Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

9-26-2018 10:00 AM

Statistical tools for assessment of spatial properties of mutations observed under the microarray platform

Bin Luo
The University of Western Ontario

Supervisor
Dean, Charmaine
The University of Western Ontario Joint Supervisor
Kulperger, Reg
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy
© Bin Luo 2018

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Biostatistics Commons

Recommended Citation

Luo, Bin, "Statistical tools for assessment of spatial properties of mutations observed under the microarray platform" (2018). *Electronic Thesis and Dissertation Repository*. 5746.
https://ir.lib.uwo.ca/etd/5746

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Mutations are alterations of the DNA nucleotide sequence of the genome. Analyses of spatial properties of mutations are critical for understanding certain mutational mechanisms relevant to genetic disease, diversity, and evolution. The studies in this thesis focus on two types of mutations: point mutations, i.e., single nucleotide polymorphism (SNP) genotype differences, and mutations in segments, i.e., copy number variations (CNVs). The microarray platform, such as the Mouse Diversity Genotyping Array (MDGA), detects these mutations genome-wide with lower cost compared to whole genome sequencing, and thus is considered for suitability as a screening tool for large populations. Yet it provides observation of mutations with high degree of missingness across the genome due to its design, which thus leads to challenges for statistical analyses. Three topics are studied in this thesis: the development of formal statistical tools for detecting the existence of point mutation clusters under the microarray platform; the evaluation of the performance of test statistics developed while accounting for various probe designs, in terms of the capabilities of detecting mutation clusters; the development of formal statistical tools for testing the existence of spatial association between point mutations and mutations in segments. Statistical models such as Poisson point processes and Neyman-Scott processes are used for the distributions of the locations of point mutations under null and alternative hypotheses. Monte Carlo frameworks are established for statistical inference and the evaluation of power performance of the proposed test statistics. Tests with desirable performance are identified and recommended as screening tools. These statistical tools can be used for the study of other genomic events in the form of point events and events in segments, as well as with other microarray platforms than the MDGA which is utilized here. Simulated probe sets based on a window-based probe design mimicing the design of the MDGA are used to study the effect of various factors in probe design on the performance of test statistics. Insights are offered for determining key features in such design, such as probe intensity, when designing a new microarray platform, in order to achieve desired power for the purpose of mutation cluster detection.

Keywords: Spatial statistics; Point processes; Cluster detection; Spatial association; Hypothesis testing; Missing data; Genomics; Mutation shower; Single nucleotide polymorphism; Copy number variation; Microarray probe design sampling; Mouse Diversity Genotyping Array

Co-Authorship Statement

This work was completed under the supervision of Dr. Charmaine Dean and Dr. Reg Kulperger, and with collaboration of Dr. Kathleen Hill. All papers resulting from this thesis will be co-authored with Drs. Dean, Kulperger and Hill.

Dedicated to my parents and wife for their love and support

Acknowledgments

First and foremost, I would like to express my deepest gratitude and appreciation to my supervisors, Dr. Charmaine Dean nad Dr. Reg Kulperger. I would like to thank them for providing me the opportunity to become their student while I was searching for supervisors, and giving me the freedom to choose a research area I am interested in. This thesis would not have been in any way possible without their mentorship, guidance, support, kindness, and patience. It is my greatest pleasure and honor to be their student. I would also like to give my most sincere gratitude to our collaborator, Dr. Kathleen Hill. I would like to thank her for her introduction to many interesting biological research questions, and her tremendous support throughout my study.

I am thankful to my M.Sc. supervisor Dr. Ian McLeod, and my previous Ph.D. supervisor Dr. Jiguo Cao, who led me to research in statistics. I would also like to thank all the graduate chairs in our department during my M.Sc. and Ph.D. studies. Special thanks to Dr. John Braun who encouraged me to join the program and provided me with valuable suggestions, and Dr. Marcos Escobar-Anel who assisted with rules of School of Graduate and Postdoctoral Studies.

I would like to express special gratitude to the lab of Dr. Charmaine Dean and Dr. Doug Woolford. I benefit greatly from the excellent academic environment built by all lab members. I would also like to thank all the lab members from Dr. Kathleen Hill's lab. The discussion and communication with them provided valuable help to my research.

I am grateful to my course instructors from both the Department of Statistical and Actuarial Sciences and the Department of Epidemiology and Biostatistics. Their courses helped me build foundation in statistical knowledge and develop an interest in research.

I would like to express special appreciation to all the assistants of Dr. Charmaine Dean, who did a fantastic job helping me arrange meeting schedules. My appreciation also goes to the administrative staff in the Department of Statistical and Actuarial Sciences.

I highly appreciate the Engage Grant and Mitacs Accelerate programs for providing me the opportunity to gain industrial experience during my study. I would like to thank the colleagues

and friends at my industrial partner, GenomeDx, for providing an enriching research environment. The research experience I gained from this collaboration became very helpful during my Ph.D. research.

Many thanks are due to all the friends I have made here in London. Their company, help and support have been invaluable to me to allow me to live a good and happy life. Special thanks to Rosie and Bilbo whose accompany made my life more colorful.

Finally, my greatest gratitude and appreciation extend to my family: my parents, Jisheng Luo and Chunying Xia, and my wife, Yue Cai. Without their unconditional love and support, I would not have been able to achieve anything during my study. I wish that all of them could see this completion of my degree.

Contents

Certificate of Examination **i**

Abstract **i**

Co-Authorship Statement **iii**

Dedication **iv**

Acknowlegements **v**

List of Figures **xi**

List of Appendices **xvi**

1 Overview **1**

2 Spatial statistical tools for genome-wide mutation cluster detection under a microarray probe sampling system **7**

- 2.1 Introduction 8
- 2.2 Methods 13
- 2.3 Small sample properties of the test statistics 19
  - 2.3.1 Proposed underlying processes for the null hypothesis 19
  - 2.3.2 Proposed underlying processes for alternative hypotheses 22
  - 2.3.3 Simulation parameter settings and results 26
- 2.4 Application 34

    - 2.4.1 Genotyping method 34
    - 2.4.2 Analyses for three biological samples of interest 35
- 2.5 Discussion 37

3 Effect of probe design in microarray type data on powers of test statistics for cluster detection 42

- 3.1 Introduction 43
- 3.2 Statistical methods 45
    - 3.2.1 Probe sets designed from simulated libraries 46
        - 3.2.1.1 Library generation 46
        - 3.2.1.2 Window-based probe selection procedure 48
    - 3.2.2 Probe sets filtered from MDGA 50
    - 3.2.3 Methodology for power study 50
- 3.3 Power study 53
    - 3.3.1 Parameter settings 53
    - 3.3.2 Power comparisons 54
- 3.4 Discussion 63

4 Nonparametric association test for spatial independence in the occurrence of single nucleotide polymorphism differences and copy number variations under a microarray probe sampling system 66

- 4.1 Introduction 67
- 4.2 Statistical methods 69
    - 4.2.1 Data structure 69
    - 4.2.2 Proposed statistics 70
- 4.3 Small sample properties of the test statistics 71
    - 4.3.1 Monte Carlo simulation approaches for inference 72
        - 4.3.1.1 Poisson null process 73

4.3.1.2 Overall block bootstrap 73
4.3.1.3 Partial block bootstrap 76
4.3.2 Statistical inference 76
4.3.3 Alternative processes for power studies 78
4.3.3.1 Step function nhPP 78
4.3.3.2 Modified non-homogeneous parent Neyman-Scott process (NPNSP) 79
4.3.4 Parameter settings and simulation results 80
4.3.4.1 Parameter settings for confidence band construction 80
4.3.4.2 Parameter settings for power study 81
4.3.4.3 Power study results, interpretation, and recommendation . . . 81
4.4 Application 91
4.5 Discussion 98

5 Summary and future work **102**

Bibliography **105**

A Supplementary information for Chapter 2 **111**
A.1 Validation of size of test statistics 111
A.1.1 Scheme a 111
A.1.2 Scheme b 113
A.2 Additional power study figures 115

B Supplementary information for Chapter 3 **118**

C Supplementary information for Chapter 4 **121**
C.1 Geometrical illustration of the partial block bootstrap 121
C.2 Parameter setting in the alternative hypothesis for the step function nhPP model 124
C.3 Examination of autocovariance for location of SNP differences on microarray data 125

D Related techniques **128**

D.1 Monte Carlo simulation for statistical inference and power estimation 128

D.2 Block bootstrap for dependent data 129

D.3 DNA data structure: discreteness 130

D.4 Spatial point process 131

D.5 Notes on numerical computation 132

Curriculum Vitae **133**

List of Figures

2.1 The inter-SNP locus distances (bp) for 493,290 SNP loci assayed by the probes on the Mouse Diversity Genotyping Array (MDGA) are summarized for each chromosome. 11

2.2 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (1) with parameter $\mu_o = 375$. 28

2.3 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (2) with parameter $\mu_o = 375$. 29

2.4 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (3) with parameter $\mu_o = 375$. 30

2.5 Power performance of test statistics $\tilde{R}(d)$ across a grid of d under alternative hypothesis (1) with parameter $\mu_o = 375$. 33

2.6 Rainfall plots portraying the SNP differences due to mixed genetic background, putative new mutations arising during development of two normal tissues of the same mouse and putative mutations arising between two cancerous tissues from the same mouse. 36

3.1 Comparison of rejection rates of $CvM_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from hPP samples. 56

3.2 Comparison of rejection rates of $KS_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from hPP samples. 57

3.3 Comparison of rejection rates of $CvM_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from NS samples. 58

3.4 Comparison of rejection rates of $KS_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from NS samples. 59

3.5 Comparison of rejection rates of $CvM_{\tilde{R}}$ under filtered probe sets based on chromosome 19, hPP library ($l_w = 100k$) and NS library ($l_w = 100k$). The blue and black dashed vertical lines indicate probe intensities required for high performance with the hPP library and filtered probe sets on chromosome 19 from MDGA respectively. 60

3.6 Comparison of rejection rates of $KS_{\tilde{R}}$ under filtered probe sets based on chromosome 19, hPP library ($l_w = 100k$) and NS library ($l_w = 100k$). 61

4.1 The first two steps of the overall block bootstrap are illustrated. Sample a point $g_1 \sim U(s_f, s_l)$ on the original chromosome. All SNP differences and probes within the segment $[g_1, g_1 + B_O - 1]$ on the original chromosome are relocated to the segment $[s_f, s_f + B_O - 1]$ on the resampled chromosome, maintaining their exact position relative to g_1, but now from s_f on the resampled chromosome. A subsequent point $g_2 \sim U(s_f, s_l)$ is taken on the original chromosome, the SNP differences and probes within $[g_2, g_2 + B_O - 1]$ are relocated to the segment $[s_f + B_O, s_f + 2B_O - 1]$ on the resampled chromosome, again maintaining their exact position relative to g_2, but now from $s_f + B_O$ on the resampled chromosome. 75

4.2 Power performance of J and C statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the Poisson Null Process. 83

4.3 Power performance of J and C statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the overall block bootstrap. 84

4.4 Power performance of J and C statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the partial block bootstrap. 85

4.5 Power performance of J and C statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the Poisson Null Process. 86

4.6 Power performance of J and C statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the overall block bootstrap. 87

4.7 Power performance of J and C statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the partial block bootstrap. 88

4.8 Rainfall plot displaying mutation landscapes. Black dots refer to SNP differences. Red markers identify CNV locations. The green dashed vertical lines show the centers of the CNVs. 92

4.9 Rainbow plot displaying mutation landscapes. Black dots refer to SNP differences. Red markers identify CNV locations. The green dashed vertical lines show the centers of the CNVs. 93

4.10 J statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the Poisson null process. 95

4.11 J statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the approach of overall block bootstrap. 96

4.12 J statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the approach of partial block bootstrap. 97

A.1 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (1) with parameter $\mu_o = 1125$. 115

A.2 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (2) with parameter $\mu_o = 1125$. 116

A.3 Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (3) with parameter $\mu_o = 1125$. 117

C.1 Geometrical illustration of the partial block bootstrap. The distances from SNP differences and probes to the nearest CNV are obtained by superposition of SNP differences and probes in a defined region on either side of each CNV. The distances within the interval of $(0, B_P]$ and $[Z, Z + B_P)$ are exchanged by transformation, maintaining their relative positions within the segments. 123

C.2 Examination of the autocovariance for the locations of SNP differences in the example in Section 4.4. 127

List of Tables

A.1 size of $D_{min}(n)$ under various argument of n based on Scheme a. 111

A.2 size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme a part 1. 111

A.3 size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme a part 2. 112

A.4 size of functional forms of the five statistics based on Scheme a. 112

A.5 size of $D_{min}(n)$ under various argument of n based on Scheme b. 113

A.6 size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme b part 1. 113

A.7 size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme b part 2. 113

A.8 size of functional forms of the five statistics based on Scheme b. 114

List of Appendices

Appendix A Supplementary information for Chapter 2 111

Appendix B Supplementary information for Chapter 3 118

Appendix C Supplementary information for Chapter 4 121

Appendix D Related techniques 128

Chapter 1

Overview

The discovery of the structure of Deoxyribonucleic acid (DNA) in the 20th century has greatly inspired the field of biological research, and has led to a more profound understanding of life and vast changes to our understanding of health leading to better health outcomes for humans. In the 21st century, numerous scientific hypotheses in biological research have been investigated through the conduct of very many observational studies and experiments. Tremendous volumes of data are generated rapidly, thanks to the advancement of measuring techniques such as sequencing. Yet data generated from various sources often require that new appropriate statistical tools be developed for the analyses. This research aims to reduce the gap of suitable statistical methods to study features of biological phenomenon using data obtained by a specific measuring technique, microarray analysis.

DNA is the fundamental molecule that stores crucial biological information and plays the role of hereditary material. Genetic instructions are encoded in DNA and used in the development and functioning of all living organisms as well as many viruses. The genome is the complete set of DNA in an organism. Genomics aims to study the mechanisms related to the entire DNA set, instead of individual pieces. Genomics research is crucial for tackling problems faced in health, environment, and agriculture.

Mutations are alterations of the DNA nucleotide sequence of the genome. During replication of DNA, there is a chance that DNA nucleotides are mismatched. There are repairing mechanisms to ensure that the genetic information is inherited stably from generation to generation. However, repairing mechanisms are not perfect, thus mutations in the DNA may occur. A mutation is an important source of genetic variability, adaptability and evolution, yet it can also lead to cancer. Mutations can refer to several different types of mismatches or changes in the DNA sequence. Two common types are single nucleotide polymorphism (SNP) genotype differences and copy number variations (CNVs), which are investigated in this thesis research.

To measure the mutations genome-wide, two options are usually available: DNA sequencing and the microarray. DNA sequencing measures the genome with a resolution of a single nucleotide, unveiling all the DNA information in the genome. With the advancement of technology in recent years, the price of sequencing is becoming more affordable for sequencing the entire genome of some organisms, including for human research. Yet for most other organisms, the price of DNA sequencing can still be prohibitively high. On the other hand, the microarray, or genotyping array, provides an alternative for genome-wide mutation measurement. A microarray contains a large set of specifically designed probes targeting different areas across the genome. Only the information on the targeted areas is obtained, yielding missing observations in other areas of the genome. Yet the microarray has substantially lower cost compared to sequencing, and is used in a wealth of genetics research. It has been adopted as a cost-effective way to measure mutation information across the genome.

In mutation related research, the relationship between genotypes and phenotypes has normally been a main focus. The phenotype is determined by DNA composition, which can be altered due to mutations. The microarray platform has been widely used to conduct genome-wide association studies, aiming to identify association between SNP genotypes and diseases [1]. In recent years, the spacing between mutations has caught the attention of some researchers. From observation by sequencing some small DNA fragments, it was found that mutations may tend to be close to each other in term of spatial locations, rather than occurring in a random pattern. Thus there raises a hypothesis that mutations may be generated in what are called patterns of showers because of some mechanisms. More evidence about this phenomenon and the mechanisms or leading factors that can cause this non-random spacing pattern are thus of high interest currently to many biologists. And research on a genome scale, rather than over small DNA segments, is now needed to give a broader perspective. Such research may help better understand fundamental mutagenesis mechanisms and provide new insights to developing therapies of cancer or other disease. To our knowledge, the statistical tools for studying the spatial association between mutations observed under the microarray platform have not been developed. Importantly, biologists have developed ad hoc graphical tools for investigating spatial patterns and, though helpful, this thesis aims to offer more rigorous tools for such investigations.

This thesis focuses on statistical methods for analyzing spatial properties of mutations detected under the microarray platform. Based on the research interest of the biologists with whom we collaborated, three questions regarding the spatial association between mutations are addressed and discussed in three chapters as discussed below. The thesis is presented as a compilation of three papers developed through this research. Each of Chapter 2, 3, and 4 represent articles prepared for submission for publication. University regulations permit that a thesis be assembled in this manner.

In Chapter 2, we develop spatial statistical tools for genome-wide mutation cluster detection under a microarray probe sampling system. Mutation cluster analysis is critical for understanding certain mutational mechanisms relevant to genetic disease, diversity, and evolution. Yet, whole genome sequencing for detection of mutation clusters is prohibitive with high cost for most organisms and population surveys. SNP genotyping arrays, like the Mouse Diversity Genotyping Array (MDGA), offer an alternative low-cost, screening for mutations at hundreds of thousands of loci across the genome using experimental designs that permit capture of *de novo* mutations in any tissue. Formal statistical tools for genome-wide detection of mutation clusters under a microarray probe sampling system are yet to be established. A challenge in the development of statistical methods is that microarray detection of mutation clusters is constrained to select SNP loci captured by probes on the array. This chapter develops a Monte Carlo framework for cluster testing and assesses test statistics for capturing potential deviations from spatial randomness which are motivated by, and incorporate, the array design. While null distributions of the test statistics are established under spatial randomness via the homogeneous Poisson process, power performance of the test statistics is evaluated under postulated types of Neyman-Scott clustering processes through Monte Carlo simulation. A new statistic is developed and recommended as a screening tool for mutation cluster detection. The statistic is demonstrated to be excellent in terms of its robustness and power performance, and useful for cluster analysis in settings of missing data. The test statistic can also be generalized to any one-dimensional system where every site is observed, such as DNA sequencing data. The chapter illustrates how the informal graphical tools for detecting clusters may be misleading. The statistic is used for finding clusters of putative SNP differences in a mixture of different mouse genetic backgrounds and clusters of *de novo* SNP differences arising between tissues with development and carcinogenesis.

In Chapter 3, we study the effect of probe design in microarray type data on powers of test statistics for cluster detection. Mutation clusters are important signatures for understanding mutational mechanisms, genetic diversity, disease, adaptation and evolution. The microarray platform detects mutations genome-wide with lower resolution and cost compared to next-generation sequencing, and thus may be considered for screening for mutation cluster detection in large populations. As well, formal statistical tools have been developed and recommended for mutation cluster detection for a specific microarray platform. In this chapter, we further evaluate properties of the tests by assessing comparisons of their performance across various probe designs in terms of their capabilities of detecting mutation clusters. The various probe designs compared here include simulated probe sets using a window-based probe selection procedure from simulated genomic variation libraries, as well as probe sets filtered from an existing mouse array design. A framework and algorithm are provided for numerical calculation of the rejection rates of the test statistics. The methods and tools developed here provide insight for determining key features such as probe intensity, when designing a new microarray platform with desired power for the purpose of mutation cluster detection.

In Chapter 4, we develop a nonparametric association test for spatial independence in occurrence of SNP differences and CNVs under a microarray probe sampling system. The study of spatial properties of mutations can help to better understand mutational mechanisms as well as genetic diversity to inform our understanding of health and disease. Investigation of the spatial association between the locations of two types of mutations, such as genotypic differences at SNP loci and CNVs, may help uncover potential relationships in these locations, including interactions between these mutational mechanisms. The microarray platform, such as the MDGA, provides a cost-effective way to assay SNP genotypes and detect CNVs and hence may allow for screening studies for large scale investigations. In this study, we propose two test statistics to test the existence of spatial association between SNP differences and CNVs. Importantly, these test statistics incorporate the microarray design, accounting then for intermittent observation over the genome. We propose three null hypotheses, with different generality, related to the association between SNP differences and CNVs, and three Monte Carlo simulation approachs for statical inference, including one based on the parametric Poisson model and two block bootstrap methods. Power performance of the test statistics is evaluated under a step function Poisson process, as well as a modified version of the Neyman-Scott parent child process. The statistics are based on neighborhood properties of SNP differences related to CNVs and are modifications of well-established association tests in the literature that are known to have good performance. One statistic, the J statistic, is demonstrated to perform well in this context of missing data and is recommended. We demonstrate the utility of the J statistic in an example that considers mutation profile differences between primary tumor and metastatic tissue of the same mouse. The methods and tools provided in this chapter can be utilized for the analysis of association for other genomic events using the microarray platform for mouse, human, and other species.

In Chapter 5, we provide a summary of the innovations in this thesis, and suggest directions for future work.

Chapter 2

Spatial statistical tools for genome-wide mutation cluster detection under a microarray probe sampling system

2.1 Introduction

Mutation signatures are useful tools for identifying mutagens and mutational mechanisms, and understanding genetic diversity, disease, adaptation and evolution. These signatures are identified by comparison of genomic sequences with a reference sequence and association with specific exogenous and/or endogenous conditions. Genome sequences can be viewed as a string in the genome alphabet, or equivalently as a time series or lattice sequence of large length. For the mouse genomic experiments discussed here, the length of a single chromosome ranges from 6.14×10^7 base pairs (bp) of nucleotides for chromosome 19 to 1.95×10^8 bp for chromosome 1.

Current genomic technologies have broadened our perspective to mutation analysis, revealing a critically important phenomenon of non-random spacing of mutations as a new mutation signature [2]. This signature is crucial for discovery of mechanisms for mutagenesis and carcinogenesis, as well as for development of cancer treatments that target effects of driver mutations. Proximal spacing of multiple mutations has been termed 'Kataegis' or thundershowers of mutations [3]. Mutation showers have been reported in genomes of yeast [4, 5], mice [6, 7] and humans [8], within genes and dispersed across the genome. To date, mutation showers have been arbitrarily defined based on cancer whole genome sequencing data as the occurrence of sequence segments containing six or more consecutive mutations with an average intermutation distance of less than or equal to 1,000 bp [8]. Another definition for mutation clusters was based on empirical data for the observation of multiple mutations within 30 kb in the context of postzygotic mutations in healthy mouse tissues[7]. The largest dataset for detection of mutation showers exists for large pan-cancer studies, where mutation showers are found with low incidence in certain cancer types [8]. A chief mechanism proposed for this signature is transient hypermutagenesis, an elusive and incompletely understood phenomenon [9, 10]. Examination of the human genome for mutation showers is restricted to a very limited number of tissues or cell types and next generation sequencing. Whole genome sequencing, although the highest resolution possible, is not affordable as a population screening approach in general.

Since complete genome sequencing is expensive and generally impractical as a screening or survey method, genotyping microarrays are a low-cost alternative which are commonly used to detect mutations at loci with single nucleotide polymorphisms (SNPs). These loci are referred to as SNP sites. Differences in a single nucleotide, referred to as SNP genotype differences (SNP differences), can be interpreted as mutations when comparing two samples. These two samples can be two biological samples of interest, or a biological sample of interest and a reference sample, which is usually B6 mouse in mouse studies. SNPs are genotyped using designed single-stranded short nucleotide probes affixed to a microarray platform. These probes complement specific locations within the genome and these locations are quite sparse in distribution across the genome relative to the genome length, yielding low cost for the array process relative to sequencing. Thus, a SNP genotype difference can be detectable or undetectable by a microarray platform, depending on whether the probes on the array are at that SNP locus. The objective we study in this chapter is the development of a population, i.e., a large sample size, screening tool for a wide variety of tissues and cell types, using the low cost SNP array data for identifying clusters of putative mutations. The challenge is that arrays provide windows of observations along the genome, which depend on probe sites, in terms of both number of sites and distribution or spacing of the sites. Hence the screening tool would need to accommodate this constraint in the experimental design with microarray platforms.

The Mouse Diversity Genotyping Array (MDGA) is a single nucleotide polymorphism (SNP) microarray [11] that detects SNP alleles at 493,290 SNP loci [12] across the mouse genome. The alleles at each SNP locus are detected by a SNP probe set on the array. A probe set consists of eight single-stranded DNA sequences (probes) 25 bp in length. The probes are fixed to a solid surface (or chip) in a known arrangement. Due to several conditions a SNP probe needs to satisfy in design, the probes are not evenly distributed along each chromosome. To illustrate the sparsity of the probes, Figure 2.1 is a boxplot of the MDGA inter-SNP locus distances for each autosome and the X chromosome. The average inter-SNP locus distance is 5,210 bp, with a maximum and minimum distance of 7,268,520 bp and 16 bp, respectively. Of the SNP loci, 83.6% (412,181 SNP loci) are within 10,000 bp of another SNP locus, and 38.7% (190,714 SNP loci) are within 1,000 bp of another SNP locus. There are 22 SNP *probe deserts*, defined as consecutive probe sites spanning more than 1 million bp; the two largest gaps between consecutive probe sites are 7,268,520 bp and 7,033,330 bp on chromosomes 7 and X, respectively.

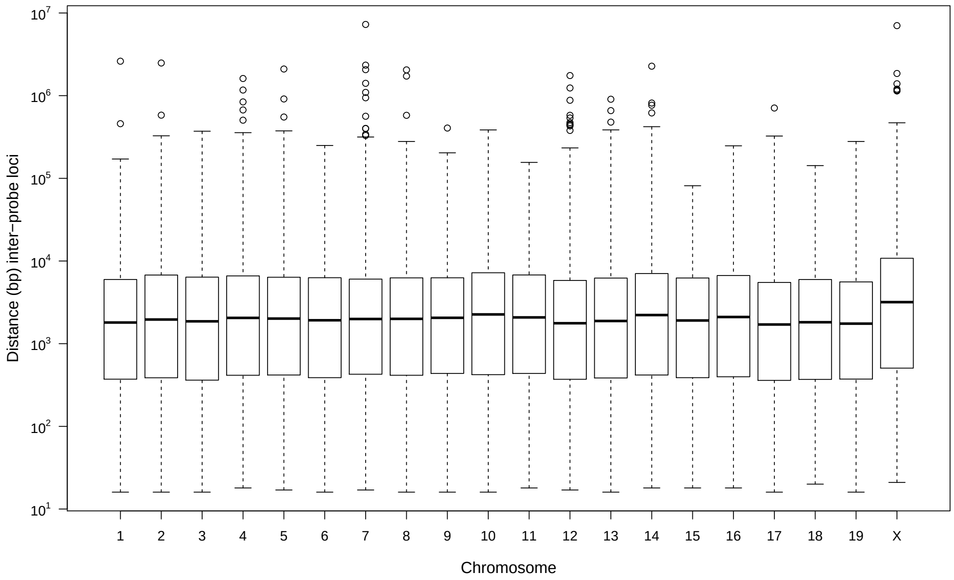


Figure 2.1: The inter-SNP locus distances (bp) for 493,290 SNP loci assayed by the probes on the Mouse Diversity Genotyping Array (MDGA) are summarized for each chromosome. A boxplot of the distribution of these inter-SNP locus distances (bp) for each autosome and the X chromosome.

With the rapid development of genotyping and sequencing techniques in recent years, more genetic studies have begun to focus on assembling, visualizing and studying the spatial information of genomic events under different scenarios such as genome-wide association studies [13]. For cluster detection, several statistical methods have been developed and applied in DNA and protein sequencing data [4, 14, 15, 16]. Despite previous efforts for detecting clusters with sequencing data, to our knowledge, there have not been formal studies attempting to detect mutation clusters under a genotyping array system. For sequencing data, the rainfall plot has been introduced recently for visualizing the landscape of mutations [8, 17]. Specifically, a rainfall plot portrays the base pair distance of intermutation spacing along the chromosome or entire genome sequence. Here, rainfall plots are adopted to visually examine the potential existence of clusters on the whole genome or individual chromosomes for data from a mouse SNP genotyping array. Mutation clusters are suggested by low intermutation spacing values in such plots; the goal of this paper is to attach rigorous statistical inference to the identification of clusters.

From the discussion above we see that the observable microarray data depend on the probe design, that is, the locations of the probes. In this chapter, we study several statistics for detecting mutation clusters: a set of non-parametric statistics based on neighbourhood measures, and a test statistic based on distances between SNP loci where mutations are detected, which is related to rainfall plots. These statistics are also studied in real-valued functional forms to summarize the cluster features. The microarray probe sampling system yields missing observations in the domain of interest. Numerical techniques have become increasingly important for the analysis of complex data structures, such as observed here. Such techniques are utilized in our analyses to incorporate the probe design constraints. The null process of complete randomness is a homogeneous Poisson process. For a natural alternative cluster process we consider the family of Neyman-Scott processes, which are a class of parent-child point processes. We evaluate the techniques through power studies which demonstrate that the tests proposed provide suitable tools for screening samples for clustering effects on the genome scale. We then apply the recommended statistical tools for finding clusters of putative SNP genotype differences (SNP differences) in a mixture of different mouse genetic backgrounds and for finding clusters of *de novo* SNP differences between tissues with development and with carcinogenesis.

2.2 Methods

To detect mutation clusters genome-wide, chromosomes are studied individually as each chromosome consists of a linear space in itself. Define the set of the probe locations, determined by design, as $S : S \subset \mathbb{R}^+$. Denote the location of the first probe target site (a SNP locus) on the chromosome as $s_f = \min_{s \in S} s$, and the location of the last probe target site on the chromosome as $s_l = \max_{s \in S} s$. Denote the locations of SNP differences detected by the probes as $X : X \subseteq S$.

The test statistics proposed below consider SNP genotype differences within the neighborhood of a known SNP genotype difference, where neighbourhood is defined by either distance d from the known SNP genotype difference, or by the number of SNP differences n, within the neighborhood. Each statistic can be considered as a function of a specific value of d or n, or alternatively, the behavior of each statistic over a range of d or n may be considered. The summary statistics for functional behaviors utilize the well-known frameworks of the Kolmogorov-Smirnov (KS) and Cramér-von Mises (CvM) tests, adapted for this missing data context. The test statistics proposed are:

(I) Mean over all sites with SNP differences of the ratio of the number of sites with SNP differences to the number of probes within fixed distance d

$$\bar{R}(d) \equiv \bar{R}_S(d) = \frac{\sum_{x \in X} \frac{N_X(x,d)}{N_S(x,d)}}{|X|} \tag{2.1}$$

where for arbitrary set A, fixed distance d, and site with a SNP genotype difference x, we have

$$N_A(x,d) = \sum_{z\in A} I(0 < |z-x| \leqslant d) \tag{2.2}$$

where $I(E)$ is the indicator function for the event E.

(II) Pooled mean detection ratio: the ratio of the total, over all $x \in X$, of $N_X(x,d)$, the number of SNP differences within distance d of each SNP genotype difference, to the total, over all $x \in X$, of $N_S(x,d)$, the number of probes within distance d of each SNP genotype difference

$$\tilde{R}(d) \equiv \tilde{R}_S(d) = \frac{\sum_{x\in X} N_X(x,d)}{\sum_{x\in X} N_S(x,d)} \tag{2.3}$$

The two statistics above are inspired by the K function introduced by Ripley [18], which tests for general clustering in a point process by measuring the number of events occuring within a certain distance of other events. The $\bar{R}(d)$ and $\tilde{R}(d)$ statistics proposed here summarize properties in the neighborhood of distance d from observed SNP differences, while adjusting for varying probe sparsity over the chromosome. The index S is used to emphasize that the statistics depend on the design of the probe set S. Comparing the two statistics, $\tilde{R}(d)$ as a pooled estimate of the mean detection ratio is more robust and numerically more stable. These two statistics can be invalid in case their denominators are zero in some data or settings, which are discussed in Appendix D.5. While we focus on the above formulations, for comparison purposes, we also consider traditional neighbourhood formulations of test statistics:

(III) Consider $D(x_0,n)$ the minimum distance to include n SNP differences around x_0

$$D(x_0,n) = \inf_d \left\{ d : \sum_{x\in X} I(|x_0 - x| \leqslant d) = n \right\} \tag{2.4}$$

The test statistic is the minimum of such distances over all SNP differences $x_0 \in X$,

$$D_{min}(n) = \min_{x_0 \in X} D(x_0, n) \tag{2.5}$$

Notice than when $n = 2$, $D_{min}(n)$ becomes the minimum of the distances between any two SNP differences. Algorithm 2.1, provided at the end of this section, describes an efficient procedure for the calculation of $D_{min}(n)$.

(IV) Maximum of the number of SNP differences within distance d of any given SNP genotype difference

$$N_{max}(d) = \max_{x \in X} N_X(x, d) \tag{2.6}$$

Another test statistic proposed is a count statistic inspired by the rainfall plot. The statistic is related to the distances between SNP loci with genotype differences, which are features shown in the rainfall plots. The count statistic is defined as follows:

(V) Count of inter-SNP locus distances for those SNP loci with different genotypes under threshold d

$$C(d) = \sum_{b \in B_X} I(b < d) \tag{2.7}$$

where $B_X = \uplus_{i=1}^{n-1}\{X_{(i+1)} - X_{(i)}\}$, and $X_{(i)}$ is denoted as the ith ordered statistic in X, where $i = 1, \cdots, |X|$. The multiset B_X contains all of the inter-SNP locus distances for those SNP loci with different genotypes for the sample X.

These five statistics, generically denoted as $G(y)$ with argument y, may be viewed as function valued statistics with a fixed argument d or n. In the statistics above, the argument is typically d or n. Instead of considering a fixed argument y, they may also be viewed as a functional form $G(\cdot)$, $G(\cdot) \equiv \{G(y), y \in R(y)\}$, where $R(y)$ is the range of y considered. Let $G^*(\cdot) = E_0(G(\cdot))$, the expectation of $G(\cdot)$ under an appropriate null hypothesis, e.g., homogeneous Poisson process, which is discussed in further detail in Section 2.3.1. Two test statistics measuring the distance of $G(\cdot)$ from $G^*(\cdot)$ as considered here are of the forms of Kolmogorov-Smirnov (KS) and Cramér-von Mises (CvM) tests [19] described as follows:

(a) Kolmogorov-Smirnov test framework

The KS test statistic is the supremum norm distance of G to G^* over a range of y:

$$KS(G, G^*) = \sup_y |G(y) - G^*(y)| \tag{2.8}$$

(b) Cramér-von Mises test framework

The CvM test statistic integrates the squared difference between G and G^* over a range of y:

$$CvM(G, G^*) = \int [G(y) - G^*(y)]^2 dy \tag{2.9}$$

The five test statistics $G(y)$ for specific argument y as described above and KS and CvM based on their functional forms $G(\cdot)$ are used to conduct inference.

To evaluate KS and CvM, the support of function $G(\cdot)$ is discretized and set as a finite grid $Y = \{y_i, i = 1, \cdots, k\}$. The grid points y_1 and y_k represent the smallest and largest values of d and n in the evaluation range respectively. Given the grid Y, the discrete versions of KS and CvM statistics are calculated as:

$$\widetilde{KS}(G, G^*) = \max_{y_i, i=1,\cdots,k} |G(y_i) - G^*(y_i)| \tag{2.10}$$

$$\widetilde{CvM}(G, G^*) = \frac{1}{2} \sum_{i=1}^{k-1} \Big\{ \{[G(y_i) - G^*(y_i)]^2 + [G(y_{i+1}) - G^*(y_{i+1})]^2\}(y_{i+1} - y_i) \Big\} \tag{2.11}$$

The parameter k controls how dense the function $G(\cdot)$ is evaluated on the support $[y_1, y_k]$. If the selected grid points are dense, $\widetilde{KS}$ and $\widetilde{CvM}$ converge to KS and CvM; yet the selection of k should also account for feasible computational load.

Algorithm 2.1: Calculation of $D_{min}(n)$

```
Let $X = \{x_i, i = 1, \cdots, K\}$ denote the set of ordered SNP differences, where $x_i$ is the $i$th ordered SNP genotype difference on the chromosome. Then there are $K - n + 1$ clusters of consecutive SNP differences of size $n$: $\{\{x_l, \cdots, x_{l+n-1}\}; l = 1, \cdots K-n+1\}$.
Define $D_l \equiv \min_{m \in [l+1, l+n-2]} \max(x_m - x_l, x_{l+n-1} - x_m)$, $l = 1, \cdots K - n + 1$. For the $l$th cluster of SNP differences, consider the set of minimum distances to include $n$ *cluster SNP differences* around each SNP genotype difference in the cluster; then $D_l$ is the minimum distance in the set. Note that *cluster SNP differences* refer to SNP genotype difference in the $l$th cluster.
$D_{min}(n) = \min_l D_l$; $l = 1, \cdots K - n + 1$.
```

2.3 Small sample properties of the test statistics

Mutations may occur at any of the 2.8 billion base positions in the mouse genome. Among these mutations some exist at the genomic loci targeted by SNP probes and are thus detectable as SNP differences by the SNP probe system, while the existence of the other mutations remains unknown. Both null and alternative hypotheses are established on underlying processes that generate all mutations, both detectable and undetectable. Since the target loci of the SNP probes are unique and non-random on each chromosome, the null and alternative distributions of the proposed test statistics are calculated conditional on the probe locations on the specific chromosome considered.

2.3.1 Proposed underlying processes for the null hypothesis

Under the null hypothesis, the locations of SNP differences are assumed to follow complete spatial randomness in this study. This model would be used for the most generic application scenario where no further genetic background information is available. Under the null hypothesis that SNP differences are located at random locations along the chromosome, the underlying process generating SNP differences can be assumed as a homogeneous Poisson process (hPP). Under such a process, every site on the chromosome, and in particular, every probe site, is independent and has an identical probability of having a SNP genotype difference. The relationship between the hPP rate parameter and the total expected number of detected SNP differences η is linear. Numerical methods are adopted to obtain the null distributions of the test statistics for testing that X_s, the observed locations of SNP differences from the sample, are randomly located along the chromosome. Algorithm 2.2 develops the Monte Carlo estimate of the null distribution of the test statistics, while algorithm 2.3 provides an inferential procedure.

Algorithm 2.2: Monte Carlo estimates of the null distributions of summary statistics

```
2.1: Set a finite grid $Y = \{y_i, i = 1, \cdots, k\}$, which defines the scale of $d$ or $n$ as the evaluation range;
2.2: Simulate $M$ replications of detected SNP differences $\{X_0^{(m)}, m = 1, \cdots, M\}$ from the hPP. At the $m$th replication, $X_0^{(m)}$ is obtained as follows:
    (a): Generate the total number of underlying SNP differences $N_{null}^{(m)} \sim Pois(\hat{\lambda})$, where $\hat{\lambda}$ is an estimate of the rate parameter from the observed sample $X_s$: $\hat{\lambda} = \frac{(s_l - s_f)\eta}{|S|}$. The parameter $\eta$ can be set as $|X_s|$, where $|A|$ is the norm of set $A$, that is the count of the number of elements in $A$;
    (b): Generate the set of underlying (both observable and unobservable) locations with SNP differences $U_{null}^{(m)} = \{u_j, j = 1, \cdots, N_{null}^{(m)}\}$, where independent and identically distributed random variables $u_j \sim U[s_f, s_l]$, and $U$ is the discrete uniform distribution on $\{s_f, \cdots, s_l\}$;
    (c): Obtain the set of observed SNP differences: $X_0^{(m)} = U_{null}^{(m)} \cap S$.
2.3: For each $m = 1, \cdots M$, obtain $G_{X_0^{(m)}}(\cdot) \equiv \{G_{X_0^{(m)}}(y_i), i = 1, \cdots, k\}$ at the grid sites $y_i, i = 1, \cdots, k$;
2.4: The Monte Carlo estimate of $G^*(\cdot)$ is $\hat{G}^*(\cdot) \equiv \{\frac{1}{M} \sum_{m=1}^{M} G_{X_0^{(m)}}(y_i), i = 1, \cdots, k\}$;
2.5: For each $m = 1, \cdots M$, calculate the $\widetilde{KS}$ or $\widetilde{CvM}$ test statistic:
    (a): $\widetilde{KS}_G^{X_0^{(m)}} = \widetilde{KS}(G_{X_0^{(m)}}, \hat{G}^*)$;
    (b): $\widetilde{CvM}_G^{X_0^{(m)}} = \widetilde{CvM}(G_{X_0^{(m)}}, \hat{G}^*)$;
2.6 The Monte Carlo estimates of the cumulative distribution functions of the test statistics $\hat{F}_{\widetilde{KS}_G}$ and $\hat{F}_{\widetilde{CvM}_G}$ are:
    (a): $\hat{F}_{\widetilde{KS}_G}(t) = \frac{1}{M} \sum_{m=1}^{M} I(\widetilde{KS}_G^{X_0^{(m)}} \leqslant t)$
    (b): $\hat{F}_{\widetilde{CvM}_G}(t) = \frac{1}{M} \sum_{m=1}^{M} I(\widetilde{CvM}_G^{X_0^{(m)}} \leqslant t)$
```

Algorithm 2.3: Hypothesis testing procedure

```
3.1 Based on the observed sample $X_s$, calculate $G_{X_s}(\cdot) \equiv \{G_{X_s}(y_i), i = 1, \cdots, k\}$. The test statistics are:
    (a): $\widetilde{KS}_G^{X_s} = \widetilde{KS}(G_{X_s}, \hat{G}^*)$;
    (b): $\widetilde{CvM}_G^{X_s} = \widetilde{CvM}(G_{X_s}, \hat{G}^*)$;
3.2 Statistical inference:
    (a): For hypothesis testing at significance level $\alpha$:
        (i): KS test: if $\widetilde{KS}_G^{X_s} > \hat{F}^{-1}_{\widetilde{KS}_G}(1 - \alpha)$, reject the null hypothesis, otherwise do not reject.
        (ii): CvM test: if $\widetilde{CvM}_G^{X_s} > \hat{F}^{-1}_{\widetilde{CvM}_G}(1 - \alpha)$, reject the null hypothesis, otherwise do not reject.
    (b): The p-values are calculated as:
        (i): KS test: $\frac{1+\sum_{m=1}^{M} I(\widetilde{KS}_G^{X_0^{(m)}} \geqslant \widetilde{KS}_G^{X_s})}{1+M}$;
        (ii): CvM test: $\frac{1+\sum_{m=1}^{M} I(\widetilde{CvM}_G^{X_0^{(m)}} \geqslant \widetilde{CvM}_G^{X_s})}{1+M}$;
```

The methods for calculation of the p-value in step 3.2(b) of Algorithm 2.3 are based on the approaches for calculating p-values for Monte Carlo simulation provided in [20], which would yield empirical p-values having correct type-I error rate.

To study the size of the statistics with fixed arguments as well as their functional forms, two schemes are adopted as follows:

a: Under a null process with the Poisson intensity $\hat{\lambda} = \frac{(s_l - s_f)|X_s|}{|S|}$, 10^4 Monte Carlo samples are simulated, and the null distributions and critical values are estimated based on these same 10^4 samples. Another 10^3 Monte Carlo samples are simulated under the null process with the same parameter setting of $\hat{\lambda} = \frac{(s_l - s_f)|X_s|}{|S|}$, and tested based on the same null distributions and critical values. The size would be calculated as the proportion rejected among these 10^3 simulated samples.

b: Under a null process with the Poisson intensity $\hat{\lambda} = \frac{(s_l - s_f)|X_s|}{|S|}$, 10^3 Monte Carlo samples, $X_{t0}^{(m)}, m = 1, \cdots, 10^3$, are simulated. For each $X_{t0}^{(m)}$, another 10^3 Monte Carlo samples are simulated under a null process with the Poisson intensity $\hat{\lambda} = \hat{\lambda}^{(m)} = \frac{(s_l - s_f)|X_{t0}^{(m)}|}{|S|}$, and $X_{t0}^{(m)}$ is tested using the null distributions and critical values estimated from these 10^3 samples. The size would be calculated as the proportion rejected among simulated samples $X_{t0}^{(m)}, m = 1, \cdots, 10^3$.

2.3.2 Proposed underlying processes for alternative hypotheses

Under the alternative hypotheses, the underlying process would generate SNP differences following a non-random spacing pattern. Here, the Neyman-Scott (NS) process is proposed as a suitable clustering process. The NS process is a parent-offspring process, where a cluster of several offspring is generated around each unobservable parent. The parent locations can be randomly spaced along the chromosome or follow some alternate spacing patterns. This parent-offspring type of underlying process is reasonable because it mimics a specific mutagenesis mechanism that one source of error may lead to a cluster of mutations nearby. The error source could be a binding site of a particular protein that leads to the generation of nearby mutations. This is an example of a transient state of an error-prone polymerase or a period in replication of biased dNTP pools or error-prone conditions associated with translesion bypass [21, 22, 23, 24, 6, 9].

Three alternative hypotheses are considered, all derived from the NS parent-offspring clustering process. Each of these three alternatives differs in the domain D_p on which parent sites are generated as discussed below. Each parent site generates a cluster of offspring sites, with the random number of offspring following the Poisson distribution with the expected number μ_o. The offspring sites are independent and identically distributed, truncated normal random variables centered at the parent site location. The standard deviation of the truncated normal distribution is denoted as σ. The half-length of the window of the truncation range is denoted as h.

(1) Parent sites with an expected number μ_p are generated along the chromosome from an hPP. The domain on which parent sites are located, D_p, is $[s_f - h, s_l + h]$. Only parents within this range can yield offspring detectable by the probe set, because of the truncation range in offspring distribution.

(2) Parent sites are constrained to SNP probe locations: $D_p = S$. There are two important reasons to constrain parent sites to probe locations. First, probes are located where the corresponding SNP differences have an occurrence of at least 1% in the population, so that the probe sites are selected based on their being favorable in terms of having SNP differences. Secondly, under this constraint, all of the test statistics will attain the highest power compared to other parent site settings. Thus this setting is helpful for eliminating some candidate tests with sub-optimal performance.

(3) The parent sites are constrained to be within a certain distance h_p of a probe; $D_p = \cup_{s \in S} [s - h_p, s + h_p]$. This setting recognizes possible errors in identifying probe locations, so parents may not be exactly placed at favorable sites for SNP differences.

In the simulation of each alternative hypothesis, as in the null hypothesis, the expected total number of detected SNP differences η is set to equal the observed total SNP differences $|X_s|$, which is achieved by adjusting the parameters in the alternative process. Algorithm 2.4 details the Monte Carlo estimates of the powers.

Algorithm 2.4: Power Study

```
4.1: Set a finite grid $Y = \{y_i, i = 1, \cdots, k\}$ the same as in Algorithm 2.2;
4.2 Simulate $M'$ replications of detected SNP differences $\{X_a^{(m)}, m = 1, \cdots, M'\}$ from a Neyman Scott process. At the $m$th replication, $X_a^{(m)}$ is generated as follows:
    (a): Generate the total number of unobservable parent points $N_p^{(m)} \sim Pois(\mu_p)$, where $\mu_p$ is the Poisson mean parameter.
    (b): Generate the set of parent points $Z^{(m)} = \{z_t^{(m)}, t = 1, \cdots, N_p^{(m)}\}$, where the iid random variable $z_t^{(m)} \sim U(D_p)$ and $U$ is the discrete uniform distribution on the domain $D_p$.
    (c): For each parent point $z_t^{(m)}$, generate the number of offspring $N_{ot}^{(m)} \sim Pois(\mu_o)$, and a set of offspring $O_t^{(m)} = \{u_{tj}^{(m)}, j = 1, \cdots, N_{ot}^{(m)}\}$, where iid random variables $u_{tj}^{(m)} \sim N(z_t^{(m)}, \sigma^2)$ with truncation interval $[z_t^{(m)} - h, z_t^{(m)} + h]$;
    (d): Obtain the set of all generated offspring $U_{alt}^{(m)} = \cup_{t=1}^{N_p^{(m)}} O_t^{(m)}$;
    (e): Obtain the set of observed SNP differences $X_a^{(m)} : X_a^{(m)} = U_{alt}^{(m)} \cap S$.
4.3: For each $m = 1, \cdots M'$, obtain $G_{X_a^{(m)}}(\cdot) \equiv \{G_{X_a^{(m)}}(y_i), i = 1, \cdots, k\}$ at the grid sites $y_i, i = 1, \cdots, k$;
4.4: For each $m = 1, \cdots M'$, using $\hat{G}^*(\cdot)$ from step 2.4 in Algorithm 2.2, calculate:
    (a): $\widetilde{KS}_G^{X_a^{(m)}} = \widetilde{KS}(G_{X_a^{(m)}}, \hat{G}^*)$;
    (b): $\widetilde{CvM}_G^{X_a^{(m)}} = \widetilde{CvM}(G_{X_a^{(m)}}, \hat{G}^*)$;
4.5 The Monte Carlo estimates of the power of the test statistics $\hat{\beta}_{\widetilde{KS}_G}$ and $\hat{\beta}_{\widetilde{CvM}_G}$ are as follows, where :
    (a) : $\hat{\beta}_{\widetilde{KS}_G} = \frac{1}{M'} \sum_{m=1}^{M'} I(\widetilde{KS}_G^{X_a^{(m)}} > \hat{F}^{-1}_{\widetilde{KS}_G}(1 - \alpha))$;
    (b) : $\hat{\beta}_{\widetilde{CvM}_G} = \frac{1}{M'} \sum_{m=1}^{M'} I(\widetilde{KS}_G^{X_a^{(m)}} > \hat{F}^{-1}_{\widetilde{CvM}_G}(1 - \alpha))$.
```

2.3.3 Simulation parameter settings and results

Chromosome 19 is selected as an illustrative example to conduct simulation studies. A mouse with a primary mammary tumor and lung metastasis with about 50 putative *de novo* SNP differences between these two tissue samples on its chromosome 19 is selected for consideration here. Based on this example, the total expected number of detected SNP differences η is chosen as 50. Under the null hypothesis, the estimate of the underlying rate parameter of the hPP, $\hat{\lambda}$, is calculated as 1.77×10^{-4} (See step 2.2(a) in Algorithm 2.2 in this chapter).

All of the statistics are evaluated using a grid of values for d or n, which are selected to be scientifically meaningful. In sequencing data, having six or more consecutive mutations with an average distance of less or equal to 1 kb is considered as a mutation shower [8]. Another definition of a mutation cluster, obtained empirically from analysis of a genic region, is having multiple mutations (2 or more) within a 30 kb region [7]. In genotyping array data, as information is missing between SNP probe sites, the evaluation range for identifying clusters would necessarily be larger than the range used in sequencing data with single base pair resolution. In this simulation study, a grid of distances $d_i, i = 1, \cdots, 20$ are set from 5000 bp to $100,000$ bp with an interval of 5000 bp, so $d_i = 5000i$; while a grid of cluster sizes $n_i, i = 1, \cdots, 7$ is set from 2 to 8 with an interval of 1, so $n_i = i + 1$.

Thus there are, in total, 97 statistics formulated: $\bar{R}(d_i), i = 1, \cdots, 20$, $\tilde{R}(d_i), i = 1, \cdots, 20$, $D_{min}(n_i), i = 1, \cdots, 7$; $N_{max}(d_i), i = 1, \cdots, 20$, $C(d_i), i = 1, \cdots, 20$, and the 10 functional forms of these statistics based on KS or CvM frameworks. The critical values for all tests are based on $\alpha = 0.05$.

The results of the study validates the size of the test statistics under both schemes are shown in Tables A.1 to A.8 in Appendix A. For statistics with a single argument, it can be seen that the sizes of $\bar{R}(d)$, $\tilde{R}(d)$, and $D_{min}(n)$ are close to the significance level of $\alpha = 0.05$; the sizes of $N_{max}(d)$ and $C(d)$ can be quite lower than the significance level for some argument settings. For the functional forms of the test statistics, the sizes of all statistics are close to the significance level in both schemes, except that the sizes of functional forms of $D_{min}(n)$ and $C(d)$ are lower than the significance level in Scheme b. That the sizes are somewhat low for $N_{max}(d)$ and $C(d)$ may due to the discreteness of the test statistics.

In power study, for each statistic, the null distribution is estimated from $M = 10^4$ replications generated under the null process. For the alternative processes, the parameters σ and h jointly reflect the spread of clusters of the SNP genotype differences. Here, the truncation range h is set as $h = 3\sigma$, as there are very low probabilities associated with the normal distribution outside this range. In the definition of D_p in alternative hypothesis (3), h_p is set as $h_p = \sigma$; note that $h_p = +\infty$ for alternative hypothesis (1), where $h_p = 0$ for alternative hypothesis (2). The simulation study evaluates power performance of all test statistics with two factors, μ_o and σ. With μ_o and σ specified, the parameter μ_p is set to ensure that the expected number of detected SNP genotype differences $\eta = 50$. The experiment adopts a full factorial design with: (i) μ_o having two levels, 375 and 1125, denoting low and high levels of offspring within a cluster in order that powers of the statistics being evaluated are away from the extremes of 0 and 1, so that the performance of the test statistics can be differentiated; and (ii) σ having levels of grid distances of 500 bp, and from 1000 bp to 10000 bp with increment of 1000 bp. These values of σ are based on the definition of a mutation cluster by [7]; i.e., the truncation range 6σ ranges from 3kb to 60kb.

Figure 2.2 to 2.4 provide power results for $\mu_o = 375$ under each of the three alternative hypotheses. The power of the statistics with relatively lower performance are not displayed. The statistics with fixed argument are only shown for the argument d_{max}, which is the optimal argument for the corresponding parameter setting. The display of the statistics with fixed optimal argument is intended to show the best power performance of the collection of the statistics with various argument settings, which is to be compared with the functional form of the statistics.

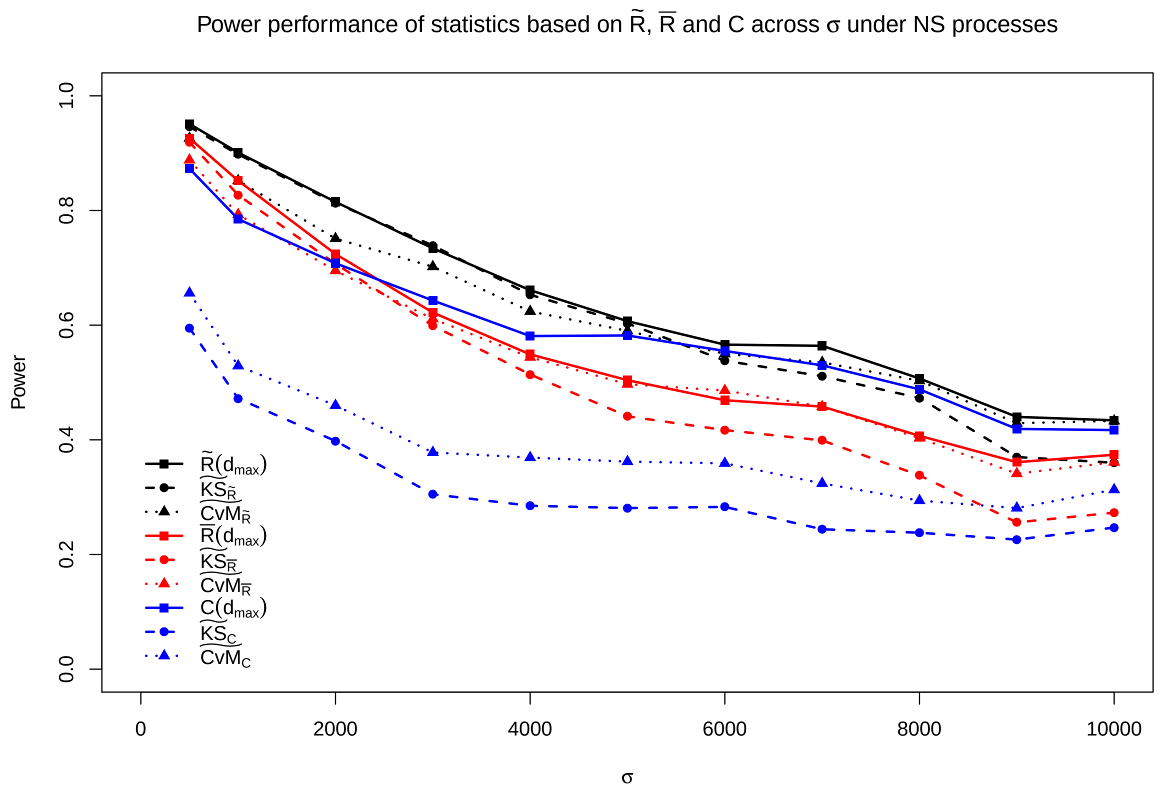


Figure 2.2: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (1) with parameter $\mu_o = 375$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.

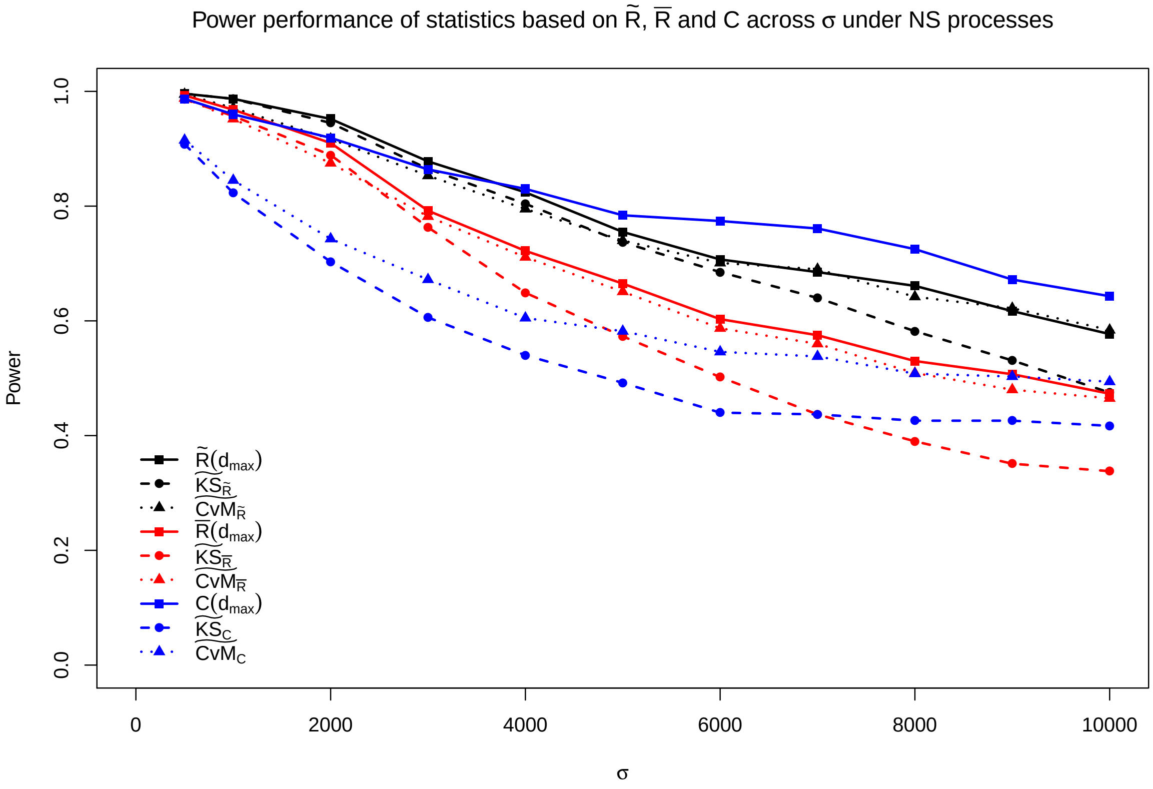


Figure 2.3: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (2) with parameter $\mu_o = 375$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.

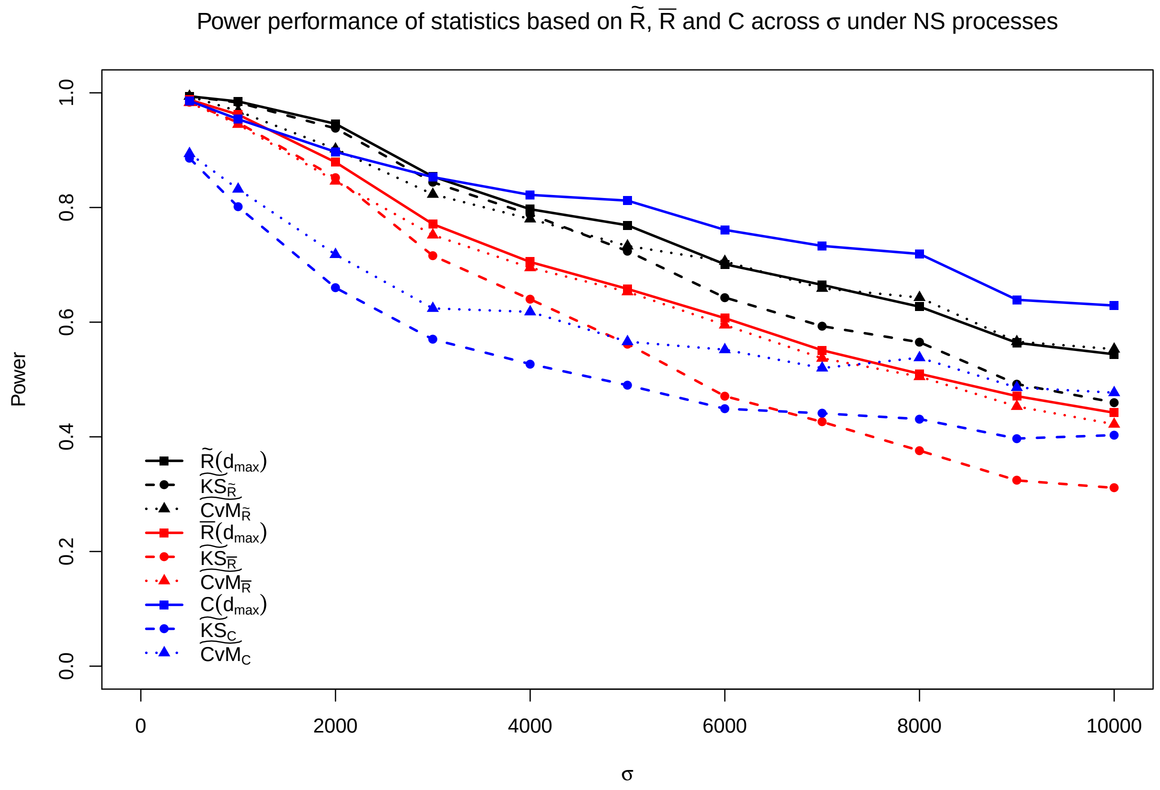


Figure 2.4: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (3) with parameter $\mu_o = 375$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.

Under the alternative hypothesis (1), for $\mu_o = 375$, in general, the power of each test statistic decreases as σ increases. The test statistics based on $\tilde{R}(d)$, $\bar{R}(d)$ and $C(d)$ generally have higher powers than those based on $N_{max}(d)$ and $D_{min}(n)$. Figure 2.2 contrasts the power performance of nine categories of statistics based on $\tilde{R}(d)$, $\bar{R}(d)$ and $C(d)$, including the statistics with fixed arguments as well as their function forms. Among these nine, $\tilde{R}(d)$ has the highest power and outperforms $\bar{R}(d)$ and $C(d)$ in all the settings of σ. The statistics related to $\tilde{R}(d)$ seem to always outperform the statistics related to $\bar{R}(d)$, which may be because $\tilde{R}(d)$ is more robust and numerically more stable than $\bar{R}(d)$. Among the six functional forms of statistics, $\widetilde{CvM}_{\tilde{R}}$ and $\widetilde{KS}_{\tilde{R}}$ outperform the other four functional forms of statistics, and $\widetilde{CvM}_{\tilde{R}}$ has better power performance than $\widetilde{KS}_{\tilde{R}}$ as σ increases.

The power performance under alternatives (2) and (3) for $\mu_o = 375$, available in Figure 2.3 and Figure 2.4, provide similar results to that described for alternative hypothesis (1). Power is generally highest under alternative (2) and lowest under alternative (1) given all the other settings remain constant. One noticeable difference from alternative hypotheses (2) and (3) compared to (1) is that the powers of $C(d)$ outperform $\tilde{R}(d)$ when σ is not small. The comparison among the six functional forms of statistics shows similar results for alternative hypothesis (1).

For $\mu_o = 1125$, the powers of the test statistics are higher than when $\mu_o = 375$. The powers are closer to 1 and decrease less dramatically over σ than for the cases where $\mu_o = 375$. The patterns of power comparisons are similar to the cases where $\mu_o = 375$. Yet the powers of $\tilde{R}(d)$ are comparable with $C(d)$ when σ is large and both are quite close to 1 under alternative hypotheses (2) and (3). The power performance of the statistics under the three alternative hypotheses for $\mu_o = 1125$ is available in Figures A.1 to A.3 in Appendix A.

The power performance of $\tilde{R}(d)$ and $C(d)$ seem to be best among the nine categories of statistics, yet they suffer the disadvantage that they require a choice of d. The optimal argument choices of d are usually unknown in application. Moreover, the optimal choices of d may change over parameter settings, particularly for σ, as seen in Figure 2.5 for $\tilde{R}(d)$. Importantly, using a sub-optimal choice of d can yield very low power.

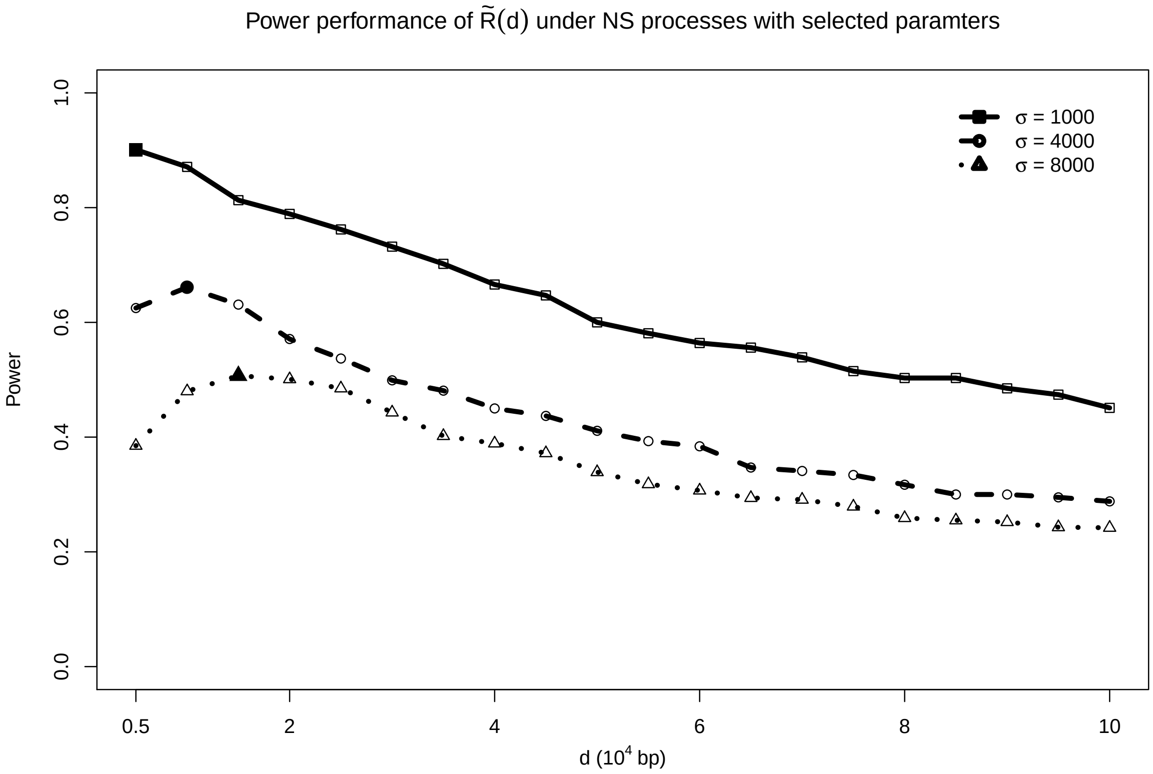


Figure 2.5: Power performance of test statistics $\tilde{R}(d)$ across a grid of d under alternative hypothesis (1) with parameter $\mu_o = 375$.
The solid points indicate the maximum power for the particular parameter setting.

In conclusion, the functional statistic $\widetilde{CvM}_{\tilde{R}}$ is the preferred test statistic in applications because it has the correct size and general high power performance, oftentimes close to the best among all statistics; importantly, with this statistic no specific choice of tuning parameter d needs to be defined.

2.4 Application

2.4.1 Genotyping method

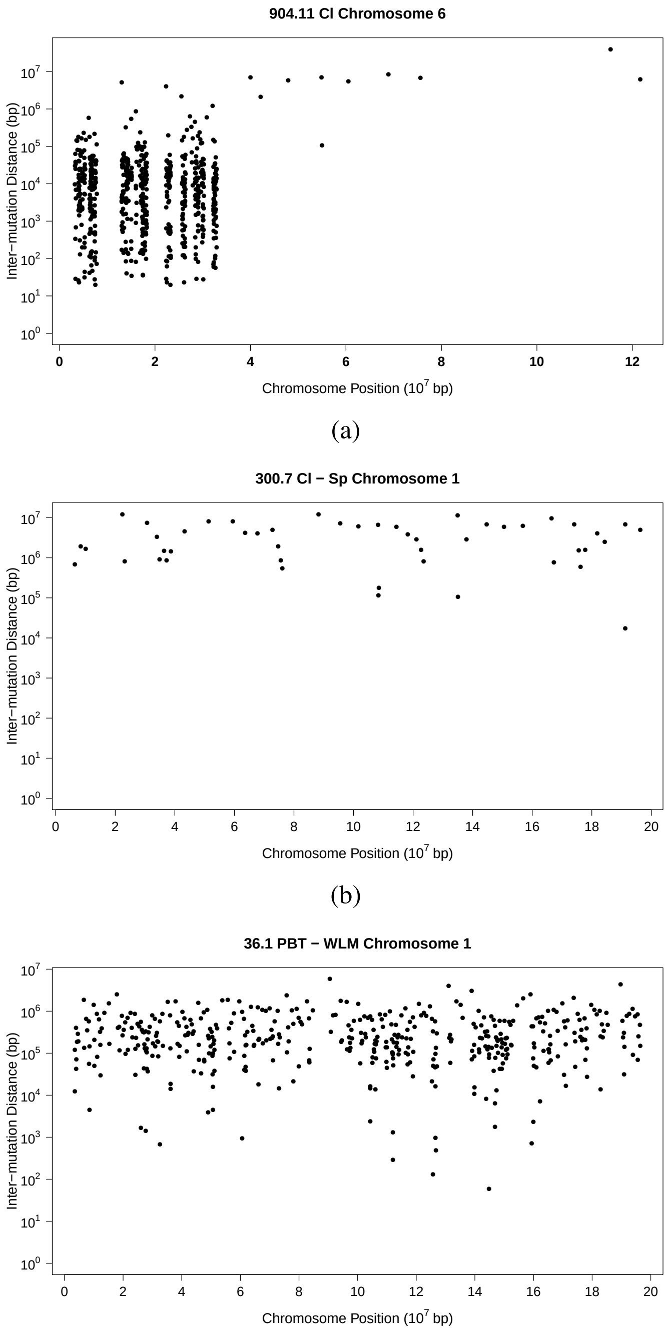
DNA was extracted from mouse tissue samples using the Wizard® Genomic DNA Purification Kit (Promega, Madison, WI). Isolated DNA was submitted to the London Regional Genomics Centre to be processed (restriction enzyme digested, amplified, fragmented and fluorescently labeled) and hybridized to the Mouse Diversity Genotyping Array (MDGA; Affymetrix®, Santa Clara, CA) [11]. Genotyping was performed for each of the three specific examples within the context of separate experimental designs with a minimum cohort size of 12 samples and a maximum of 351 samples. Genotyping Console (Affymetrix®, Santa Clara, CA) was used to call genotypes at the 493,290 SNP loci represented by the MDGA, using the fluorescence intensity data. The Genotyping Console software uses a clustering algorithm, Birdseed v2, and assigns each SNP locus as 1 of 4 possible calls: AA (homozygous for the most common allele), AB (heterozygous, one of each allele), BB (homozygous for the less common allele), or no call if the SNP genotype calls did not cluster well with any of the three possible genotypes. The resulting data for each biological sample used for further analysis consist of a list of SNP genotype calls, their locations in the genome (chromosome number and base pair number) and the genotyping call given by Genotyping Console for each sample. In the data sets utilized for testing for existence of clusters in this research, the events are defined as SNP differences, which are the binary indicators of differences at SNP loci when contrasting two biological samples. The genotyping call and the consequent SNP differences are putative until the genotyping is confirmed by an alternate technology. All animal work was conducted according to relevant national and international guidelines. Western University's Animal Use Subcommittee approved the study. All guidelines were followed including those approved standard operating procedures for euthanasia.

2.4.2 Analyses for three biological samples of interest

Three specific examples are considered here.

1. Detection of known clusters of putative SNP differences in a mouse with a known mixed genetic background;
2. Test for the existence of clusters of putative SNP differences arising postzygotically between two healthy tissues from a C57BL/6J mouse;
3. Test for the existence of clusters in comparison of two cancerous tissues from a MMTV-PyMT transgenic mouse [25].

Rainfall plots portraying the mutation landscapes of the three samples are provided in Figure 2.6. On a rainfall plot, each point represents a single mutation with its distance (in base pairs) to the previous mutation in log scale plotted on the y axis, and the base pair location in the genome is plotted on the x axis. Rainfall plots display mutations detected along a single chromosome or potentially across the entire genome. Although the plots offer a helpful visualization of the data including potential clustering, they do not provide formal evidence of clustering [17].



(a)

(b)

(c)

Figure 2.6: Rainfall plots portraying the SNP differences due to mixed genetic background, putative new mutations arising during development of two normal tissues of the same mouse and putative mutations arising between two cancerous tissues from the same mouse. (A) Rainfall plot for chromosome 6 from a mouse (identifier: 904.11) with mixed genetic background (75% C57BL/6J and 25% CBA/CaJ). (B) Rainfall plot for chromosome 1 for a comparison of normal cerebellum and spleen tissue from the same mouse (identifier: 300.7). (C) Rainfall plot for chromosome 1 for comparison of primary mammary tumor and lung tissue with metastases from a MMTV-PyMT transgenic mouse (mouse identifier 36.1). (Legend: Cl cerebellum, Sp spleen, PMT primary mammary tumor, WLM whole lung with metastases)

As an example of a positive control for known clustered putative SNP differences in a genome, the recommended $\widetilde{CvM}_{\tilde{R}}$ test statistic was used to analyze SNP differences in normal cerebellar tissue from a mouse with a known mixed genetic background of two common inbred mouse strains (75% C57BL/6J and 25% CBA/CaJ), example 1. For chromosome 6 (Figure 2.6a), the test statistic rejects the null hypothesis at a significance level of 0.05, indicating existence of mutation clusters along the chromosome.

In example 2, the $\widetilde{CvM}_{\tilde{R}}$ test statistic was used to analyze SNP differences along chromosome 1 between cerebellar and splenic tissue from a healthy C57BL/6J inbred mouse (Figure 2.6b). The SNP differences detected are hypothesized to have arisen by spontaneous mutation mechanisms resulting in somatic mutations propagated with cell division during development. The test statistic failed to reject the null hypothesis at the significance level of 0.05, indicating no existence of clusters of putative SNP differences along the chromosome.

In the third example, the $\widetilde{CvM}_{\tilde{R}}$ test statistic was used to analyze SNP genotype differences observed along chromosome 1 for a comparison of primary mammary tumor and lung tissue with metastases from the same MMTV-PyMT transgenic mouse (Fig 2.6c). The test statistic rejects the null hypothesis at a significance level of 0.05, indicating existence of mutation clusters along the chromosome. As mentioned in [17], the interpretation of rainfall plots is difficult and subject to pitfalls. The example in Fig 2.6c shows that when a subjective judgment from a visual examination of the rainfall plot is ambiguous and inconclusive, the rigorous statistical tool developed here can provide an objective decision-making approach for detecting the existence of mutation clusters.

2.5 Discussion

In order to perform rigorous statistical testing to detect existence of clusters of putative SNP differences identified by genotyping array probe systems, 97 candidate test statistics are proposed and evaluated. Conditional null distributions of test statistics are obtained by Monte

Carlo simulations. The powers of all the test statistics are studied under three different types of Neyman-Scott processes, intended to mimic the unknown underlying mutation generation mechanisms. Various choices of parameters for alternative hypotheses are used to evaluate the power performance of the candidate statistics. Among all of the parameter settings, the Cramér-von Mises version of the pooled ratio estimate ($\widetilde{CvM}_{\tilde{R}}$) has high power among all candidate tests and lacks dependence on optimal argument choices. It also possesses the desirable property of having correct size and power performance degrade less over various parameter settings as the cluster range becomes larger. The functional form of the $C(d)$ statistic based on the rainfall plot performs substantially poorer. Therefore $\widetilde{CvM}_{\tilde{R}}$ is recommended as an effective statistic for detection of clustering.

The test statistics are developed conditional on the probe design and total number of detected SNP genotype differences. When applied to a new scenario, the null distributions of all the statistics need to be established according to the specific probe design on a chromosome and total number of detected SNP genotype differences using Algorithm 2.2. Depending on the total number of observed SNP differences and probes, the computational time using recommended testing procedure can be in the order of minutes on a PC with a four core Intel i7 CPU. The rate parameter of hPP under the null hypothesis can be estimated from a single chromosome of interest without the need of extra information from other chromosomes in the same biological sample or any other replicates. However, it can also be estimated from several chromosomes under a justified experimental setting. For example, the rate parameter can be estimated from certain replicates which can be assumed to share a common underlying mutation rate under certain experimental conditions. When the objective is to carry out the mutation cluster detection genome-wide, all of the chromosomes in a sample should be tested separately. Multiple testing issues arise when the statistic is applied to multiple chromosomes from either one or a number of biological samples. These multiple tests can be independent or correlated depending on the biological context. In order to achieve a desirable overall type I error rate or false discovery rate (FDR), statistical methods such as the Bonferroni correction or by [26] may be applied to achieve desirable testing properties, depending on the goal of the research.

The methods developed in this article are designed for cluster detection under a genotyping array probe design. The probe design provides a cost-effective way for mutation detection compared to sequencing every base pair of the entire genome. Instead of a high resolution of mapping of mutations in the genome, the probe system usually only reveals a small proportion of information on a chromosome, leaving the regions outside of the probe sites unknown. As mutations in regions where probes are absent are undetectable by design, any mutation clusters occurring in such regions are correspondingly undetectable. The test statistics are established based on the information on the probe system, so they can only identify clustering when the probe system is capable of detecting potential clusters. The power performance of the test statistics in this study is evaluated under the alternatives that there exist underlying clusters generated from a known clustering mechanism. This mechanism does not necessarily guarantee that clusters are detectable by the specific probe system. If all the samples evaluated in the power studies contained clusters detectable by the probe systems, the power performances of the tests would most likely be higher. One of the reasons for some low power performances in certain alternative parameter settings may be that clusters generated are not detected by the probe system. Designing an array with a larger number of probes or switching to an existing array system with a larger number of probes will augment the probability of detecting existing clusters.

In studies involving known genetic backgrounds, prior information on detected SNP differences may be utilized to improve the power of testing for mutation clusters. For example, information on SNP differences in high linkage disequilibrium (LD) with more unobserved SNP differences in their neighborhood may be given greater weight in the testing procedure. Alternatively, information on SNP genotypes undetectable by the microarray platform may be inputed based on other information such as known haplotypes[27]. However, for studies with *de novo* mutations, such as in healthy somatic tissues and in cancer studies, the imputation based on LD or known haplotypes may not be appropriate; even so, other prior knowledge may become helpful. Extensions of the methods discussed in this paper could incorporate improvements based on such prior knowledge.

The statistical tool recommended in this research can test for the existence of clusters of SNP differences, but cannot identify the locations and the sizes of the clusters. The tool is ideal for large scale experimental designs, which are usually intended to compare various exposures. The feature of existing clusters can be compared among experimental groups with different exposures. After mutation clusters have been detected, different downstream analyses are possible. The nature of the mutation types in clusters can be used to identify mutation signatures and to infer the underlying mutational mechanisms. Alternatively, the mutation clusters can be linked to functional annotations for the genome and inferences can be made about the functional impact of the mutation clusters.

The SNP differences here in this study are used as an example of genetic events that take place on one base pair on the chromosome. The method developed here can be applied to cluster detection of any single site event along any one dimensional system. These events can be defined by biological researchers based on their genetic contexts and study interest. An example is the distribution of DNA methylation locations detected by the CpG site probe system as described by [13]. The method provided here is not only applicable to the mouse, but also to other organisms, e.g. human, with microarray designed for SNP genotype detection. The method can be generalized to any one dimensional system where every site is observed, such as DNA or protein sequencing data, with probes designated as having length one at each site of the system.

The arbitrary and informal graphical tools and definitions for portraying and detection mutation clusters can now be replaced with a formal statistic test for mutation cluster detection. The recommended test statistics in this study provide tools for genome-wide detection of mutation clusters under the genotyping probe system. Due to the cost-effectiveness of array systems, larger scales of experimental designs can be adopted compared to those possible with next generation sequencing techniques. Certain samples with putative mutation clusters can be further confirmed and investigated by sequencing techniques.

Chapter 3

Effect of probe design in microarray type data on powers of test statistics for cluster detection

3.1 Introduction

Mutations are alterations of DNA composition, arising during the regular course of cellular development or due to environmental factors. The study of signatures of mutations may help provide a better understanding of mutational mechanisms, genetic diversity, disease, adaptation and evolution. Although the allocations or spatial properties of the mutations on the chromosomes have not been the focus of biological research in the past, the recent discovery of the phenomenon that mutations may form clusters in space has attracted a lot of attention [2], because of potential connections with known or unknown mutagens or mechanisms. Understanding the mechanisms that generate mutation clusters may also be crucial for research on mutageneses and carciogenesis.

Currently there are two main technologies to measure mutations on the genome scale in DNA samples: next-generation sequencing and the microarray. Next-generation sequencing, also known as high throughput sequencing, has been developed in recent years to measure DNA composition in single nucleotide resolution, the highest resolution possible, in a much faster and cheaper way than the previously used sequencing technologies developed in the 1970s. With the efforts of researchers worldwide, the efficiency of whole next-generation sequencing has been improved even while the cost has been reduced, especially for commonly studied organisms. For example, the price for sequencing the whole genome of a human has decreased to the order of a thousand dollars. Even so, the price still remains very high, especially for less commonly studied organisms. For example, for a mouse, although a widely studied model organism, the cost of whole genome sequencing could be on the order of thousands of dollars. Though price reductions are still on the horizon, substantial changes in pricing may take some effort yet. Importantly, for human screening purposes or for studies with a relatively large populations, next-generation sequencing is still not cost efficient or even feasible.

On the other hand, the microarray platform provides an alternative approach at a lower cost to measure mutations. A type of mutation can be genotypic alteration at a single nucleotide polymorphism (SNP) locus, referred to as a ‘SNP genotype difference’, or in short as ‘SNP difference’ between two samples. The SNP genotypes are measured by a collection of probes, specially designed for specific purposes for each organism. Only the SNP genotypes at specific loci on the probes can be detected by the microarray platform, while those outside the probe sites remain unknown. The number of targeted probe sites usually covers far less than the length of the entire genome, leading to an observation of the SNP genotypes with potentially substantial missing data.

For example, in the Mouse Diversity Genotyping Array (MDGA) [11], the total number of SNP probe sites is 493,290, which is much fewer than the mouse genome size of 2.8 billion base pairs (bp). The number of SNP probe sites and their locations in the microarray are selected by a design procedure. Genomic variations in the form of SNP differences from 7 categories of mouse samples were used to develop the design of MDGA. From samples in each category, probe site locations were selected. For example, in the first category of 25 widely used classical laboratory mouse strains, the genome was divided into non-overlapping 40kb intervals. In each interval, the top three SNP loci with the highest minor allele frequency and low missing rate were selected. In the second category of 15 strains from National Institute of Environmental Health Sciences (NIEHS), probe sites were selected based on local phylogenetic trees. Other constraints to ensure feasibility of the physical experiment, such as annealing temperature, GC content, also need to be considered in the selection of probe sites. With the observation of a small proportion of nucleotides spanning the entire genome, the microarray platform provides a much cheaper and feasible approach to detect mutations and become a screening tool in a study with a large population, if it can be shown to be effective.

SNP differences detected from the microarray platform can be used for the detection of the existence of mutation clusters. Formal statistical tools have been developed for such a purpose in Chapter 2. The power performance of several statistics were studied based on a real probe setting on one chromosome. One of the statistics emerged as having excellent power and is further utilized here to study the properties of the tests across a variety of probe designs.

This information and the approaches considered here are useful for the design of microarray platforms.

In this investigation, we compare various microarray platforms in terms of their capabilities for detecting the existence of clusters. Here we develop procedures to obtain: (1) simulated probe sets using a window-based probe selection procedure from two types of simulated genomic variation libraries: the homogeneous Poisson process library and the Neyman-Scott library; (2) probe sets filtered from existing MDGA probe design. We also develop a framework and algorithm to numerically calculate the rejection rates of the test statistics developed in Chapter 2 under these various probe designs. We compare the power of these proposed test statistics under (i) the simulated probe settings with two varying factors: a parameter defining the probe selection procedure, and the type of genomic variation library; as well as (ii) the probe settings that are based on MDGA. The framework developed provides valuable insights to help determine key features such as what level of probe intensity would provide for high power in detecting clusters under a variety of mutation levels, for the design of microarray platforms for the purpose of mutation cluster detection for population screening.

In Section 3.2 of this chapter, we first introduce the data structure of the SNP differences detected under the microarray platform as well as the process for the construction of simulated probe settings considered here. Section 3.3 describes the study design and the results for comparison of various probe designs. Section 3.4 discusses the roles of various factors in the simulation study, and insights the study framework provides on the effectiveness of microarray platforms as screening tools.

3.2 Statistical methods

For genome-wide mutation cluster detection with information collected from the microarray platform, when focusing on an individual chromosome, the data consist of two components: the set of probe locations on the chromosome $S : S \subset \mathbb{R}^+$ and the set of locations of SNP differences detected by the probes $X : X \subseteq S$. The probe locations S are determined by the design of the mircroarray platform. Only SNP differences falling inside the probe regions can potentially be detected, while those falling outside of the probe regions cannot. Denote all the underlying SNPs on the sample chromosome as U, then $X = U \cap S$. Thus, given the same DNA sample with underlying SNP differences U, an alternative probe system S^* may yield a different observation of the set observed SNP differences $X^* = U \cap S^*$. The statistical inference for cluster detection of SNP differences relies on both the probe set S and the corresponding observation X. Thus altering the probe design S may also alter inference.

The goal of this study is to evaluate the effect of probe design on powers of test statistics for cluster detection. Various alternative probe designs are required to represent different probe features, such as probe intensity and probe allocation. There are two strategies adopted in this research to achieve probe design alterations - the use of: (1) probe sets designed from simulated libraries of genomic variations; (2) probe sets filtered from the MDGA probe design.

3.2.1 Probe sets designed from simulated libraries

The probe design procedure follows the essence of the design process for the Mouse Diversity Genomic Array (MDGA) [11] and can be divided into two main parts. The first is generation of a library that stores all the genomic variation information in a population. The second is the probe selection procedure, which selects the locations of the probes on the chromosome according to the genomic variation information in the library.

3.2.1.1 Library generation

The library of genomic variation is a collection of DNA information of the species of interest in the study. It provides sources of genetic variability, which indicates plausible locations of SNP probe sites and consists of genetic information from a number of samples depending the array being constructed. For example, in the design of MDGA, the sources of genetic variations are 25 widely used classical laboratory mouse strains.

The information on genetic variability in a library is determined by the group of biological samples of interest, which may have different properties in terms of their genetic variation behavior. To consider a broad coverage of genetic variations, the library is obtained through simulations of genetic variations under different models.

For a target chromosome of a simulated biological sample, consider the two following models: the homogeneous Poisson model and the Neyman-Scott model. In the homogeneous Poisson model, genetic variation follows a homogeneous Poisson process (hPP). Under such a process, every base pair on the chromosome acts independently and has an identical probability of having a genetic variation. The Poisson intensity in this model is denoted as λ. In the Neyman-Scott (NS) model, the genetic variation follows a parent-offspring type of clustering process. Under this clustering process, unobservable parent sites with an expected number μ_p are generated along the chromosome from an hPP. Each parent site generates a cluster of offspring sites with expected number μ_o. The offspring sites are independent and identically distributed truncated normal random variables centered at the parent site. The standard deviation of the truncated normal distribution is denoted as σ. The half-length of the window of the truncation range is denoted as h. Neither the hPP model nor the NS model perfectly portrays genetic variations in real world libraries. They are utilized as references for two ‘extreme’ conditions of complete spatial randomness and clustering process. The genetic variations in the real world libraries likely lie somewhere in between these two extremes.

Thus two libraries storing genetic variability information from N biological samples can be constructed: one with all samples independently following the homogeneous Poisson model, another with the Neyman-Scott model; these are named the hPP library and the NS library respectively. The hPP library represents a population with randomly spaced genetic variations, while the NS library represents a population with clustered genetic variations.

3.2.1.2 Window-based probe selection procedure

Probe sites are selected using a window based design procedure mimicking the design of the MDGA array. As well, this design offers even coverage of the genome which is important for cluster analyses. In this window based design, the chromosome is divided into non-overlapping windows of size l_w. In each window, n_w probe sites are selected based on similar criteria for selections as for the MDGA design. The main criteria are high frequency of genetic variation observed amongst samples in the library at a particular site, and the importance of sites for the construction of local phylogenetic trees. The dominance of each of these two criteria depends on the goal of the probe design, with the former being dominant where sites with higher mutation rates are critical to capture, and the latter considered when the focus is identification of the difference amongst samples. For this study, high frequency of genetic variation is the criterion used to select probe sites because of the goal of mutation cluster detection. For each window, the n_w sites with the highest count of genetic variation are selected. When ties occur, random selection from the ties is conducted to achieve the total as n_w. The design parameters l_w and n_w are adjusted to vary the overall total number of probes. Algorithm 3.1 specifies the window based probe selection procedure.

Algorithm 3.1: Window-based probe selection procedure

1.1 For a chromosome of length L, denote the chromosomal position of each base pair as $\{1, \cdots, L\}$;

1.2 Denote the set of chromosomal positions of genomic variations in each of the N samples in the library as $T_c, c = 1, \cdots, N$;

1.3 Divide the chromosome into $m_w = \lceil L/l_w \rceil$ consecutive non-overlapping windows, each of size l_w. If $m_w l_w > L$, then the number of probe sites chosen for the last window that then has length less than l_w would be adjusted proportionate to the length of that window. In our studies, $m_w l_w = L$. Denote the windows as $\{W_{(i)}\}, i = 1, \cdots, m_w$. For the ith window $W_{(i)}$, the chromosomal positions are $\{b_{ij} = (i-1)l_w + j, j = 1, \cdots, l_w\}$.

1.4 In the ith window $W_{(i)}$, for each chromosomal position b_{ij}, obtain the count of samples in the library that have a genomic variation at this location $C_{ij} = \sum_{c=1}^{N} I(b_{ij} \in T_c)$, $j = 1, \cdots, l_w$, where $I(\cdot)$ is the indicator function.

1.5 The chromosomal position of the largest n_w elements in the vector $(C_{i1}, \cdots, C_{il_w})$ and their corresponding chromosomal positions $s_{i1}, \cdots, s_{in_w}$ are selected as elements in the probe set S. In case of occurrence of ties, random selection from the ties is conducted to ensure the number of total selected sites is n_w.

1.6 Repeat step 1.4 to 1.5 for every window $W_{(i)}, i = 1, \cdots, m_w$. Obtain the probe set $S = \cup_{i=1}^{m_w} \cup_{k=1}^{n_w} s_{ik}$.

3.2.2 Probe sets filtered from MDGA

Probes from the MDGA microarray probe design are filtered by randomly selecting sites without replacement. Denote the target probe intensity for the filtered probe set as λ_s. The number of probes is then calculated as $[\lambda_s L]$, where $[x]$ is the nearest integer of x and L is the length of the chromosome. Then the filtered probe set is obtained by selecting a random sample of size $[\lambda_s L]$ without replacement from the probe set for the corresponding chromosome on MDGA.

3.2.3 Methodology for power study

The emphasis here is the study of the effect of probe designs on the performance of test statistics for detecting the existence of mutation clusters. Once a probe set S is determined either from a simulated library or by filtering from MDGA, the corresponding observed SNP differences can be obtained for a given sample as $X = U \cap S$, where recall U is the complete set of SNP differences for test sample over the whole chromosome.

We utilize statistics developed in Chapter 2 that have been demonstrated as having good performance for cluster detection. The test statistics are defined as follows:

$$KS_{\tilde{R}} = \sup_{y} |\tilde{R}(d) - \tilde{R}^{*}(d)|, \tag{3.1}$$

$$CvM_{\tilde{R}} = \int [\tilde{R}(d) - \tilde{R}^{*}(d)]^{2} dy, \tag{3.2}$$

where

$$\tilde{R}(d) \equiv \tilde{R}_S(d) = \frac{\sum_{x \in X} N_X(x, d)}{\sum_{x \in X} N_S(x, d)}, \tag{3.3}$$

and where for arbitrary set A, fixed distance d, and specific SNP difference x,

$$N_A(x, d) = \sum_{z \in A} I(0 < |z - x| \leqslant d), \tag{3.4}$$

and $\tilde{R}^*(d) = E_0(\tilde{R}(d))$, the expectation being taken under an appropriate null hypothesis.

To numerically approximate the test statistics, we evaluate the supremum in equation 3.1 and integral in equation 3.2 over D, $D = \{d_{(1)}, \cdots, d_{(k)}\}, k = |D|$, D being a discrete set of grid points for distance argument d. A finer choice of grid points leads to a better approximation. Suitable choices were discussed in Chapter 2, and adopted here.

When D is not well specified, some elements of D may lead to invalid calculation of $\tilde{R}(d)$, as the denominator in (2.3) can be zero. This is because when d is selected small enough, there is no probe within distance d to an observed SNP difference x except for the one that detects x itself. A subset of D, however, usually exists such that calculation of $\tilde{R}(d)$ is valid. Given the grid set D, suppose for a sample X, $\exists D_v \equiv D_v(X) \subset D$ and $D_v \neq \phi$, such that $\forall d \in D_v(X)$, $\tilde{R}(d) \in \mathbb{R}$ and $\forall d \in D_v(X)^C \cap D$, $\tilde{R}(d) \notin \mathbb{R}$. The calculation of discretized functional statistics to approximate 3.1 and 3.2 are as follows:

$$\widetilde{KS}^*(\tilde{R}, \tilde{R}^*) = \max_{d \in D_v(X)} |\tilde{R}(d) - \tilde{R}^*(d)|, \tag{3.5}$$

$$\widetilde{CvM}^*(\tilde{R}, \tilde{R}^*) = \frac{\sum_{i=1}^{k_v-1} \left\{\{[\tilde{R}(d_{v(i)}) - \tilde{R}^*(d_{v(i)})]^2 + [\tilde{R}(d_{v(i+1)}) - \tilde{R}^*(d_{v(i+1)})]^2\}(d_{v(i+1)} - d_{v(i)})\right\}}{2(\max_{d \in D_v(X)}(d) - \min_{d \in D_v(X)}(d))}, \tag{3.6}$$

where $D_v = \{d_{v(1)}, \cdots, d_{v(k_v)}, k_v = |D_v|\}$.

Note that $\widetilde{KS}^*(\tilde{R}, \tilde{R}^*)$ and $\widetilde{CvM}^*(\tilde{R}, \tilde{R}^*)$ are referred to as the CvM and KS test statistics respectively in subsequent discussion.

To study the rejection rate under any given probe setting S, the homogeneous Poisson process is adopted as the null process while the Neyman-Scott cluster process is adopted as the alternative. Algorithm 3.2 identifies the calculation of the rejection rate given any specific probe setting S.

Algorithm 3.2: Calculate rejection rate given probe setting S

2.1 Generate M_a replications of underlying SNP differences $U_a^{(r)}, r = 1, \cdots, M_a$ based on an alternative NS process with the parameter setting of μ_{ap}, μ_{ao}, σ_a and h_a;

2.2 For the sample of $U_a^{(r)}$, generate the detected SNP differences $X_S^{(r)} = U_a^{(r)} \cap S$ based on the probe design S;

2.3 For detected SNP differences $X_S^{(r)}$, obtain Monte Carlo estimations of the null distributions of the KS and CvM statistics and hypothesis testing procedure based on Algorithm 2.2 and 2.3 in Chapter 2, replicated as shown in Algorithm S1 and S2 in Appendix B;

2.4 Denote $R_{KS}^{(r)}/R_{CvM}^{(r)}$ as an indicator for rejection of the null hypothesis for sample $X_S^{(r)}$ based on the KS/CvM statistics, where $R_{KS}^{(r)}/R_{CvM}^{(r)} = 1$ indicates rejection of the null hypothesis and 0 otherwise.

2.5 Repeat the procedure from 2.2 to 2.4 for $r = 1, \cdots, M_a$. Then the rejection rates based on KS and CvM statistics are $R_{KS}^S = \sum_{r=1}^{M_a} R_{KS}^{(r)}/M_a$ and $R_{CvM}^S = \sum_{r=1}^{M_a} R_{CvM}^{(r)}/M_a$ respectively.

3.3 Power study

3.3.1 Parameter settings

For the construction of probe sets designed from simulated libraries, a pseudo chromosome of length $L_0 = 10^7$bp is considered in this simulation study, without loss of generality, to maintain feasible computational load. Though this length is close to the order of a typical mouse chromosome length of 10^8bp, it is important to note that different organisms have different chromosome lengths and these may vary substantially in scale.

For library generation, to obtain the genomic variation library based on a hPP, the Poisson intensity λ is set as 0.002, which is within the range of typical rates discussed in the genomic literature [28]. For the library based on the NS process, the parameters are set as $\mu_p = 100$, $\mu_o = 200, \sigma = 10^4, h = 3 \times 10^4$, close to values reported in Chapter 2 for the analysis of mouse data. For each library, $N = 100$ samples are generated for selecting probe sites as described in the next paragraph.

For the window-based design procedure, for each type of library, 3 window lengths l_w are chosen as $20,000$bp, $100,000$bp and $200,000$bp. For $l_w = 20k$bp, the number of sites selected in each window n_w is set as $1, 2, 3$, or 4, which would yield a total number of probes as $500, 1000, 1500, 2000$, respectively, with corresponding overall probe intensities as 5×10^{-5}, 1×10^{-4}, 1.5×10^{-4}, 2×10^{-4}. For $l_w = 100k$bp, n_w is set as 1 through 10 with increments of 1, as well as 15 and 20. These settings of n_w would yield a total number of probes as 100 to 1000 with increment of 100 and additionally 1500 and 2000. The corresponding overall probe intensities are from 10^{-5} to 10^{-4} with increment of 10^{-5}, 1.5×10^{-4} and 2×10^{-4}. For $l_w = 200k$bp, n_w is set from 2 to 20 with increments of 2 as well as 30 and 40. These settings yield the same number of probes as well as overall probe intensities corresponding to the $l_w = 100k$bp settings. Note that when the length of the chromosome is 10^7 and the total number of probes is 2000, the overall probe density in the simulation is comparable to the average density observed for the mouse SNP microarray (MDGA).

For generation of probe sets filtered from MDGA, probe set on chromosome 19 from MDGA array is utilized for filtering. The probe intensity on chromosome 19 is 1.938×10^{-4} probe/bp, which is comparable to the highest probe intensity of 2×10^{-4} probe/bp in the probe sets designed from simulated libraries. The probe set on chromosome 19 from MDGA is then filtered, targeting the probe intensities of 10^{-5} to 10^{-4} with increment of 10^{-5}, as well as 1.5×10^{-4} from those simulated probe sets with window length $l_w = 100k$bp and $l_w = 200k$bp.

For the power study, for both simulated probe sets and the probe sets filtered from MDGA, the number of samples generated under the alternative process M_a is set as 100. For the simulated probe sets, the parameters in the alternative NS process are set as $\mu_o = 100$, $\mu_p = 200$, $\sigma = 5 \times 10^4$, $h = 1.5 \times 10^5$. For the probe sets filtered from MDGA, in order that the alternative process is comparable with that for the simulated probe sets, the parameter μ_p in the NS process for the filtered probe sets is multiplied by a factor of $\frac{L_M}{L_0}$ due to the change of chromosome length, where L_M here is the length of chromosome 19. By using this multiplier, the expected number of mutations per unit length is the same between the simulated probe settings and the filtered probe set from chromsome 19. The remaining parameters in NS process are kept the same as previously in order to have the identical behaviour of offsprings given a parent location in a cluster. That is, we set $\mu_p = 100\frac{L_M}{L_0}$, $\mu_o = 200$, $\sigma = 5 \times 10^4$, $h = 1.5 \times 10^5$.

For each sample from an alternative process for either simulated probe sets or the probe sets filtered from MDGA, for each specific probe set, a Monte Carlo estimate of the null distribution of summary statistics is calculated as in Algorithm S1 in the Appendix, with the number of replicates M(See Appendix B) set as 1000. The grid points D utilized in this algorithm are set from 10^3 to 10^5 with increments of 10^3.

3.3.2 Power comparisons

The rejection rates under the alternative process based on hPP mouse library for the CvM and KS statistics are shown in Figures 3.1 and 3.2, while corresponding rates for the NS library are displayed in Figures 3.3 and 3.4. In Figures 3.5 and 3.6, we provide a comparison of rejection rates based on the filtered probe sets of chromosome 19 with those from probe sets based on the simulated hPP and NS libraries with $l_w = 100k$ for the two test statistics.

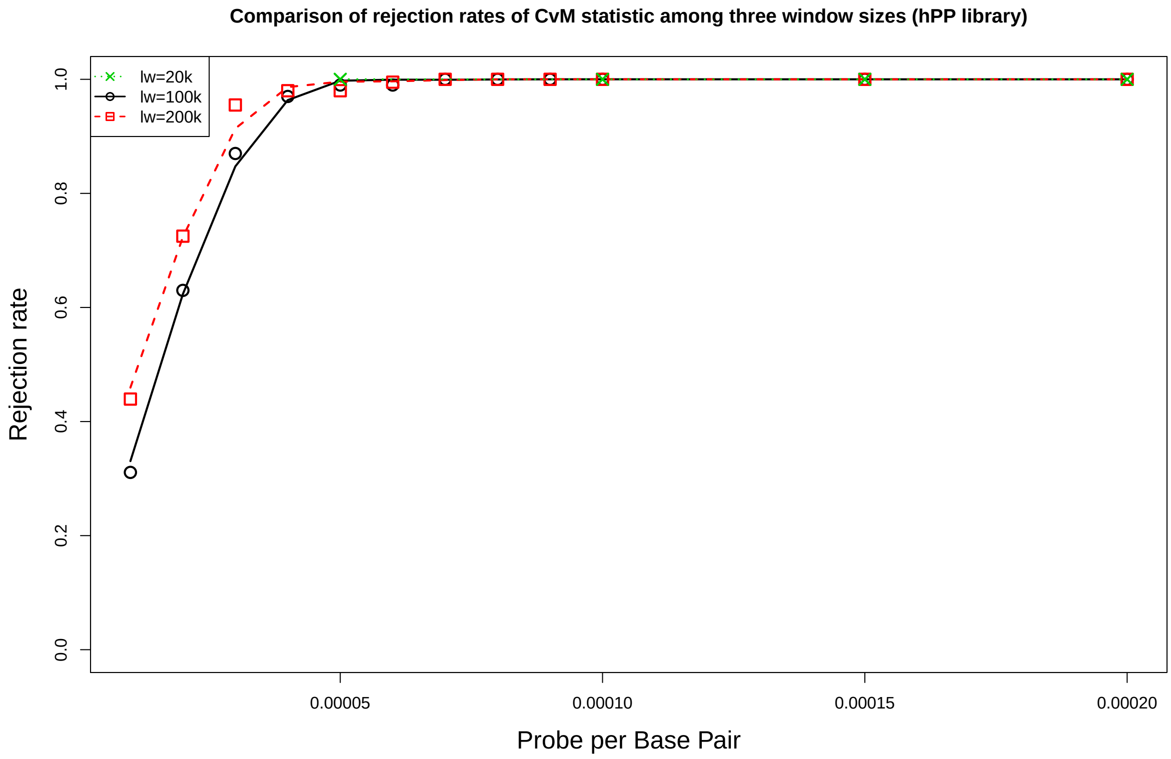


Figure 3.1: Comparison of rejection rates of $CvM_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from hPP samples.

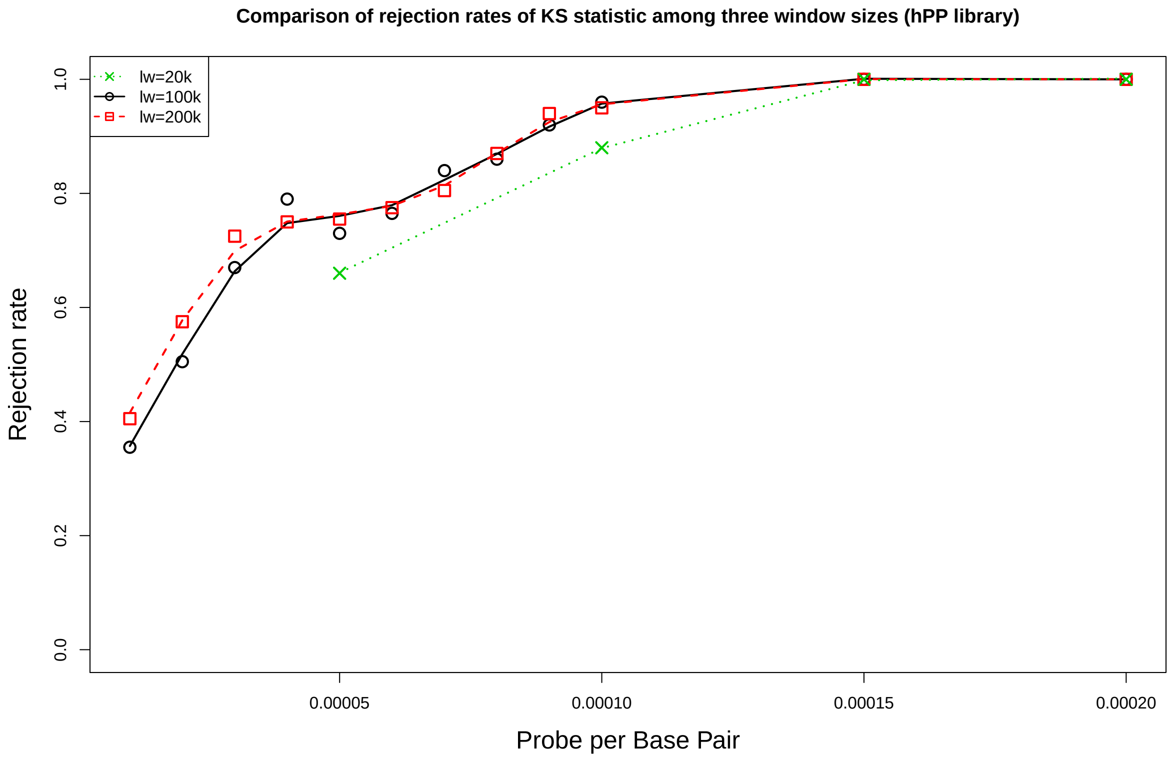


Figure 3.2: Comparison of rejection rates of $KS_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from hPP samples.

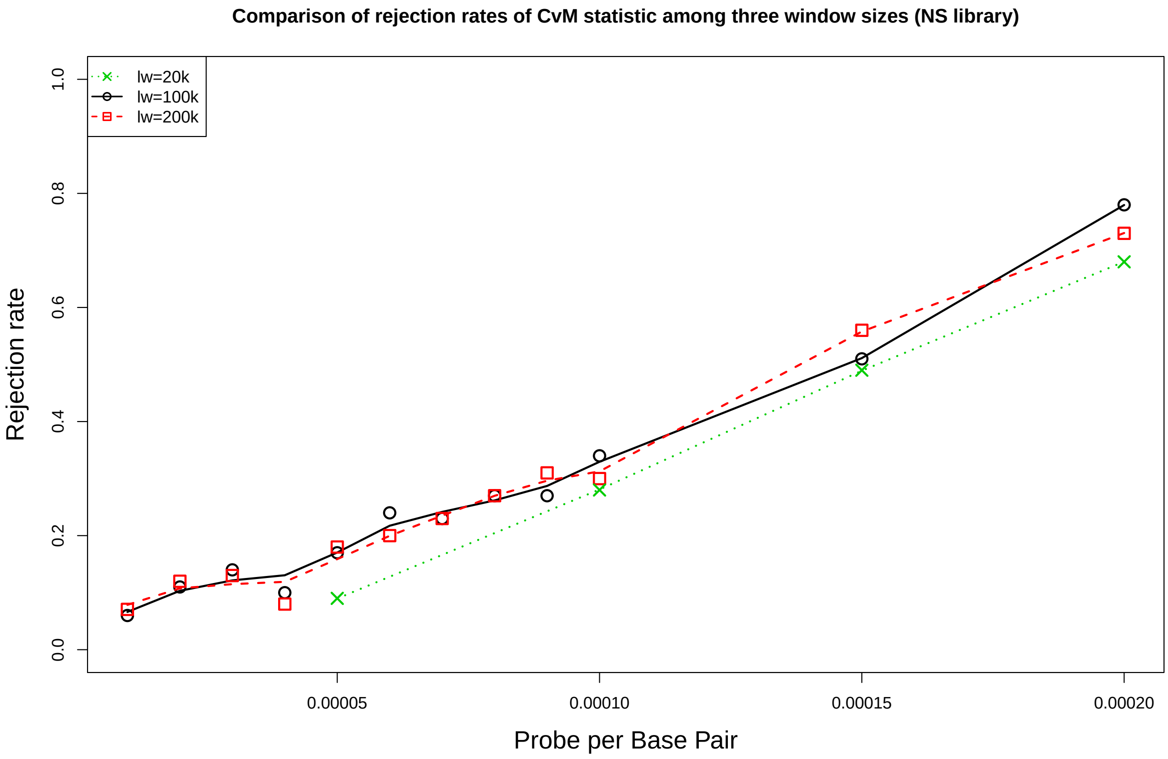


Figure 3.3: Comparison of rejection rates of $CvM_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from NS samples.

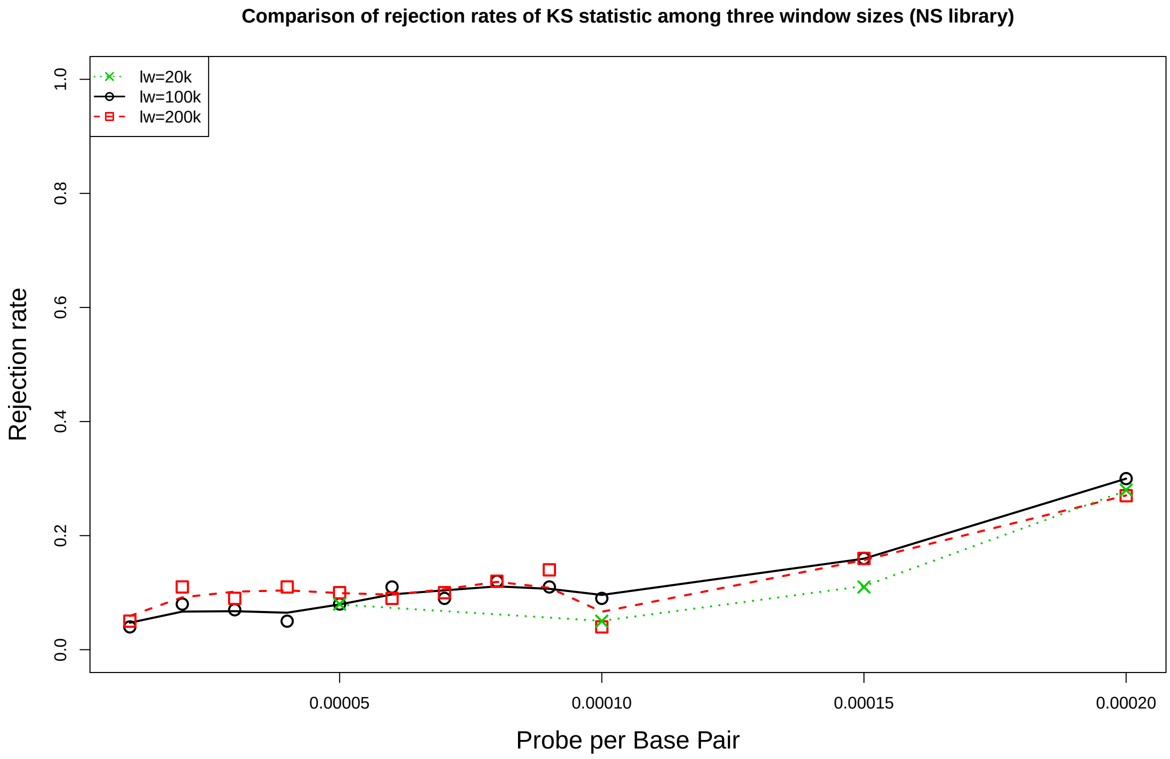


Figure 3.4: Comparison of rejection rates of $KS_{\tilde{R}}$ under various probe settings among three window sizes. The mouse library is constructed from NS samples.

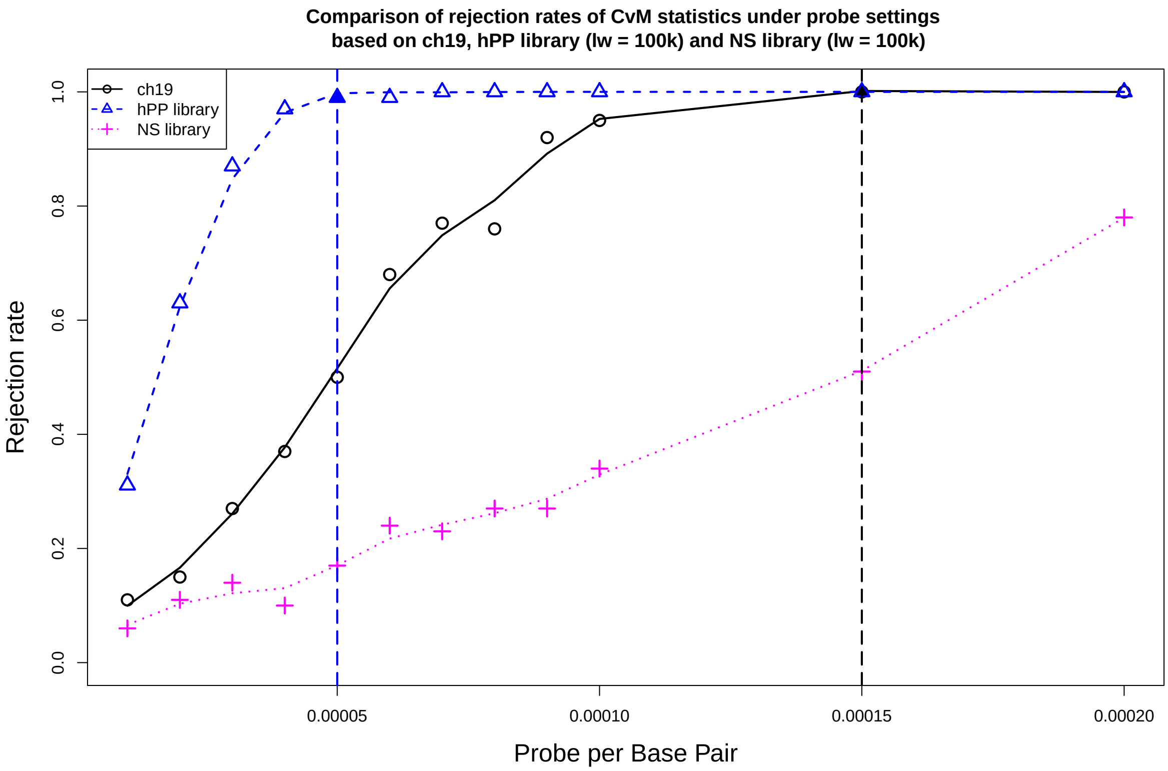


Figure 3.5: Comparison of rejection rates of $CvM_{\tilde{R}}$ under filtered probe sets based on chromosome 19, hPP library ($l_w = 100k$) and NS library ($l_w = 100k$). The blue and black dashed vertical lines indicate probe intensities required for high performance with the hPP library and filtered probe sets on chromosome 19 from MDGA respectively.

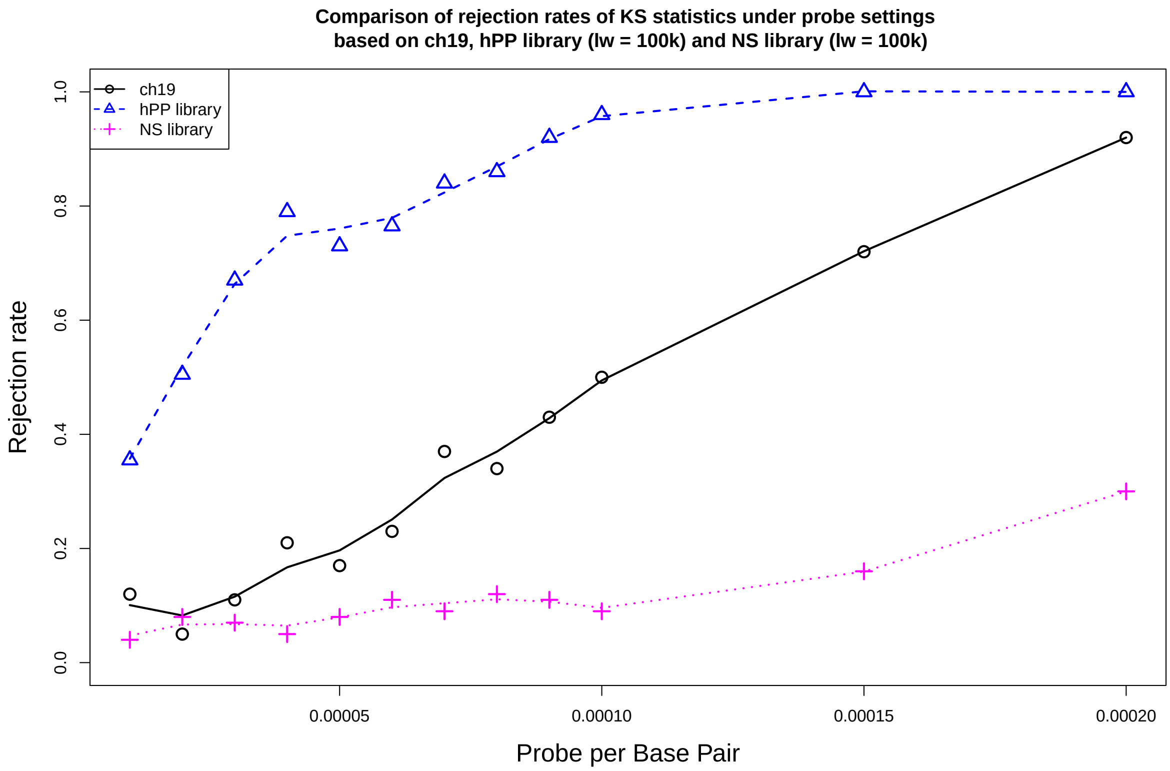


Figure 3.6: Comparison of rejection rates of $KS_{\tilde{R}}$ under filtered probe sets based on chromosome 19, hPP library ($l_w = 100k$) and NS library ($l_w = 100k$).

For any probe design, or alternative process, and for any window size, as expected, the rejection rates tend to increase as the probe intensity increases over the range considered with more information being available from higher intensities. Rejection rates under the NS library tend to be considerably lower than those under the hPP library, again as expected because of the increased variability in event locations corresponding to the NS library. The difference in rejection rates for these two libraries is striking.

Though window lengths chosen for the study were considerably different, no substantial differences in rejection rates were observed over the window lengths considered, with $l_w = 20k$ having generally lowest power, and largest power seen with $l_w = 200k$. Under any setting, the CvM statistic has better performance in terms of rejection rate than the KS statistic.

Comparing probe sets designed from simulated libraries and those filtered from the MDGA design, Figures 3.5 and 3.6 show that the rejection rates from the filtered probe sets are in between those from the hPP and NS libraries. When the intensity is high, the filtered probe set is closer to the hPP library. In particular, at the probe intensity of 2×10^{-4}, which is close to the actual probe intensity of MDGA in chromosome 19, the rejection rates of the filtered and hPP libraries are quite close, and the rejection rates of CvM statistic is close to 1. Hence, good performance of the CvM statistic is expected with MDGA on chromosome 19.

From Figure 3.5, we see that the power of the CvM statistic achieves a value of about 1 at probe intensity of 5×10^{-5}, while that from the filtered MDGA probe set achieves the peak of 1 only with a much higher probe intensity over 1.5×10^{-4}. We also investigate the power performance of the test statistics under the sequencing platform, which can be viewed as an extreme case of microarray platform with designated probes for every base pair and thus having probe intensity of 1/bp. In this case, both the CvM and KS statistics yield power of 1 (results not shown). This indicates that for the mutation level corresponding to the alternative process a high power would have been achieved with a lower probe intensity by using 75% of the probes associated with MDGA. Thus changing the probe intensity from about 2×10^{-4} to 1.5×10^{-4}, the high performance of the test for detecting the existence of mutation clusters could have been maintained, indeed as good as sequencing. Under the same mutation level, for probe sets designed from a hPP library, a quarter of the probes would be required to maintain the performance.

3.4 Discussion

To study the effect of different probe designs on the powers of test statistics for mutation cluster detection, this research adapts real word microarray probe design strategies as well as develops a framework to develop probe designs from simulated genomic variation libraries. Using a selection of probe designs with different intensities, power curves of the test statistics for mutation cluster detection are obtained by testing on Monte Carlo simulated samples under the alternative Neyman-Scott process. Factors in the probe design simulation, such as types of libraries, window-size choices are considered and compared in the simulation study to demonstrate their effects on the probe design process.

The genomic variations in the hPP library are completely randomly spaced, while those in the NS library are clustered, because of the parent-child clustering mechanism for that library. The probe designs based on the hPP library thus lead to more randomly spaced probes, while those based on the NS library tend to yield more clustered probes. In the study, it is apparent that the former probe designs result in higher rejection rates than the latter ones. This indicates that the more randomly spaced allocation of the probe sites yield higher rejection rates than clustered allocation of the probe sites. It is also worth noting that the rejection rates are under the assumption of existing underlying clusters. The probe sets in certain designs may not be able to detect those mutations in clusters. Given that same test statistics are used in this study, it seems that the probe sets designed from a randomly spaced library tend to be able to detect those mutation in clusters than the probe sets designed from a library with clustering genetic variations. The power curves for the simulated probe designs from both hPP and NS libraries are also compared with the probe settings filtered from the probe design on chromosome 19 of MDGA. Contrary to the simulated libraries, the sources of genomic variation used in the design of MDGA are from a collection of various types of real world mouse samples. The comparisons of power curves based on the three genomic variation sources in this study show that the rejection curve from probe setting on chromosome 19 lies lower than that from the hPP library and higher than that from the NS library. This indicates that the genomic variation sources from the real samples in the MDGA design, as well as the degree of clustering of the probe sites, may lie between being completely random spaced versus a Neyman-Scott type of clustering.

In the construction of the hPP library, the Poisson intensity is the only parameter to be considered to achieve a certain fixed mutation level. In the construction of the NS library, there are four parameters involved, and one constraint to define the fixed mutation level. The parameter μ_p can be regarded as the expected number of sources when genomic variations were generated and μ_o as the expected number of genomic variations that each source generated. The product $\mu_p \mu_o$ corresponds to the fixed mutation level. The other two parameters σ and h determine the range of genomic variation clusters. For developing power curves, the choices of the parameters in the NS process in library generation should be related to scientifically meaningful values if certain background knowledge could apply for their choices.

In this study, three different window sizes in the probe selection procedure are adopted and compared to investigate the effect of window size choices on the probe design and thus the corresponding power curves. Given the same genomic variation source, with a smaller window size, the resulting probe sites tend to be more evenly spaced, which would presumably provide higher rejection rate. However, from the numerical results in study, the window size $l_w = 200k$ and $l_w = 100k$ seem to have similar rejection rates, while the window size $l_w = 20k$ provides seemingly slightly lower rejection rates than the other two sizes. Considering the limited replication size in this study, the window size choices in the probe selection procedure do not seem to have large impact on the rejection rates. We note that a study with increased number of replications in each setting may be able to provide better insights regarding whether the differences seen here with window size choices are really impactful. In any event, window size choice seems to be less relevant in terms of impact on power. Comparably, the source of genetic variations and probe intensity seem to have much more substantial an effect on the rejection rate.

The $CvM_{\tilde{R}}$ and $KS_{\tilde{R}}$ statistics used for testing existence of mutation clusters in this chapter were developed from the study in Chapter 2. The power studies in the previous chapter showed that the KS version of the statistic seems to be less stable when the standard deviation parameter σ in the alternative NS process is increasing, and recommend that the CvM statistic is more desirable. Here we compare the two statistics with respect to probe design. Our results show that the CvM statistic remains better than the KS statistic with various probe settings, including factors in probe design such as probe intensity and type of genomic variation source.

The framework for understanding power as developed in this study can be used to help determine key features when designing new microarray platforms for any species, either with or without a previous microarray design already in place, for mutation cluster detection. For any chromosome with known length in the entire genome of an organism of interest, the minimum required total number of probes, in order to achieve a minimum power to detect existence of mutation clusters under a certain mutation level, needs to be determined when designing the microarray platform. By applying the framework in simulated hPP and NS libraries, power curves could provide upper and lower bounds for the rejection rates based on the potential new probe designs with various probe intensity. By referring to the lower bound based on the NS library and the required minimum power, the corresponding probe intensity would then be the minimum probe intensity required for the new probe design. Thus the total number of probes required can be calculated given the total length of the chromosome. This procedure could help determine and thus control the total number of probes required in the microarray platform design for mutation cluster detection, making it more cost-effective and feasible, compared to sequencing techniques, to be applied as screening tools for mutation cluster detection with large sample sizes.

Chapter 4

Nonparametric association test for spatial independence in the occurrence of single nucleotide polymorphism differences and copy number variations under a microarray probe sampling system

4.1 Introduction

Single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) are both important types of genetic variation in genomic studies. The SNP genotype refers to the genetic content at a SNP locus, a single nucleotide position on the genome. SNP genotype differences (SNP differences) refer to the change of the genetic content at the SNP locus over two samples. Alterations may occur on the chromosome in the form of duplications and deletions of DNA segments ranging from several hundred base pairs up to millions of base pairs [29]. CNVs refer to such regions of structural alterations. CNVs may be generated from inheritance, or from diverse mechanisms at various stages of development of the organism [30]. In organisms with so-called diploid cells, such as the mouse, although not distinguished here, CNVs can be classified into several categories based on the number of copies in the gain or loss of a DNA segment, such as single deletion, double deletion, single duplication, double duplication, or duplication with greater number of copies. As CNVs are structural changes of genetic content on the chromosome, the prevalence of CNVs in the genome is much smaller in magnitude than that of SNP differences [31], which makes them rare events in the genome.

The signatures of mutations are especially useful for discovery and understanding of biological principles such as identifying mutational mechanisms, understanding genetic diversity and using these elements to understand diseases. In recent years, a new mutation signature of non-random spacing on the chromosome has been revealed [3]. These findings have broadened our perspective encouraging further analysis of spatial properties of mutations in order to gain knowledge of the mechanisms for mutageneses and carcinogeneses, as well as for development of cancer treatments. To study mutational mechanisms, while it is important to analyze clustering of a single type of mutation such as SNP differences, identification of the association between two types of mutations, such as SNP differences and CNVs, could contribute to understanding the relationship between the mutational mechanisms that generate these two types of mutations.

To identify these mutations in biological samples, sequencing techniques have advantages that they yield precise measurements, yet they suffer from high cost and often become infeasible in large scale studies, for example, for screening for multiple samples. A more cost-effective way for genome-wide mutation identification may be through the use of the genotyping array, such as Mouse Diversity Genotyping Array (MDGA). However challenges in inference occur because the inferential setup is strikingly different from typical spatial association analysis frameworks with MDGA data. MDGA utilizes two sets of probes to identify both SNP genotypes and CNVs in the same sample under the same experimental process. There are about 500,000 SNP probe sites for detecting SNP alleles across the mouse genome. There are, additionally, about 400,000 invariant genomic probes along with the SNP probe sites, for detecting CNVs. The SNP genotype assay relies on the trigger of fluorescence on the probe, either carrying a nucleotide the same as the reference, or a mutant type. On the other hand, the identification of CNVs depends on the simultaneous high or low fluorescent level of a certain number of consecutive probes. Abundance of DNA measured by consecutive probes determines the type of CNV detected. The start and end of a CNV identified by the genotyping array system are estimated based on these fluorescent levels.

To study the spatial association between two objects, Foxall and Baddeley [32] proposed a method for a nonparametric measure of association between a point process X and a random set Y in two-dimensional space. They proposed the so-called J function, a ratio statistic, that compares the distribution function of (1) the distance from a random point in the point process X, to the random set Y, to (2) the distribution of the distance from a random point in the domain to the random set Y. However, with this statistic, information on the entire domain of interest is required. On the other hand, the microarray platform provides only intermittent observations, as measurements on only certain specific regions of the chromosome are available. In this chapter, we adopt some elements of the conceptual framework of the J function to develop a test for association. Inferential methods are developed that account for the MDGA setting of substantial missing data. A Monte Carlo simulation approach with a homogeneous Poisson process as the null process is proposed in this study for statistical inference. Our nonparametric approaches are optimized for the microarray probe design. A block bootstrap approach [33] developed for the time series setting that randomly selects segments from the original series for hypothesis testing is also adapted here for our missing data setting.

The tools provided in this study aim to test three null hypotheses related to association between SNP differences and CNVs, each with a different level of generality. We describe the data structure and the two proposed non-parametric test statistics for detecting spatial association in Section 4.2. In Section 4.3, we describe three null hypotheses and three Monte Carlo simulation approaches for statistical inference. The small sample properties of the two statistics are studied by simulation under the three approaches developed: a parametric approach utilizing a Poisson null process, as well as nonparametric overall and partial block bootstrap approaches. A step function Poisson process and a modification of the Neyman-Scott parent-child process, are adopted as the alternative processes for studying the power performance of the two test statistics. In Section 4.4, we introduce our motivating example and utilize graphical tools to help visualize landscapes of the two types of mutations. We examine the association between SNP differences and CNVs under three null hypotheses in this example using our recommended test statistic.

4.2 Statistical methods

4.2.1 Data structure

Consider a specific chromosome with known SNP probe design. Denote the domain of the chromosome as $L \subset \mathbb{R}^+$. Define the set of the probe locations as $S : S \subset L$. Denote the location of the first probe on the chromosome as $s_f = \min_{s \in S} s$, and the location of the last probe on the chromosome as $s_l = \max_{s \in S} s$. Denote the locations of SNP differences detected by the probes as $X : X \subseteq S$. The probes at which no SNP differences are detected are denoted as $V \equiv S \setminus X$. Denote the regions of CNVs as R. The regions outside of CNVs are thus denoted as R^C.

4.2.2 Proposed statistics

To assess the spatial relationship between SNP differences and CNVs, statistics are considered that accommodate our the scenario of the microarray probe sampling system. Here we consider X as a point process and R as a random set, and compare the distribution of the distance from $X \cap R^C$ to R and that from $V \cap R^C$ to R. The definitions of the two test statistics are as follows.

Define the bivariate J-function

$$J(r) = \frac{1 - G_{X \cap R^C,\ R}(r)}{1 - G_{V \cap R^C,\ R}(r)}.$$

where

$$G_{Y,R}(r) = P\{d(x, R) \leqslant r | x \in Y\}$$

is the distribution function of the distance from a typical point of Y to the nearest point of R, and

$$d(z, A) = \inf_{a \in A} \|z - a\|$$

denotes the shortest distance from a point $z \in L$ to a subset A of L.

The function J can be used as a nonparametric measure of association between X and R. The distribution functions of $G_{X \cap R^C,\ R}$ and $G_{V \cap R^C,\ R}$ can be estimated empirically. The J function considers only the SNP differences and probes outside of all CNV regions, since the focus of study is the behaviour of SNP differences outside CNV regions.

We also consider a count statistic which summarizes information about SNP differences in the regions nearby CNVs. The count statistic $C(r)$ specifically takes into account the regions outside of CNVs, and is defined as following:

$$C(r) \equiv \sum_{x \in X} I(0 < d(x, R) \leqslant r) \tag{4.1}$$

The J statistic is estimated as follows,

$$\hat{J}(r) = \frac{1 - \hat{G}_{X \cap R^C,\, R}(r)}{1 - \hat{G}_{V \cap R^C,\, R}(r)}, \tag{4.2}$$

where $\hat{G}_{Y,R}(r)$ is

$$\hat{G}_{Y,R}(r) = \frac{1}{|Y|} \sum_{x \in Y} \left(I(d(x, R) \leqslant r) \right),$$

an empirical estimator of $G_{Y,R}(r)$.

The statistic $T(r)$, $T(r) \in \{\hat{J}(r), C(r)\}$ can be evaluated at any fixed r of interest in a specific application, for testing for association. The distance r would reflect the neighborhood of the CNVs that has potentially different behavior than regions further away. Except when a specific value of r is of interest to study, it is often more desirable to study a range of r instead of a pre-determined fixed one. Hence, we also consider a functional form of the T statistic below to test for association between SNP differences and CNVs.

4.3 Small sample properties of the test statistics

Monte Carlo simulation is adopted to obtain sample paths in order to construct confidence bands and assess power of the test statistics. To generate the sample paths, we consider CNV locations as fixed on the chromosome. In most experimental scenarios, including the data sets available in our study, the observed CNV counts are usually very small, especially compared to the count of SNP differences.

Biologists are interested in the scientific questions of whether the existence of CNVs could influence the spacing of SNP differences outside of CNVs or whether there tend to be more or less SNP differences in regions near the CNVs compared to those farther away. Based on these scientific hypotheses of interest to the biologists, three specific null hypotheses with increased level of generality are proposed as follows:

(1) the SNP differences outside of CNV regions follow complete spatial randomness;

(2) the SNP differences outside of CNV regions have similar properties in clearly defined CNV nearby regions as for regions further away;

(3) the SNP differences outside of CNV regions have similar properties everywhere on the chromosome.

Note that for SNP differences outside of CNVs, the process in the first null hypothesis is stationary and with the restriction of being hPP; in the second null hypothesis, the process in the two regions are stationary and the same without the restriction of being hPP; in the third, the process is stationary everywhere and without the restriction of being hPP.The first null hypothesis is a commonly adopted one in spatial analysis. For example, Foxall and Baddeley [32] adopted it in their analysis of the association between ore deposits and lineaments in an earth science investigation. However, rejection of this null hypothesis does not imply rejection of the second or the third null hypothesis. Given a null hypothesis, specific inferential methods need to be adopted.

For statistical inference, three Monte Carlo approaches are developed, one based on the Poisson null process, a partial block bootstrap and an overall block bootstrap. The first approach adopts parametric modeling of the conditional distribution of the underlying SNP differences given the CNV locations; while the latter two utilize nonparametric bootstrap methods. The approach based on the Poisson null process, the partial and the overall block bootstraps are suitable for the first, second and third null hypothesis respectively. In the power study, two parametric models would be specified for the distribution of underlying SNP differences given the locations of the CNVs. In considering parametric models, once the locations of the underlying SNP differences are generated through a specified parametric process, the observed SNP differences would then be determined by overlaying the probe locations.

4.3.1 Monte Carlo simulation approaches for inference

We describe below the three Monte Carlo simulation approaches.

4.3.1.1 Poisson null process

The approach based on the Poisson null process aims to test the first null hypothesis, where the SNP differences outside of CNV regions follow complete spatial randomness. In the Poisson null process, SNP differences are distributed according to a non-homogeneous Poisson process (nhPP) such that the SNP differences follow different homogeneous Poisson processes (hPP) in CNV regions R and non-CNV regions R^C, with a constant Poisson intensity that may differ over these two regions. Under the null process, the Poisson intensity in each region, R or R^C, is estimated as the the empirical rate of SNP differences in the corresponding region in the data sample. Thus the expected number of observed SNP differences in the entire chromosome regions observed is the same as that in the data sample. The reason that the Poisson intensities are allowed to be different in R and R^C is that only SNP differences outside the CNV regions are of interest in the hypothesis. Thus SNP differences inside the CNV regions are not included in the estimation of the intensity outside the CNV regions.

4.3.1.2 Overall block bootstrap

For the overall block bootstrap, a resample of the SNP differences and probes over the entire chromosome is constructed by connecting block samples from the original chromosome as described below. Sampled blocks have fixed length B_O. Consider the entire chromosome L. Sample a point $g_1 \sim U(s_f, s_l)$ on the original chromosome, where $U(\cdot)$ is the uniform distribution, and s_f and s_l are the locations of the first and last probe on the chromosome respectively; then the corresponding block sample is obtained as $[g_1, g_1 + B_O - 1]$ on the chromosome. All SNP differences and probes within this segment are relocated to the segment $[s_f, s_f + B_O - 1]$ on the resampled chromosome maintaining their exact position relative to g_1, but now from s_f on the resampled chromosome. A subsequent point $g_2 \sim U(s_f, s_l)$ is taken on the original chromosome, the SNP differences and probes within $[g_2, g_2 + B_O - 1]$ are relocated to the segment $[s_f + B_O, s_f + 2B_O - 1]$ on the resampled chromosome, again maintaining their exact position relative to g_2, but now from $s_f + B_O$ on the resampled chromosome (See Figure 4.1).

This procedure is repeated with the mth point $g_m \sim U(s_f, s_l)$ leading to all SNP differences and probes within segment $[g_m, g_m + B_O - 1]$ relocated to the segment $[s_f + (m-1)B_O, s_f + mB_O - 1]$ on the resampled chromosome, until s_l is within the last resampled segment. Data beyond s_l on the resampled chromosome are omitted. Note that for the procedure described here, the blocks sampled from the original chromosome $[g_m, g_m + B_O - 1], m = 1, \cdots$ can have overlap between each other.

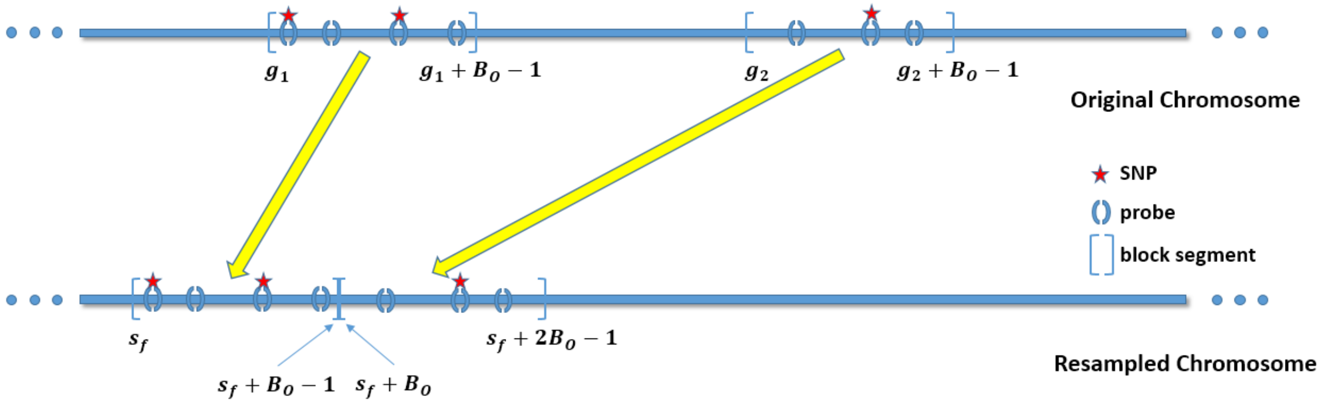


Figure 4.1: The first two steps of the overall block bootstrap are illustrated. Sample a point $g_1 \sim U(s_f, s_l)$ on the original chromosome. All SNP differences and probes within the segment $[g_1, g_1 + B_O - 1]$ on the original chromosome are relocated to the segment $[s_f, s_f + B_O - 1]$ on the resampled chromosome, maintaining their exact position relative to g_1, but now from s_f on the resampled chromosome. A subsequent point $g_2 \sim U(s_f, s_l)$ is taken on the original chromosome, the SNP differences and probes within $[g_2, g_2 + B_O - 1]$ are relocated to the segment $[s_f + B_O, s_f + 2B_O - 1]$ on the resampled chromosome, again maintaining their exact position relative to g_2, but now from $s_f + B_O$ on the resampled chromosome.

4.3.1.3 Partial block bootstrap

In the partial block bootstrap, the resampling of the SNP differences and probes is not based on their positions on the chromosome, but their distances to the nearest CNV. The set of distances of SNP differences and probes to the nearest CNV in the original sample are denoted as $D_X, D_X = \{d(x, R), x \in X\}$ and $D_S, D_S = \{d(s, R), s \in S\}$ respectively. These sets of distances D_X and D_S are sufficient for calculating both $\hat{J}(r)$ and $C(r)$ statistics. With this resampling procedure, these distances of SNP differences to CNVs are transformed such that small distances are translated by a random positive value, whereas large distances are translated by the negative of that value. A heuristic conceptual explanation is that information about SNP differences and probes near CNVs are exchanged with information further away. The translation function $p(y; Z, B_P)$ satisfies:

$$p(y; Z, B_P) = \begin{cases} y + Z & \text{if } y \in (0, B_P] \\ y - Z & \text{if } y \in [Z, Z + B_P) \\ y & \text{otherwise} \end{cases} \tag{4.3}$$

where $Z \sim U[B_P + 1, B_L]$, and B_P is the block length defining what is considered as a small distance and should satisfy $B_P < B_L - 1$ in order to have a valid definition for the uniform distribution from which Z is generated. The upper limit B_L is defined as $B_L = \max_{s \in S} d(s, R) - B_P$ to ensure $Z + B_P$ does not exceed the maximum distance observed between probes and the nearest CNV. The resampled distances D_X^* are defined as $D_X^* = \{p\,(d; Z, B_P)\,, d \in D_X\}$, and D_S^* are defined as $D_S^* = \{p\,(d; Z, B_P)\,, d \in D_S\}$.

4.3.2 Statistical inference

For hypothesis testing, the 95% confidence bands of the expected values of the statistics can be constructed through any of the Monte Carlo simulation approaches presented in Section

4.3.1. There are two types of confidence bands that could apply in our context: point-wise and global confidence bands. Given any Monte Carlo simulation approach, in each Monte Carlo sample, the function $T(r)$ can be evaluated at a given r. The sampling distribution of $T(r)$ can be obtained by Monte Carlo methods and thus point wise 95% confidence bands can be calculated as the estimates of the 2.5% and 97.5% quantiles from the sampling distribution.

For the functional form of $T(r)$ over a range of values of r, global $100(1-\alpha)\%$ confidence bands are obtained to ensure that the entire $T(r)$-function stays within the bands at the controlled $1-\alpha$ rate under the null hypothesis. For this property to hold, the confidence bands are constructed in the following way. Suppose $\exists c$ such that

$$P\left(\sup_r \left|\frac{T(r)-E(T(r))}{\sqrt{Var(T(r))}}\right| > c\right) \leqslant \alpha$$

Thus $\forall r$, there is

$$P\left(T(r) \in \left[E(T(r)) - c\sqrt{Var(T(r))}, E(T(r)) + c\sqrt{Var(T(r))}\right]\right) \leqslant \alpha$$

Then $E(T(r))-c\sqrt{Var(T(r))}$ and $E(T(r))+c\sqrt{Var(T(r))}$ can be estimated to be the $1-\alpha$ global confidence lower and upper boundaries. respectively. The value of $\hat{E}(T(r))$ and $\widehat{Var}(T(r))$ can be obtained by Monte Carlo simulation. Denote Q as the quantile function of the statistic:

$$\max_r \left|\frac{T(r)-\hat{E}(T(r))}{\sqrt{\widehat{Var}(T(r))}}\right|.$$

The constant c can be estimated by

$$\hat{c} = \hat{Q}(1-\alpha)$$

From the Monte Carlo replicates, there are several ways to estimate the sample quantile function $\hat{Q}$. The method used in this research follows the Definition 7 in [34], which is the default method implemented for a continuous sample in the quantile function in R.

Then the global confidence bands are obtained by

$$GB(r) = \left[\hat{E}(T(r)) - \hat{c}\sqrt{\widehat{Var}(T(r))}, \hat{E}(T(r)) + \hat{c}\sqrt{\widehat{Var}(T(r))}\right]$$

The null hypothesis would be rejected if there $\exists r : T_S(r) \notin GB(r)$, where $T_S(r)$ is the functional statistic calculated from the sample being tested.

4.3.3 Alternative processes for power studies

For the alternative process, we consider two parametric models for the conditional distribution of locations of underlying SNP differences given the CNV locations: a step function nhPP and a modified non-homogeneous parent Neyman-Scott process (NPNSP). The first process is a nhPP with different intensity rates for locations of SNP differences near the CNVs versus locations farther away. The second process uses a NPNSP for generating SNP differences outside CNV regions, whereas SNP differences within CNV regions are generated by a hPP.

The entire chromosome can be divided into three regions based on $d(z, R), z \in L$, which is the distance from a generic location to the nearest CNV. For all the locations in regions within CNVs, the property $d(z, A) = 0$ holds. The regions within a certain distance D to the nearest CNV have the property that $d(z, A) \in (0, D]$, and are termed here as regions *nearby* CNVs or *nearby* regions. The regions further than D from the CNV have the property that $d(z, A) \in (D, +\infty)$, and are termed here as regions *faraway* from CNVs or *faraway* regions. Hence D is a parameter defining the boundary between the *nearby* and *faraway* regions.

4.3.3.1 Step function nhPP

Under the alternative hypothesis of a step function nhPP, the Poisson intensities in three different regions on the chromosome are constant, but allowed to be different. Denote γ_s as the ratio of the intensities of SNP differences in the regions *nearby* CNVs and SNP differences in the regions *faraway* from CNVs. When $\gamma_s = 1$, the intensities of SNP differences in both *nearby* and *faraway* regions are the same, and the process reduces to the Poisson null process mentioned in Section 4.3.1.1.

4.3.3.2 Modified non-homogeneous parent Neyman-Scott process (NPNSP)

The NPNSP is an modified version of the parent-child Neyman-Scott process (NSP). To generate underlying SNP differences under NPNSP, the parent locations follow a nhPP. The parent intensity of the NPNSP is allowed to be different in the three different regions. For regions within CNVs, the intensity of parents is set as 0. The ratio of the intensity of parents in the CNV *nearby* regions to the intensity of parents in *faraway* regions is defined as γ_p. When $\gamma_p \neq 1$, properties of clusters in the regions *nearby* and *faraway* from CNVs are different. When the parameter γ_p is greater than a value of b, there tend to be more offspring yielded by the parents hence more clustering in the *nearby* regions compared to the *faraway* regions; in this case there tend to be positive association between SNP differences and CNVs. The parameter b is a constant that is greater than 1 and depends on other parameters in the process. When $\gamma_p < b$, there are less offspring yielded by the parents hence less clustering in the *nearby* regions compared to the *faraway* regions; there tends to be negative association between SNP differences and CNVs. When $\gamma_p = 1$, the offspring yielded in the *nearby* region is slightly less than that in the *faraway* region, as there are no parents inside of the CNV regions. Slightly increasing γ_p would offset this effect.

For a given parent site, a cluster of offspring sites are generated around it, with expected number μ_o. Centered at the parent site location, the offspring sites are independent and identically distributed truncated normal random variables. The standard deviation of the truncated normal distribution is denoted as σ. The half-length of the window of the truncation range is denoted as h. All underlying SNP differences outside of CNV regions are offspring generated through the clustering process.

Although there are no parents in NPNSP inside CNV regions, the offspring generated from the parents may lie inside the CNV regions. In the censored NPNSP, the set of underlying SNP differences generated from the NPNSP located inside CNV regions are censored. The SNP differences inside the CNV regions are generated from a hPP, whose intensity is the same as that from the null process. The overall underlying SNP differences are then the combination of the SNP differences inside the CNV regions generated from the hPP and SNP differences outside the CNV regions produced from the NPNSP.

4.3.4 Parameter settings and simulation results

A mouse with cancer is used as an example to study power performance of the J and C statistics under three Monte Carlo simulation approaches. Specifically, the chromosome 1 of a wild caught mouse is selected. There are 5 CNVs and 4906 SNP differences that are different between primary tumor and metastatic tissues. Among the 4906 SNP differences, there are 27 inside CNV regions and 4879 outside of the CNV regions. The length of the chromosome 1 is approximately 1.9×10^8bp.

For the calculation of the J and C test statistics, the argument r defines the CNV neighborhood being considered and needs to be specified. Prior biological knowledge about a certain neighborhood of interest can be used to set the values of r. Yet without such knowledge, a relatively large r can be specified, as the functional form of the J or C statistic is evaluated between 0 (not included) to the set value r. In the example of this study, a grid of values of r is set from 100 to 10^6 with increment of 100, 1.0005×10^6 to 10^7 with increment of 500 and 1.0001×10^7 to 2.5×10^7 with increment of 1000. As a result, there are in total 43000 values of r evaluated.

4.3.4.1 Parameter settings for confidence band construction

For the approach utilizing the Poison null process, the Poisson intensity inside and outside of CNV regions is set as 2.34×10^{-5} and 2.47×10^{-5} respectively based on the observed number of SNP differences in each region in the data example. For this null process, 10^4 Monte Carlo samples are generated for the construction of confidence bands.

For the overall and partial block bootstrap, both B_O and B_P are set as 2×10^6. Although these parameters can be set arbitrarily, depending on researcher's requirement, a general guidance for setting these parameter is through examination of the autocovariance structure from the data sample, as illustrated in the Appendix C.3. With either of these two approaches, a total of 10^3 resamples are generated for each sample to be tested for the construction of confidence bands.

The significance level of the tests is set as $\alpha = 0.05$ for all three approaches.

4.3.4.2 Parameter settings for power study

With both alternative processes, we set $D = 10^6$. For the alternative as a step function nhPP, the range of γ_s is specified as 0.4 to 1.5 with values in the range having increments of 0.1; this includes both positive ($\gamma_s > 1$) and negative ($\gamma_s < 1$) association between SNP differences and CNVs.

For the alternative process of a modified NPNSP, the parameter γ_p is set from 0.4 to 1.5 with an increment of 0.1. The parameters for offspring generation are $\mu_o = 10^3$, $\sigma = 5 \times 10^5$, and the truncation range h set as 3σ. The values of the parameter μ_p are set so as to ensure that the number of detected SNP differences across the chromosome are close to the observed number of SNP differences in the data sample. The Poisson rate for SNP differences inside the CNV regions is set the same as the rate in the CNV regions for the Poisson null process, that is 2.34×10^{-5}.

For each parameter setting of γ_s or γ_p, 10^2 replications were generated for power calculation by Monte Carlo simulation.

4.3.4.3 Power study results, interpretation, and recommendation

The results of the power study comparing J and C statistics are displayed in Figures 4.2 through 4.4 for the alternative process of a step function nhPP, and Figures 4.5 through 4.7 for the alternative process of a modified NPNSP. In these figures, dots or triangles representing the powers of J or C statistics are jittered horizontally to enhance readability. The vertical dashed lines are plotted to clearly separate dots and triangles between consecutive γ_s or γ_p values. The blue horizontal dashed line represents the significance level of the test procedure, which is $\alpha = 0.05$.

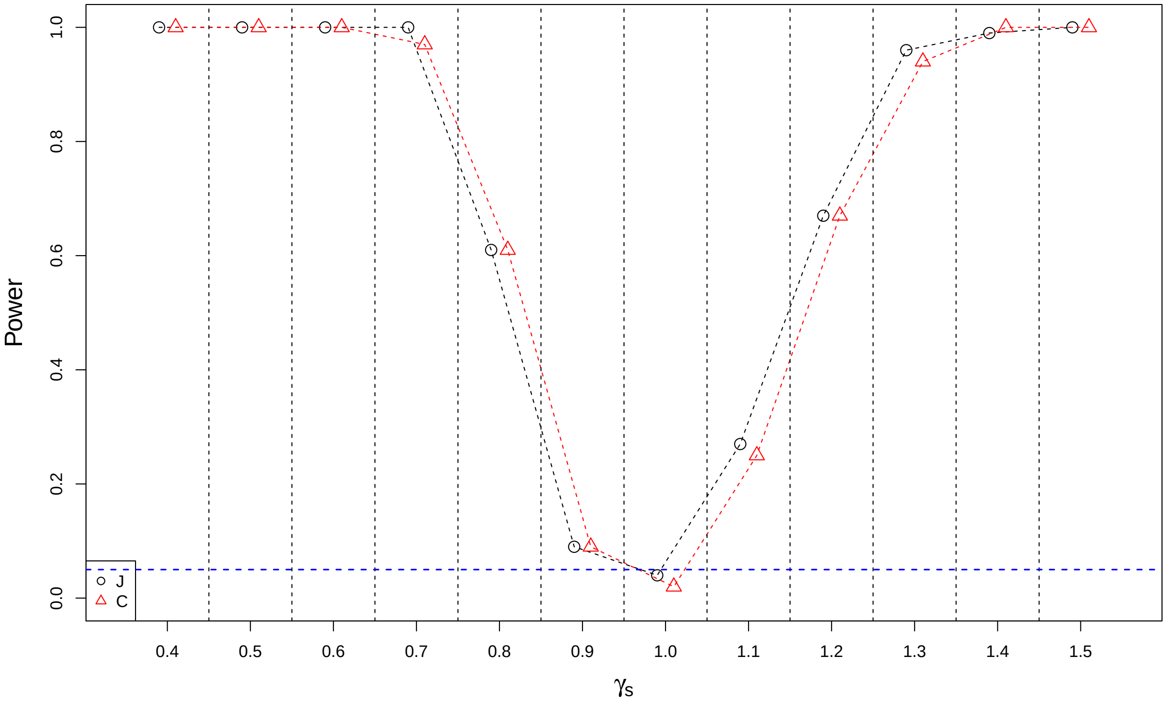


Figure 4.2: Power performance of *J* and *C* statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the Poisson Null Process.

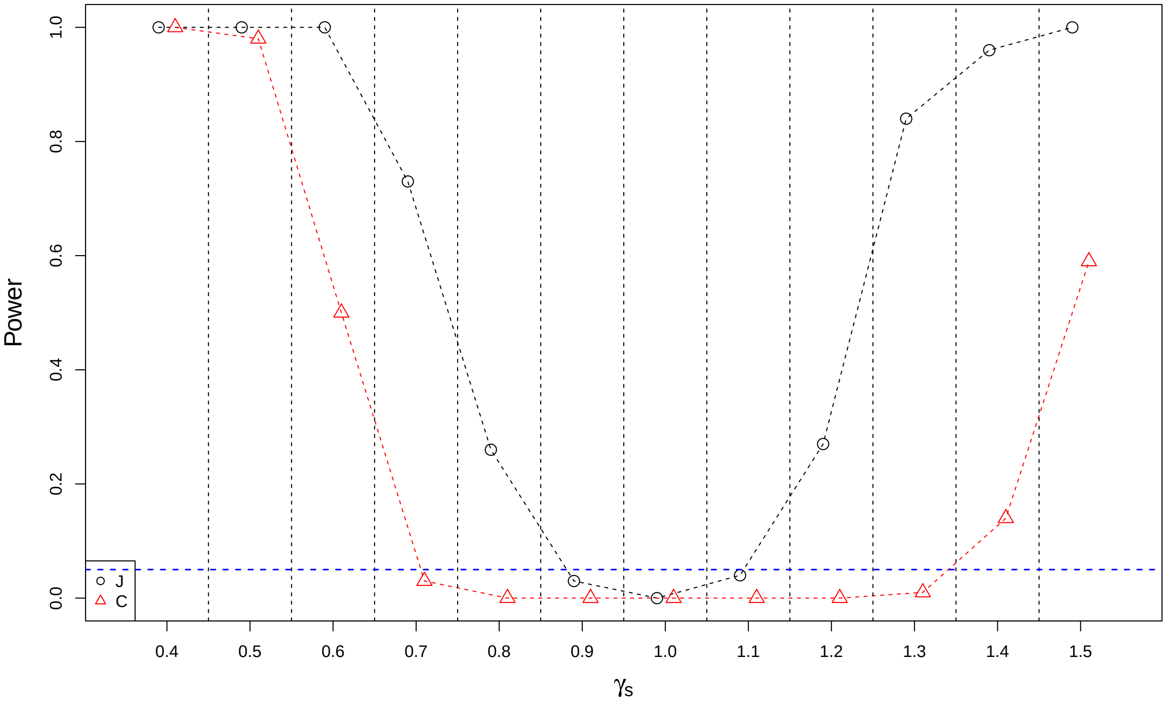


Figure 4.3: Power performance of J and C statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the overall block bootstrap.

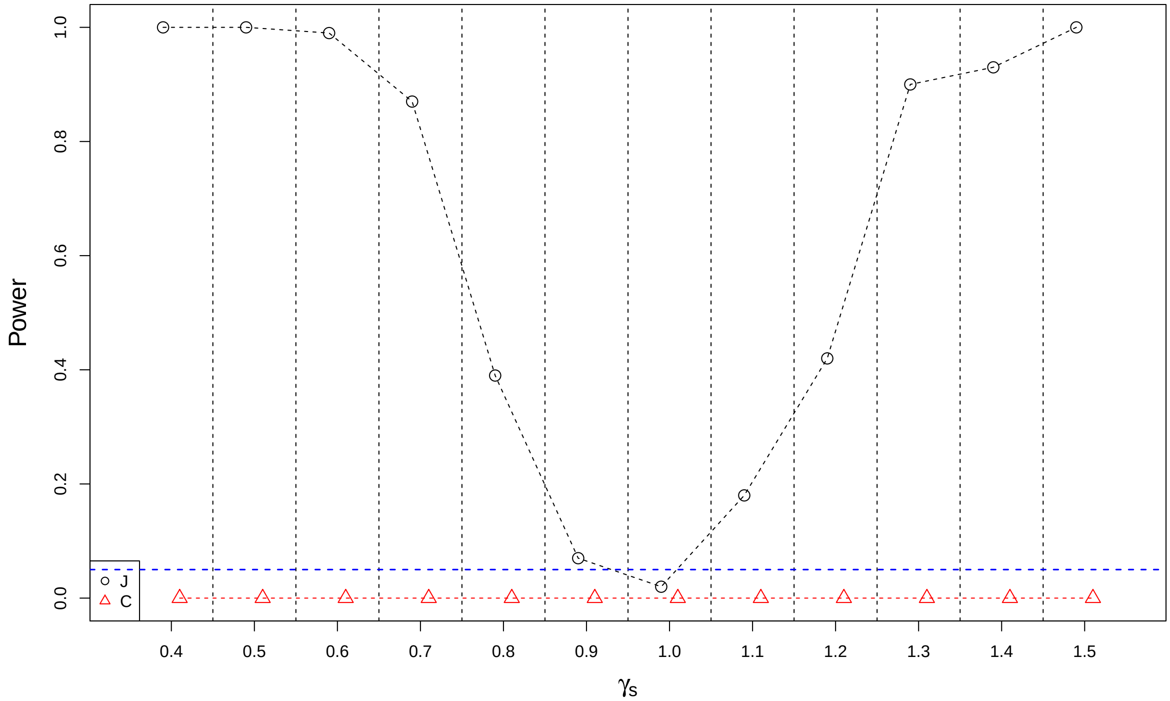


Figure 4.4: Power performance of J and C statistics under the alternative hypothesis of the step function nhPP using confidence bands constructed from the partial block bootstrap.

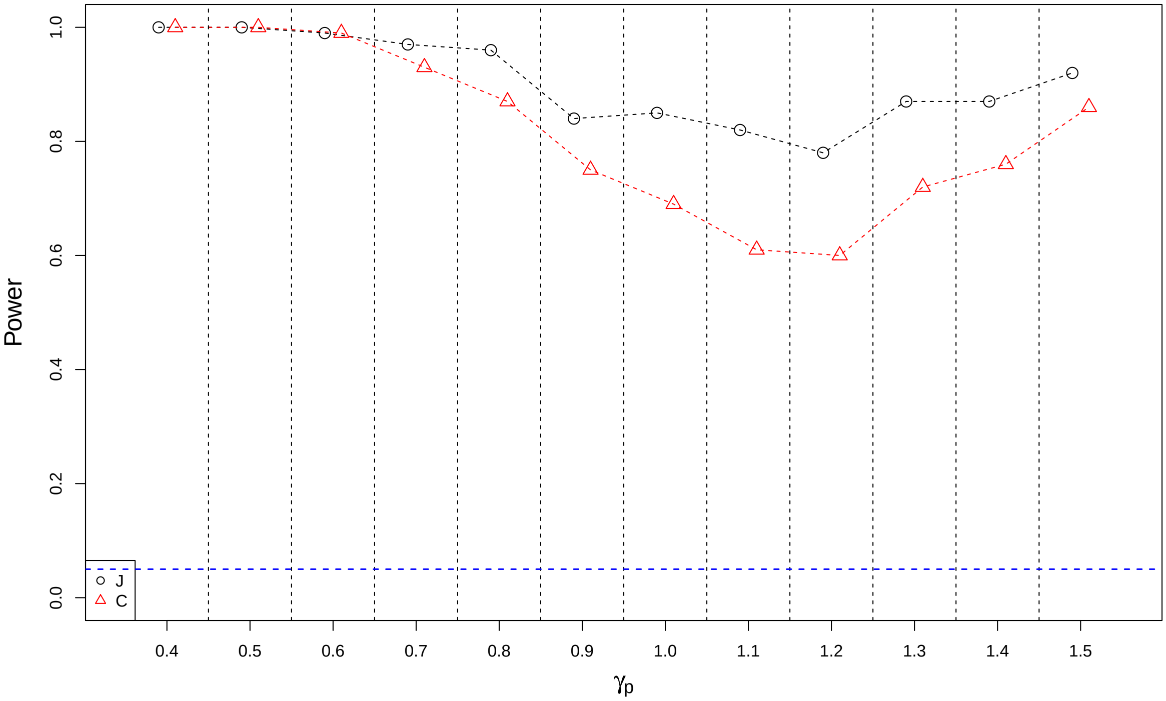


Figure 4.5: Power performance of *J* and *C* statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the Poisson Null Process.

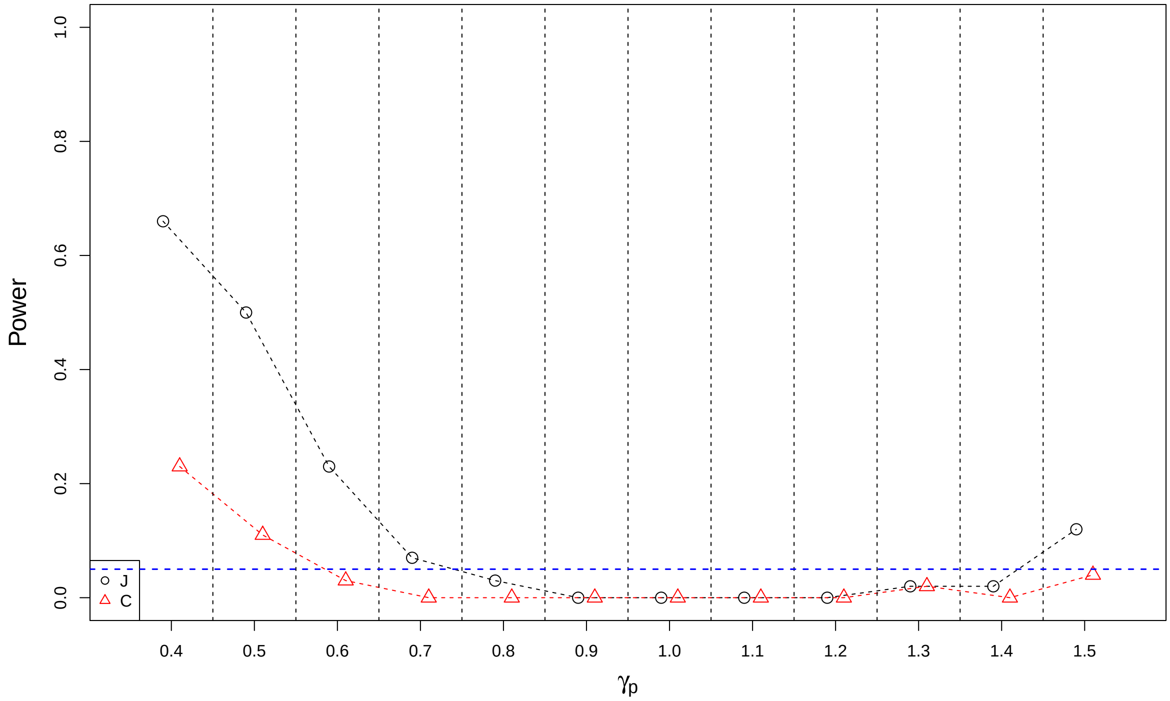


Figure 4.6: Power performance of J and C statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the overall block bootstrap.

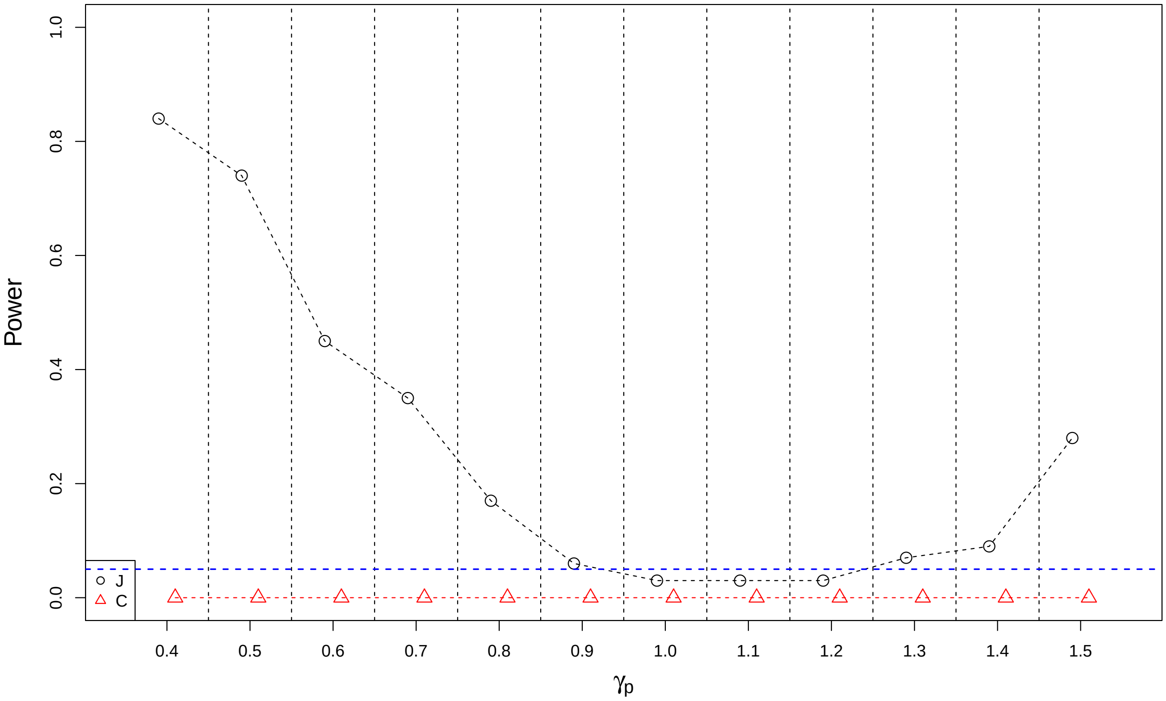


Figure 4.7: Power performance of J and C statistics under the alternative hypothesis of the modified NPNSP using confidence bands constructed from the partial block bootstrap.

In many parameter settings in each scenario, the J statistic has higher or at least comparable power to the C statistic. When testing the simulated samples from the step function nhPP against the first null hypothesis using the Poisson null process approach, both J and C statistics have size about as expected when $\gamma_s = 1$, when the first null hypothesis is true. When testing simulated samples from the step function nhPP and the modified NPNSP against the second and the third null hypothesis using the partial and overall block bootstrap approach, the rejection rate for both J and C are fairly close to, but lower than, the significance level of the test, which is $\alpha = 0.05$, when the null hypotheses are true. It is worth noting that for both alternative processes for the C statistic using the partial block bootstrap where power is 0. This may due to the discreteness of the C statistic, which may lead to wide confidence bands in the *nearby-D* regions, that is, when r is small.

A seemingly strange result appears in the scenario of testing the simulated samples from the modified NPNSP against the first null hypothesis using the Poisson null process approach. The lowest powers of both J and C statistics in this situation are reaching above 0.8 and 0.6 respectively and substantially higher than the significance level of 0.05. However, it should be noted that when γ_p is close to 1 under the modified NPNSP process, even though the regions outside of the CNVs have the same properties everywhere, there still exists clustering. The statistical inference procedure using the Poisson null process identifies that there is deviation from a hPP process outside of the CNV regions leading to rejection of the null hypothesis. If the hypothesis is to test whether the SNP differences outside the CNVs behave similarly in *nearby-D* or *faraway-D* region, it would be problematic to use the approach as for the Poisson null, as it would tend to reject the null hypothesis for the cases where the SNP differences outside of CNV regions behave similar everywhere, but the process itself deviates from a hPP. On the other hand, the overall and partial block bootstrap have appropriate size. The partial block bootstrap tends to have higher power than the overall bootstrap for the J statistic. The partial block bootstrap is more appropriately used for testing a null hypothesis with a well defined boundary between *nearby-D* and *faraway-D* regions. The overall block bootstrap would be suitable for handling a null hypothesis without such a clear definition of the boundary between *nearby-D* and *faraway-D* regions.

In this power study, for the alternative process of the modified NPNSP, the varying factor in the parameter setting is γ_p. As the power study is time consuming, there is only one parameter setting for the offspring generation. More simulation studies could be conducted in the future to investigate how sensitive the power performance of the test statistics are with various parameter settings of offspring generation mechanism.

We summarize here our recommendations on the statistics and approaches for construction of confidence bands based on the different types of null hypothesis. For testing the null hypothesis that the SNP differences outside of the CNV regions follow complete spatial randomness, both J and C statistics are recommended using the Poisson null process. For testing the null process that SNP differences outside the CNV regions behave the same in *nearby* and *faraway* regions without a clearly defined boundary, the J statistic is recommended using the inferential approach of the overall block bootstrap. For testing the null process that SNP differences outside the CNV regions behave the same in *nearby* and *faraway* regions with clearly defined boundary, the J statistic is recommended using the inferential approach of the partial block bootstrap.

4.4 Application

Our motivating analysis considers whether there is association between SNP differences and CNVs on a chromosome of a mouse. This is an important first step in determining how these features are related to genetic backgrounds as well as environmental effects that link such features to cancer occurrence. The mutation profile used is the difference between primary tumor and metastatic tissues from chromosome 1 of a wild mouse. As mentioned earlier, there are 5 CNVs and 4906 SNP differences that are different between the primary tumor and metastatic tissues. The mutation landscape can be seen in the rainfall plot in Figure 4.8 and the rainbow plot in Figure 4.9. In both plots, the CNV locations are indicated by the red markers and the green dashed vertical lines, where width of a red marker identifies the start and end base pairs of the CNV location and the green line shows the center of the corresponding CNV region. The black dots in both plots convey information regarding the SNP differences, with the x axis showing the locations of a SNP difference; in the rainfall plot, the y axis shows the logarithm of the distance from a SNP difference to the previous one; while in the rainbow plot the y axis shows the logarithm of the distance from the SNP difference to the closest CNV. The rainfall and rainbow plots are helpful for visualizing the mutation landscape. For example, by drawing a horizontal line in the rainfall plot, one can see how many pairs of SNP differences are proximal within the distance defined by the placement of the horizontal line. In the rainbow plot, a horizontal line can help one see how many SNP differences are within a certain distance to the nearest CNV. However, these plots cannot provide formal judgment on the question of whether the SNP differences are CNVs are spatially associated. Thus, our statistical tools become useful for finding evidence to answer this question.

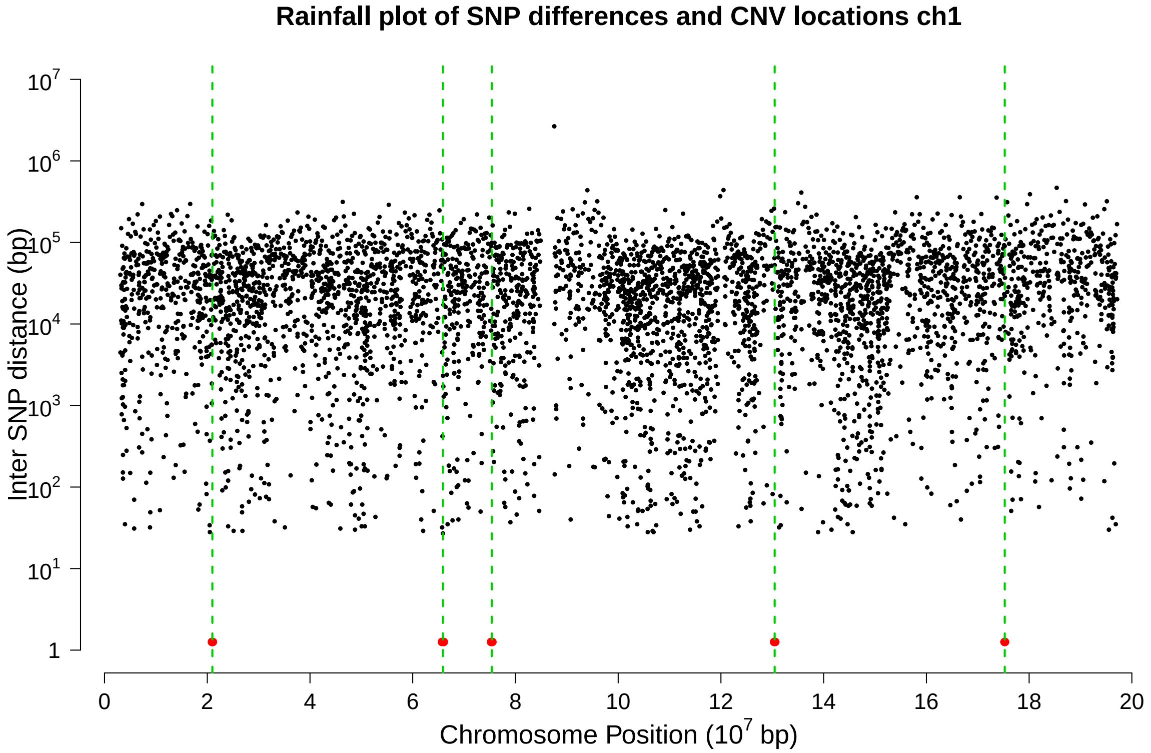


Figure 4.8: Rainfall plot displaying mutation landscapes. Black dots refer to SNP differences. Red markers identify CNV locations. The green dashed vertical lines show the centers of the CNVs.

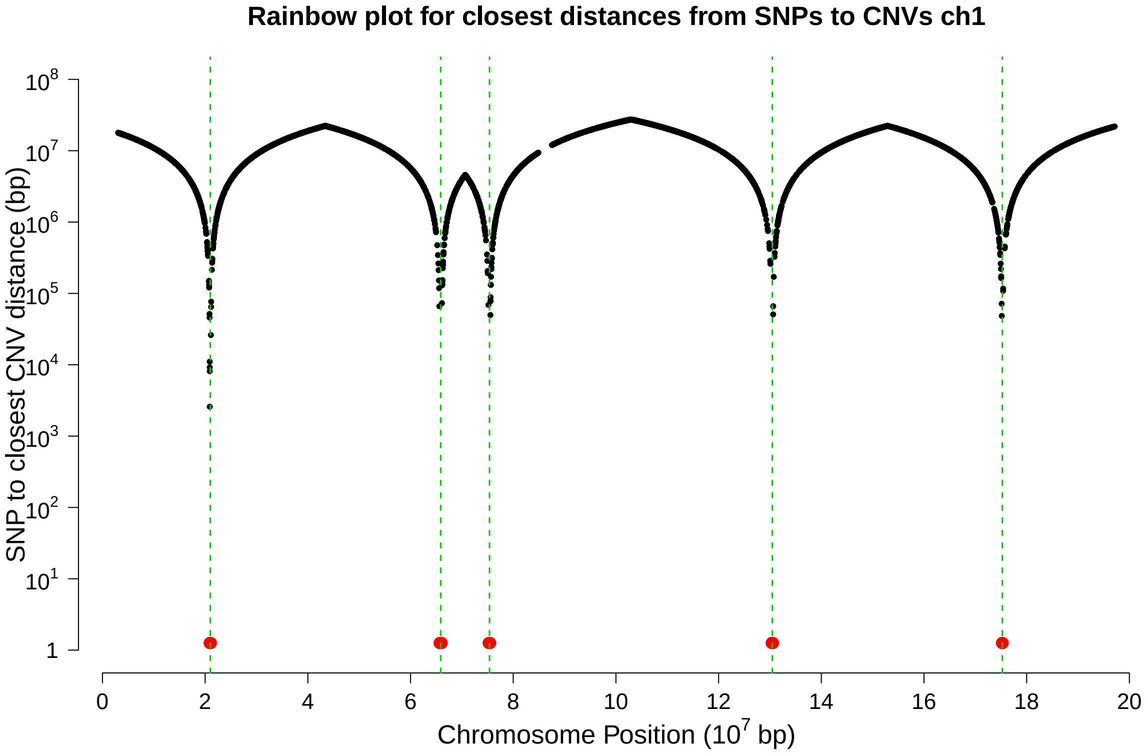


Figure 4.9: Rainbow plot displaying mutation landscapes. Black dots refer to SNP differences. Red markers identify CNV locations. The green dashed vertical lines show the centers of the CNVs.

To detect association between SNP differences and CNVs in the data example, we need to set the null hypothesis and thus corresponding Monte Carlo simulation approach to link with the requirement of the scientific research. If the research interest is to test whether the SNP differences outside of CNV regions follow spatial randomness, the Poisson null process approach of section 4.3.1.1 should be adopted; if the research interest is to test that SNP differences outside of CNV regions have similar properties everywhere, the overall block bootstrap approach of section 4.3.1.2 should be adopted; finally if the research interest is to test that SNP differences outside of CNV regions have similar properties in clearly defined CNV *nearby-D* and *faraway-D* regions of CNV, the partial block bootstrap approach of section 4.3.1.3 should be adopted.

All three of these approaches are considered for the mouse data analysis. The values of r and the number of replications in the Monte Carlo simulation are set to be the same over all these approaches. The block length is set as 3×10^6 as suggested from the examination of the autocovariance structure; details of this examination are available in Figure C.2 in Appendix C.3. The significance level is set as 0.05 for all testing procedures.

The functional J statistic and three pairs of confidence bands for the three approaches are plotted in Figure 4.10 through 4.12. As the J statistic for the data example is clearly outside of the global confidence band constructed from the Poison null process (see Figure 4.10), the the null hypothesis that the SNP differences outside the CNV regions follows completely spatial randomness is rejected. The J statistic lies completely inside the confidence bands constructed from the overall and partial block bootstrap, as seen in Figure 4.11 and 4.12. Thus there is no evidence against the null hypothesis that 1) SNP differences outside the CNV regions behave the same everywhere and hence logically also no evidence against 2) SNP differences within the distance of 3×10^6 to the CNVs and farther away than this distance threshold have similar properties.

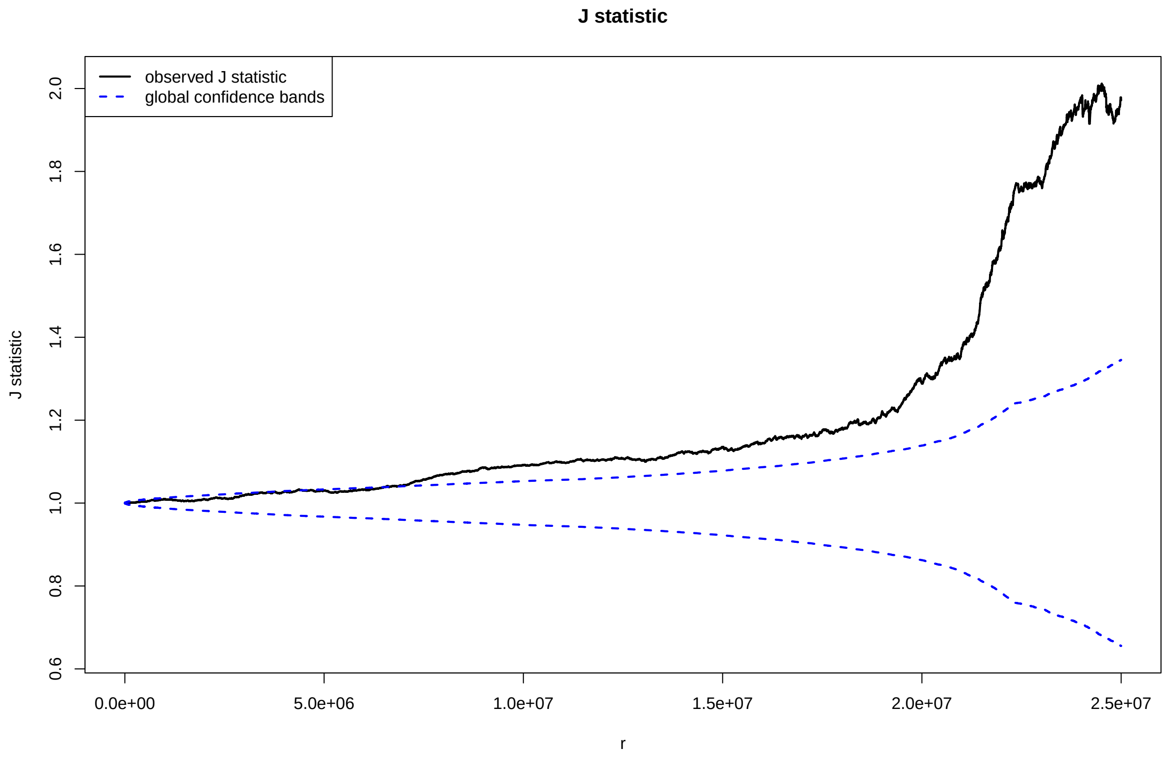


Figure 4.10: *J* statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the Poisson null process.

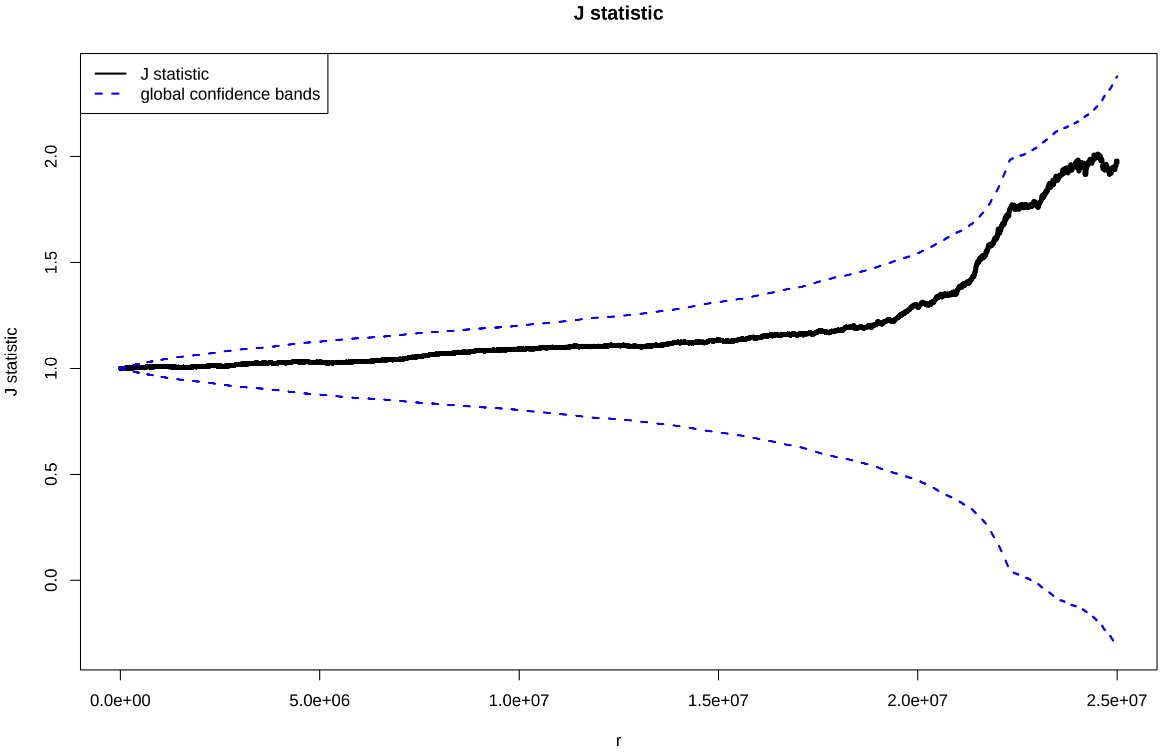


Figure 4.11: *J* statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the approach of overall block bootstrap.

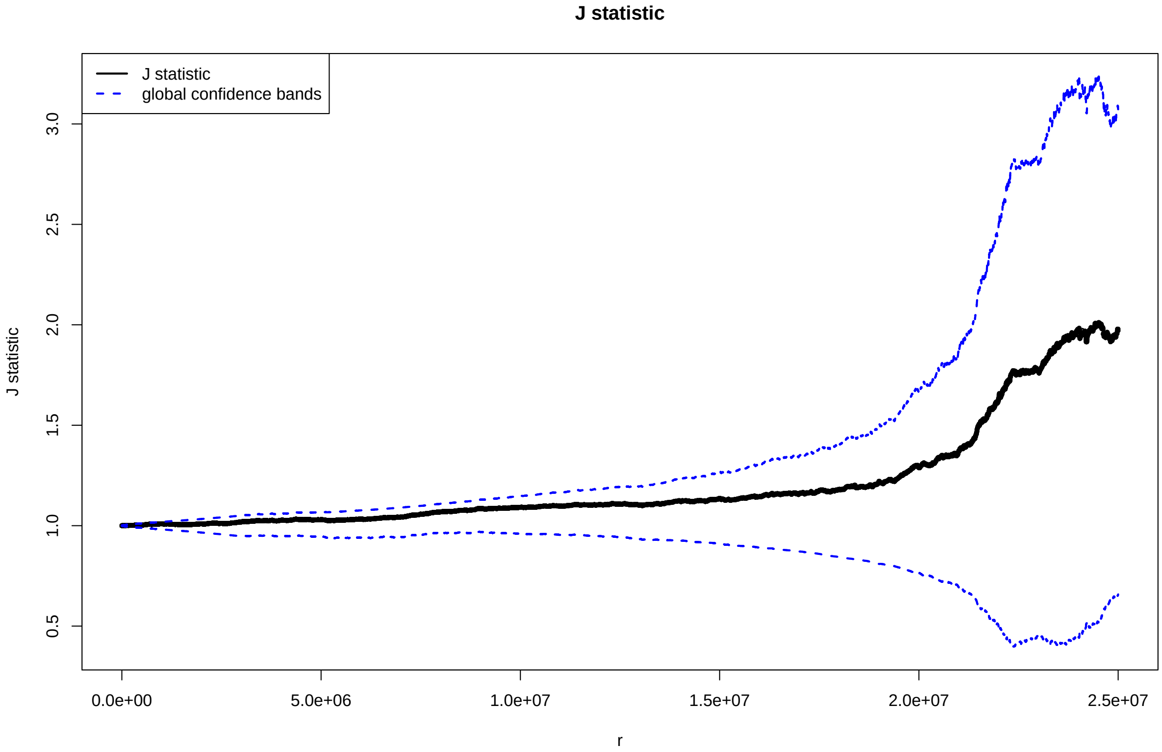


Figure 4.12: *J* statistic applied to the SNP and CNV difference profiles between primary tumor and metastatic tissue in chromosome 1 of a mouse. Confidence bands are constructed using the approach of partial block bootstrap.

To conclude, there is evidence that the SNP differences outside of the CNV regions in this data example do not follow complete spatial randomness. But there is no evidence against the hypothesis that SNP differences with distance less than 3×10^6 to CNVs have the same properties as those farther than 3×10^6 from CNVs. There is also no evidence against that the SNP differences outside of the CNV regions have the same properties everywhere.

4.5 Discussion

In order to perform statistical testing for spatial association between SNP differences and CNVs observed from the microarray platform, we propose to use two types of statistics, the J statistic (equation 4.2) and the more conventional neighborhood count C statistic (equation 4.1). The SNP differences can be viewed as following a spatial point process, and the CNV regions can be viewed as random sets existing in one dimensional spatial space. Foxall and Baddeley's J statistic [32] is designed to test association between a spatial point process and random sets in two dimensional space, and thus adapted in this study. Based on the conceptual framework of the Foxall and Baddeley's J statistic, the J statistic utilized here is developed for the one dimensional scenario while accounting for the probe sampling design. In particular, the location of a typical point in the point process is restricted to the probe sites, a subset of the entire chromosomal space, instead of the entire domain in [32]. The C statistic is developed in the spirit of the rainbow plot, which takes count of the SNP differences that are close to a certain distance to the nearest CNV. Both statistics are studied in their functional form over a large range of distance arguments. We propose three null hypotheses with increased generality over these hypotheses, related to the association between SNP differences and CNVs. We develop three Monte Carlo simulation approaches to construct confidence bands for statistical inference, one based on the parametric Poisson model and two block bootstrap methods. Our simulation studies show that the two block bootstrap methods seem to work reasonably well to test their corresponding null hypothesis, especially for the J statistic. The overall block bootstrap adapts the idea of overlapping moving block bootstrap as discussed in [35], which is well studied. However the partial block bootstrap is developed in this research, specifically targeting to test the second null hypothesis. The partial block bootstrap method still needs to be validated in the further study. A step function nhPP and a modified NPNSP allowing excessive or prohibitive clustering in the CNV *nearby* regions are proposed as the alternative processes to study the powers of the two test statistics. All three approaches for confidence band construction using the J statistic are recommended for testing the three null hypothesis.

The first null hypothesis and its corresponding approach of the Poisson null process are usually adopted in spatial analyses, as in [32]. The second and the third null hypotheses are potentially more related to the scientific questions from biologists and thus more of interest. We have included all the three null hypotheses for completeness, and provided corresponding approaches for inference. As illustrated and explained in our study, the rejection of the first null hypothesis cannot lead to the rejection of the second and the third null hypothesis. Misusing the inferential approach could lead to false conclusion.

The null distributions of the test statistics are evaluated conditional on the probe design. This design would then vary by any new scenario for a different chromosome setting or some other probe sets; however the framework for developing confidence bands would be the same as discussed here. The null distributions would also depend on the total number of SNP differences outside of the CNVs regions. The implementation for obtaining null distributions and power study is achieved by algorithms we developed using R software. For each chromosome, the computational time for testing procedure can be in the order of minutes on a PC with a four core Intel i7 CPU, depending on other parameters such as total number of SNP differences, CNVs and probes. Given the Monte Carlo samples, the global confidence bands constructed in this study are symmetric given any argument r. In case the distribution of $J(r)$ is highly skewed, the symmetric global confidence bands may not be the most efficient, yet an easy way to work on. The method for constructing more efficient asymmetric global confidence bands may be achieved by estimating global upper and lower quantile functions of the functional J

statistic, which still remains to be further explored.

The conclusion reached by the testing procedure in this study are based on mutation features on a single chromosome, and cannot infer to the individual organism or a certain population. Proper study design should be adopted when inference on the population level is of interest. The statistical tool in this study is developed under the microarray platform, and can be used to summarize the dependence between SNP differences and CNVs. The tool is ideal for large scale experimental designs, which are usually intended to compare various exposures. Dependence of mutations can be compared among experimental groups with different exposures. When considering a barrage of these tests for different chromosomes or multiple biological samples, such multiple tests should be adjusted in order to achieve either desirable overall type I error rate or false discovery rate (FDR). Methods such as Bonferroni correction or that provided by Benjamini [26] may be applied. Desirable FDR can be controlled to achieve higher discovery, which could lead to further detailed investigations and downstream analyses.

Both test statistics require arguments be set. Prior biological knowledge could help determine the choice of the argument r, which would define the range of the CNV neighborhood to be examined. Without prior knowledge, a relatively large argument r can be set, to incorporate a wide range of r values that are then visualized in the plot of the J statistic. The plot of the J statistic can help determine the regions where there is rejection of no association between SNP differences and CNVs, and the direction of the association, positive or negative.

In this chapter, the spatial models for evaluating the test statistics focus on the distributions of SNP differences conditional on CNVs. These tests could also incorporate the distribution of CNVs. However, as the prevalence of CNVs in typical data examples is very low, properly specifying the distribution of CNVs becomes infeasible. Naturally, there are more mutations in a single nucleotide base pair than those occurring in large segments of DNA. However, future data sets with more observation of CNVs in a single chromosome may help establish distributional models for CNVs, and as well joint distributions of SNP differences and CNVs may be specified. The J and C statistic would also be applicable in such a joint modeling framework.

The methods provided in this chapter are not restricted to utility with the mouse microarray, but can be adopted to any microarray data from other organisms, including human. The methods can also be extended to consider other distance distributions. Importantly, for example, the distances between a SNP and the 5′ or 3′ end of CNVs can be compared to investigate if the behaviors of SNP differences are symmetric on both side of the CNVs. The method can also be applied to study other genetic elements that are in the form of segments, such as genes, or extrons.

Chapter 5

Summary and future work

The research in this thesis develops formal statistical tools for the analysis of spatial properties of mutations detected under the microarray platform. Nonparametric statistics are developed for testing the existence of clusters of single point mutations and for association between point mutations and mutations in segments. We also study how the design of probe sites in microarray studies can affect the performance of recommended tests for clustering and we compare the performance of the test under the microarray platform to that for data from sequencing studies. Power performance of various statistics under different scenarios and different microarrays are contrasted via simulation experiments. Statistics with preferred power performance are identified and their utility is demonstrated with specific applications of interest. The $CvM_{\tilde{R}}$ statistic is recommended for testing the existence of clusters of single point mutations. The J statistic is recommended together with three inferential approaches to test three null hypotheses regarding the association between point mutations and mutations in segments.

The statistical tools developed in this thesis are delivered to the biologists to be applied in their research. It is worth emphasizing that the test statistics recommended are designed to provide a measure of spatial association between mutations on an entire single chromosome. The statistical tools are not capable of answering questions regarding the local properties such as where the clusters exist, what the size of the clusters are, or how the SNP differences are associated with the CNV regions. Further methods are required to be developed to answer these questions using the data observed under the microarray platform. It is also worth noting that the conclusion based on testing a single chromosome cannot infer to the individual organism or a certain population. Biologists would conduct proper study designs in order to study the mutation features on a population level. Biologists could use the statistical tools to compare various exposures in different population in large scale experimental designs. The measures the statistical tools provide on each individual can be compared among experimental groups with different exposures. In these application scenarios, a barrage of these tests are usually requested for different chromosomes and multiple biological samples. Multiple tests should be adjusted in order to achieve either desirable overall type I error rate or false discovery rate.

The innovations in this thesis are summarized below:

(1) A microarray platform has highly missing observations of mutations by design and thus offers challenges in using current statistical methods to test for spatial association between mutations. The statistical tools provided in this thesis incorporate the properties of the probe detection system and offer solutions with good performance for testing the existence of clusters of point mutations and association between two types of mutations.

(2) Microarray platforms are compared with sequencing in terms of their capability to study spatial association between mutations. How various probe designs can affect the ability to study spatial association between mutations is studied. The framework developed here can help determine key features when designing new microarray platforms.

(3) In the study of testing for the existence of spatial association between two types of mutational events, such as SNP differences and CNVs, three specific null hypotheses are proposed with increased level of generality in terms of spatial association. Overall and partial block bootstraps are developed as suitable inferential approaches for the two more general null hypotheses of no association respectively. Both block bootstrap approaches are adapted for the data under the microarray platform, using the idea of the moving

block bootstrap usually used in the analysis of time series data. The two more general null hypotheses are compared to the more traditionally used nulls of complete spatial randomness. The block bootstrap approaches are shown to be capable of providing reasonable type I error rates for testing the more general null hypotheses and has fairly good power performance.

Future work, inspired by the research in this thesis regarding the analysis of spatial association between mutations under the microarray platform or other biological measuring platform, is described as below.

(1) Spatial statistical tools can be developed to investigate some other pressing questions regarding the spatial association between mutations under the microarray platform. For example, since DNA has a direction, an extension to the study of association between two mutations would be to consider if the SNP differences on one side of the CNV have the same spatial properties as those on the other side. It is also of interest to study if the CNVs on the chromosome are randomly spaced or clustered. Since CNVs are mutations in segments with a certain length instead of point mutations, approaches would thus require a different strategy than considered here for studying clustering of SNP differences. When clustering is evident, it would also be useful to develop approaches and methods for identifying the location and the size of the clusters of point mutations based on microarray data.

(2) RNA sequencing is a technique that can provide information on the DNA on specific regions in genes. Instead of having information on one nucleotide as occurs under the microarray platform, RNA sequencing measures every nucleotide within the targeted regions. Statistical tools could be developed to test for clustering with the RNA-seq data, and investigations would need to be conducted to understand how such tests perform and whether they are effective for various types of hypotheses.

(3) For the probe design question, it would also be of interest to develop a strategy for

selecting a target number of probets from a predetermined pool of potential probe loci, where the goal is to achieve optimal capability of detecting existence of clusters with that fixed number of probes.

Statistical tools are capable of providing far greater insights over traditional visualization tools that have been developed in biological research. These tools need to reflect the properties of biological data gathered from various sources. With the development of suitable statistical tools, the tremendous biological data that are generated cost-effectively in scientific research can be well utilized to produce better understanding of human health.

Bibliography

[1] Struan FA Grant and Hakon Hakonarson. Microarray technology and applications in the arena of genome-wide association. *Clinical chemistry*, 54(7):1116–1124, 2008.

[2] Peilin Jia, William Pao, and Zhongming Zhao. Patterns and processes of somatic mutations in nine major cancers. *BMC medical genomics*, 7(1):11, 2014.

[3] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[4] Steven A Roberts, Joan Sterling, Cole Thompson, Shawn Harris, Deepak Mav, Ruchir Shah, Leszek J Klimczak, Gregory V Kryukov, Ewa Malc, Piotr A Mieczkowski, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand dna regions. *Molecular cell*, 46(4):424–435, 2012.

[5] Artem G Lada, Elena I Stepchenkova, Irina SR Waisertreiger, Vladimir N Noskov, Alok Dhar, James D Eudy, Robert J Boissy, Masayuki Hirano, Igor B Rogozin, and Youri I Pavlov. Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an apobec deaminase. *PLoS genetics*, 9(9):e1003736, 2013.

[6] Kathleen A Hill, Jicheng Wang, Kelly D Farwell, William A Scaringe, and Steve S Sommer. Spontaneous multiple mutations show both proximal spacing consistent

with chronocoordinate events and alterations with p53-deficiency. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 554(1):223–240, 2004.

[7] Jicheng Wang, Kelly D Gonzalez, William A Scaringe, Kimberly Tsai, Ning Liu, Dongqing Gu, Wenyan Li, Kathleen A Hill, and Steve S Sommer. Evidence for mutation showers. *Proceedings of the National Academy of Sciences*, 104(20):8403–8408, 2007.

[8] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[9] John W Drake. Too many mutants with multiple mutations. *Critical Reviews in Biochemistry and Molecular Biology*, 42(4):247–258, 2007.

[10] Wen-Cheng Chou, Wei-Ting Chen, Chia-Ni Hsiung, Ling-Yueh Hu, Jyh-Cherng Yu, Huan-Ming Hsu, and Chen-Yang Shen. B-myb induces apobec3b expression leading to somatic mutation in multiple cancers. *Scientific Reports*, 7:44089, 2017.

[11] Hyuna Yang, Yueming Ding, Lucie N Hutchins, Jin Szatkiewicz, Timothy A Bell, Beverly J Paigen, Joel H Graber, Fernando Pardo-Manuel de Villena, and Gary A Churchill. A customized and versatile high-density genotyping array for the mouse. *Nature methods*, 6(9):663–666, 2009.

[12] M EO Locke, Maja Milojevic, Susan T Eitutis, Nisha Patel, Andrea E Wishart, Mark Daley, and Kathleen A Hill. Genomic copy number variation in mus musculus. *BMC genomics*, 16(1):497, 2015.

[13] Wai-Ki Yip, Heide Fier, Dawn L DeMeo, Martin Aryee, Nan Laird, and Christoph Lange. A novel method for detecting association between dna methylation and diseases using spatial information. *Genetic epidemiology*, 38(8):714–721, 2014.

[14] Iuliana Ionita-Laza, Vlad Makarov, Joseph D Buxbaum, ARRA Autism Sequencing Consortium, et al. Scan-statistic approach identifies clusters of rare disease variants in lrp2, a gene linked and associated with autism spectrum disorders, in three datasets. *The American Journal of Human Genetics*, 90(6):1002–1013, 2012.

[15] Jingjing Ye, Adam Pavlicek, Elizabeth A Lunney, Paul A Rejto, and Chi-Hse Teng. Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC bioinformatics*, 11(1):11, 2010.

[16] Jose M Muiño, Ercan E Kuruoğlu, and Peter F Arndt. Evidence of a cancer type-specific distribution for consecutive somatic mutation distances. *Computational biology and chemistry*, 53:79–83, 2014.

[17] Diana Domanska, Daniel Vodák, Christin Lund-Andersen, Stefania Salvatore, Eivind Hovig, and Geir Kjetil Sandve. The rainfall plot: its motivation, characteristics and pitfalls. *BMC bioinformatics*, 18(1):264, 2017.

[18] Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212, 1977.

[19] Donald A Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957.

[20] Bernard V North, David Curtis, and Pak C Sham. A note on the calculation of empirical p values from monte carlo procedures. *The American Journal of Human Genetics*, 71(2):439–441, 2002.

[21] Bernard A Kunz and Susanne E Kohalmi. Modulation of mutagenesis by deoxyribonucleotide levels. *Annual review of genetics*, 25(1):339–359, 1991.

[22] Errol C Friedberg. Why do cells have multiple error-prone dna polymerases? *Environmental and molecular mutagenesis*, 38(2-3):105–110, 2001.

[23] Errol C Friedberg, Paula L Fischhaber, and Caroline Kisker. Error-prone dna polymerases: novel structures and the benefits of infidelity. *Cell*, 107(1):9–12, 2001.

[24] Myron F Goodman. Error-prone repair dna polymerases in prokaryotes and eukaryotes. *Annual review of biochemistry*, 71(1):17–50, 2002.

[25] CT Guy, RD Cardiff, and WJ Muller. Induction of mammary tumors by expression of polyomavirus middle t oncogene: a transgenic mouse model for metastatic disease. *Molecular and cellular biology*, 12(3):954–961, 1992.

[26] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[27] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499, 2010.

[28] Martin J Lercher and Laurence D Hurst. Human snp variability and mutation rate are higher in regions of high recombination. *Trends in genetics*, 18(7):337–340, 2002.

[29] Andy Itsara, Gregory M Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M Krauss, Richard M Myers, Paul M Ridker, Daniel I Chasman, et al. Population analysis of large copy number variants and hotspots of human genetic disease. *The American Journal of Human Genetics*, 84(2):148–161, 2009.

[30] Martin F Arlt, Thomas E Wilson, and Thomas W Glover. Replication stress and mechanisms of cnv formation. *Current opinion in genetics & development*, 22(3):204–210, 2012.

[31] Iuliana Ionita-Laza, Angela J Rogers, Christoph Lange, Benjamin A Raby, and Charles Lee. Genetic association analysis of copy-number variation (cnv) in human disease pathogenesis. *Genomics*, 93(1):22–26, 2009.

[32] Rob Foxall and Adrian Baddeley. Nonparametric measures of association between a spatial point process and a random set, with geological applications. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(2):165–182, 2002.

[33] Jens-Peter Kreiss and Soumendra Nath Lahiri. Bootstrap methods for time series. In *Handbook of statistics*, volume 30, pages 3–26. Elsevier, 2012.

[34] Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[35] Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013.

[36] James E Gentle. *Computational statistics*, volume 308. Springer, 2009.

[37] Bradley Efron. Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, 21(4):460–480, 1979.

Appendix A

Supplementary information for Chapter 2

A.1 Validation of size of test statistics

A.1.1 Scheme a

n	2	3	4	5	6	7	8
$D_{min}(n)$	0.053	0.064	0.058	0.05	0.056	0.054	0.046

Table A.1: size of $D_{min}(n)$ under various argument of n based on Scheme a.

d	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
$\bar{R}(d)$	0.054	0.058	0.058	0.055	0.051	0.048	0.05	0.052	0.051	0.053
$\tilde{R}(d)$	0.055	0.051	0.051	0.051	0.054	0.05	0.052	0.059	0.049	0.053
$N_{max}(d)$	0.001	0.005	0.01	0.022	0.034	0.05	0.002	0.004	0.004	0.006
$C(d)$	0.004	0.029	0.056	0.017	0.04	0.022	0.032	0.014	0.023	0.045

Table A.2: size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme a part 1.

d	55000	60000	65000	70000	75000	80000	85000	90000	95000	100000
$\bar{R}(d)$	0.057	0.053	0.052	0.056	0.057	0.052	0.045	0.045	0.051	0.047
$\tilde{R}(d)$	0.05	0.045	0.05	0.048	0.052	0.05	0.045	0.046	0.05	0.048
$N_{max}(d)$	0.007	0.007	0.009	0.012	0.015	0.015	0.017	0.02	0.022	0.029
$C(d)$	0.022	0.029	0.037	0.024	0.031	0.042	0.021	0.025	0.038	0.052

Table A.3: size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme a part 2.

	KS	CvM
$D_{min}(n)$	0.046	0.046
$\bar{R}(d)$	0.051	0.057
$\tilde{R}(d)$	0.058	0.053
$N_{max}(d)$	0.047	0.048
$C(d)$	0.051	0.048

Table A.4: size of functional forms of the five statistics based on Scheme a.

A.1.2 Scheme b

n	2	3	4	5	6	7	8
$D_{min}(n)$	0.053	0.046	0.042	0.031	0.043	0.049	0.045

Table A.5: size of $D_{min}(n)$ under various argument of n based on Scheme b.

d	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
$\bar{R}(d)$	0.05	0.053	0.05	0.048	0.051	0.05	0.054	0.059	0.062	0.051
$\tilde{R}(d)$	0.042	0.043	0.049	0.043	0.05	0.05	0.05	0.056	0.051	0.047
$N_{max}(d)$	0	0.006	0.015	0.021	0.029	0.026	0.022	0.013	0.009	0.008
$C(d)$	0.018	0.019	0.018	0.014	0.022	0.023	0.021	0.024	0.019	0.021

Table A.6: size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme b part 1.

d	55000	60000	65000	70000	75000	80000	85000	90000	95000	100000
$\bar{R}(d)$	0.056	0.057	0.057	0.056	0.054	0.052	0.056	0.046	0.047	0.046
$\tilde{R}(d)$	0.051	0.048	0.062	0.061	0.053	0.051	0.05	0.047	0.04	0.037
$N_{max}(d)$	0.01	0.009	0.01	0.009	0.009	0.009	0.013	0.015	0.015	0.017
$C(d)$	0.021	0.026	0.017	0.015	0.021	0.016	0.015	0.011	0.01	0.01

Table A.7: size of $\bar{R}(d)$, $\tilde{R}(d)$, $N_{max}(d)$, and $C(d)$ under various argument of d based on Scheme b part 2.

	KS	CvM
$D_{min}(n)$	0.008	0.005
$\bar{R}(d)$	0.046	0.043
$\tilde{R}(d)$	0.047	0.052
$N_{max}(d)$	0.041	0.043
$C(d)$	0.02	0.03

Table A.8: size of functional forms of the five statistics based on Scheme b.

A.2 Additional power study figures

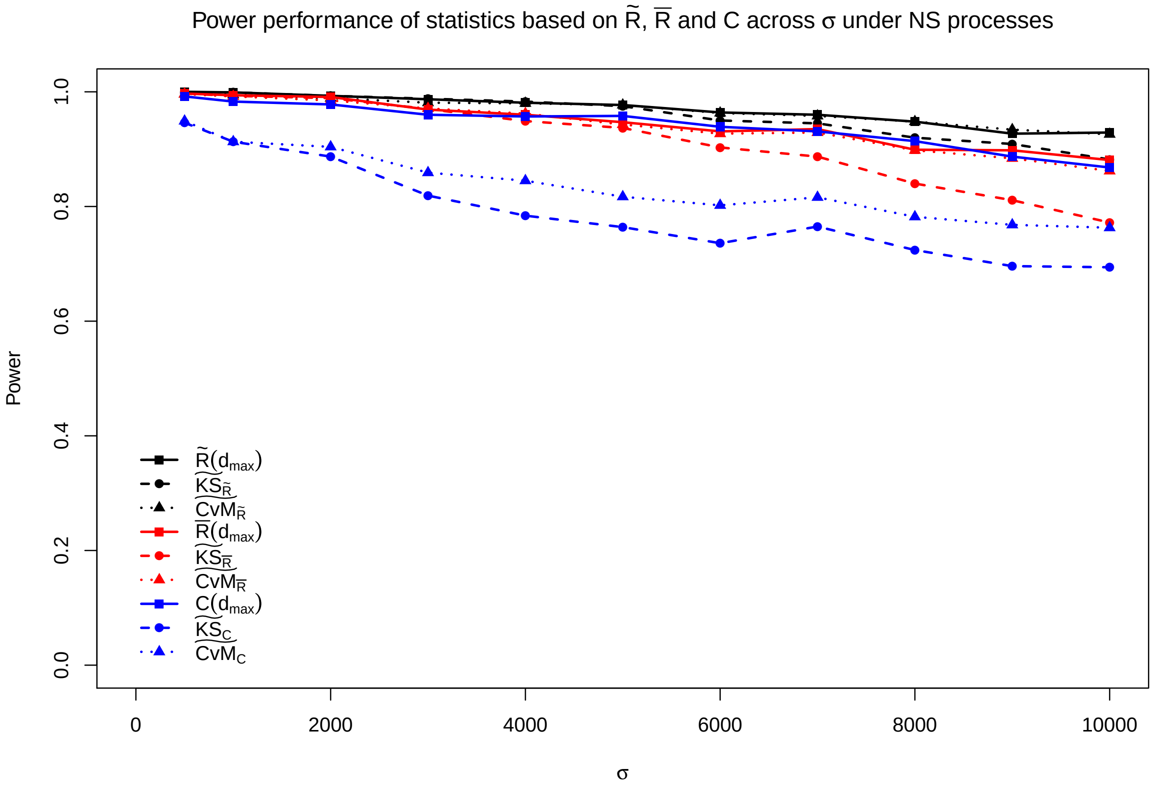


Figure A.1: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (1) with parameter $\mu_o = 1125$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.

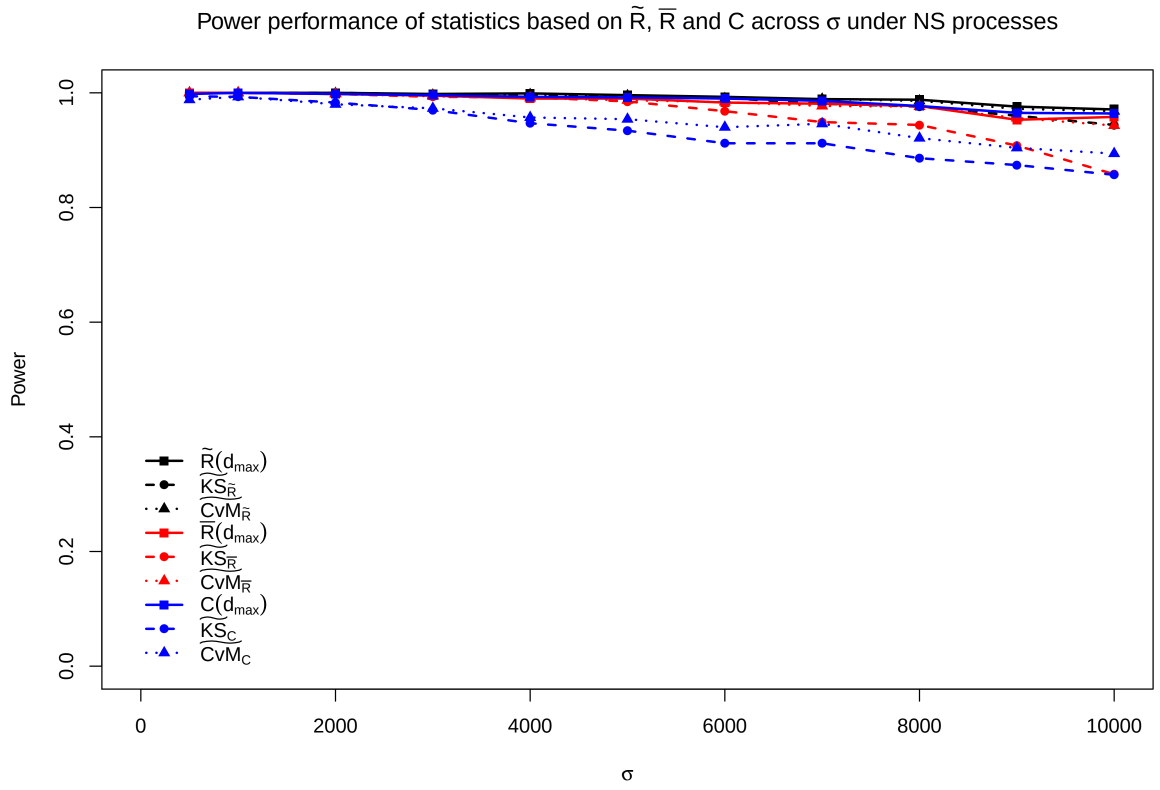
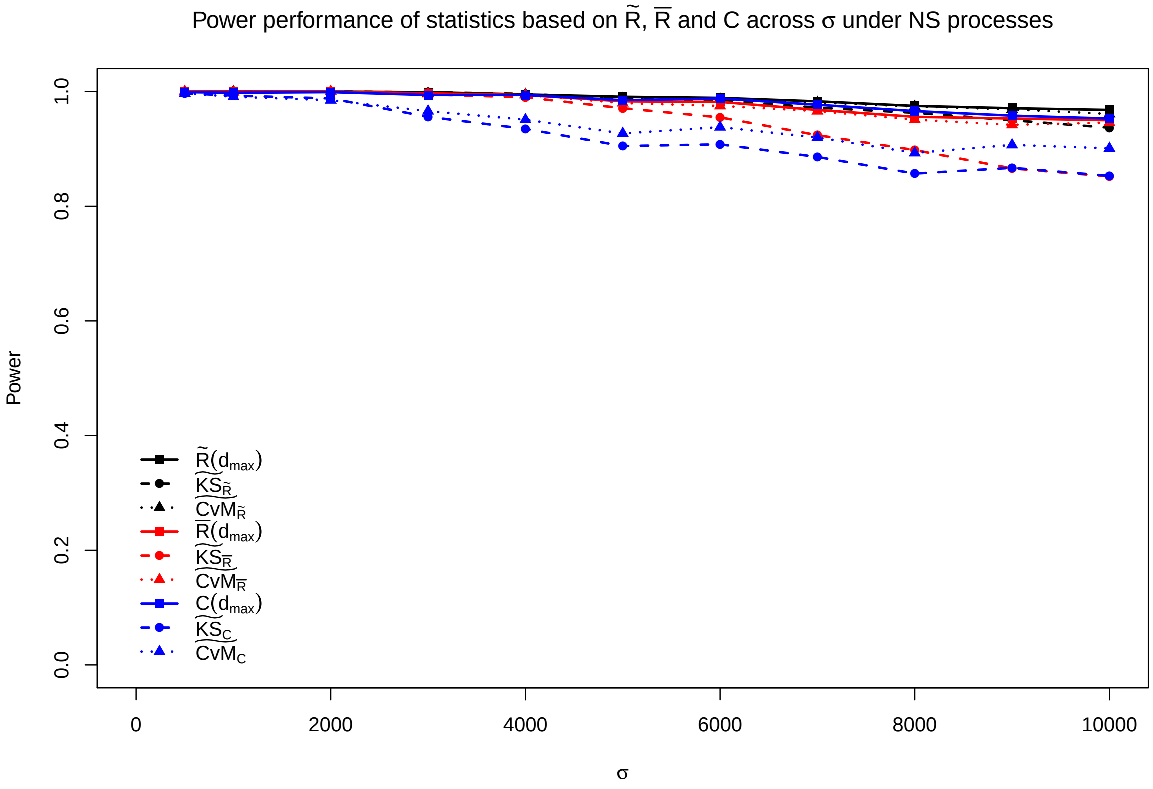


Figure A.2: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (2) with parameter $\mu_o = 1125$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.



Power performance of statistics based on R̃, R̄ and C across σ under NS processes

Figure A.3: Power performance of statistics related to $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ under alternative hypothesis (3) with parameter $\mu_o = 1125$.
Only maximum powers of $\bar{R}(d)$, $\tilde{R}(d)$, and $C(d)$ over values of d considered are displayed; d_{max} refers to the value of d yielding the largest power.

Appendix B

Supplementary information for Chapter 3

Algorithm S1: Monte Carlo estimates of the null distributions of summary statistics

2.1: Set a finite grid $D = \{d_i, i = 1, \cdots, k\}$, which defines the scale of d as the evaluation range;

2.2: Simulate M replications of detected SNP differences $\{X_0^{(m)}, m = 1, \cdots, M\}$ from the hPP. At the mth replication, $X_0^{(m)}$ is obtained as follows:

- (a): Generate the total number of underlying SNP differences $N_{null}^{(m)} \sim Pois(\hat{\lambda})$, where $\hat{\lambda}$ is an estimate of the rate parameter from the observed sample $X_s^{(r)}$: $\hat{\lambda} = \frac{(s_l - s_f)\eta}{|S|}$. The parameter η can be set as $|X_s^{(r)}|$, where $|A|$ is the norm of set A, that is the count of the number of elements in A;
- (b): Generate the set of underlying (both observable and unobservable) SNP difference locations $U_{null}^{(m)} = \{u_j, j = 1, \cdots, N_{null}^{(m)}\}$, where iid random variables $u_j \sim U[s_f, s_l]$, and U is the discrete uniform distribution on $\{s_f, \cdots, s_l\}$;
- (c): Obtain the set of observed SNP differences: $X_0^{(m)} = U_{null}^{(m)} \cap S$.

2.3: For each $m = 1, \cdots M$, obtain $\tilde{R}_{X_0^{(m)}}(\cdot) \equiv \{\tilde{R}_{X_0^{(m)}}(d_i), i = 1, \cdots, k\}$ at the grid sites $d_i, i = 1, \cdots, k$;

2.4: The Monte Carlo estimate of $\tilde{R}^*(\cdot)$ is $\hat{\tilde{R}}^*(\cdot) \equiv \{\frac{1}{M}\sum_{m=1}^{M} \tilde{R}_{X_0^{(m)}}(d_i), i = 1, \cdots, k\}$;

2.5: For each $m = 1, \cdots M$, calculate the $\widetilde{KS}$ or $\widetilde{CvM}$ test statistic:

(a): $\widetilde{KS}_{\tilde{R}}^{X_0^{(m)}} = \widetilde{KS}(\tilde{R}_{X_0^{(m)}}, \hat{\tilde{R}}^*)$;

(b): $\widetilde{CvM}_{\tilde{R}}^{X_0^{(m)}} = \widetilde{CvM}(\tilde{R}_{X_0^{(m)}}, \hat{\tilde{R}}^*)$;

2.6 The Monte Carlo estimates of the cumulative distribution functions of the test statistics $\hat{F}_{\widetilde{KS}_{\tilde{R}}}$ and $\hat{F}_{\widetilde{CvM}_{\tilde{R}}}$ are:

(a): $\hat{F}_{\widetilde{KS}_{\tilde{R}}}(t) = \frac{1}{M}\sum_{m=1}^{M} I(\widetilde{KS}_{\tilde{R}}^{X_0^{(m)}} \leqslant t)$

(b): $\hat{F}_{\widetilde{CvM}_{\tilde{R}}}(t) = \frac{1}{M}\sum_{m=1}^{M} I(\widetilde{CvM}_{\tilde{R}}^{X_0^{(m)}} \leqslant t)$

Algorithm S2: Hypothesis testing procedure

```
3.1 Based on the observed sample $X_s^{(r)}$, calculate $\tilde{R}_{X_s}(\cdot) \equiv \{\tilde{R}_{X_s^{(r)}}(d_i), i = 1, \cdots, k\}$. The test statistics are:
    (a): $\widetilde{KS}_{\tilde{R}}^{X_s^{(r)}} = \widetilde{KS}(\tilde{R}_{X_s^{(r)}}, \hat{\tilde{R}}^*)$;
    (b): $\widetilde{CvM}_{\tilde{R}}^{X_s^{(r)}} = \widetilde{CvM}(\tilde{R}_{X_s^{(r)}}, \hat{\tilde{R}}^*)$;
3.2 Statistical inference:
    (a): For hypothesis testing at significance level $\alpha$:
        (i): KS test: if $\widetilde{KS}_{\tilde{R}}^{X_s} > \hat{F}^{-1}_{\widetilde{KS}_{\tilde{R}}}(1-\alpha)$, reject the null hypothesis, otherwise do not reject.
        (ii): CvM test: if $\widetilde{CvM}_{\tilde{R}}^{X_s} > \hat{F}^{-1}_{\widetilde{CvM}_{\tilde{R}}}(1-\alpha)$, reject the null hypothesis, otherwise do not reject.
    (b): The p-values are calculated as:
        (i): KS test: $\frac{1+\sum_{m=1}^{M} I(\widetilde{KS}_{\tilde{R}}^{X_0^{(m)}} \geqslant \widetilde{KS}_{\tilde{R}}^{X_s^{(r)}})}{1+M}$;
        (ii): CvM test: $\frac{1+\sum_{m=1}^{M} I(\widetilde{CvM}_{\tilde{R}}^{X_0^{(m)}} \geqslant \widetilde{CvM}_{\tilde{R}}^{X_s^{(r)}})}{1+M}$;
```

Appendix C

Supplementary information for Chapter 4

C.1 Geometrical illustration of the partial block bootstrap

The partial block bootstrap procedure transforms two sets of distances, D_X and D_S into resampled sets of distances, D_X^* and D_S^* with the translation function $p(y; Z, B_P)$ (see equation 4.3). The transformation is geometrically illustrated as follows.

Consider a CNV on the chromosome as illustrated on the top panel in Figure C.1; find the neighborhood regions on its left side ($5'$ side in genetic terminology) and right side ($3'$ side in genetic terminology) that contain SNP differences and probes with nearest distances to this particular CNV. The boundaries of these regions are either the mid-point between two consecutive CNVs or either end of the chromosome, which is in practice the starting or ending probe on the chromosome. The shortest distance from any location in a CNV to the boundary of its neighborhood regions on the left or right sides is termed the distance from the CNV to its left or right boundary, respectively. The set of distances from CNVs to their left and right boundaries can be denoted as $D_B = \{D(b), b = 1, \cdots B\}$ and assume $D(1) < D(2) < \cdots < D(B)$. As the set D_B contains only unique elements, we have $B = |R| + 1$, where $|R|$ is the total number of CNVs on the chromosome.

Select all SNP differences and probes on both left and right side neighborhoods of all CNV and superpose them. Then, without lost of generality, perform a reflection of the left side SNP differences and probes so they land on the right side with the same distance to the nearest CNV. Through this superposition, the distances from SNP differences and probes to the nearest CNV can be displayed on a new coordinate system, all now being to the right of the CNV. The bootstrap procedure swaps all SNP differences and probes in the two distance region $(0, B_P]$ and $[Z, Z + B_P)$. The procedure is illustrated graphically with three CNVs on a chromosome in Figure C.1.

Given the data, in the bootstrap procedure, the information in the segment $(0, B_P]$ is fixed; however, information in the segment $[Z, Z + B_P)$ is random as it depends on the random point Z. Consider all the neighborhood regions, as Z increases and $Z + B_P > D(b)$ for some $b \in B$, the number of the neighborhood regions contributing information to the random interval of $[Z, Z + B_P)$ decreases. From a distributional perspective in the Monte Carlo simulation, the larger the corresponding distance to the boundary, the more information a neighborhood of the CNV would contribute to the bootstrap randomness.

Given parameter B_P, for a certain random number Z in a bootstrap resample, the two block segments to be exchanged, $(0, B_P]$ and $[Z, Z + B_P)$, may therefore arise from different numbers of neighborhood regions of the CNVs. Yet, the statistic $\hat{J}(r)$ contrasts the proportions of SNP differences and probe sites in the form of a ratio. Although not rigorously proved here, heuristically, under the null hypotheses, having the two segments contain information from different number of neighborhood regions would not affect the first order property of the ratio. The J statistic exchanges and contrasts information of SNP differences and probes in CNV *nearby* and *faraway* regions. The statistic is specially constructed for testing the second null hypothesis, where clearly defined CNV *nearby* and *faraway* regions are of interest for comparison. Further work is required for validation of this method. It is also worth noting that the global confidence bands constructed using the Monte Carlo samples from the partial block bootstrap can be instable, as seen in Fig 4.12 when r is relatively large. The reason for this behavior is yet to be investigated in the future.

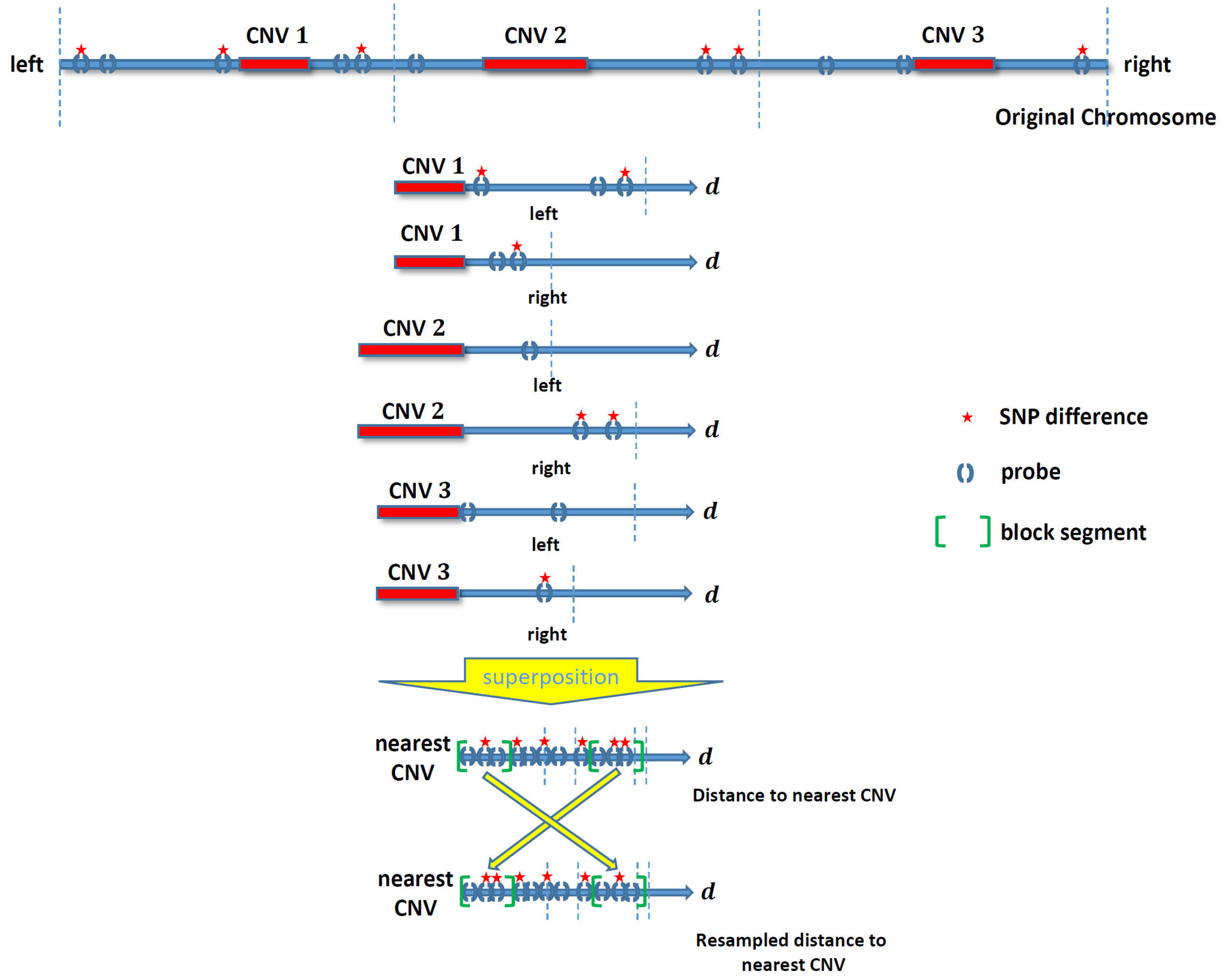


Figure C.1: Geometrical illustration of the partial block bootstrap. The distances from SNP differences and probes to the nearest CNV are obtained by superposition of SNP differences and probes in a defined region on either side of each CNV. The distances within the interval of $(0, B_P]$ and $[Z, Z + B_P)$ are exchanged by transformation, maintaining their relative positions within the segments.

C.2 Parameter setting in the alternative hypothesis for the step function nhPP model

Define the shortest distance from a probe site $s \in S$ to CNV regions R by

$$d_s = \inf_{a \in R} \|s - a\|$$

Then the probability of the probe detecting a SNP difference under the step function nhPP is given by

$$P(s \in X) = \tau I(d_s = 0) + \theta I(d_s \geqslant D_s) + \gamma_s \theta I(0 < d_s < D_s)$$

where τ corresponds to the probability of a probe inside of CNV detecting a SNP difference, D_s defines the neighborhood that is the *nearby-D* region of a CNV, θ corresponds to the probability that a probe in the *faraway-D* region from a CNV detects a SNP difference. The parameter τ is estimated from the sample:

$$\hat{\tau} = \frac{\sum_{s \in S} I(s \notin X) I(d_s = 0)}{\sum_{s \in S} I(d_s = 0)}$$

since

$$E\left[\sum_{s \in S} I(s \in X)\right] = m_0$$

where m_0 is the total number of SNP differences observed in the data, the parameter γ_s and θ have the constraint that

$$\hat{\tau} \sum_{s \in S} I(d_s = 0) + \theta \sum_{s \in S} I(d_s \geqslant D_s) + \gamma_s \theta \sum_{s \in S} I(0 < d_s < D_s) = m_0$$

Since γ_s is a parameter that determines how different the SNP difference intensities are in the *nearby-D* locations from *faraway-D* locations, it is altered arbitrarily to study power under different situations. The parameter D_s is also arbitrarily altered to reflect different ‘effect ranges’.

Fixing γ_s, then

$$\theta = \frac{m_0 - \hat{\tau} \sum_{s \in S} I(d_s = 0)}{\sum_{s \in S} I(d_s \geqslant D_s) + \gamma_s \sum_{s \in S} I(0 < d_s < D_s)}$$

C.3 Examination of autocovariance for location of SNP differences on microarray data

The autocovariance structure for the location of SNP differences in microarray data can be examined to provide guidance on selecting the block length parameters B_O and B_P. For a stationary process, the covariance can be expressed as follows:

$$Cov(X_t, X_{t+\delta}) = E(X_t X_{t+\delta}) - \mu^2,$$

where

$$\mu = E(X_t) = E(X_{t+\delta}),$$

$t \in L$ is an arbitrary location on the chromosome L, and $\delta \in Z^+$. A value of δ yielding that $E(X_t X_{t+\delta}) = \mu^2$, and thus $E(X_t X_{t+\delta}) - \mu^2 = 0$ would indicate no association between locations with δ distance. Since the microarray data is sparse and the number of probes is large, a special method is used to estimate the autocovariance structure, described as follows:

(1) randomly sample a set $Y \subset S$, $|Y| = N_1$, where $|Y|$ is the norm for set Y and S is the probe set;

(2) for each $y \in Y$, randomly sample a set $W(y) \subset \{g : g \in S, |g - y| > h_a\}$, where $|W(y)| = N_2$ and h_a is a sampling distance parameter;

(3) Sample pairs $Q_n = \bigcup_{y \in Y} \bigcup_{q \in \{g : g \in S, 0 < |g-y| \leqslant h_a\}} (y, q)$ and $Q_f = \bigcup_{w \in W(y)} (y, w)$. Obtain $Q = Q_n \cup Q_f$;

(4) Obtain $V = \bigcup_{(y,z) \in Q} \{(I(y \in X)I(z \in X), |y - z|)\}$. The elements in pairs $(\rho, \delta) \in V$ are samples of $(X_t X_{t+\delta}, \delta)$.

(5) Use kernel smoothing with a uniform kernel (bandwidth h_k) to estimate the function $\bar{\rho}(\delta)$. This estimate of $E(X_t X_{t+\delta})$ is to be compared with $\bar{\mu}^2$, where

$$\bar{\mu} = \frac{\sum_{(y,z)\in Q}[I(y \in X) + I(z \in X)]}{2|Q|}.$$

In the example in Section 4.4, the autocovariance structure is examined as seen in Figure C.2. The parameters are set as follows: $N_1 = N_2 = 10^3$, $h_a = 5\times10^6$, and $h_k = 4\times10^5$. Around $\delta = 3\times10^6$, the horizontal line seems to cross with the confidence bands estimated in the kernel smoothing, thus 3×10^6 is selected for the block length parameter B_O and B_P.

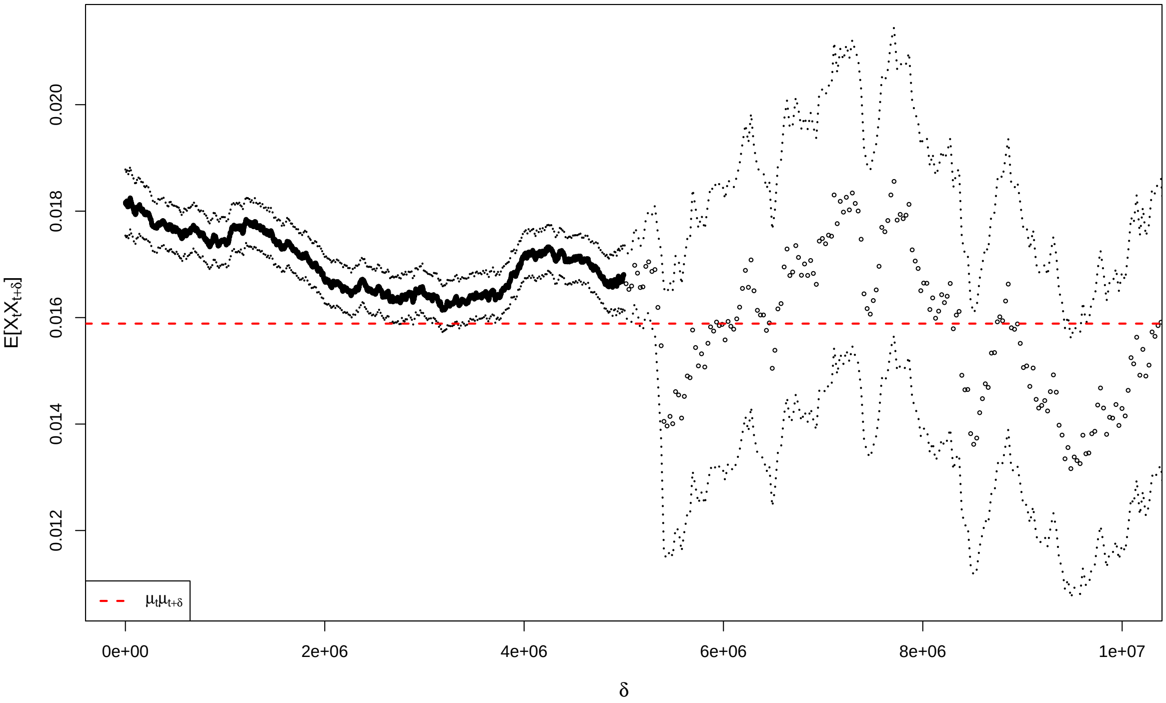


Figure C.2: Examination of the autocovariance for the locations of SNP differences in the example in Section 4.4.

Appendix D

Related techniques

D.1 Monte Carlo simulation for statistical inference and power estimation

For hypothesis testing, Monte Carlo simulation offers an effective approach for estimating the distribution of a test statistic T [36]. This method is especially convenient when T does not have closed form, as for the spatial statistics proposed in this thesis. Monte Carlo methods can be used to estimate the quantiles of T. With this method, m random samples are generated under the null hypothesis, with the parameter m selected to be equal to the size of the original data sample. The test statistic is calculated for each of the generated random samples, yielding a sample of test statistics, $t_1^*, \cdots, t_m^*$. The empirical cumulative distribution function P_m^* of these sample test statistics can be used as an estimate of the CDF of the test statistic P_T, used to estimate the critical region or p-value of the observed statistic. Denote r as the number of sample test statistics that are greater than or equal to the observed statistic. An unbiased estimator of the p-value is given as $\frac{r}{m}$. Another estimator of the p-value, $\frac{(r+1)}{(m+1)}$, introduces a slight bias, yet provides the correct type 1 error interpretation, which is often a desirable property [20]. Note that the bias introduced by the latter estimator diminishes when m increases.

In this thesis, Monte Carlo simulation is used to estimate the distribution of all test statistics for hypothesis testing as well as their power under alternative hypotheses. Monte Carlo samples are generated from parametric models (such as the hPP or NS) with specified parameters, or from bootstrap methods. The number of Monte Carlo samples is set arbitrarily. Although it is more accurate when large, increasing the number of samples increases computational resources required. The p-value is estimated in the simulation studies through the thesis by $\frac{r}{m}$, as the main goal here is to compare power between various test statistics under different scenarios, rather than having strictly correct interpretation of type I error. In real applications, it is recommended that the estimator $\frac{(r+1)}{(m+1)}$ be used.

D.2 Block bootstrap for dependent data

For simple random sampling of the data, Efron's bootstrap method [37] uses a nonparametric sampling scheme to approximate the distribution of a certain random variable without requiring model assumptions on the CDF. The bootstrap approach is described as follows, as taken from [35]. Assume $X_1, X_2, \cdots$ is a sequence of iid random variables with common CDF F. Suppose $\mathcal{X}_n = \{X_1, \cdots, X_n\}$ are the data at hand and let $T_n = t_n\{\mathcal{X}_n; F\}, n \geq 1$ be the random variable of interest. A simple random sample $\mathcal{X}_n^* = \{X_1^*, \cdots, X_n^*\}$ of size n is drawn with replacement from $\mathcal{X}_n$. Then conditional on $\mathcal{X}_n$, $\{X_1^*, \cdots, X_n^*\}$ are iid random variables with common distribution $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, where δ_y denotes the probability measure putting unit mass at y. The bootstrap version T_n^* of T_n is defined as $T_n^* = t_n(\mathcal{X}_n^*; F_n)$, with $\mathcal{X}_n^*$ and F_n replacing $\mathcal{X}_n$ and F respectively. Denote $\hat{G}_n$ as the conditional distribution of T_n^* given $\mathcal{X}_n$ and G_n as the unknown distribution of T_n. Then the bootstrap principle advocates $\hat{G}_n$ as an estimator of G_n. For functional $\varphi(G_n)$, a bootstrap estimator is given by $\varphi(\hat{G}_n)$. Monte Carlo simulation can be used to approximate $\hat{G}_n$.

However, the method described above is not suitable for dependent data, and other bootstrap techniques have been developed, such as the moving block bootstrap, the nonoverlapping block bootstrap, and a generalized block bootstrap. This thesis adapts the moving block bootstrap approach. Other methods can be investigated in future studies. Instead of resampling a single observation each time, the moving block bootstrap method resamples blocks of consecutive observations. Thus the dependence structure of the original observations within each block is preserved. The moving block bootstrap method is described as follows. Let $X_1, X_2, \cdots$ be a sequence of stationary random variables, and $\mathcal{X}_n = \{X_1, \cdots, X_n\}$ be the observations in the data. Suppose $l \in [1, n]$ is an integer. Let $\mathcal{B}_i = (X_i, \cdots, X_{i+l-1})$ denote the block of length l starting with X_i, $1 \leqslant i \leqslant N$, where $N = n - l + 1$. Let $\mathcal{B}_1^*, \cdots, \mathcal{B}_k^*$ denote a simple random sample drawn with replacement from $\{\mathcal{B}_1, \cdots, \mathcal{B}_N\}$. Denote the elements in $\mathcal{B}_i^*$ by $(X^*_{(i-1)l+1}, \cdots, X^*_{il}), i = 1, \cdots, k$. The sample $X_1^*, \cdots, X_m^*$ is a moving block bootstrap sample of size $m \equiv kl$. The empirical distribution of $(X_1^*, \cdots, X_m^*)$, $F^*_{m,n}$, can then be used for estimation of other statistics, such as $\hat{\theta}_n = T(F_n)$.

In Chapter 4, block bootstrap strategies are adopted to generate bootstrap samples in order to estimate the variance of test statistics under the null hypotheses. Simple random samples of blocks containing probe site locations and their corresponding outcomes of SNP differences are generated and used to form resamples. Monte Carlo simulations are used to estimate the variance of the test statistics based on these resamples.

D.3 DNA data structure: discreteness

The nucleotides on a DNA forms a linear structure. Each nucleotide has a defined location determined by its order on the chromosome based on a pre-defined starting location. This can be regarded as its coordinate on a one dimensional coordinate system (number line). As determined by its order, the locations of the nucleotides are always integer and cannot be real numbers. However, the scale of the chromosomes in some organisms is in the order of 10^8 bp and a nucleotide is of length 1 bp. It is thus often reasonable to treat the locations as continuous rather than discrete. When generating Monte Carlo samples from certain point processes, such as the Neyman-Scott process, the locations of generated points,i.e., parents and offspring, can be continuous. In this case, the locations of these points are rounded to the nearest integer. The rounding here should in principle not affect the goal of this study and interpretability.

D.4 Spatial point process

In this thesis, a spatial point process is used to describe the behavior of the underlying mutations, either detected or undetected. A spatial point process X is a random countable subset of a space S. In this scenario, S is the chromosome space $L \in \mathbb{R}^+$, a one dimensional space. Two major types of point processes, the Poisson process and the Neyman-Scott process, are considered in this thesis, and described below.

The Poisson point process is described as follows. For bounded $B \subseteq S$, the intensity measure μ is given by

$$\mu(B) = \int_B \rho(\xi) \mathrm{d}\xi, B \subseteq S;$$

and the count function $N(B) = |B|$, which is the random number of points falling in B. A point process X consisting of $n, n \in \mathbb{Z}$, i.i.d. points with a density function f on a set $B \subseteq S$ is called a binomial point process, denoted as $X \sim \text{binomial}(B, n, f)$. A point process X on S is a Poisson point process with intensity function ρ if the following properties are satisfied:

(1) for any $B \subseteq S$ with $\mu(B) < \infty$, $N(B) \sim \text{Pois}(\mu(B))$; if $\mu(B) = 0$ then $N(B) = 0$.

(2) for any $n \in \mathbb{Z}$ and $B \subseteq S$ with $0 < \mu(B) < \infty$, conditional on $N(B) = n$, $X_B \sim \text{binomial}(B, n, f)$, where $X_B = X \cap B$ is the restriction of X to B.

Note that if $\rho(\xi) = \rho$ is a constant in a Poisson process, the process is a homogeneous Poisson process; otherwise it is an non-homogeneous Poisson process.

The Neyman-Scott process is described as follows. Let C be a stationary Poisson process with intensity $\kappa > 0$. Conditional on C, let $X_c, c \in C$ be an independent Poisson process with intensity function $\rho_c(\xi) = \alpha k(\xi - c)$, where $\alpha > 0$ and k is a kernel function. Then $X = \cup_{c \in C} X_c$ is a special case of a Neyman-Scott process, where C are cluster centers or parents and $X_c, c \in C$

are the clusters or offspring. In a more general definition, $N(X_c)$ given C is not restricted to be Poisson distributed. In this thesis, the kernel function is chosen to be the Gaussian kernel. The cluster centers C can be a homogeneous Poisson process or a non-homogeneous Poisson process, where the intensity of C depends on the distance to the nearest mutation event of another type.

D.5 Notes on numerical computation

In this thesis, some definitions of statistics involve ratios between two numbers. However, the denominator of the ratio is not always guaranteed to be non-zero as it is often depends on the data and the choice of argument.

For example, the statistic $\bar{R}(d)$ has $|X|$ and $N_S(x, d)$ in the denominator. If $|X| = 0$, that is, there is no SNP difference observed at all, the calculation of $\bar{R}(d)$ would be invalid. Meanwhile if for any SNP difference $x \in X$, there is no probe site for which the distance to x is less than or equal to a given distance d, the calculation of $\bar{R}(d)$ would be invalid as well. Thus for the former issue, it is required that a check is made that there exists at least one SNP difference in the data set. For the latter issue, the argument d needs to be specified large enough for all of the SNP differences to have at least one nearby probe site in the neighborhood. The same issue exists in the recommended statistic $\tilde{R}(d)$, where the denominator $\sum_{x \in X} N_S(x, d)$ needs to be non-zero. This constraint only requires that there are probe site within distance d to at least one SNP difference, which is a much easier condition to satisfy compared to that needed for valid computation of $\bar{R}(d)$. In Chapter 2, that the denominator is zero is not an issue, as d is specified large enough, relative to the inter-probe distances in MDGA; this is not a major constraint. In Chapter 3, a subset of values of $d \in D$ could lead to invalid calculation of $\tilde{R}(d)$, and we discussed suitable methods to handle the situation in Section 3.2.3.

In the study of association between SNP differences and CNVs, the calculation of the J statistic also has the potential issue that the denominator may theoretically be zero. The denominator, $1 - \hat{G}_{V \cap R^C,\, R}(r)$, can be zero when r is chosen to be larger than $\max_{x \in (V \cap R^C)} d(x, R)$, which is the largest distance from a SNP difference outside of CNVs to the nearest CNV. Thus r should be specified small enough for the denominator not to take value zero, and this is a trivial constraint to impose.

Curriculum Vitae

Name: Bin Luo

Post-Secondary Education and Degrees: Huazhong University of Science and Technology
Wuhan, Hubei, China
2004 - 2008 B.Eng. in Bioinformatics

Western University
London, ON
2011 - 2012 M.Sc. in Biostatistics

Western University
London, ON
2012 - 2018 Ph.D. in Biostatistics

Grants: Engage Grant
2013-2014

Mitacs Accelerate Grant
2014 - 2015

Related Work Experience:	Teaching Assistant Western University 2011 - 2013
	Research Assistant Western University 2015 - 2018

Publications:

1. Luo, Bin, Alanna K. Edge, Cornelia Tolg, Eva A. Turley, C. B. Dean, Kathleen A. Hill, and R. J. Kulperger. ”Spatial statistical tools for genome-wide mutation cluster detection under a microarray probe sampling system.” PloS one 13, no. 9 (2018): e0204156.