of copy number estimate in segment 2.

Table 2.10: Evaluation of Multiple Change Point Detection by CBS

Indices	Gain	Loss	Control
PRCPI	0.808	0.904	0.019
Mean(k)	49.956	149.910	295.446
MSE(k)	0.460	0.166	222.656
Mean(λ_1)	40.021	40.426	39.768
MSE(λ_1)	0.813	0.947	0.471
Mean(C_1)	2.001	2.021	1.988
MSE(C_1)	0.002	0.002	0.001
Mean(λ_2)	59.573	20.461	39.768
MSE(λ_2)	1.301	0.656	0.471
Mean(C_2)	2.979	1.023	1.988
MSE(C_2)	0.003	0.002	0.001

PRCPI: positive rate of change point identification; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(λ_1):mean read count estimate in segment 1; MSE(λ_1): mean squared error of read count estimates in segment 1; Mean(C_1):mean copy number estimate in segment 1 (before the change point k); MSE(C_1): mean squared error of copy number estimate in segment 1; Mean(λ_2): mean read count estimate in segment 2 (after the change point); MSE(λ_2): mean squared error of read count estimate in segment 2; Mean(C_2): mean copy number estimate in segment 2; MSE(C_2): mean squared error of copy number estimate in segment 2.

Table 2.11: Evaluation on Human NGS+Simulation Data by CBS

Indices	Gain	Loss
PRCPI2	0.915	0.875
FCPR	0.030	0.007
Mean(Δk)	-0.215	-7.665
MSE(k)	2.545	34362.725
Mean($\Delta\lambda$)	0.307	1.563
MSE($\Delta\lambda$)	4.898	55.092
Mean(ΔC)	0.012	0.063
MSE (ΔC)	0.008	0.088

PRCPI2: positive rate of change point identification for k$\pm$2; FCPR: False Change Point Rate; Mean(Δk): Mean Deviation of Change Point Estimate; MSE (k): mean squared error of change point estimate; Mean($\Delta\lambda$): mean deviation of read count estimate; MSE($\Delta\lambda$): mean squared error of copy number estimate; Mean(ΔC): mean deviation of copy number estimate; MSE(ΔC): mean squared error of copy number estimate;

2.8 Conclusion

We developed a normal approximation chang point algorithm and modified a Bayesian approach to Poisson change point algorithm, and built up BayNormal and BayGamma R packages to identify change points and estimate copy number variants in human NGS read count data from single genome. BayNormal and BayGamma have been optimized in the NGS read count data analysis. The evaluation on single change point, multiple change point and human NGS + simulation data showed BayGamma and BayNormal are sensitive and specific tools in CNV detection compared to CBS, a popular approach in CNV detection. The significances of these tools in CNV estimate using human NGS data are to be further explored in future.

Although several steps such as dividing big data calculation into several parts and setting speed parameters have been applied in BayGamma and BayNormal to overcome the limitations of R language in handling with big data such as slower computation speed and lower storage size, further improvements with big and faster computation facility and settings such as parallel calculation are expected to increase the potential of these R programs.

Chapter 3

NORMAL APPROXIMATION BATESIAN CHANGE POINT MODEL FOR PAIRED GENOMES

3.1 Rationale

Copy number variants have been associated with the occurrence of cancer. The identification of CNVs between cancer cells and healthy cells either from same subject or from a control subject is crucial for deciphering genetic roles in cancer development. The current detection strategy and algorithms are mostly based on comparison of ratio or difference of NGS read counts in paired genomes between cancer and control cells. The normal approximation for the ratio and difference has been widely used in these algorithms. We developed a new Bayesian approach based on more accurate normal approximation algorithms from Anscomb's work to identify the change points in paired genomes and built up R package PairedBayNormal to simulate Poisson data, estimated and display copy number variants in read count data. The evaluation result of the R package has been produced.

3.2 Models

Let Y_i be the read count of the ith bin of a cancer genome, and let Z_i be the read count of the ith bin of a control genome at the same locus i = 1, ..., n along the reference genome sequence. Then $Y_i \sim$ independent Poisson (λ_i), and $Z_i \sim$ independent Poisson (μ_i), i=1, ..., n, where $Y_i \perp\!\!\!\perp Z_i$. According to Anscombe

(1948),

$$X_i = \sqrt{Y_i + \frac{3}{8}} \qquad \sim approxiN\left(\sqrt{\lambda_i + \frac{1}{8}}, \frac{1}{4}\right), \qquad if\ \lambda_i\ is\ large \tag{3.1}$$

$$t_i = \sqrt{Z_i + \frac{3}{8}} \qquad \sim approxiN\left(\sqrt{\mu_i + \frac{1}{8}}, \frac{1}{4}\right), \qquad if\ \mu_i\ is\ large. \tag{3.2}$$

The independence of Y_i and Z_i leads to the independence of X_i with t_i since X_i is a function of only Y_i and t_i is a function of only Z_i , and the difference of X_i and t_i follows an independent normal distribution.

$$D_i = X_i - t_i \sim N\left(\sqrt{\lambda_i + \frac{1}{8}} - \sqrt{\mu_i + \frac{1}{8}}, \frac{1}{2}\right), \tag{3.3}$$

Let us think about one change point k among n bins. The kth change point separates the genome into two segments. Our research hypothesis is about the significant difference of the ith bin D_i between two segments. Thus, we assume

$$D_i \sim N\left(\sqrt{\lambda_1 + \frac{1}{8}} - \sqrt{\mu_1 + \frac{1}{8}}, \frac{1}{2}\right), \qquad for\ i = 1, \ldots, k \tag{3.4}$$

and

$$D_i \sim N\left(\sqrt{\lambda_2 + \frac{1}{8}} - \sqrt{\mu_2 + \frac{1}{8}}, \frac{1}{2}\right), \qquad for\ i = k+1, \ldots, n \tag{3.5}$$

Statistical inference for one change point analysis is basd on the following hypothesis testing. The null hypothesis is given by:

$$H_0 : \qquad F(D_1, \cdots, D_n|\lambda_1, \mu_1) = F(D_1, \cdots, D_n|\lambda_2, \mu_2) \tag{3.6}$$

vs. alternative hypothesis:

$$H_1 : \qquad F(D_1, \cdots, D_n|\lambda_1, \mu_1) \neq F(D_1, \cdots, D_n|\lambda_2, \mu_2) \tag{3.7}$$

3.2.1 No Change Point Model

Under the null hypothesis, we assume that:

$$\begin{aligned}\sqrt{\lambda_1 + \frac{1}{8}} - \sqrt{\mu_1 + \frac{1}{8}} &= \sqrt{\lambda_2 + \frac{1}{8}} - \sqrt{\mu_2 + \frac{1}{8}} \\ &= \sqrt{\lambda + \frac{1}{8}} - \sqrt{\mu + \frac{1}{8}}\end{aligned} \tag{3.8}$$

All the difference of read counts between cancer and control genomes $D_1, \cdots, D_n$ follows the same normal distribution with variance $\frac{1}{2}$ and mean $\sqrt{\lambda + \frac{1}{8}} - \sqrt{\mu + \frac{1}{8}}$. For Bayesian approach, the pdf of λ and μ prior distribution is assumed as in the following under the belief that it is related to the addition of the parameters:

$$\pi_0(\lambda, \mu) \propto \left\{ exp(-n(\sqrt{\lambda + \frac{1}{8}} + \sqrt{\mu + \frac{1}{8}})^2) \right\} \tag{3.9}$$

Then, the likelihood function is given by

$$\begin{aligned}L_0(\lambda, \mu|D_i, i = 1, \cdots, n) &= f(D_1, \ldots, D_n|\lambda, \mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\frac{1}{2}}} e^{-\frac{(D_i - \sqrt{\lambda+\frac{1}{8}} + \sqrt{\mu+\frac{1}{8}})^2}{2\frac{1}{2}}} \\ &\propto \pi^{-\frac{n}{2}} e^{-\sum_{i=1}^{n}\left(D_i - \sqrt{\lambda+\frac{1}{8}} + \sqrt{\mu+\frac{1}{8}}\right)^2}\end{aligned} \tag{3.10}$$

The joint posterior distribution of the parameters is derived from equation (3.9) and (3.10) :

$$\begin{aligned}\pi_0(\lambda,\mu|D_i,i=1,\ldots,n) &= L_0(\lambda,\mu|D_i,i=1,\ldots,n)\cdot\pi_0(\lambda,\mu)\\ &\propto exp\{-\sum_{i=1}^{n}\left(D_i-\sqrt{\lambda+\frac{1}{8}}+\sqrt{\mu+\frac{1}{8}}\right)^2\}\\ &\quad exp\left\{-n\left(\sqrt{\lambda+\frac{1}{8}}+\sqrt{\mu+\frac{1}{8}}\right)^2\right\}\\ &= g(\theta), \qquad \theta=(\lambda,\mu) \end{aligned} \tag{3.11}$$

The posterior distribution of no change is :

$$\pi_0(\underline{D}) \propto \int g(\theta)\,\mathrm{d}\theta = \int_0^\infty\int_0^\infty g(\lambda,\mu)\,\mathrm{d}\lambda\,\mathrm{d}\mu \tag{3.12}$$

Let

$$\sqrt{\lambda+\frac{1}{8}} = w_1 \Longrightarrow \lambda+\frac{1}{8} = w_1^2 \Longrightarrow \mathrm{d}\lambda = 2w_1\,\mathrm{d}w_1 \tag{3.13}$$

$$\sqrt{\mu+\frac{1}{8}} = w_2 \Longrightarrow \mu+\frac{1}{8} = w_2^2 \Longrightarrow \mathrm{d}\mu = 2w_2\,\mathrm{d}w_2 \tag{3.14}$$

Then, the equation (3.12) becomes

$$\pi_0(\underline{D}) \propto \int_{\sqrt{\frac{1}{8}}}^{\infty}\int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-\sum_{i=1}^{n}(D_i - w_1 + w_2)^2 - n\,(w_1 + w_2)^2\}\; 2w_1\,\mathrm{d}w_1 2w_2\,\mathrm{d}w_2$$

$$= \int_{\sqrt{\frac{1}{8}}}^{\infty}\int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-\sum_{i=1}^{n}(D_i^2 + w_1^2 + w_2^2 - 2D_i w_1 + 2D_i w_2 - 2w_1 w_2)$$

$$- n(w_1^2 + w_2^2 + 2w_1 w_2)\}4w_1 w_2\,\mathrm{d}w_1\,\mathrm{d}w_2$$

$$= \int_{\sqrt{\frac{1}{8}}}^{\infty}\int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-\sum_{i=1}^{n} D_i^2 - 2nw_1^2 - 2nw_2^2 + 2n\bar{D}w_1 - 2n\bar{D}w_2\}$$

$$4w_1 w_2\,\mathrm{d}w_1\,\mathrm{d}w_2$$

$$= e^{-\sum_{i=1}^{n} D_i^2} \int_{\sqrt{\frac{1}{8}}}^{\infty}\int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-2nw_1^2 - 2nw_2^2 + 2n\bar{D}w_1 - 2n\bar{D}w_2\}$$

$$4w_1 w_2\,\mathrm{d}w_1\,\mathrm{d}w_2$$

$$= e^{-\sum_{i=1}^{n} D_i^2} \int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-2nw_1^2 + 2n\bar{D}w_1\}2w_1\,\mathrm{d}w_1$$

$$\cdot \int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-2nw_2^2 - 2n\bar{D}w_2\}2w_2\,\mathrm{d}w_2$$

$$= e^{-\sum_{i=1}^{n} D_i^2} \int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-2n(w_1 - \frac{1}{2}\bar{D})^2 + \frac{n}{2}\bar{D}^2\}2w_1\,\mathrm{d}w_1$$

$$\cdot \int_{\sqrt{\frac{1}{8}}}^{\infty} exp\{-2n(w_2 + \frac{1}{2}\bar{D})^2 + \frac{n}{2}\bar{D}^2\}2w_2\,\mathrm{d}w_2 \qquad (3.15)$$

where $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$.

Introducing $w_1 - \frac{\bar{D}}{2} = s_1,\ \mathrm{d}w_1 = \mathrm{d}s_1$ and $w_2 + \frac{\bar{D}}{2} = s_2,\ \mathrm{d}w_2 = \mathrm{d}s_2$ results in the following:

$$
\begin{aligned}
\pi_0(\underline{D}) \propto & \, exp\{-\sum_{i=1}^{n} D_i^2 + n\bar{D}^2\} \cdot \int_{\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2}}^{\infty} exp\{-2ns_1^2\}\ 2(s_1 + \frac{\bar{D}}{2})\,\mathrm{d}s_1 \\
& \cdot \int_{\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2}}^{\infty} exp\{-2ns_2^2\}\ 2(s_2 - \frac{\bar{D}}{2})\,\mathrm{d}s_2 \\
& = exp\{-\sum_{i=1}^{n} D_i^2 + n\bar{D}^2\} \left[\int_{\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2}}^{\infty} e^{-2ns_1^2}\ \mathrm{d}s_1^2 + \bar{D} \int_{\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2}}^{\infty} e^{-2ns_1^2}\,\mathrm{d}s_1 \right] \\
& \cdot \left[\int_{\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2}}^{\infty} e^{-2ns_2^2}\ \mathrm{d}s_2^2 - \bar{D} \int_{\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2}}^{\infty} e^{-2ns_2^2}\,\mathrm{d}s_2 \right] \\
& = exp\{-\sum_{i=1}^{n} D_i^2 + n\bar{D}^2\} \\
& \cdot \left[-\frac{1}{2n} e^{-2ns_1^2} \,|_{\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2}}^{\infty} + \bar{D} \int_{\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2}}^{\infty} e^{-\frac{(2\sqrt{n}s_1)^2}{2}} \frac{1}{2\sqrt{n}}\,\mathrm{d}(2\sqrt{n}s_1) \right] \\
& \cdot \left[-\frac{1}{2n} e^{-2ns_2^2} \,|_{\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2}}^{\infty} - \bar{D} \int_{\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2}}^{\infty} e^{-\frac{(2\sqrt{n}s_2)^2}{2}} \frac{1}{2\sqrt{n}}\,\mathrm{d}(2\sqrt{n}s_2) \right] \\
& = exp\{-\sum_{i=1}^{n} D_i^2 + n\bar{D}^2\}
\end{aligned}
$$

$$\cdot \left[\frac{1}{2n} e^{-2n(\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2})^2} + \bar{D} \int_{\sqrt{\frac{n}{2}}-\sqrt{n}\bar{D}}^{\infty} e^{-\frac{z^2}{2}} \frac{1}{2\sqrt{n}} \, \mathrm{d}z \right]$$

$$\cdot \left[\frac{1}{2n} e^{-2n(\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2})^2} - \bar{D} \int_{\sqrt{\frac{n}{2}}+\sqrt{n}\bar{D}}^{\infty} e^{-\frac{z^2}{2}} \frac{1}{2\sqrt{n}} \, \mathrm{d}z \right] \tag{3.16}$$

where $z = 2\sqrt{n}s_1$ or $z = 2\sqrt{n}s_2$. Therefore,

$$\begin{aligned} \pi_0(\underline{D}) \propto & \, exp\{-\sum_{i=1}^{n} D_i^2 + n\bar{D}^2\} \\ & \cdot \left[\frac{1}{2n} e^{-2n(\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2})^2} + \bar{D}\sqrt{\frac{\pi}{2n}}(1 - \phi(\sqrt{\frac{n}{2}} - \sqrt{n}\bar{D})) \right] \\ & \cdot \left[\frac{1}{2n} e^{-2n(\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2})^2} - \bar{D}\sqrt{\frac{\pi}{2n}}(1 - \phi(\sqrt{\frac{n}{2}} + \sqrt{n}\bar{D})) \right] \end{aligned} \tag{3.17}$$

where $\phi(z) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, \mathrm{d}z$.

3.2.2 One Change Point Model

Under our research hypothesis, the genome sequence is separated at change point k into two segments so that the differences of read counts between cancer and control at the ith bin follow normal distribution as showed in equation (3.4) and (3.5). The differences of read counts between case and control genomes follow same distribution within each segment but follow different distributions between segments.

For Bayesian approach, we assume $k \sim$ independent discrete uniform (n-1) since k can be any point between 1 and n-1 with equal probability. The pdf of k as a prior distribution is:

$$\pi_0(k) = \begin{cases} \frac{1}{n-1} & , \; for \; k = 1, \ldots, n-1 \\ 0 & , \; otherwise. \end{cases} \tag{3.18}$$

and the prior probability for $\lambda_1, \mu_1, \lambda_2, \mu_2$ is as in the following according to the

independence between two segments:

$$\begin{aligned}\pi_0(\lambda_1,\mu_1,\lambda_2,\mu_2|k) \propto \pi_0(\lambda_1,\mu_1|k)\pi_0(\lambda_2,\mu_2|k) \\ \left\{exp(-k(\sqrt{\lambda_1+\frac{1}{8}}+\sqrt{\mu_1+\frac{1}{8}})^2)\right\} \\ \cdot\left\{exp(-(n-k)(\sqrt{\lambda_2+\frac{1}{8}}+\sqrt{\mu_2+\frac{1}{8}})^2)\right\} \qquad (3.19)\end{aligned}$$

The likelihood function of the compound one change point model is, by the independence assumption discussed above, just the product of the probabilities of the two segments considered separately.

$$\begin{aligned}L(\lambda_1,\mu_1,\lambda_2,\mu_2|D_1,\cdots,D_n,k) &= f(D_1,\cdots,D_n|\lambda_1,\mu_1,\lambda_2,\mu_2,k) \\ &= f(D_1,\cdots,D_k|\lambda_1,\mu_1,k) \\ &\cdot f(D_{k+1},\cdots,D_n|\lambda_2,\mu_2,n-k) \qquad (3.20)\end{aligned}$$

The joint distribution for parameters $\lambda_1,\mu_1,\lambda_2,\mu_2,k$ under alternative hypothesis H_1 can be derived as

$$\begin{aligned}\pi_1(\lambda_1,\mu_1,\lambda_2,\mu_2,k|D_1,\cdots,D_n) &= f(D_1,\cdots,D_n|\lambda_1,\mu_1,\lambda_2,\mu_2,k) \\ &\cdot \pi_0(\lambda_1,\mu_1,\lambda_2,\mu_2|k)\pi_0(k) \qquad (3.21)\end{aligned}$$

The posterior distribution for the kth change point is:

$$\pi_1(k) \propto \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty \pi_1(\lambda_1, \mu_1, \lambda_2, \mu_2, k|D_1, \cdots, D_n)\, \mathrm{d}\lambda_1\, \mathrm{d}\mu_1\, \mathrm{d}\lambda_2\, \mathrm{d}\mu_2$$

$$\propto \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty f(D_1, \cdots, D_k|\lambda_1, \mu_1, k)\pi_0(\lambda_1, \mu_1|k)$$

$$\cdot f(D_{k+1}, \cdots, D_n|\lambda_2, \mu_2, n-k)\pi_0(\lambda_2, \mu_2|k)\pi_0(k)\, \mathrm{d}\lambda_1\, \mathrm{d}\mu_1\, \mathrm{d}\lambda_2\, \mathrm{d}\mu_2$$

$$\propto \int_0^\infty \int_0^\infty f(D_1, \cdots, D_k|\lambda_1, \mu_1, k)\pi_0(\lambda_1, \mu_1|k)\, \mathrm{d}\lambda_1\, \mathrm{d}\mu_1$$

$$\cdot \int_0^\infty \int_0^\infty f(D_{k+1}, \cdots, D_n|\lambda_2, \mu_2, n-k)\pi_0(\lambda_2, \mu_2|k)\pi_0(k)\, \mathrm{d}\lambda_2\, \mathrm{d}\mu_2 \qquad (3.22)$$

Since the read counts within segment from 1 to k or from k+1 to n are assumed to follow no change point distribution, the derivation for the posterior distribution under null hypothesis (3.17) can be applied to the above equation (3.22) for both segments separately and result in the following posterior probability of one change point at k.

$$\pi_1(k) \propto exp\{-\sum_{i=1}^{k} D_i^2 + k\bar{D_1}^2\} \left[\frac{1}{2k} e^{-2k(\sqrt{\frac{1}{8}} - \frac{\bar{D_1}}{2})^2} + \bar{D_1}\sqrt{\frac{\pi}{2k}}(1 - \phi(\sqrt{\frac{k}{2}} - \sqrt{k}\bar{D_1}))\right]$$
$$\cdot \left[\frac{1}{2k} e^{-2k(\sqrt{\frac{1}{8}} + \frac{\bar{D_1}}{2})^2} - \bar{D_1}\sqrt{\frac{\pi}{2k}}(1 - \phi(\sqrt{\frac{k}{2}} + \sqrt{k}\bar{D_1}))\right]$$

$$
\begin{aligned}
&\cdot exp\{-\sum_{i=k+1}^{n} D_i^2 + (n-k)\bar{D_n}\} \\
&\cdot \left[\frac{1}{2(n-k)} e^{-2(n-k)(\sqrt{\frac{1}{8}}-\frac{\bar{D_n}}{2})^2} + \bar{D_n}\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}} - \sqrt{n-k}\bar{D_n}))\right] \\
&\cdot \left[\frac{1}{2(n-k)} e^{-2(n-k)(\sqrt{\frac{1}{8}}+\frac{\bar{D_n}}{2})^2} - \bar{D_n}\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}} + \sqrt{n-k}\bar{D_n}))\right] \\
&= exp\{-\sum_{i=1}^{n} D_i^2 + k\bar{D_1}^2 + (n-k)\bar{D_n}\} \\
&\cdot \left[\frac{1}{2k} e^{-2k(\sqrt{\frac{1}{8}}-\frac{\bar{D_1}}{2})^2} + \bar{D_1}\sqrt{\frac{\pi}{2k}}(1-\phi(\sqrt{\frac{k}{2}} - \sqrt{k}\bar{D_1}))\right] \\
&\cdot \left[\frac{1}{2k} e^{-2k(\sqrt{\frac{1}{8}}+\frac{\bar{D_1}}{2})^2} - \bar{D_1}\sqrt{\frac{\pi}{2k}}(1-\phi(\sqrt{\frac{k}{2}} + \sqrt{k}\bar{D_1}))\right] \\
&\cdot \left[\frac{1}{2(n-k)} e^{-2(n-k)(\sqrt{\frac{1}{8}}-\frac{\bar{D_n}}{2})^2} + \bar{D_n}\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}} - \sqrt{n-k}\bar{D_n}))\right] \\
&\cdot \left[\frac{1}{2(n-k)} e^{-2(n-k)(\sqrt{\frac{1}{8}}+\frac{\bar{D_n}}{2})^2} - \bar{D_n}\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}} + \sqrt{n-k}\bar{D_n}))\right]
\end{aligned}
\tag{3.23}
$$

where $\bar{D_1} = \frac{1}{k}\sum_{i=1}^{k} D_i$, $\bar{D_n} = \frac{1}{n-k}\sum_{k+1}^{n} D_i$ and $\phi(z) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,\mathrm{d}z$.

3.2.3 H_0 vs.H_1 and Change Point

Similarly, we define the posterior probability odds ratio (OR) of one change point at k vs. no change point as:

$$
OR_k = \frac{\pi_1(k)}{\pi_0(\underline{D})} \tag{3.24}
$$

where $\pi_1(k)$ and $\pi_0(\underline{D})$ can be obtained from equation (3.17) and (3.23). The natural logarithm of OR_k can be subsequently deduced by the substitution of equation (3.17) and (3.23) as:

$$
\begin{aligned}
lnOR_k = & \sum_{i=1}^{n} D_i^2 - n\bar{D}^2 \ - \sum_{i=1}^{n} D_i^2 + k{\bar{D}_1}^2 + (n-k)\bar{D}_n\} \\
& + log(\left[\frac{1}{2k}e^{-2k(\sqrt{\frac{1}{8}}-\frac{\bar{D}_1}{2})^2} + \bar{D}_1\sqrt{\frac{\pi}{2k}}(1-\phi(\sqrt{\frac{k}{2}}-\sqrt{k}\bar{D}_1))\right] \\
& \cdot \left[\frac{1}{2k}e^{-2k(\sqrt{\frac{1}{8}}+\frac{\bar{D}_1}{2})^2} - \bar{D}_1\sqrt{\frac{\pi}{2k}}(1-\phi(\sqrt{\frac{k}{2}}+\sqrt{k}\bar{D}_1))\right] \\
& \cdot \left[\frac{1}{2(n-k)}e^{-2(n-k)(\sqrt{\frac{1}{8}}-\frac{\bar{D}_n}{2})^2} + \bar{D}_n\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}}-\sqrt{n-k}\bar{D}_n))\right] \\
& \cdot \left[\frac{1}{2(n-k)}e^{-2(n-k)(\sqrt{\frac{1}{8}}+\frac{\bar{D}_n}{2})^2} - \bar{D}_n\sqrt{\frac{\pi}{2(n-k)}}(1-\phi(\sqrt{\frac{n-k}{2}}+\sqrt{n-k}\bar{D}_n))\right]) \\
& - log(\left[\frac{1}{2n}e^{-2n(\sqrt{\frac{1}{8}}-\frac{\bar{D}}{2})^2} + \bar{D}\sqrt{\frac{\pi}{2n}}(1-\phi(\sqrt{\frac{n}{2}}-\sqrt{n}\bar{D}))\right] \\
& \cdot \left[\frac{1}{2n}e^{-2n(\sqrt{\frac{1}{8}}+\frac{\bar{D}}{2})^2} - \bar{D}\sqrt{\frac{\pi}{2n}}(1-\phi(\sqrt{\frac{n}{2}}+\sqrt{n}\bar{D}))\right])
\end{aligned}
\tag{3.25}
$$

Following the same steps for single genome, the maximum of log posterior probability odds ratio (mlnOR) test statistics is compared to a threshold OR.level for determination of significance in favor of $H_1 vs. H_0$.. The OR.level is inferred from the empirical cumulative distribution of mlnORs in Monte Carlo simulations with the assumption that one copy number difference (gain or loss) or no copy number difference is found between control and test sample. The read counts from NGS read count data in a two copy number genome will be used for the simulations. The optimization of OR.level is performed so that high sensitivity and low false positive rate can be achieved in the tested data.

Once the significant result is found, the change point k will be identified based on mlnOR: $\hat{k} = argmax_k\ mlnOR$.

If the test statistics is less than or equal to OR.level, it is claimed as no significance

for one change point model in favor of no change point model. Therefore, no change point in the segment is identified and end of the data sequence is outputed.

3.3 Multiple Change Point Decomposition

Similarly, we adapted a sliding window algorithm for identification of change points and segmentation. The identification of change point is conducted within the window to determine significance for the test and for the change point identification with the maximum posterior odds (mlnOR). The identified change points or end of the data sequences are outputted based on positive or negative result from comparison of test statistics to the chosen OR.level. Then, a new window is chosen along the genome sequence with certain distance (e.g. sp=10) and the same procedure is applied for identification of next change point. The segment between two adjacent change points is considered following homogeneous distribution. The average read counts per bin within the segment is calculated as estimated parameter of the expected homogeneous distribution in the segment. The copy number can be calculated from read counts as described in the following.

The sliding window procedure allow optimization of algorithm to significantly detect the read counts change sensitively and specifically and efficiently based on detection resolution requirement and computation speed. The number of multiple comparisons can be controlled so that inflated type I error can be minimized in multiple comparisons.

3.4 Data

3.4.1 One Change Point Read Count Data Simulation in Paired Genomes

No DNA Copy Change Control We follow the same data simulation procedures as that for single genome to generate nc=100 control read count data Z_{ij} with parameter

$\mu_1 = \mu_2 = 40$ and then generate another set of nt=100 control read count data Y_{ij} with parameter $\lambda_1 = \lambda_2 = 40$ for m=1,000 times.The difference between Y_{ij} and Z_{ij} is expected to indicate no DNA copy change in the second segment.

One DNA Copy Gain We follow the same procedures as above to generate nc=100 control read count data Z_{ij} with parameter $\mu_1 = \mu_2 = 40$, and then generate nt=100 read count data for one DNA copy gain Y_{ij} with parameters $\lambda_1 = 40$, and $\lambda_2 = 60$ for m=1,000 times. The difference between Y_{ij} and Z_{ij} should reflect one extra DNA copy gain in the second segment. The assumed change point should be at the k=50th bin position.

One DNA Copy Loss We follow the same procedures as above to simulate nt=100 control read count data Z_{ij} with parameter $\mu_1 = \mu_2 = 40$, and then generate nt=100 read count data for DNA copy loss Y_i with parameters $\lambda_1 = 40$ and $\lambda_2 = 20$ for m=1,000 times. The difference between Y_{ij} and Z_{ij} is expected to reflect one DNA copy loss in the second segment. The assumed change point should be at the k=50th bin position .

3.4.2 Multiple Change Point Data Simulation in Paired Genomes

We followed the same procedure as the above multiple change point data simulation for single genome read count to generate six segments with one copy gain at k_1, one copy loss at k_3 and no copy change control at k_5 following R function MultiPois.Data as test sample. Then we produce similar six segments but with $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6 = 40$ as control sample. The Positive Rate of Change Point Identification for one copy gain was based on identification of k_1, the positive ration of change point identification for loss was based on identification of k_3

and the false positive rate of change point identification was based on identification of any change point in the region from k_4 to n in m=1000 simulations.

3.4.3 Next Generation Sequencing Datasets

NGS DNA copy control data: we used the low coverage read count data with bin size 1kb in chromosome 6 of NA19239 from the 1000Genome Sequencing Project as the NGS DNA copy control data as described in the section for single genome. The read counts Z_{ij} in a segment defined by the random location corresponding to $j = 1, \cdots, m$ location from test sample (either gain or loss simulation data) are used for evaluation.

NGS DNA copy gain data: We followed the same procedure of NGS DNA copy gain data for single genome to generate NGS DNA copy gain data for test sample in paired genomes. We obtained m=100 segments with one copy gain for consecutive l=50 bins and randomly located in NGS human genome sequence. Total 2m=200 change points which define one DNA copy gain randomly around the genome sequence are expected to be identified by the proposed approaches. The identification of these change points and estimation of copy number variation based on read counts in these regions Y_{ij} were evaluated for sensitivity and specificity of the proposed approaches on one copy gain.

NGS DNA copy loss data: Similarly we got m=100 segments with one copy loss for consecutive l=50 bins and randomly located in NGS human genome sequence. Total 2m=200 extra change points in addition to existing change points in NGS read count data are expected to be identified. The identification of these change points and estimation of copy number based on read counts in these regions Y_{ij} are evaluated for sensitivity and specificity of the above approaches on one copy loss.

3.5 Evaluation

3.5.1 Evaluation Indices for One Change Point and Multiple Change Point Data

Sensitivity and Specificity

Positive Rate of DNA one copy gain (or loss) detection (PR, also called sensitivity) based on m one DNA copy gain (or loss) data simulations was obtained by:

$$PR = \frac{1}{m}\sum_{j=1}^{m} I(mlnOR_j > OR.level), \tag{3.26}$$

where I is an indicator function with 1 for a true event (at least one mlnOR> OR.level), 0 for a false event (not any mlnOR >OR.level, mlnOR is the maximum posterior odds ratio and OR.level is a chosen threshold level used for positive inference.

False Positive Rate of DNA one copy gain (or loss) detection (FPR) was calculated similarly except that the evaluation was based on m no DNA copy change control simulations. The specificity is 1- false positive rate.

Change Point Identification

The Positive Rate of Change Point Identification(PRCPI) for the qth change point based on m one DNA copy gain (or loss) data simulations was computed by :

$$PRCPI(k_q) = \frac{1}{m}\sum_{j=1}^{m} I(\hat{k}_q = k_q), \qquad j = 1, \cdots, m, \tag{3.27}$$

where $I(\hat{k}_{jq} = k_q)$ is an indicator function with 1 for a true event and 0 for a false event, $\hat{k}_{jq}$ is the closest estimated change point position to k_q in the j th gain (or loss) simulation or in a defined range (e.g, $k \pm 2$) for PRCPI2.

The False Positive Rate of Change Point Identification(FPRCPI) was obtained similarly except that the evaluations was based on m times of no DNA copy change

control simulations. The change point k_q is at the position which two separate segments were simulated but with no change point assumed.

Mean Change Point Estimate was obtained by

$$\bar{k}_q = \frac{1}{m}\sum_{j=1}^{m}\hat{k}_{jq}, \tag{3.28}$$

Estimated Mean Squared Error (MSE) of Change Point Estimate was obtained by

$$MSE(k_q) = \frac{1}{m}\sum_{j=1}^{m}(\hat{k}_{jq} - k_q)^2 \tag{3.29}$$

Read Count Estimate

Read Count Estimate was calculated as the average read counts per bin in the segment refined between estimated change points. If no change point was identified, end position of the data sequence is assigned as the change point. The estimated read count in the segment before ($\hat{\lambda}_{jq}$) change point $\hat{k}_{jq}$ at the qth change point in the jth simulation of test sample was computed as

$$\hat{\lambda}_{jq} = \frac{1}{\hat{k}_{jq} - \hat{k}_{j(q-1)}}\sum_{i=\hat{k}_{j(q-1)}+1}^{\hat{k}_{jq}} Y_{ij} \tag{3.30}$$

The $\hat{\mu}_{jq}$ for segment before change point in control sample was calculated as

$$\hat{\mu}_{jq} = \frac{1}{\hat{k}_{jq} - \hat{k}_{j(q-1)}}\sum_{i=\hat{k}_{j(q-1)}+1}^{\hat{k}_{jq}} Z_{ij} \tag{3.31}$$

and *difference of read count estimate* between control and test sample $\hat{\Delta}_{jq}$ in the segment before the change point $\hat{k}_{jq}$ in the jth simulation were obtained by:

$$\hat{\Delta}_{jq} = \hat{\lambda}_{jq} - \hat{\mu}_{jq} \tag{3.32}$$

where $j \in (1, \cdots, m)$, $i \in (1, \cdots, n)$, $q \in (1, 2)$ for single change point data and $q \in (1, \cdots, 6)$ for six segment multiple change point data, and $n = nt = nc$

Mean difference of read count estimate ($\bar{\Delta}_q$) in the segment before the estimated change point q was obtained by

$$\bar{\Delta}_q = \frac{1}{m} \sum_{j=1}^{m} \hat{\Delta}_{jq} \tag{3.33}$$

Estimated Mean Squared Error(MSE) of difference in read count estimates ($\hat{\Delta}_q$) for segment before the estimated change point q by

$$MSE(\hat{\Delta}_q) = \frac{1}{m} \sum_{j=1}^{m} (\hat{\Delta}_{jq} - \Delta_q)^2 \tag{3.34}$$

Copy Number Estimation

Copy Number Estimate in the segment before the qth estimated change point in the test sample was obtained by the following:

$$\hat{C_{Yjq}} = \frac{2\hat{\lambda}_{jq}}{\lambda_1} \tag{3.35}$$

and that in control sample by

$$\hat{C_{Zjq}} = \frac{2\hat{\mu}_{jq}}{\mu_1} \tag{3.36}$$

and the difference between test and control genome by

$$\hat{\Delta} C_{jq} = \hat{C}_{Yjq} - \hat{C}_{Zjq} \tag{3.37}$$

Mean Difference of Copy Number Estimates between test and control genome was obtained by

$$\bar{\Delta}C_q = \frac{1}{m}\sum_{j=1}^{m}\hat{\Delta}C_{jq} \tag{3.38}$$

Estimated Mean Squared Error of Copy Number Estimate of the copy number estimate difference between test and control genome was obtained by

$$MSE(\bar{\Delta}C_q) = \frac{1}{m}\sum_{j=1}^{m}(\hat{\Delta}C_{jq} - \Delta C_q)^2 \tag{3.39}$$

3.5.2 Evaluation Indices for NGS in Paired Genomes

Assumed Change Points: The NGS plus one copy gain or loss simulations in test genome have Q=2m assumed change points in addition to unknown change points in NA19329 NGS data sets (Q_0) in control genome. The change points k_q are randomly located in $k_1, k_{1+l}, \cdots, k_j, k_{j+l}, \cdots, k_m, k_{m+l}$. The evaluation indexes are based on identification of these 2m change points and estimation of read counts and copy number in the segments refined by two neighboring change points k_j and k_{j+l} in test genome by comparing to the control genome.

Assumed Read Counts Let Y_{ij} represent read counts in bin i which is classified in segment $j \in (1, \cdots, m)$ in test genome and Z_{ij} in control genome. We assume that read counts in segments between two change points are homogeneous and follow Poisson distribution with parameter $\lambda_1, \cdots, \lambda_m$ for test sample and $\mu_1, \cdots, \mu_m$ for control genome respectively. Assumed read count in test genome is estimated by:

$$\lambda_j = \bar{Y_j} = \frac{1}{l}\sum_{i=k_j+1}^{k_j+l} Y_{ij}, \tag{3.40}$$

and in control by

$$\mu_j = \bar{Z_j} = \frac{1}{l}\sum_{i=k_j+1}^{k_j+l} Z_{ij}, \tag{3.41}$$

Expected read count (λ) for one copy gain or loss :We used trimmed mean of NGS read count data with 10 percent of extreme values removed as described in chapter 2 for single genome.

Assumed copy number in test genome was estimated by:

$$C_{Yj} = \frac{\lambda_j}{\lambda} \tag{3.42}$$

and that in control genome by

$$C_{Zj} = \frac{\mu_j}{\lambda} \tag{3.43}$$

Similar to single genome, we expect to generate multiple estimated change points $\hat{k_q}$ by the proposed change point algorithms. Then each assumed change point was searched for matching of estimated change points based on either exact match ($\hat{k_q} = k_q$) or with certain range of assumed change points (e.g. $k_q \pm 2$). When no matched change point was found, the closest change point was assigned as the corresponding estimated change point. The following indices are used for the evaluation of test efficiency and accuracy.

Change Point Identification

Positive Rate of Change Point Identification (PRCPI, also called sensitivity) was obtained by

$$PRCPI = \frac{1}{2m} \sum_{q=1}^{2m} I(\hat{k_q} = k_q), \tag{3.44}$$

where 2m is the number of assumed change points and I is an indicator function with 1 for a true event that estimated change points matched assumed change points or within certain range and 0 for a false event.

False Change Point Rate (FCPR) was obtained by:

$$FCPR = \frac{1}{\hat{Q}}(\hat{Q} - \sum_{q=1}^{2m} I(\hat{k}_q = k_q)) \tag{3.45}$$

where $I(\hat{k_q} = k_q)$ is an indicator function with 1 for true event and 0 for false event, $\hat{Q}$ is total number of change points identified by the proposed algorithm comparing NGS plus simulation data to NGS data.

Mean Deviation of Change Point Estimate was obtained by

$$\Delta k = \frac{1}{2m} \sum_{q=1}^{2m} (\hat{k}_q - k_q) \tag{3.46}$$

Estimated Mean Squared Error of Change Point Estimation was obtained by:

$$MSE(\bar{\Delta k}) = \frac{1}{2m} \sum_{q=1}^{2m} (\hat{k_q} - k_q)^2 \tag{3.47}$$

Read Count Estimation

Read Count Estimate was calculated as mean of the read counts in a segment of test genome obtained by:

$$\hat{\lambda_j} = \frac{1}{l} \sum_{i=k_j+1}^{k_j+l} Y_{ij} \tag{3.48}$$

and that for control genome by

$$\hat{\mu_j} = \frac{1}{l} \sum_{i=k_j+1}^{k_j+l} Z_{ij} \tag{3.49}$$

and that for difference between test and control by

$$\hat{\Delta_j} = \hat{\lambda}_j - \hat{\mu}_j \tag{3.50}$$

Mean Read Count Estimate Difference between test and control genome was ob-

tained by

$$\bar{\Delta} = \frac{1}{m}\sum_{j=1}^{m}(\hat{\Delta}_j) \tag{3.51}$$

Estimated Mean Squared Error (MSE) of Read Count Estimation difference between test and control genome was obtained by

$$MSE(\bar{\Delta}) = \frac{1}{m}\sum_{j=1}^{m}(\hat{\Delta}_j - \Delta_j)^2 \tag{3.52}$$

Copy Number Estimation

Copy number estimate in the test sample was obtained by:

$$\hat{C_{Yj}} = \frac{\hat{\lambda}_j}{\lambda} \tag{3.53}$$

and that in the control sample by

$$\hat{C_{Zj}} = \frac{\hat{\mu}_j}{\mu} \tag{3.54}$$

and the difference between test and control genome by

$$\Delta\hat{C}_j = \hat{C}_{Yj} - \hat{C}_{Zj} \tag{3.55}$$

Mean Copy Number Estimate difference between test and control sample by

$$\Delta\bar{C} = \frac{1}{m}\sum_{j=1}^{m}(\Delta\hat{C}_j) \tag{3.56}$$

Estimated Mean Squared Error of Copy Number Estimate difference between test and control sample was obtained by:

$$MSE(\Delta\bar{C}) = \frac{1}{m}\sum_{j=1}^{m}(\Delta\hat{C}_j - \Delta C_j)^2 \tag{3.57}$$

3.6 Results

3.6.1 PairedBayNormal

R programs called PairedBayNormal have been developed using the normal approximation Bayesian change point model for paired genome. The optimization of test statistics and OR.level has been conducted in PairedBayNormal to identify change points and estimate copy numbers from single change point, multiple change point and human NGS read count data simulations. The indices as described above have been computed for the evaluation purpose.

3.6.1.1 Optimization of Test Statistics and OR.level for PairedBayNormal

Control Correction of lnOR The log posterior odds (lnOR) based on two control simulation (top), one copy gain vs. one control simulation (middle) and one copy loss vs. control simulation as paired samples according to the description above are showed in Figure 3.1. Although mlnOR can be found around the assumed change point at k=50 for both one copy gain and one copy loss, lnOR in two control simulations are not distributed randomly from change point k=1 to 99 and higher lnOR are found at both ends of the segment. This is expected to influence sensitivity and specificity of significant test and change point identification by mlnOR. In our PairedBayNormal, lnOR was corrected by lnOR from control samples at each location where difference of mlnOR (m Δ lnOR)between one paired test genome and one paired control was used. The corrected lnOR along the segment sequence clears the noise from control sample and improved the lnOR for identification of change point in one copy gain

and loss data simulations(see Figure 3.2).

Window Size Paired_BayNormal was run on two control simulations, one copy gain vs. control simulations and one copy loss vs. control simulations with segment size n=10 to 200 for m=1000 times to understand the impact of segment length on the power of change point detection. The empirical cumulative probability of lnOR showed that the increased segment size results in increased lnOR value for one copy gain and one copy loss but slightly decreased lnOR value in control simulation (see Figure 3.3). The results also suggest that we can have enough power to detect assumed change points but meanwhile to avoid false positive results given certain window size and OR.level are set. We choose n=100 in our following read count analysis through Paired_BayNormal programs.

Location of Change Point After running Paired_BayNormal program in m=1000 times of control, one copy gain and one copy loss simulations with segment size n=100 and location of assumed change point at k=5 to 50, the empirical cumulative distributions showed that the mlnORs are increased with increased change point location to 50 in one copy gain and one copy loss simulations but not in control simulations (see Figure 3.4). By setting appropriate mΔlnOR=OR.level, we can determine the significance level (e.g., 2.0) for identification of change point at high resolution (e.g. < 5 bins) with high power and low false positive rate.

OR.level The empirical distributions of mlnOR on controls, one copy gain vs. control and one copy loss vs. control are not consistent with normal distribution based on Shapiro-Wilk normality test . More dispersions at tails of the empirical distributions than normal distribution are observed as seen in an example of normal Q-Q plot for control data (top), one copy gain data (middle) and one copy loss data (bottom) with assumed single change point at k=50 and segment length n=100 (see Figure 3.5). Based on empirical distributions of mlnOR, type I error for identifying a change point is less than 5%, 0.01%, and 0.001% when OR.level is set at 1.0, 2.0

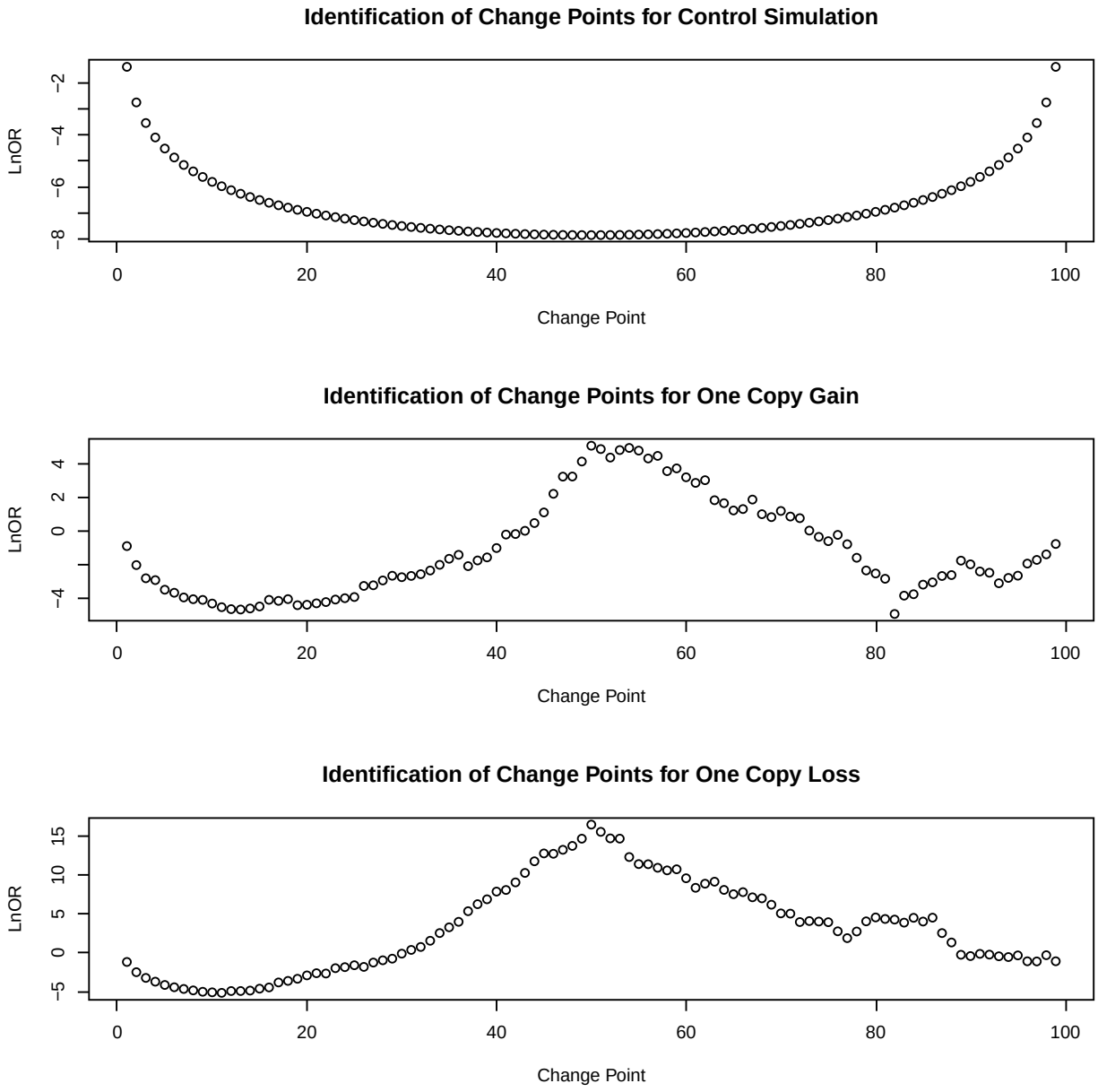

Figure 3.1: Log Posterior Odds ratio (lnOR) generated by normal approximation Bayesian change point model for paired genome at each possible change point along the segment based on a control, one copy gain and one copy loss single change point simulation.

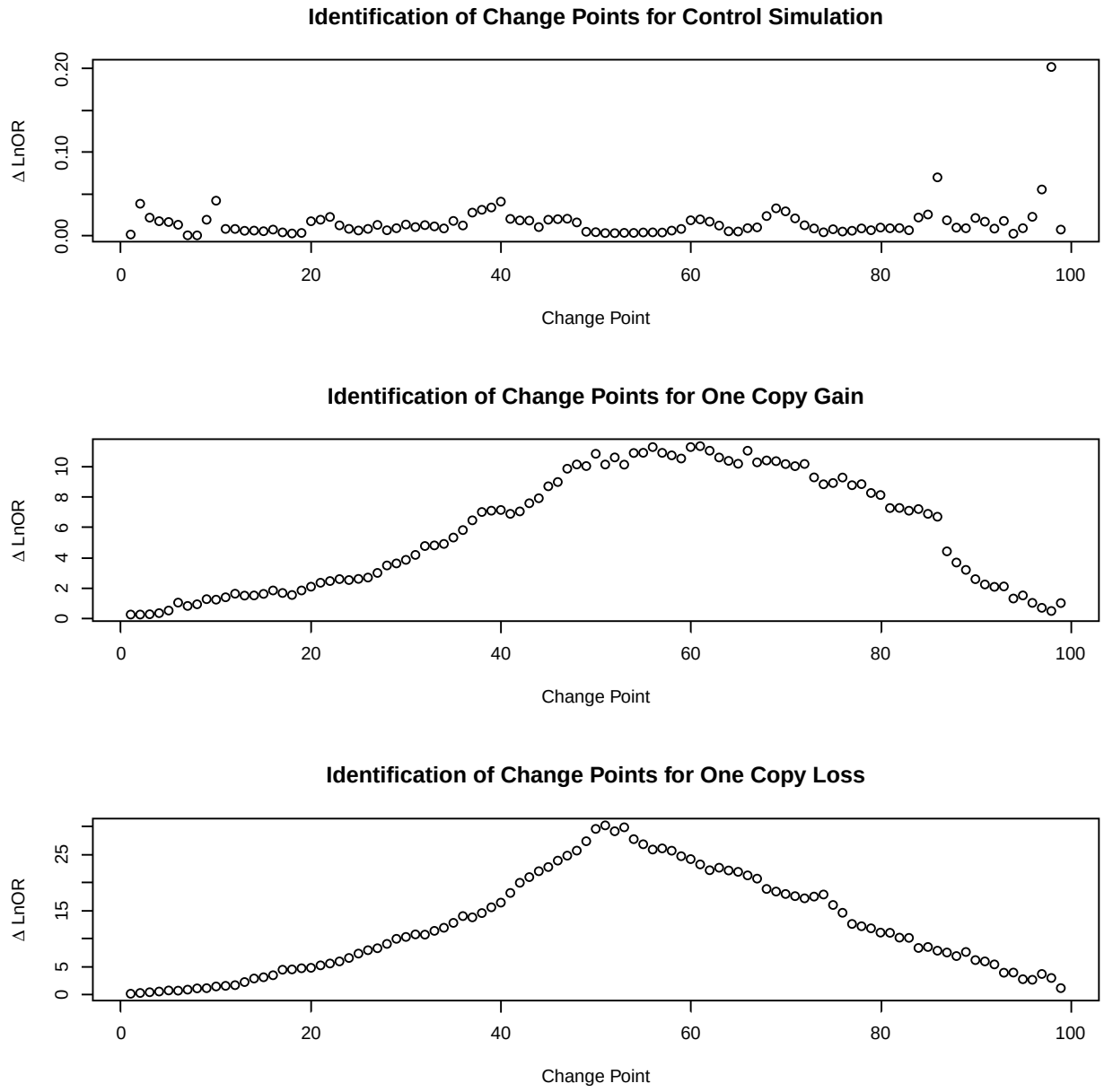

Figure 3.2: Control Adjusted Log Posterior Odds ratio (mΔ lnOR) generated by PairedBayNormal at each possible change point along the segment based on a control, one copy gain and one copy loss single change point simulation after control correction.

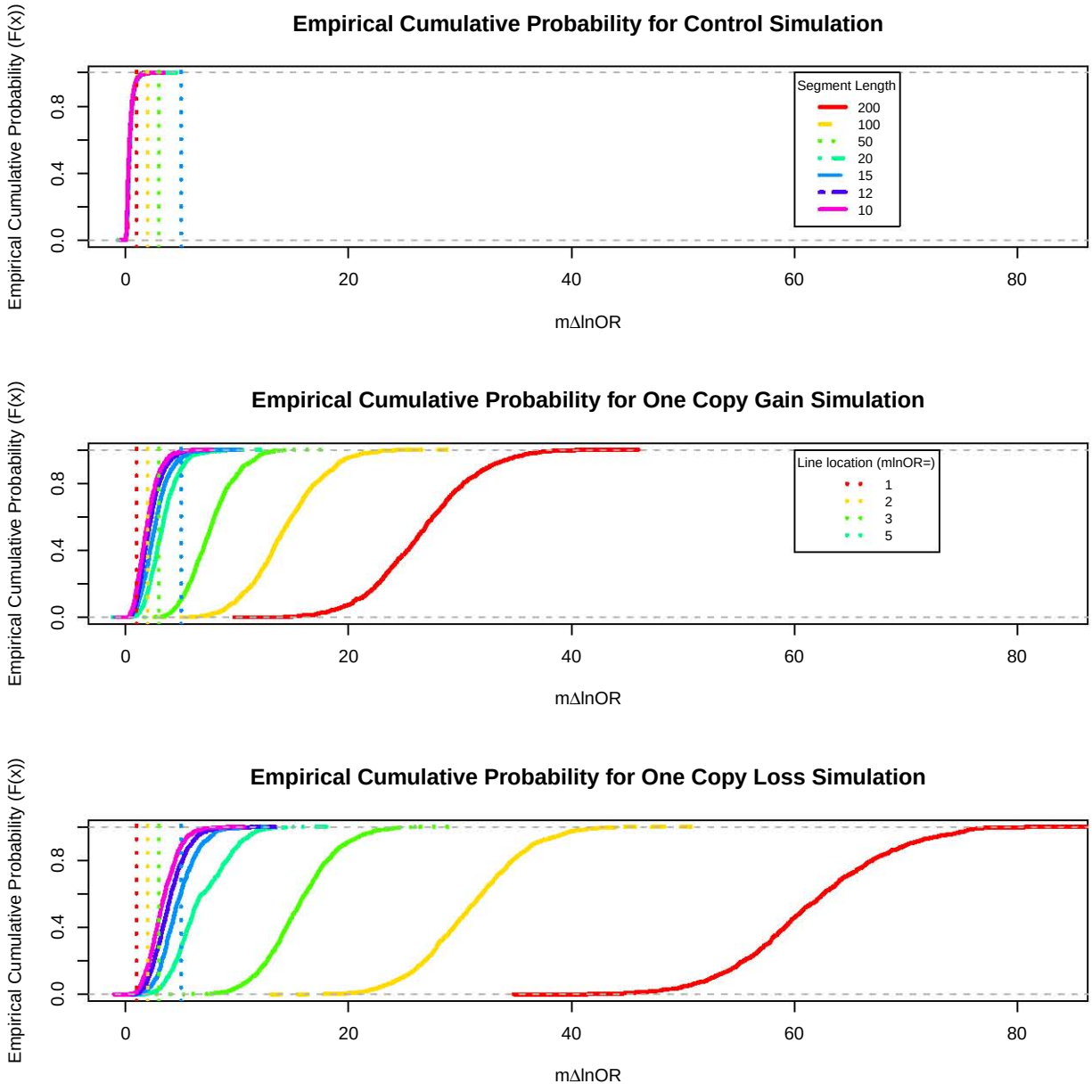

Figure 3.3: The impact of segment length on the empirical distribution of maximum log posterior odds ratio (m Δ lnOR) generated by PairedBayNormal based on m=1000 times of control (top), one copy gain (middle) and one copy loss (bottom) single change point simulations. Segment length n=10, 12, 15, 20, 50, 100, 200 with change point k in the middle.

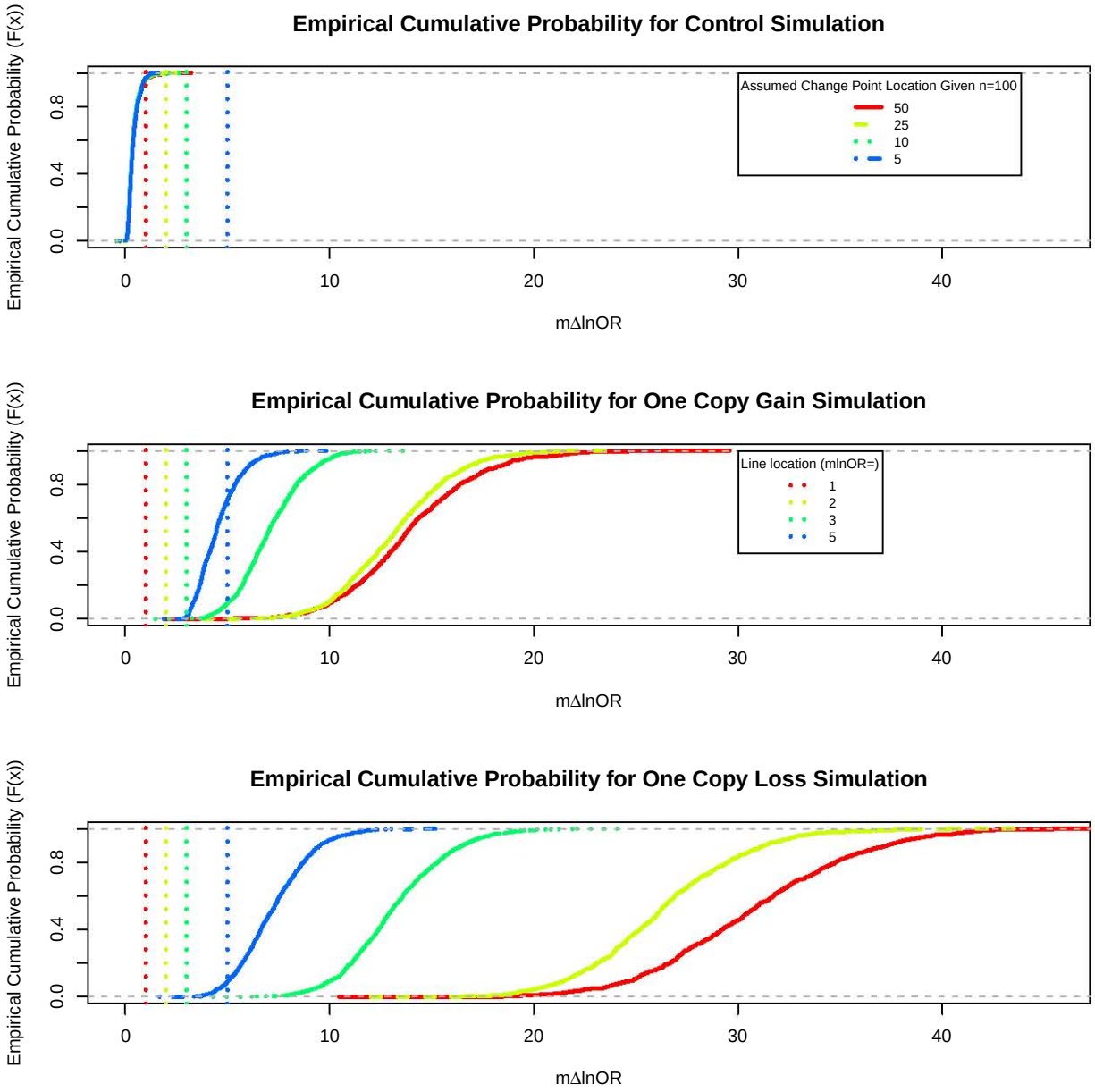

Figure 3.4: The impact of change point location on the empirical cumulative distribution of mΔlnOR generated by PairedBayNormal based on m=1000 times of control (top), one copy gain (middle) and one copy loss single change simulations. Segment length n=100 and change points k=5, 10, 25 and 50.

and 3.0 respectively (see table 3.1). We choose OR.level=3.0 with respect to possible type I inflated error in multiple comparisons when applied in sliding algorithm for mulitple change point and NGS read count analysis.

Table 3.1: Empirical Cumulative Distribution of Maximum Posterior Odds Ratio (mΔ lnOR) by Paired_BayNormal

Prob	**mΔ lnOR**		
	Control	**Gain**	**Loss**
0.00%	0.009	6.537	17.097
0.01%	0.011	6.561	17.118
0.05%	0.021	6.670	17.216
0.06%	0.023	6.697	17.240
0.08%	0.026	6.737	17.277
0.10%	0.032	6.805	17.338
0.50%	0.055	7.143	18.060
1%	0.066	7.719	20.223
5%	0.105	9.170	23.014
50%	0.317	14.004	30.367
95%	0.917	19.418	39.025
97.50%	1.136	20.617	40.823
98%	1.218	21.137	41.079
98.50%	1.285	22.031	42.496
99%	1.373	22.762	43.614
99.90%	1.847	25.583	47.247
99.99%	1.997	27.635	49.104
100.00%	2.012	27.840	49.290

3.6.1.2 Evaluation on Single Change Point Data

The evaluation of Paired_BayNomoral was based on m=1000 times of simulations of two control, one copy gain vs. control and one copy loss vs. control with segment length n=100 at change point k=50 according to descriptions in data section (see table 3.2). The existence of one copy gain and one copy loss was identified 100% in one copy gain and one copy loss vs. control simulations with 3.0 $<$ mlnOR . No one significance was ever found in controls. About 78% and 94% of assumed change points were identified within two bin deviation of the corresponding estimated change

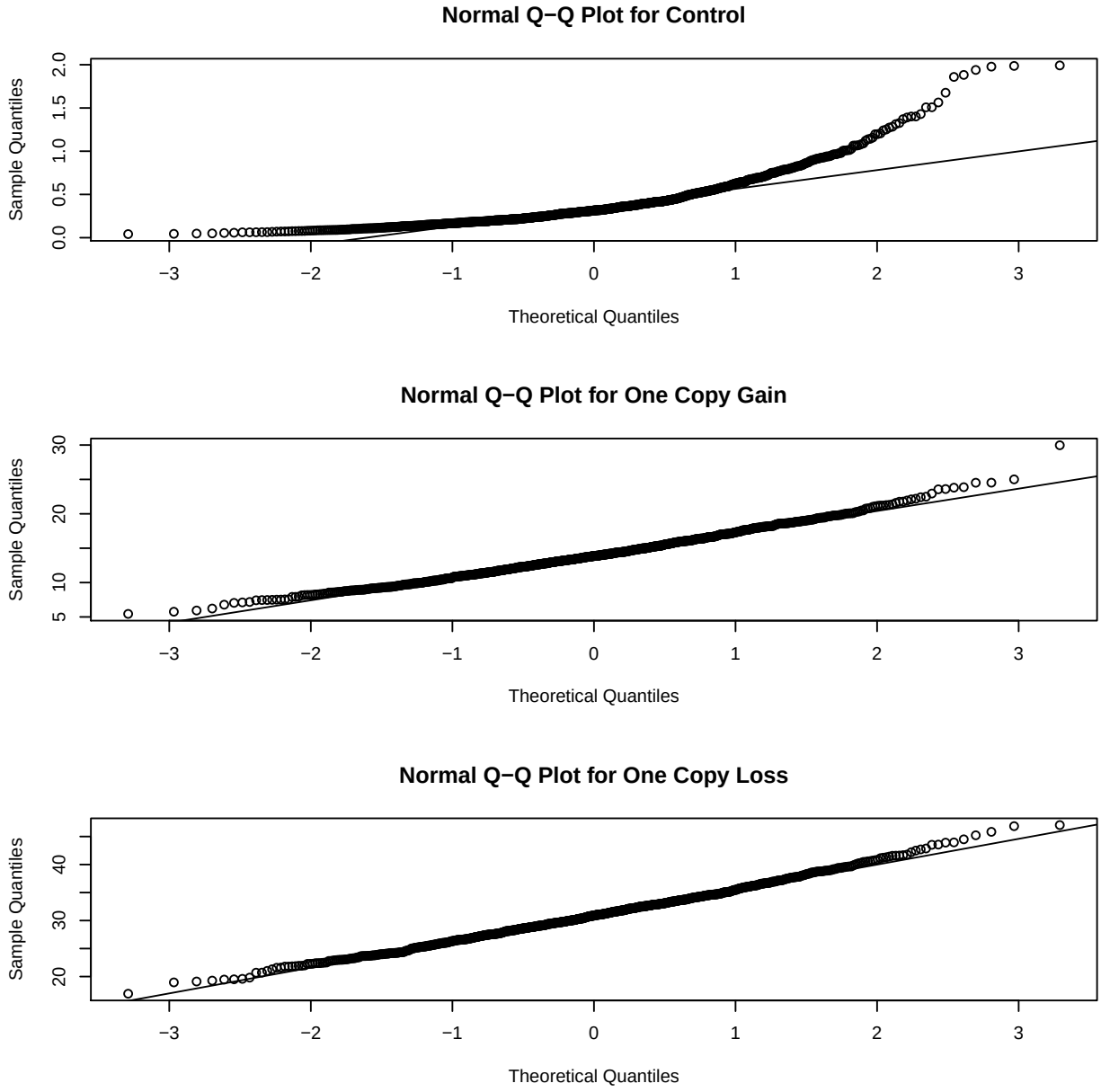

Figure 3.5: Normal Q-Q plot for mΔlnOR by PairedBayNormal in m=1000 times of control, one copy gain and one copy loss single change point simulations with segment length n=100 and change point k=50.

point location. The mean change point locations are close to the assumed change point location (51.84 vs. 50.0 and 50.53 vs 50). The mean differences of read count estimates and copy number estimates between test and control sample in segments before and after the change point reflect our simulation assumption accurately. The detection of one copy gain seems to have a lower accuracy in change point identification and higher deviation in read count estimates than that for one copy loss simulations.

Table 3.2: Evaluation of Single Change Point Detection at OR.level=3.0 by Paired_BayNormal

Indices	Gain	Loss	Control
PR	1.000	1.000	0.000
PRCPI2	0.777	0.938	0.000
Mean(k)	51.835	50.527	100.000
MSE(k)	3367.225	277.729	0.000
Mean(Δ_1)	0.357	-0.131	0.045
MSE(Δ_1)	127.178	17.214	2.060
Mean(ΔC_1)	0.018	-0.007	0.002
MSE(ΔC_1)	0.000	0.000	0.000
Mean(Δ_2)	20.319	-20.074	0.045
MSE(Δ_2)	101.718	5.413	2.060
Mean(ΔC_2)	1.016	-1.004	0.002
MSE(ΔC_2)	0.000	0.000	0.000

PR: Positive Rate, Sensitivity; PRCPI: positive rate of change point identification; PRCPI1: positive rate of change point identification for k $\in k \pm 1$; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(Δ_1): mean read count estimate difference between test and control in segment 1; MSE(Δ_1): mean squared error of read counts estimate difference in segment 1; Mean(ΔC_1): mean copy number estimate difference between test and control in segment 1 (before the change point k); MSE(ΔC_1): mean squared error of copy number estimate difference in segment 1; Mean(Δ_2): mean read count estimate difference in segment 2 (after the change point); MSE(Δ_2): mean squared error of read count estimate in segment 2; Mean(ΔC_2): mean copy numberestimate difference in segment 2; MSE(ΔC_2): mean squared error of copy number estimate in segment 2.

3.6.1.3 Evaluation on Multiple Change Point Data

The evaluation of multiple change point detection as described in data section indicates that 95% of assumed change points at k=50 for one copy gain change and 98.7% at k=150 for one copy loss change were identified within two bins around the estimated change point. Only 0.5% of change point was showed in the segment between k=200 and 299 as in the false positive rate of change point identification (see table 3.3). The accuracy and precision of change point identification could be reflected by the mean change point and MSE of the change point estimates. The mean difference in read count estimate and copy number estimates between one copy gain and control or between one copy loss and control or between two controls are consistent with our assumption in data simulations for segments before and after change points. The only 10% of change points on average were identified among total identified change points in all simulated regions as defined for false change point rate.

3.6.1.4 Evaluation on Human NGS Data

The evaluation of Paired_BayNormal was based on human NGS plus one copy gain and one copy loss simulations with λ=25 (see table 3.4). About 96% of randomly one copy gain change points and 98.5% of randomly one copy loss change points with deviation of two bins around were identified with OR.level=3.0, sliding window size end=100. This is consistent with our findings on mean change point and MSE of the estimated change points. The estimation of read counts and copy numbers in the simulated regions showed accurate prediction.

3.7 Conclusion

PairedBayNormal which utilized Bayesian solution of normal approximation of read counts based on Anscombe'work has been demonstrated to be a powerful tool to identify change points in NGS read count analysis with high sensitivity and specificity. Since our data format can not match the required format in current packages such as

Table 3.3: Evaluation of Multiple Change Point Detection by Paired_BayNormal

Indices	Gain	Loss	Control
PRCPI2	0.928	0.987	0.005
Mean(k)	50.491	150.132	294.916
MSE(k)	2.303	0.398	254.082
Mean(Δ_1)	0.076	0.423	-0.212
MSE(Δ_1)	1.723	1.770	0.820
Mean(ΔC_1)	0.004	0.021	-0.011
MSE(ΔC_1)	0.004	0.004	0.002
Mean(Δ_2)	19.913	-19.696	-0.212
MSE(Δ_2)	2.031	1.175	0.820
Mean(ΔC_2)	0.996	-0.985	-0.011
MSE(ΔC_2)	0.005	0.003	0.002

PRCPI2: positive rate of change point identification for k $\in k \pm 2$; Mean(k): Mean Change Point; MSE (k): mean squared error of change point estimate; Mean(Δ_1):mean difference of read count estimates between test and control samples in segment 1 (before the change point k); MSE(Δ_1): mean squared error of read count estimate difference in segment 1; Mean(ΔC_1):mean difference of copy number estimates between test and control samples in segment 1; MSE(ΔC_1): mean squared error of copy number estimate difference in segment 1; Mean(Δ_2): mean difference of read count estimates between test and control samples in segment 2 (after the change point); MSE(Δ_2): mean squared error of read count estimate difference in segment 2; Mean(ΔC_2): mean difference of copy number estimates between test and control samples in segment 2; MSE(ΔC_2): mean squared error of copy number estimate difference in segment 2;

Table 3.4: Evaluation on Human NGS+Simulation Data by Paired_BayNormal

Indices	Gain	Loss
PRCPI2	0.960	0.985
Mean(Δk)	0.300	-5.305
MSE(Δk)	35.530	2773.415
Mean(Δ)	24.098	-24.289
MSE(Δ)	13.374	6.124
Mean(ΔC)	0.945	-0.971
MSE (ΔC)	0.022	0.010

PRCPI2: positive rate of change point identification for k$\pm$2; Mean(Δk): Mean Change Point deviation; MSE (Δk): mean squared error of change point estimate; Mean(Δ):mean difference of read count estimates between test and control samples; MSE(Δ): mean squared error of copy number estimate difference; Mean(ΔC): mean difference of copy number estimates between test and control samples; MSE(ΔC): mean squared error of copy number estimate difference;

CNVseq, segseq etc, the comparison to other packages remain to be explored further in near future. The application of PairedBayNormal in copy number estimation from more human NGS data remains to be explored more in future.

Chapter 4

CONCLUSION AND FUTURE WORK

4.1 Conclusion

Copy Number Variation is one kind of important genetic variation located in human genome. CNVs have been associated with the development of genetic evolution and various diseases esp. cancers. Identification of CNVs is the crucial step to understand CNVs and their roles in the development of diseases.

With the rapid development of Next Generation Sequencing technology we could obtain unprecedented big data about the detailed genomic sequence on one human subject. We expect to see the wide use of NGS technology in medical research and clinical care practice with expected reducing cost and increasing speed for sequencing a genome in future. The detection of CNVs through NGS data will be an important task for future genetic studies before they can be applied in clinical care.

The current CNV detection strategies are based on normal assumption or normal approximation of read count data from NGS which has been demonstrated to follow dispersed Poisson distribution across the genome sequence. The possible existing CNVs are believed to be the main reason accounting for the dispersion. The CNVs detected are not consistent among detection packages. Most statistical packages in detection of CNVs have relatively low sensitivity and high false positive rate. An improved algorithm with more accurate and precise detection of CNVs is needed for

deciphering the genetic roles of CNVs in disease development.

We assumed that read count data from NGS follow non homogeneous Poisson distribution. The read count data from a segment with same copy number is believed to follow homogeneous Poisson distribution in a single genome. The first step for identification of copy numbers is actually to find change points in a series of read count data along the genome. The copy number within the segment between two neighboring change points can then be calculated from mean of the read count data.

We first developed a Bayesian normal approximation algorithm for detection of change points based on Anscombe' variance stabilizing algorithm called BayNormal. The posterior probability for no change point model and the maximum posterior probability of one change point at k were derived from the Bayesian normal approximation algorithm. The posterior odds ratio between one change point at k and no change point was used for statistical inference in favor of one change point vs. no change point model in comparison to a chosen posterior odds ratio level. The threshold level of maximum natural logarithm of posterior odds ratio (mlnOR) was derived from empirical cumulative distribution of mlnOR based on Monte Carlo simulations with null hypothesis. The chosen level makes it possible to detect one change point with about 100% positive rate but 0 percent of false positive rate in one change point simulations for both one copy gain or one copy loss. Meanwhile the positive rate of identification of change points matched at the exact assumed change points can reach 83.3% for one copy gain and 92.8% for one copy loss. The positive rate of identification of change points matched at one deviated from the assumed change points can reach 96.0% for one copy gain and 99.3% for one copy loss.

We extended a Bayesian Poisson algorithm with prior assumption of Gamma distribution for Poisson distribution intensity parameter (BayGamma). The posterior odds ratio between the posterior probability for one change point at k and the posterior probability for no change point model was utilized too. Following the same

procedure in comparison to a chosen mlnOR level from Monte Carlo simulations, it reach the same range of sensitivity and false positive rate as the Bayesian normal approximation approach (BayNormal).

We also found that segment length within which the change point analysis was conducted impact the empirical distribution of mlnOR. The longer segment length will cause the slight left shift of empirical distribution of null hypothesis and significant right shift of empirical distribution of one copy gain and one copy loss simulations. Especially in normal approximation algorithm, short segment length (e.g. n=10) may cause significant right shift of empirical distribution of mlnOR for null hypothesis. These indicated that sensitivity and specificity in binary segmentation procedure will be changed since segment length changes based on the previous identification of change points. This implies that a fixed window size or segment length will be helpful to get a consistent and accurate identification of change points. In our simulations, segment length greater than 12 can have enough power to detect change points with high specificity by BayNormal and BayGamma. We have utilized sliding window algorithm with fixed segment length to keep the analysis more consistent in terms of high power and low false positive rate.

The change point location within the segment also impacts the empirical distribution of mlnOR. The closer to end of the segment is the assumed change point, the further left shift for the empirical distribution of mlnOR for one copy gain or loss. The sensitivity or specificity will be reduced for change point closer to end of the segment. The change point at 5 bins close to end of the segment can be detected by our Bayesian approaches with high sensitivity and specificity based on Monte Carlo simulations.

The application of BayNormal and BayGamma in multiple change point simulations indicated that they can detect the change point exactly with sensitivity greater than 85.0% and 95.0 % for one copy gain and one copy loss respectively while the

sensitivity of circular binary segmentation algorithm is 81.0% and 91.0%. The false positive rate can be lower than 5.0%.

The application of BayNormal and BayGamma in human NGS data plus one copy or minus one copy simulations showed that more than 90.0% of one copy gain simulations and 94.% of one copy loss simulations can be detected at 2 deviation from the true change points while those are about 91.0% and 87.5% for CBS respectively. The false discovery rate is less than 10.0%.

We also developed Bayesian normal approximation algorithms for finding the difference of copy number variants between cancer and control genomes called PairedBayNormal. The posterior odds ratio between one change point at k and no change point model in terms of difference between cancer and control was derived. We used the posterior odds ratio after correction for posterior odds ratio between control and control at each change point k to determine whether the change point at k exists or not in comparison to a chosen mΔ lnOR level. A chosen level of mΔlnOR allows us to detect one change point with about 100.0% positive rate but 0.0% false positive rate for both one copy gain and loss. The positive rate of change point identification rate for 2 bins match deviated from the assumed change point can reach to 77.8% for one copy gain and 93.8% for one copy loss.

The impact of segment length and change point location on CNV detection has been observed too in PairedBayNormal algorithm.

The application of PairedBayNormal in multiple change point simulations indicated that the positive rate of change point identification for 2 bins deviated from the assumed change point can reach to 91.5% for one copy gain and 98.7% for one copy loss.

The application of PairedBayNormal in human NGS data plus one copy or minus one copy simulations showed that more than 96.0% and 98.5% of positive change points for 2 bins deviated from the assumed change points in one copy gain and one

copy loss respectively can be obtained.

4.2 Future Work

Although we have developed powerful approaches to address the CNV detection, there are more questions that need to be answered in future which has not been addressed in this dissertation due to time limit. Information about these questions will help to clarify arguments in theory and practices of statistical analysis as well as accurate CNV detection.

Comparisons of Bayesian Change Points to Likelihood Models with Poisson Assumption

CBS has been demonstrated to be the most powerful algorithm in detecting CNVs. Our results indicated that our Bayesian approaches improved power and specificity in detection of CNVs in comparison to CBS. Our Bayesian approaches take advantage of both Bayesian inference and better normal approximation of Poisson distribution data. We don't have direct evidences to support whether the prior consumption with Gamma distribution is helpful or not in improving the power and specificity of CNV detection from Poisson read count data. We can compare the likelihood ratio model and our Bayesian Poisson change point algorithms in terms of sensitivity and specificity for detection of copy number variants.

Comparison of Bayesian Normal Approximation To Mean Change Point Model and Mean Variance Change Point Model

Mean Change Point Model and Mean Variance Change Point Model have been widely used in CNVs detection especially from aCGH. A R package with these change point models included has been built up. Although normal approximation adapted in current approaches is expected to be better than these algorithms, we still need more direct evidences to support the conclusion.

Extension of CNV Detection To Diverse Coverage of NGS Data? Sensitivity and

Specificity for Smaller Window Size and High Coverage Data?

Human NGS read data can be binned to read count with 100bp or 10kb bp as bin size in literatures with respect to coverage and CNV detection resolution. High coverage data (100X) has also been utilized in literatures. In current work we have used read count data in bin size with 1kb to develop the algorithm. Whether our approaches is still sensitive and specific for other coverage and data remains to be explored in future. Although the performance of our algorithms has been evaluated on one kind of simulated data condition, their performance on more noised data simulations and various data conditions remains to be explored further in future.

Integration of Other Algorithms With Our Program To Detect CNVs

Since GC content correction and Mappability error could account for some dispersion in read count Poisson distribution, the integration of other information into our algorithms such as GC content correction and Mappability error and algorithm in repeated regions is expected to improve the detection of CNVs in human NGS data. Whether paired end sequencing information will be helpful remains to be explored in future.

Implication of the Established Bayesian Change Points to More Human NGS Data to Verify the CNVs Detected in Population

Our approaches have been evaluated based on simulations. The application of these approaches in more human NGS data will provide more information in comparison to current available packages in terms of sensitivity and specificity in detection of CNVs. The confirmation of our approaches in detection of CNVs under various circumstances will provide basis to conduct population study so that the CNVs in disease development can be understood in future. The application of our approaches to more human NGS data is expected to help us understand the roles of CNVs in genetic development.

Bibliography

Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21, 974-984.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S.C., Gibbs, R.A. and Eichler, E.E. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41, 1061-1067.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12, 363-376.

Anscombe, F.J. (1948) The Transformation of Poisson, Bionomial and Negative-Bionomial Data, Biometrika, 35, 246-254.

Bellos, E., Johnson, M.R. and LJ, M.C. (2012) cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome biology*, 13, R120.

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

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.P., Janoueix-Lerosey, I., Delattre, O. and Barillot, E. (2011) Control-free calling of copy number alterations in deep-

sequencing data using GC-content normalization. *Bioinformatics* (Oxford, England), 27, 268-269.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., Teague, J.W., Menzies, A., Goodhead, I., Turner, D.J., Clee, C.M., Quail, M.A., Cox, A., Brown, C., Durbin, R., Hurles, M.E., Edwards, P.A., Bignell, G.R., Stratton, M.R. and Futreal, P.A. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40, 722-729.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L. and Mardis, E.R. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6, 677-681.

Chen, J. and Gupta, A.K. (2012). Parametric Statistical Change Point Analysis - With Applications to Genetics, Medicine, and Finance, second edition, Birkhauser: New York.

Chiang, D.Y. and McCarroll, S.A. (2009) Mapping duplicated sequences. *Nature biotechnology*, 27, 1001-1002.

Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C. and Ragoussis, J. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research*, 35, 2013-2025.

Conrad, D.F. and Hurles, M.E. (2007) The population genetics of structural variation. *Nature genetics*, 39, S30-36.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, et al. (2010) Origins and functional impact of copy number variation in the human genome, Nature, 464, 704-712.

Wellcome Trust Case Control Consortium, W., Craddock, N., et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464, 713-720.

Gregory, P.C., Loredo, T.J. (1992) A method for the detection of a periodic signal of unknown shape and period. *The Astrophysical Journal*, 398, 146-168.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. and McVean, G. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44, 226-232.

Ivakhno, S., Royce, T., Cox, A.J., Evers, D.J., Cheetham, R.K. and Tavare, S. (2010) CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* (Oxford, England), 26, 3051-3058.

Kim, T.M., Luquette, L.J., Xi, R. and Park, P.J. (2010) rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC bioinformatics*, 11, 432.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U. and Hochreiter, S. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research*, 40, e69.

Koboldt, D.C., Larson, D.E., Chen, K., Ding, L. and Wilson, R.K. (2012) Massively parallel sequncing approaches for characterization of structural variation. *Methods in molecular biology (Clifton, N.J*, 838, 369-384.

LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37, 4181-4193.

Lai, W.R., Johnson, M.D., Kucherlapati, R. and Park, P.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* (Oxford, England), 21, 3763-3770.

Langer-Safer, P.R., Levine, M. and Ward, D.C. (1982) Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79, 4381-4385.

Lesch, S.M., Jeske, D. R. (2009) Some Suggestions for Teaching About Normal Approximation to Poisson and Binomial Distribution Functions The American Statistician, 63, 274-277.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012, 251364.

Magi, A., Tattini, L., Pippucci, T., Torricelli, F. and Benelli, M. (2012) Read count approach for DNA copy number variants detection. *Bioinformatics* (Oxford, England), 28, 470-478.

Margulies, M., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376-380.

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 560-564.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) Computational methods for dis-

covering structural variation with next-generation sequencing. *Nature methods*, 6, S13-20.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. and Brudno, M. (2010) Detecting copy number variation with mated short reads. *Genome research*, 20, 1613-1622.

Miller, C.A., Hampton, O., Coarfa, C. and Milosavljevic, A. (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PloS one*, 6, e16327.

Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics (Oxford, England), 5, 557-572.

Pichard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.J. (2005) A statistical approach for array CGH data analysis, BMC bioinformatics, 6, 27.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13, 341.

Raftery, A.E., Akman, V.E. (1986) Bayesian analysis of a Poisson process with a change-point, Biometrika, 73, 85-89.

Raphael, B.J., Volik, S., Collins, C. and Pevzner, P.A. (2003) Reconstructing tumor genome architectures. *Bioinformatics* (Oxford, England), 19 Suppl 2, ii162-171.

Ren, H., Francis, W., Boys, A., Chueh, A.C., Wong, N., La, P., Wong, L.H., Ryan, J., Slater, H.R. and Choo, K.H. (2005) BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints. *Human mutation*, 25, 476-482.

Scargle, J.D. (1998) Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data, The Astrophysical Journal, 504, 405-418.

Segmund, D. (1986) Boundary Crossing Probabilities And Statistical Applications, The Annals of Statistics, 14, 361-404.

Sen, A., Srivastava, M.S. (1975) On tests for detecting change in mean, The Annals of Statistics, 3, 98-108.

Simpson, J.T. and Durbin, R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22, 549-556.

Smith, A.F.M. (1975) A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62, 407-416.

Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009) The cancer genome. *Nature*, 458, 719-724.

Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S. and Salim, A. (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* (Oxford, England), 28, 2711-2718.

The 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56-65.

Urban, A.E., Korbel, J.O., Selzer, R., Richmond, T., Hacker, A., Popescu, G.V., Cubells, J.F., Green, R., Emanuel, B.S., Gerstein, M.B., Weissman, S.M. and Snyder, M. (2006) High-resolution mapping of DNA copy alterations in human

chromosome 22 using high-density tiling oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 4534-4539.

Valsesia, A., Stevenson, B.J., Waterworth, D., Mooser, V., Vollenweider, P., Waeber, G., Jongeneel, C.V., Beckmann, J.S., Kutalik, Z. and Bergmann, S. (2012) Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. *BMC genomics*, 13, 241.

Valsesia, A., Mace, A., Jacquemont, S., Beckmann, J.S. and Kutalik, Z. (2013) The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Frontiers in genetics*, 4, 92.

Venkatraman, E.S. and Olshen, A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* (Oxford, England), 23, 657-663.

Vostrikova, L.Y. (1981) On the detection of ¡¡discordance¿¿ of Wiener process. *Teor. Veroyatnost. i Primenen.*, 26, 362-368.

Wald, A., Wolfowitz, J. (1950) Bayes Solutions of Squential Decision Problems, the Annals of Mathematical Statistics, 21, 82-99.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. and Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*, 17, 1665-1674.

West, W.R., Ogden, T.R. (1997) Continuous-time estimation of a change point in a poisson process. *Journal of Statistical Computation and Simulation*, 56, 293-302.

Worsley, K.J. (1986) Confidence regions and tests for a change point in a sequence of expoential family random variables, Biometrika, 73, 91-104.

Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.M., Lee, E., Zhang, J., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Gibbs, R.A., Kucherlapati, R. and Park, P.J. (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America*, 108, E1128-1136.

Xie, C. and Tammi, M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC bioinformatics*, 10, 80.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* (Oxford, England), 25, 2865-2871.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research*, 19, 1586-1592.

Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E.S. and Liu, C. (2011) Accuracy of CNV Detection from GWAS Data. *PloS one*, 6, e14511.

Zhu, M., Need, A.C., Han, Y., Ge, D., Maia, J.M., Zhu, Q., Heinzen, E.L., Cirulli, E.T., Pelak, K., He, M., Ruzzo, E.K., Gumbs, C., Singh, A., Feng, S., Shianna, K.V. and Goldstein, D.B. (2012) Using ERDS to infer copy-number variants in high-coverage genomes. *American journal of human genetics*, 91, 408-421.

VITA

Jianfeng Meng was born in LuoYang, HeNan province, China in 1965. He received his M.D. degree from West China University of Medical Sciences in 1988. He received his M.P.H degree from the same University in 1992. He received his M. S. degree in Pharmacology and Toxicology from Indiana University and Purdure University at Indianapolis in 2001. He received his M.S aster degree in Biostatistics from University of Colorado School of Public Health in 2007.