ASSOCIATION ANALYSIS OF GENOMIC SEQUENCES

Abdulrahman Obaid Alshammari

Thesis submitted for the degree of Doctor of Philosophy

Department of Mathematics and Statistics

May 2022

University of Reading

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Abdulrahman Obaid Alshammari

Abstract

Studying genetic variations can help improve understanding of cancer aetiology and provide scientists with inspirational perspectives of tumour cells growth. Somatic mutations play a significant role in the development of cancer. Therefore, substantial effort has been expanded in order to identify somatic mutations. In light of this, in this research, we develop a novel method for detecting the impact of somatic mutations by matching tumour and normal sequences taken from an individual based on the score test and implementing the generalised higher criticism (GHC) test correction. The proposed score test is appraised and compared to the binomial exact test by utilising simulations. Results of a wide range of simulations show that our method controls type I error and is more effective than the binomial exact test.

Another way we propose with regard to association analysis of somatic mutations is to account for the uncertainty of discovering mutations. Since standard association methods do not take into account possible calling errors for somatic mutations, they are limited in their suitability for investigating functional consequences of such mutations. A recent somatic mutation association test with measurement errors (SAME) that addresses this issue via the likelihood ratio test has shown that taking account of uncertainty in somatic mutation calling improves power for detecting an association. In the spirit of SAME, the proposed score test procedure in this thesis models actual somatic mutation as an unobservable variable and uses read-depth to increase the mutation calls. The score test is computationally efficient as only optimisation under the null model is required for each genetic variant. Additionally, the risk of non-convergence of optimisation routines is reduced. These computational advantages are particularly beneficial in genomewide settings. The score test is evaluated using simulations. Results of extensive evaluations and comparisons with the SAME procedure and GLM that does not consider mutation calling errors reveal that our proposed approach preserves type I error and is more efficient than the SAME and GLM methods.

Acknowledgements

First and foremost, I would like to praise Allah the Almighty, the Most Gracious and the Most Merciful for His blessing given to me during my study and in completing this thesis.

I would like to express my sincere gratitude and most profound appreciation to my supervisor, Dr Fazil Baksh, for his continuous support and unlimited help and for his patience, advice and encouragement during this thesis. His guidance and thoughtful suggestions helped me throughout my research. Without him, the research would not have been possible.

I would also like to thank my Higher Degree Committee members, Dr Jing Hua Zhao and Dr Jeroen Wouters, for their helpful comments and guidance.

To all my colleagues and friends in the Department of Mathematics and Statistics and to my friends in Reading, thank you for your outstanding support and help.

To all members of my family, I am incredibly grateful to my beloved mother for her infinite love and ceaseless prayers. I am very much thankful to my dear wife, Afnan, for her unending inspiration and constant encouragement. A special thanks to my lovely daughter, Hoor. Thank you for the love, joy and smiles you give me every day. You are the light of our life. Also, I am deeply indebted to my brothers and sisters for their tremendous support and valuable prayers.

Finally, I would like to take this opportunity to express my immense thanks to the government of Saudi Arabia and Jouf University for the financial support and assistance through funding my PhD studies.

Contents

List of Tables

List of Figures

Introduction

Many complex diseases, including cancer, are believed to include genetic variations in the causal pathways. Consequently, numerous scientific and therapeutic research efforts are dedicated to discovering the functions of genetic mutations. With the fast development and substantial advances in technologies, it is possible to align and analyse the whole genome sequences in order to investigate the impact of genetic variants. Even though genomewide association study approaches have discovered a considerable number of diseases that are thought to be associated with particular genetic mutations by applying single-variant association tests, GWAS typically searches only for common variants, defined as variants whose frequency is higher than 5% in a population (Hindorff et al., 2009).

For this reason, methods have been proposed for the purpose of dealing with low-frequency mutations, described as variants whose frequency is between 1% and 5%, and rare mutations that have a frequency of $\leq$ 1%. Association approaches of low-frequency and rare genetic variants include the cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007), combined and multivariate collapsing (CMC) test (Li and Leal, 2008), weighted-sum statistic (WSS) (Madsen and

Browning, 2009), Morris and Zeggini (MZ) test (Morris and Zeggini, 2010), variable allele-frequency threshold (VT) test (Price et al., 2010), sum of square score (SSU) test (Pan, 2009), sequence kernel association test (SKAT) (Wu et al., 2011), C-alpha test (Neale et al., 2011), SKAT-O (Lee et al., 2012), Fisher method (Derkach et al., 2013) and MiST (Sun et al., 2013). These approaches, discussed in Chapter Three, adopt the set-based technique and aggregate the entire genetic information from multiple variants within a region, such as a gene, due to the fact that considering each single low-frequency variant may lead to a loss of power in statistical association procedures.

In terms of genetic association studies of cancer, it is expected that somatic mutations have a considerable effect on cancer outcomes, and their roles are much more critical than germline mutations. As a result, exploring somatic mutations and studying the extent of the influence of somatic mutations have become a notable phase in cancer aetiology. In spite of the fact that somatic mutations have a low-frequency and rare frequency, the developed approaches of low-frequency and rare genetic variants mentioned above are not appropriate for dealing with somatic mutations. It is assumed because these approaches do not account for somatic mutation calling errors. The uncertainty of calling somatic mutations can affect the accuracy of the association testing analysis.

The aim of this thesis revolves around investigating the effect of somatic mutations. We present a development of a score test procedure based on applying the generalised higher criticism (GHC) test (Barnett et al., 2017) in order to compare cancerous and healthy samples sequenced from the same patients. Another

approach we present in this research with the objective of somatic mutation association analysis is to evaluate somatic mutations' consequences while taking mutation calling errors into consideration. We propose a novel association method for the purpose of testing the association between a single somatic mutation or multiple somatic mutations combined within a whole gene and a cancer subtype outcome.

The composition of the thesis is as follows. Chapter One provides a basic introduction to human genetics and gives a background of cancer genetics and terminologies of epidemiology. Then, it presents an overview of the 100,000 Genomes Project as the work in the thesis is motivated by the objectives of this project. Chapter Two exhibits common statistical approaches used in this thesis. One of the included methods is the GHC test as we apply it to the proposed score test in Chapter Four.

Chapter Three is a literature review on the analysis of association and detection of genetic variants. It begins to give details of the association analysis procedures used to test the effect of a single mutation or multiple mutations grouped within a genetic construct, such as a gene, on a disease outcome. Since this thesis aims at studying the impact of somatic mutations on a cancer-related outcome, an overview of detecting approaches for calling somatic mutations is presented in the second part of Chapter Three. Finally, the chapter concludes by offering an introduction to a recent method produced in order to evaluate the relationship between somatic mutations and a cancer subtype outcome. In Chapter Four, a novel methodology with the objective of assessing a set of somatic mutations gathered within a gene

is proposed and evaluated. Our developed score method compares diseased and disease-free strings sampled from the same person. A simulation study in a wide range of scenarios is presented in order to evaluate our proposed test in terms of type I error and power.

In Chapters Five and Six, we develop a methodology for somatic mutation association analysis. Novel score tests are constructed in order to study the relationship between a single somatic mutation and gene-level somatic mutations and a cancer subtype outcome while taking the somatic mutation calling errors into consideration. Simulation studies showing a range of effect sizes and mutation rates are provided at the end of both chapters to assess our proposed single somatic mutation and gene-based somatic mutations procedures regarding type I error and power.

Chapter 1

Cancer genetics

This chapter introduces human genetics and the role of genes in the development of cancer. Genetic principles and terminologies used in this thesis are showed. The focus is on genetic variations, including somatic mutations, and the role genes play in cancer development. Finally, the chapter is concluded with an overview of the 100,000 Genomes Project.

1.1 Introduction to human genetics

It is widely recognised that the human body is composed of trillions of cells, and they are the fundamental unit of all organisms, including humans. The cells have miscellaneous functions, such as providing structure for the body, creating metabolic reactions and supporting the body with energy by converting nutrients obtained from food. In addition to these tasks, cells strain the human body's genetic materials, which are encoded in chromosomes within a nucleus. Thus, the nucleus is considered the control room of a cell and is located in the middle of it, as shown in Figure 1.1.

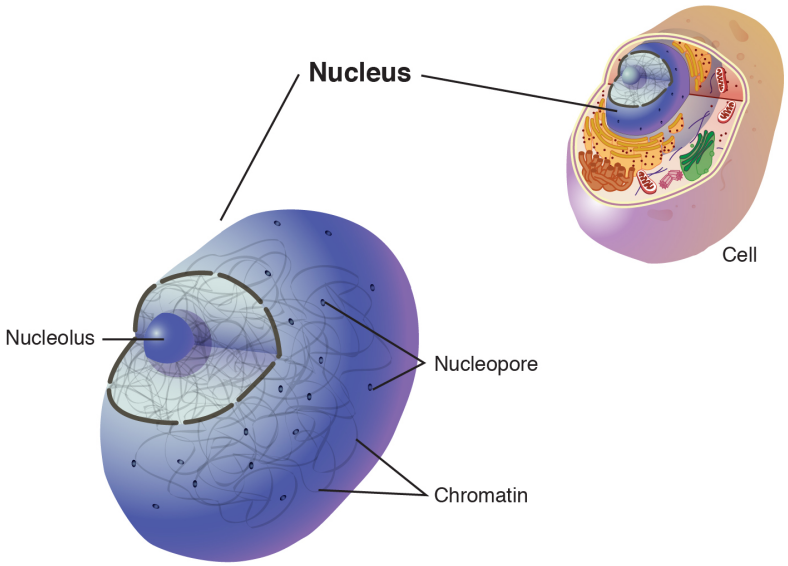

Figure 1.1: A nucleus in a cell. (National Human Genome Research Institute (NHGRI))

In humans, the genome comprises 23 pairs of chromosomes, and this means that the human genome has 46 chromosomes in total. The lengths of the chromosomes vary as displayed in Figure 1.2. The largest chromosome is Chromosome 1, and it includes around 248 million base pairs. By contrast, the smallest chromosome is Chromosome 21 that contains about 47 million base pairs (Richards and Hawley, 2011). With regard to chromosome classifications, there are two types of chromosomes. Specifically, 22 pairs of chromosomes are assorted as autosomes, and a pair of chromosomes is classified as sex chromosomes. Males and females have the same 22 pairs of autosomes, but they are different in the pair of sex chromosomes. Males' sex chromosomes are XY while females' chromosomes are XX, as illustrated in Figure 1.3. One set of 22 pairs of chromosomes is inherited from a father, whilst the other is from a mother. In terms of the sex chromosomes, a male obtains the Y chromosome from his father and the X chromosome from his mother, whereas a female takes a pair of the X chromosome from each parent. In the sex chromosomes, the pseudoautosomal region (PAR) is a particular element which has known functions in male meiosis and fertility (Helena Mangs and Morris, 2007).

Every chromosome is composed of deoxyribonucleic acid (DNA). DNA is defined as a duple chain of nucleotides, while a nucleotide consists of three elements. The components are a five-carbon sugar, a phosphate group and a base of four nitrogenous bases. The four chemical bases of nitrogen are classified as adenine (A), thymine (T), cytosine (C), and guanine (G). Figure 1.4 shows a close image of the DNA structure and demonstrates that the nitrogenous bases come with each other. Adenine goes with thymine, and guanine pairs with cytosine in order

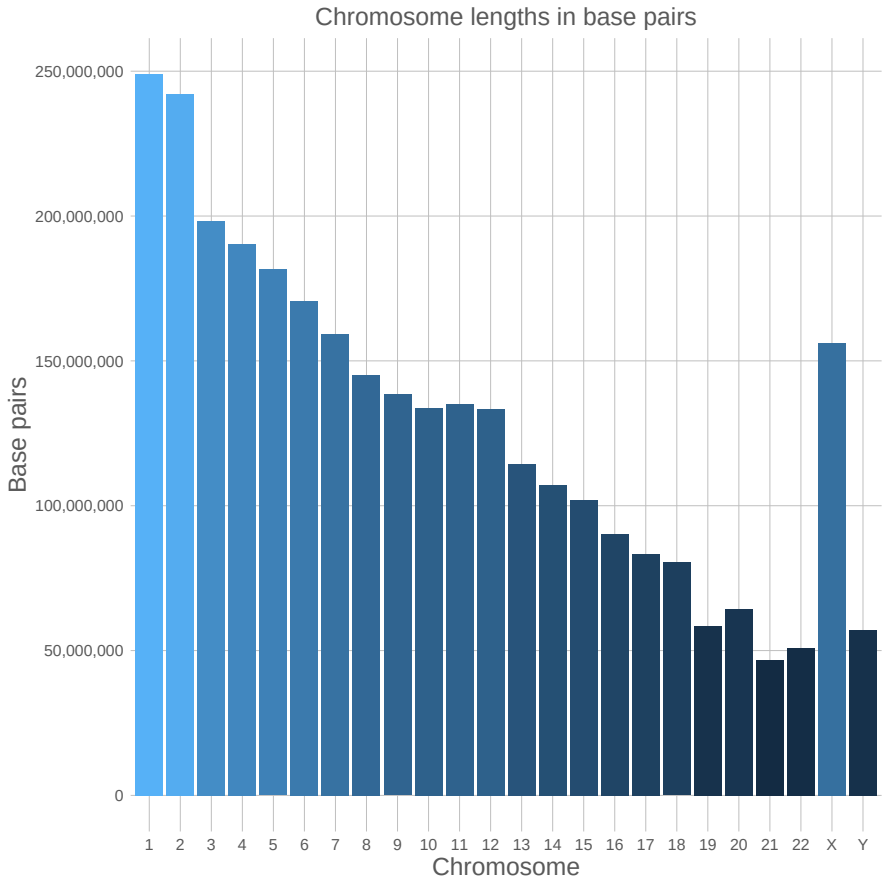

Figure 1.2: Length of the chromosomes in base pairs in the human genome.

to build the double-strand. There are approximately 3 billion base pairs in the human genome.

A specific sequence in the DNA string that encodes for one protein is described as a gene, as shown in Figure 1.5. Coiled DNA to form chromosomes in a cell nucleus provides the complete information for the cell, and genes are the functional subunits of DNA. Each gene carries a unique set of instructions, usually coding for a particular protein or specific function. There are around $20,000 - 25,000$

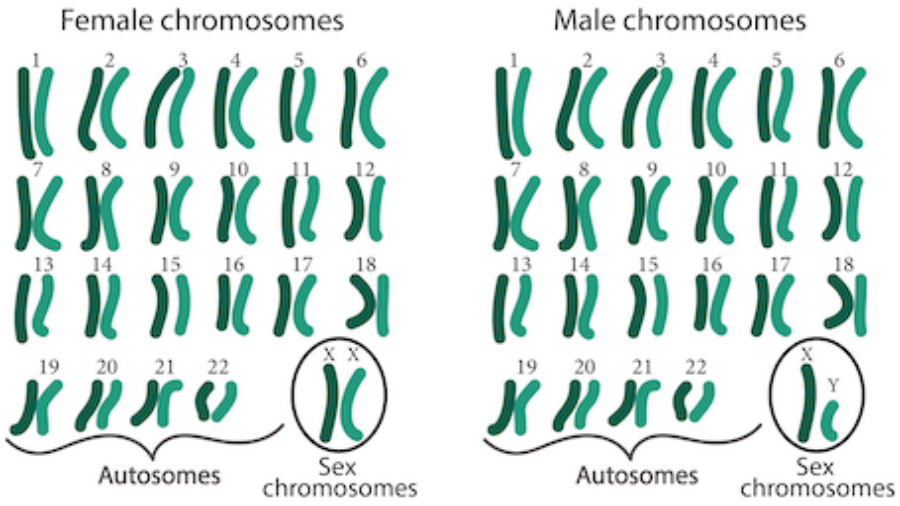

Figure 1.3: Pairs of chromosomes. 22 pairs are autosomes and a pair is sex chromosomes. (The Tech Interactive)

genes distributed in the 23 pairs of chromosomes with diverse sizes. Figure 1.6 demonstrates that Chromosome 1 is composed of the most number of genes as it contains approximately 3000 genes, while Chromosome 2 is the second-largest human chromosome and contains around 2500 genes. In contrast, Chromosome Y is the least chromosome in genes since it encompasses around 200 genes (Richards and Hawley, 2011). Genes differ in length from hundreds of base pairs to more than 2 million base pairs.

A location of a gene or any genetic marker is termed a locus (plural loci). A genetic marker is characterised as any part of the DNA sequence that can be a single base pair or a DNA series, such as a gene.

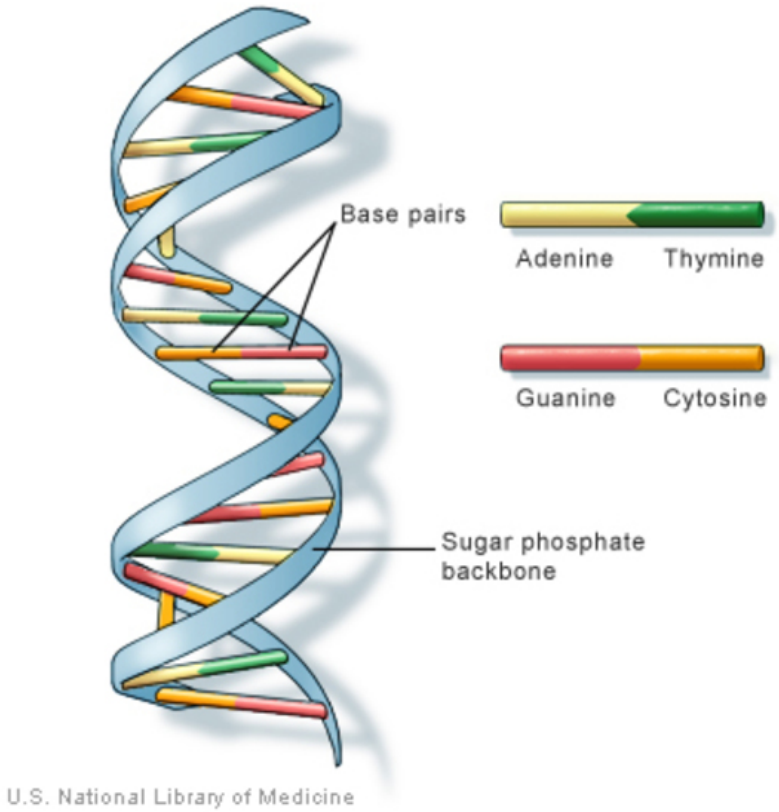

Figure 1.4: The chemical structure of deoxyribonucleic acid (DNA). (US National Library of Medicine)

1.1.1 Genetic variation

The genetic information of a genetic marker in a specific locus is called a genotype, and it produces a physical expression called a phenotype. Thus, even though the human genome is relatively the same among people, genetic variations can cause diversity in phenotypes and make different characteristics. On the other side, it is believed that some genetic variations are natural and harmless; therefore, they

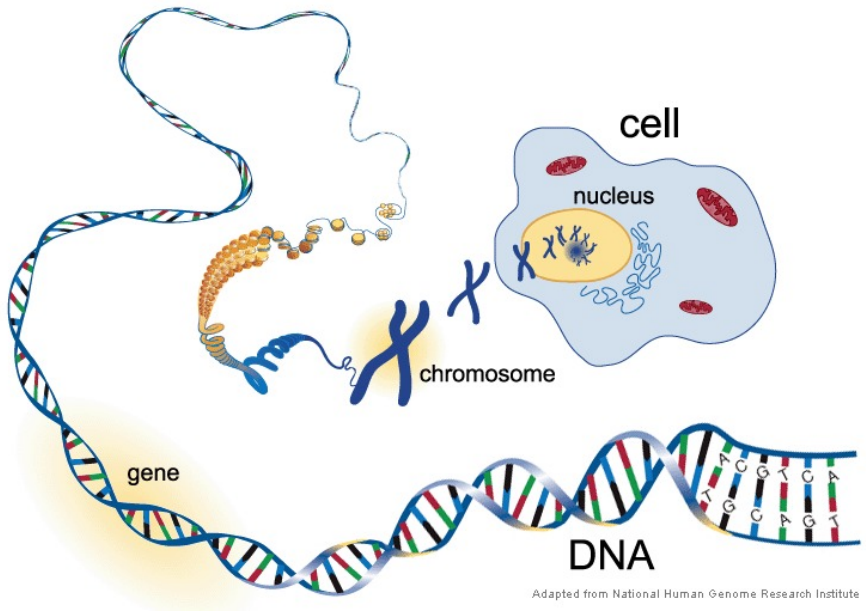

Figure 1.5: Illustration of a chromosome in the nucleus of a cell and its DNA sequences. (National Human Genome Research Institute (NHGRI))

might not affect phenotypes.

A genetic variation can occur in a single nucleotide as a change of the nucleotide to another. In this case, the alteration of a genotype is called a point variant or mutation. This genetic variant can be recognised as a single nucleotide polymorphism (SNP) if it presents at least 1% or a single nucleotide variant (SNV) if its frequency is less than 1% in a population (He et al., 2014). Generally, based on the frequency of a genetic variant in a population, it can be divided into three classes. They are a common variant when its frequency is greater than 5%, a low-frequency variant when its frequency is between 1% and 5% and rare variants when its frequency is less than 1%. A population determines the frequency of variants, and this means that a variant can be common in a particular population, but it might be rare in

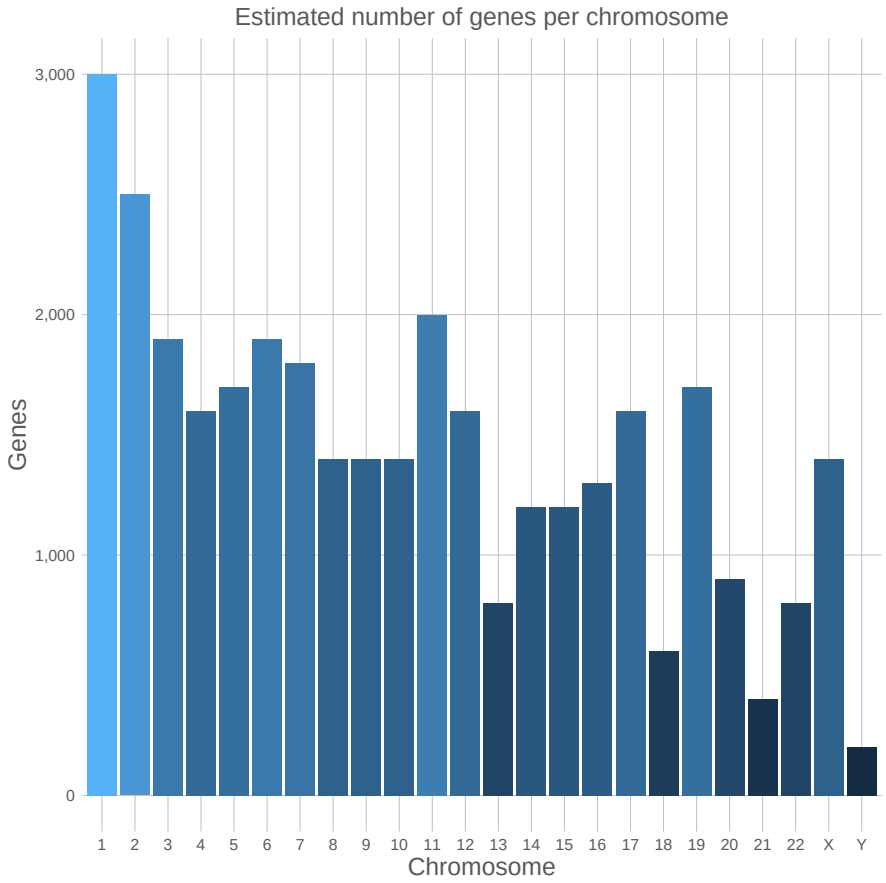

Figure 1.6: Estimated number of genes in every chromosome in the human genome. Numbers are obtained from (Richards and Hawley, 2011).

another population.

In addition to a point mutation, a genetic variation may appear as an insertion or deletion of a DNA sequence. This kind of variation is named insertion/deletion polymorphism (indel). As displayed in Figure 1.7, unlike a point mutation which is a substitution that replaces one of the nucleotides without changing the number in a DNA sequence, an indel can insert some nucleotides into a DNA sequence or delete nucleotides from it. Lastly, another possible type of genetic variation is

known as a structural variant. This form includes several genetic modifications, such as a copy number variation(CNV), duplication, and translocation.

A gene can have one or more genetic variations, and different forms of a gene are called alleles. It is possible to have different alleles for the same gene, and they contribute to various observed phenotypes, such as eye colour, hair colour and skin colour.

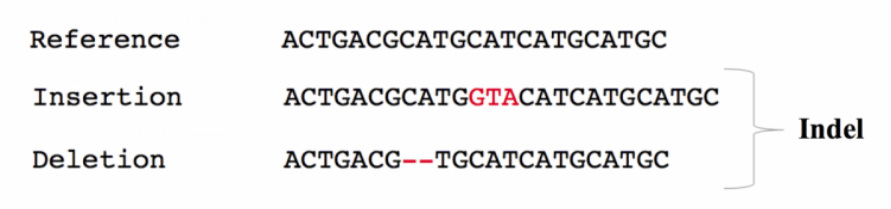

Figure 1.7: An example of a genetic variation called an insertion/deletion polymorphism (indel). (EMBL-EBI)

1.2 The genetics of cancer

Cancer is often a severe and fatal disease. Because it has become a common disease, the conducted studies in cancer research increase every year. In the UK, just fewer than 160,000 people died from cancer in 2011, with over 330,000 new cases reported every year (Caulfield et al., 2017). Cancer is deemed a genetic disease, and mutations in the DNA sequences are believed to be a possible cause of cancer. Several factors have been discovered, and they are thought to be the cause of the DNA mutations such as tobacco, age, viruses and bacteria, Ultraviolet (UV) radiation, organic and inorganic chemicals and family history (Parsa, 2012). Therefore,

studying the DNA mutations is a prime interest to expand the knowledge of cancer causes and treatments.

1.2.1 Types of genetic mutations in cancer

An average number of sequencing reads at each nucleotide on the genome is defined as the read-depth or coverage. In order to call or detect a genetic variant at a certain degree of confidence, it requires a sufficient read-depth rate. In other words, the higher the read-depth value, the more certainty specialists can have in discovering a genetic variant. For the whole genome sequencing, the typical value of read-depth can be 20x to 40x (from 20 to 40 times) (Liu et al., 2018). The read-depth value can be computed from the length of the genome (G), the number of reads (N), and the average read length (L) as $N \times L/G$. For instance, if a genome size is 100 Mbp (100 million base pairs) and 20 M (20 million) reads have been sequenced of 100 bp (100 base pairs) size, then the read-depth value at genome level would be 20x. A related term for the read-depth that should be considered in the genetic variant discovery method is the alternative reads number. The alternative reads number is the total number of reads covering the alternative genotype alleles in the same genomic position.

In cancer studies, matching normal and tumour DNA sequences, which have been taken from the same patient, is a reliable approach to identify the differences between them. The detection process of variants leads to exploring several models of genetic mutations. There are multiple types of genetic mutations. They are described as acquired mutations and inherited mutations. Acquired mutations are

defined as when a causal factor makes a change in the DNA sequences, so genetic mutations occur in a tumour cell but not in a normal cell. These acquired mutations are called somatic mutations. Inherited mutations happen when genetic mutations occur in both the healthy and diseased cells. In this case, these mutations are called germline mutations. There is a scenario that is not impossible, but its possibility for rising is rare (Roth et al., 2012). It is that when genetic mutations happen in a normal cell but not in a tumour cell. In this situation, it is believed that these mutations may not be related to cancer outcomes. Therefore, they are probably considered as an error of the machine when mutations are being detected.

In cancer aetiology, as cancer is considered a complex disease and is accompanied by other related traits, it is a good idea to study the association between genetic mutations and cancer subtype outcomes (He et al., 2018). For instance, in the diagnosis of liver cancer, α-fetoprotein (AFP) and prothrombin time are regularly evaluated as they can relate to liver cancer, so measuring them is necessary for the disease diagnosis (Lai et al., 1995). Furthermore, it is believed that all types of cancer have subtypes (Liu et al., 2021).

A genetic mutation in cancer can be found as a point mutation, described as an exchange in a single nucleotide. Alternatively, it can happen as other types of variations, such as indels, duplications and copy number variations. The type of duplication can appear when one or more nucleotides are copied and repeated. A copy number variation refers to abnormalities among individuals in the number of a particular gene for a specific trait in a genome. The most common genetic

mutation that scientists focus on to investigate the cause of cancer and examine the relationship between mutations and associated traits of cancer is a point mutation (Liu et al., 2021).

This thesis concentrates on a point mutation and considers this type of genetic mutation to be of interest. Moreover, we focus on somatic mutations rather than germline mutations, as somatic mutations are thought to play the most important role in the development of cancer (Luzzatto, 2011).

1.2.1.1 Somatic mutations

Somatic mutations, unlike germline mutations, can not be transmitted by parents to offspring as they happen in bodily tissues and are produced by physical factors. Figure 1.8 explains the difference between somatic mutations and germline mutations. It shows that a germline mutation happens in a sperm cell or an egg cell, and as the embryo grows, the mutation can be copied into every cell in the body. On the other hand, an embryo can be affected by a somatic mutation due to an environmental factor, but it will not spread to the whole body when the embryo grows. Instead, the mutation will remain in a specific area in the body.

On the subject of cancer, exploring somatic mutations and studying their impact are deemed a significant action in the studies of cancer treatment for the reason that somatic mutations are speculated to be a substantial stimulus for cancer (Liu et al., 2018). However, the principal challenge in investigating somatic mutations is the low frequency of them (He et al., 2018).

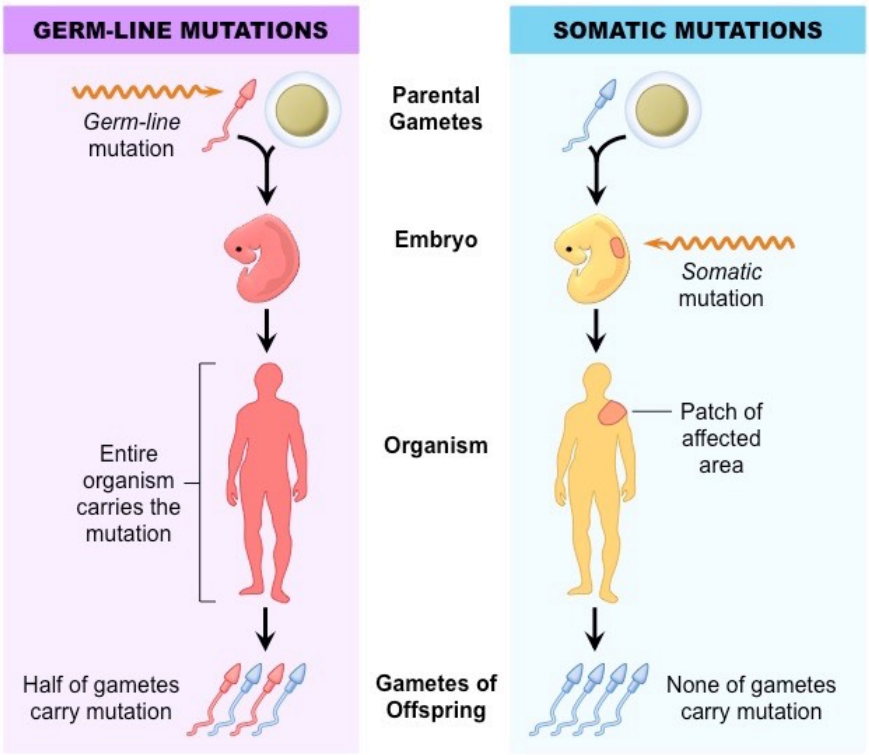

Figure 1.8: The difference between somatic mutations and germline mutations. (BioNinja Website)

1.3 The 100,000 Genomes Project

The 100,000 Genomes Project (100kGP) (Caulfield et al., 2017) is a government initiative established in order to sequence whole genomes from National Health Service (NHS) patients in England, Wales, Scotland and Northern Ireland. The project aims to study genomes of patients with rare diseases and their blood relatives and patients with cancer seeking to help scientists and doctors understand the underlying causes of diseases and provide them with new insights into the

predictions and prevention of diseases. Also, it aims to help patients by creating a new genomic medicine service, as the project pursues to track patients' conditions, drive the development of new drugs and find new uses for existing drugs.

The study of rare diseases, defined as diseases that affect fewer than one in $2,000$ people in a population, has become one of the project's focuses due to the considerable number of rare genetic diseases in the UK. Precisely, it is estimated that there are around $7,000$ different conditions of rare diseases in the UK, which means that there are around 3 million affected people (Taylor and Frankl, 2012). For each participant in the project, two family members are sequenced to assist in detecting disease causal mutations. It means that more than $50,000$ individuals are included in the branch of the rare disease. More than 120 rare diseases are considered in the 100,000 Genomes Project based on the need for better clinical therapies (Caulfield et al., 2017).

Concerning cancer, the genomic project applies the method of matching the tumour and normal sequencing data. There are $50,000$ sequenced genomes from $25,000$ independent individuals. Two genomes are sequenced from each patient; one sample is from tumour cells, while the second sample is from normal cells. Comparing cancerous and healthy cells can identify genetic regions likely to be a cause of cancer development. The project aims to investigate all types of cancer.

1.4 Discussion

This chapter discussed preliminary information to assist in understanding cancer genetics. It introduced a summary of human genetics, including genetic notions and terminologies and genetic variations. Next, as the focus of this thesis is on cancer data, an introduction to the genetics of cancer and types of cancer genetic variants is exhibited. The chapter concluded by offering a short sight of the 100,000 Genomes Project as its data is of interest.

The next chapter concerns a number of primary statistical methods employed in this research.

Chapter 2

Statistical principles

This chapter is constituted of some critical, fundamental statistical techniques that are utilised in the thesis. It starts with illustrating the logistic regression model as in this research; we concentrate on dichotomous responses. Then, the chapter moves on to introduce matched pairs analysis as this study design is used in the research for analysing two matched pairs of the genome for detecting an association. The logistic regression model of matched data and binomial exact test are introduced in the framework of matched pairs. By the same token, the chapter continues to present some of the genetic association procedures utilised for categorical outcomes. Introduced methods are the chi-square test, Cochran-Armitage Trend Test (CATT) and Wilcoxon rank-sum test. For large samples, three tests are presented. They are the likelihood ratio test, the Wald test and the score test. Conducting multiple hypotheses increases the chance of type I error; therefore, the Bonferroni correction procedure adopted in many places in this thesis and GHC test used in somatic mutation association analysis are also introduced in this

chapter. Finally, the chapter ends with brief coverage of Bayesian statistics and machine learning methods that have been adopted for somatic mutation analysis.

2.1 Generalised linear models

The logistic regression model is presented in this section within the context of a generalised linear model. The generalised linear model (GLM) is a generalisation of a linear model (LM), and it contains three elements which are a response variable, predictor variables and a link function (Nelder and Wedderburn, 1972). In the GLM, response values $Y_i, i = 1, \ldots, n$, for a sample size n, are deemed independent, and can be indicated as a binary outcome. In other words, Y_i can be coded 0 or 1 depending on whether or not an individual carries a trait of interest. In this case, Y_i follows a Bernoulli distribution.

The second element that the GLM includes is the set of predictors or explanatory variables. A predictor variable is a factor that can influence the response outcome. An example of a predictor is a genetic variant. Lastly, the third component in the GLM is a link function, which expresses the relationship between the expected value or mean of the response variable and explanatory variables in the linear predictor, and can be a nonlinear function.

2.1.1 Logistic regression model

Logistic regression is a predictive analysis approach adopted in many fields, including medical and biological fields and machine learning techniques, to interpret the

effect of predictor variables on a binary outcome. The effect is expressed in terms of the odds, defined as the ratio between the probability of the occurrence the event and the probability that the event will not occur. If P denotes the probability of an event occurring, the probability of not occurring is $(1 - P)$. Consequently, the odds can be expressed as

$$\text{Odds} = \frac{P}{1-P}.$$

The link function for the logistic regression model is the logit link. Then, the log of the odds is modelled with a linear combination of explanatory variables, each of which can be a discrete or continuous variable.

Consider a study with n independent individuals and a set of j variants gathered within a particular gene of the ith individual indicated as $\mathbf{z}_i = (z_{i1}, \ldots, z_{ij}), i = 1, \ldots, n$, and let the binary response variable Y_i denote outcome of a trait of interest. The outcome Y_i can be modelled by a generalised linear model as

$$\text{logit}P(Y_i = 1) = \beta_0 + \boldsymbol{\beta}\mathbf{z}_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where β_0 is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j)'$ are the regression coefficients of the genotypes. The association test is regarding $\boldsymbol{\beta}$, so the hypothesis test is $H_0 : \boldsymbol{\beta} = 0$ that is $\beta_1 = \beta_2 = \cdots = \beta_j = 0$.

2.2 Matched pairs analysis

The study of matched pairs analysis is an observational study employed to discover factors related to outcomes. In the case-control study design, it is most generally performed as a 1 – 1 design, where a case, defined as an individual who has the development of a particular disease, is matched or compared to a control, defined as an individual who is similar to the case sample but does not carry the disease outcome. This is probably due to its practicality, since a 1:1 matched design may not have the optimal power. However, the matched pairs case-control study is to conduct matching pairs analysis where the case and control samples are obtained from the same individual. This study design is introduced in this section as we utilise it in Chapter Four when proposing a novel association method based on comparing dependant pairs in order to evaluate the effect of somatic mutations that occur in only tumour cells. Presented methods in this section are logistic regression analysis of matched data and the binomial exact test.

2.2.1 Logistic regression analysis of matched data

Conditional logistic regression, originated by Breslow et al. (1978), is a development of logistic regression that allows taking the stratification and matching into consideration. It is mainly used in epidemiology and is deemed the most adopted method in matched data as it has the feature of controlling for covariates.

Suppose a study where two genetic sequences are sequenced from each individual.

One set is taken from tumour cells, while the second is from normal cells. Denote the ith patient's tumour cells by $y_{1i} = 1$ and the healthy cells by $y_{2i} = 0$. The conditional likelihood function for the ith individual is given by

$$\begin{aligned}
&P(Y_{1i} = 1, Y_{2i} = 0 \mid Z_{1i}, Z_{2i}, Y_{1i} + Y_{2i} = 1) \\
=&\frac{P(Y_{1i} = 1 \mid Z_{1i})P(Y_{2i} = 0 \mid Z_{2i})}{P(Y_{1i} = 1 \mid Z_{1i})P(Y_{2i} = 0 \mid Z_{2i}) + P(Y_{1i} = 0 \mid Z_{1i})P(Y_{2i} = 1 \mid Z_{2i})} \\
=&\frac{\left[\frac{e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{1i}}}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{1i}}} \times \frac{1}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{2i}}}\right]}{\left[\frac{e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{1i}}}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{1i}}} \times \frac{1}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{2i}}}\right] + \left[\frac{1}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{1i}}} \times \frac{e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{2i}}}{1+e^{\beta_0+\boldsymbol{\beta}\mathbf{z}_{2i}}}\right]} \\
=&\frac{e^{\boldsymbol{\beta}\mathbf{z}_{1i}}}{e^{\boldsymbol{\beta}\mathbf{z}_{1i}} + e^{\boldsymbol{\beta}\mathbf{z}_{2i}}},
\end{aligned}$$

where $\mathbf{z}_{1i} = (z_{1i1}, \ldots, z_{1ij})$ is a vector of the ith patient's genotypes for the j mutations within a gene of tumour cells, and $\mathbf{z}_{2i} = (z_{2i1}, \ldots, z_{2ij})$ is a vector of the ith patient's genotypes for the j mutations within a gene of normal cells. Finally, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_j)'$ is a vector of the genotypes coefficients. The association test is regarding $\boldsymbol{\beta}$, so the hypothesis test is $H_0 : \boldsymbol{\beta} = 0$ that is $\beta_1 = \beta_2 = \cdots = \beta_j = 0$.

2.2.2 Exact tests for binary responses

Several exact tests can be applied to dichotomous outcomes. The binomial exact test is a type of these tests that is used for rate comparison. It is demonstrated here within the context of study comparing normal and tumour cells because we utilise it in Chapter Four in order to estimate the validity of our proposed score test.

2.2.2.1 Binomial exact test

Here the binomial exact test is introduced within the context of a comparison of normal and tumour samples within an individual. Consider that z_{1ix} is the genotype of the ith patient for the xth genetic marker within a gene of tumour cells, and z_{2ix} is the genotype of the ith patient for the xth genetic marker within a gene of normal cells. Given that the genetic mutation is observed in the xth marker, if there is no genetic effect on the disease outcome, the probability that the variant occurs in the tumour cells and does not occur in the disease-free cells is 0.5.

		Normal cells	
		A mutation present	A mutation not present
Tumour cells	A mutation present	N_{11}	N_{10}
	A mutation not present	N_{01}	N_{00}

For the xth mutation, let $N = N_{10} + N_{01}$ be the discordant pairs where N_{10} is the total number of individuals with the variant in the tumour cells and N_{01} is the total number of individuals with the variant in the disease-free cells. Assume N is fixed,

$$N_{10} \sim b(N, \omega),$$

where ω is the conditional probability that the genetic variant occurs in the diseased cells and does not occur in the normal cells, given that the genetic variant is observed in either the tumour or healthy cells. The exact p-value for the test of

no genetic effect at the xth mutation is

$$2pr\left(N_{10} \geqslant n_{10} | \text{no genetic effect}\right) = 2\sum_{k=n_{10}}^{N} \binom{N}{k} \times (0.5)^k \times (0.5)^{N-k},$$

where $n_{10} \geqslant N/2$ is the observed number of individuals with the variant in the diseased cells but not in healthy cells.

2.3 Association procedures of genetic studies for categorical outcomes

In association analysis of genetic studies, many statistical approaches aimed at investigating the impact of genetic variation on categorical phenotype responses can be applied. This section introduces three association methods used in the genomewide association study to compare the frequency of genotypic alleles at a specific locus, usually single-nucleotide polymorphisms (SNPs). Presented methods are the chi-square test, CATT and Wilcoxon rank-sum test.

2.3.1 Chi-square test

The chi-square test (Pearson, 1900) is a hypothesis testing procedure performed to evaluate the relationship between two categorical variables. On the subject of genotypic association analysis, the chi-square test can be used to determine whether or not the presence of a mutation within a genetic marker is associated

with a particular disease outcome by comparing the observed distribution to the expected distribution if there is no association.

Consider a case-control study, and we are interested in detecting if a genetic variant at a specific locus is related to a binary disease outcome. Let z_x be the genotype of the xth genetic marker where z_x follows a genetic additive model and can be $0, 1$ or 2. The contingency table in this study design can be given as

	$z_x = 0$	$z_x = 1$	$z_x = 2$	Sum
Diseased individuals	N_{d0}	N_{d1}	N_{d2}	R_1
Healthy individuals	N_{h0}	N_{h1}	N_{h2}	R_2
Sum	C_1	C_2	C_3	N

where N_{d0} is the total number of diseased individuals who do not carry a variant in the xth genetic marker, N_{d1} is the total number of diseased individuals who have one copy of variant in the xth marker and lastly N_{d2} is the total number of diseased individuals who have two copies of variant in the xth marker. In the same manner, N_{h0} indicates the total number of healthy individuals who do not carry a variant in the xth marker, N_{h1} is the total number of healthy individuals who have one copy of variant in the xth marker and N_{h2} is the total number of healthy individuals who have two copies of variant in the xth genetic marker. The test statistic is given by

$$\chi^2 = \sum \frac{(\text{observed values - expected values})^2}{\text{expected values}}.$$

The chi-square test statistic follows an asymptotic chi-square distribution with (R – 1)(C – 1) degrees of freedom.

2.3.2 *Cochran-Armitage Trend Test (CATT)*

The second procedure adopted in categorical data analysis is CATT (Armitage, 1955). Similarly to the chi-square test, the CATT method aims to measure the impact of a genetic mutation on a binary disease outcome. However, this approach is explicitly used when there is an ordinal predictor variable with more than two categories. By applying the 2×3 above contingency table, the trend test statistic T is given by

$$T = \sum_{k=1}^{3} t_k(N_{dk}R_2 - N_{hk}R_1),$$

where t_k is the genotype weight of the kth category, and weights are selected by a user depending on the genetic model. Since an additive genetic model is considered here, $t_k = 0, 1, 2$ for $k = 1, 2, 3$, respectively, as there are three categories. The standardised test is $T/\sqrt{Var(T)} \sim N(0, 1)$ where

$$Var(T) = \frac{R_1R_2}{N}\Big(\sum_{k=1}^{3} t_k^2 C_k(N - C_k) - 2\sum_{k=1}^{2}\sum_{l=k+1}^{3} t_k t_l C_k C_l\Big).$$

2.3.3 *Wilcoxon rank-sum test*

The Wilcoxon rank-sum test, also called Mann–Whitney U test (Mann and Whitney, 1947), is introduced here as some set-based approaches, discussed in Chapter Three, use this procedure in order to discover the relationship between a collapsing genetic score of a candidate region, such as a gene, and disease outcome. The Wilcoxon rank-sum test can be applied to assess whether there is a difference in

a predictor variable for two independent samples. In the context of detecting the genetic association, it can be used to test whether there is a difference in a genetic variant within a particular marker for diseased and healthy individuals. The Wilcoxon rank-sum test statistic U is the smaller value of the statistics of the diseased individuals sample U_d and the healthy individuals sample U_h where

$$U_d = Rank\ Sum_d - \frac{n_d(n_d+1)}{2},$$

and

$$U_h = Rank\ Sum_h - \frac{n_h(n_h+1)}{2},$$

where $Rank\ Sum_d$ is the sum of the ranks in the diseased individuals' sample, $Rank\ Sum_h$ is the sum of the ranks in the healthy individuals' sample and n_d and n_h are the sample sizes of the diseased and healthy individuals, respectively.

For large sample sizes (sample sizes above 20), U is approximately normally distributed, and the standardised value is given by

$$z = \frac{U - \frac{n_d n_h}{2}}{\sqrt{\frac{n_d n_h (n_d+n_h+1)}{12}}}.$$

2.4 Large sample tests

When the sample size is very large, particular statistical hypothesis tests are used to test the null hypothesis H_0. In this section, three important statistical tests, which are applied through the thesis, are discussed. The tests considered are the

likelihood ratio test, the Wald test and the score test.

2.4.1 Likelihood ratio test

The likelihood ratio test is a tool to assess the goodness of fit of two competing statistical models. The test is based on the ratio of their likelihoods. Let $Y_1, \ldots, Y_n$ be independent identical distribution variables with a probability density function $f(y;\theta)$ where the scalar parameter θ is unknown. The null hypothesis test is $H_0 : \theta = \theta_0$ while the alternative hypothesis is $H_1 : \theta \neq \theta_0$. The likelihood ratio test statistic is given as

$$LR = -2[\ell(\theta_0) - \ell(\hat{\theta})], \tag{2.2}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ under the alternative hypothesis, and θ_0 is the value of θ under the null hypothesis. The distribution of the likelihood ratio test statistic is asymptotically chi-square distribution with 1 degree of freedom.

2.4.2 Wald test

The Wald test statistic is defined as

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})}, \tag{2.3}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ under the alternative hypothesis, θ_0 is the value of θ under the null hypothesis, and $var(\hat{\theta})$ is the variance of $\hat{\theta}$.

The variance can be calculated from the inference of Fisher Information. Under the null hypothesis, the Wald test statistic asymptotically follows the chi-square distribution with 1 degree of freedom.

2.4.3 Score test

The score test, also called the Lagrange multiplier test (LM test), is a procure that does not require an estimate of parameters under the alternative hypothesis; therefore, the score test has this advantage compared to the likelihood ratio test and Wald test. The test is based on the score function which is given by

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta},$$

and the observed Fisher Information which is expressed as

$$I(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}.$$

Finally the score test statistic is given by

$$V = \frac{U(\theta_0)^2}{I(\theta_0)}. \tag{2.4}$$

Under the null hypothesis, the distribution of the score test statistic is asymptotically chi-square distribution with 1 degree of freedom.

2.5 Testing multiple hypothesis

In this section, we discuss the ways of managing the multiple hypothesis tests. The well-known Bonferroni correction procedure is introduced due to its popular usage when many statistical tests are conducted. The Bonferroni adjustment is utilised in many places in the thesis. In addition to the Bonferroni method, the GHC test (Barnett et al., 2017) is demonstrated because it is adopted in our proposed technique for studying gene-based somatic mutation association analysis explained in Chapter Four.

2.5.1 The Bonferroni correction

In the association studies of the whole genome, many studies are conducted by utilising single-variant tests. However, when there are multiple hypotheses tested, the chance of the incorrect decision of rejecting the null hypothesis increases (Mittelhammer et al., 2000). For adjusting the multiple testing burden, an adjustment procedure is applied. Therefore, determining the correct p-value threshold for the significance is crucial, and there are several techniques used in GWAS. They include the Bonferroni correction, Sidak correction and permutation-based approaches. Permutation-based methods that are based on resampling from observed data are computationally intensive. The most conservative method for selecting a threshold for the statistical significance is the Bonferroni correction procedure (Kaler and Purcell, 2019).

The Bonferroni correction approach divides the significance level at each locus by the number of tests. In other words, it tests each hypothesis at a significance level of α/j, where α is the selected significance level in a genetic marker, and j is the number of genetic markers within a set such as a whole genome. For example, for a significance level set at 0.05, and $100,000$ SNPs were tested, this would yield an $\alpha = 5.0 \times 10^{-7}$. Correspondingly to the Bonferroni correction method, Sidak correction, where $\alpha = 1 - (1 - 0.05)^{1/j}$, gives a close result to the Bonferroni correction when j is large (Gao et al., 2008).

Even though both the Bonferroni and Sidak procedures preserve the rate of false-positive discoveries, they are inappropriate for some GWAS studies (Ziegler et al., 2010). That is believed due to the fact that some genetic markers are thought to be related. However, the Bonferroni and Sidak corrections do not account for the correlation between some SNPs and deal with them independently.

2.5.2 The generalised higher criticism test

This part introduces a recently proposed test procedure called the GHC test (Barnett et al., 2017). It computes the test statistic for each of the individual variables, called marginal test statistics, and combines information over these statistics, allowing for the correlation between these marginal test statistics. The GHC test is a generalisation of the higher criticism (HC) test (Donoho et al., 2004) which assumes that the marginal test statistics are independent.

The higher criticism test statistic

Let V_x be the test statistic for the xth genetic marker, where $x = 1, 2, \ldots, j$. Assuming independence, and under the null of no genetic association, the number of marginal tests significant at a selected significance level α follows a binomial distribution with mean $j\alpha$ and variance $j\alpha \times (1 - \alpha)$. The HC statistic is defined as

$$HC = \sup_{\alpha \geqslant 0} \left\{ \frac{[(\text{observed number of tests significant at } \alpha) - j\alpha]}{\sqrt{j\alpha \times (1 - \alpha)}} \right\}, \qquad (2.5)$$

and the test statistic asymptotically follows a Gumbel distribution (Barnett and Lin, 2014).

In recognition of the fact that the HC test has restricted applications in genetic association studies because variants within a genetic structure such as a gene are possible to be correlated, a new statistical approach has been developed by Hall et al. (2010), called the innovated higher criticism (iHC) test. The iHC procedure (Hall et al., 2010) was proposed in order to deal with the structure of the arbitrary correlation among the individual marginal tests. The idea was to transform the correlated test statistics V to independent test statistics V^* using Cholesky decomposition of the correlation matrix and then apply the higher criticism method directly.

The innovated higher criticism statistic

Let $V^* = (V_1^*, \ldots, V_j^*)$ be a transformed vector of $V = (V_1, \ldots, V_j)$ given by

$$V^* = \Sigma^{\frac{1}{2}} V,$$

where Σ is the variance-covariance matrix for $(V_1, \ldots, V_j)$. For example, if the V_x's, $x = 1, \ldots, j$, are score test statistics then Σ is the expected Fisher Information matrix. Assume V_x follows a standard normal distribution and for some given t, let

$$R^*(t) = \sum_{x=1}^{j} I_{(|V_x^*| \geqslant t)},$$

be the number of tests for which $|V_x^*| \geqslant t$. Because the V_x^*'s are independent, under the null hypothesis H_0, it follows that $R^*(t)$ is a binomial random variable with probability of success $2\Phi^c(t)$, where $\Phi^c(t) = 1 - \Phi(t)$ is the distribution function of the standard normal random variable. It follows that $E\left[R^*(t)\right] = 2j\Phi^c(t)$, $Var\left[R^*(t)\right] = 2j\Phi^c(t)(1 - 2\Phi^c(t))$ and the standardised value of $R^*(t)$ is

$$\frac{R^*(t) - 2j\Phi^c(t)}{\sqrt{2j\Phi^c(t)(1 - 2\Phi^c(t))}},$$

for a given value of t. Following the HC test, the innovated higher criticism test statistic is defined as

$$iHC = \sup_{t \geqslant 0} \left\{ \frac{R^*(t) - 2j\Phi^c(t)}{\sqrt{2j\Phi^c(t)(1 - 2\Phi^c(t))}} \right\}. \tag{2.6}$$

The iHC approach asymptotically follows a Gumbel distribution (Barnett and Lin, 2014).

When genetic variants located in a genetic unit are correlated or even moderately correlated, the iHC test can commit a significant loss of power that may happen due to diluting the sparse signals after being mixed in the process of the transformation (Barnett et al., 2017).

Due to the possibility of the correlation among the mutations within a genetic set, such transformation in the association analysis is not optimal. This motivation leads to develop the GHC test (Barnett et al., 2017). The GHC test is an association test that uses single marker statistics and their correlation matrix to construct a new test statistic and its distribution. The GHC test was developed to liberate the limitations of applying the higher criticism test in the genetic association approached and use the original marginal test statistics V.

The GHC statistic

Suppose instead of using the transformed test statistics V_j^*, we define

$$R(t) = \sum_{x=1}^{j} I_{(|V_x| \geqslant t)}.$$

$R(t)$ is no longer binomial distributed because the correlation among genetic mutations will increase the variance. The generalised higher criticism test statistic is

defined as

$$GHC = \sup_{t \geqslant 0} \left\{ \frac{R(t) - 2j\Phi^c(t)}{\sqrt{\hat{\text{var}}(R(t))}} \right\}, \tag{2.7}$$

where $\hat{\text{var}}(R(t))$ is computed from the estimated correlation $\hat{\Sigma}$ between the test statistics V_x's, $x = 1, \ldots, j$. To do so, the covariance of $R(t_x)$ and $R(t_{x'})$ is shown to be (Barnett et al., 2017),

$$cov\left\{R(t_x), R(t_{x'})\right\} = j[2\Phi^c(max\left\{t_x, t_{x'}\right\}) - 4\Phi^c(t_x)\Phi^c(t_{x'})] + 4j(j-1)\phi(t_x)\phi(t_{x'}) \times \sum_{i=1}^{\infty} \frac{H_{2i-1}(t_x)H_{2i-1}(t_{x'})\overline{r^{2i}}}{(2i)!},$$

where $\phi(t)$ is the dispersion parameter evaluated at t, $H_i(t)$ is the Hermite polynomials and $\overline{r^{2i}}$ is given by this expression $\overline{r^n} = \frac{2}{j(j-1)}\Sigma_{1 \leq x \leq x' \leq j}(\Sigma_{xx'})^n$. The generalised higher criticism test statistic asymptotically follows a normal distribution (Barnett et al., 2017).

The GHC approach was implemented and evaluated by Barnett et al. (2017) in terms of type I error and power through simulation studies while considering correlation structures among SNPs. They showed that the GHC test controls the type I error at different significance levels. In contrast, the type I error of the original higher criticism is anticonservative. Regarding power, the GHC test was compared to the iHC, SKAT and MinP approaches. It was shown that the performance of the GHC test is better than iHC in all of the various correlation settings and sparsity situations. High sparsity in a genetic set, such as a gene, means that when the number of causal variants is less than $\sqrt{j}$, where j is the number

of variants in the set. Compared to SKAT, the GHC test performs better than SKAT, particularly in sparse situations with low correlation rates between causal and noncausal variants. The GHC's performance is similar to the MinP approach; However, the GHC test outperforms MinP when the correlation among SNPs is high, and sparsity is decreased (Barnett et al., 2017).

2.6 Bayesian statistics

In this section, a brief of Bayesian statistics principles is introduced. Concepts that are discussed here include conditional probability, the law of total probability and Bayes' theorem.

Conditional probability

Consider two events B and E with $P(E) > 0$. The conditional probability of B given E written as $P(B \mid E)$ is the probability that B occurred given that E has already occurred. Mathematically,

$$P(B \mid E) = \frac{P(B \cap E)}{P(E)},$$

if B and E are not independent, then the probability of both events occurring simultaneously is given as $P(B \cap E) = P(B \mid E)P(E)$, and it is called the multiplicative rule. However, if B and E are independent, the conditional probability $P(B \mid E) = P(B)$.

Law of total probability

A collection of disjoint events $B_1, \ldots, B_n$ where $P(B_i) > 0$, $i = 1, \ldots, n$ is said to be a partition of a sample space S if $\cup_{i=1}^{n} B_i = S$. If E is any event within S then

$$E = (E \cap B_1) \cup (E \cap B_2) \cup \ldots \cup (E \cap B_n).$$

Since events $E \cap B_i$, $i = 1, \ldots, n$ are disjoint,

$$P(E) = P(E \cap B_1) + P(E \cap B_2) + \ldots + P(E \cap B_n).$$

Applying the multiplicative rule gives

$$P(E) = P(E \mid B_1)P(B_1) + P(E \mid B_2)P(B_2) + \ldots + P(E \mid B_n)P(B_n),$$

which is called the law of total probability.

Bayes' theorem

For a partition $B_1, \ldots, B_n$ of S and any event E with $P(E) > 0$, the Bayes' theorem is given as

$$P(B_i \mid E) = \frac{P(E \mid B_i)P(B_i)}{\sum_{x=1}^{n} P(E \mid B_x)P(B_x)}, \quad i = 1, \ldots, n.$$

2.7 Machine learning methods

Statistical learning or machine learning methods are of two kinds, supervised learning and unsupervised learning approaches. Unsupervised learning techniques primarily focus on grouping or clustering observations with similar characteristics and discovering hidden patterns. Common methods include K-means clustering and hierarchical clustering. Supervised learning techniques focus on predicting one or more outcome variables based on independent or predictor variables. Two general kinds of supervised learning approaches are regression and classification for continuous and categorical outcomes, respectively. While many standard statistical modelling techniques (multiple linear regression, spline regression, logistic regression) are used in supervised learning, there are also methods such as Regression and Classification Trees, Support Vector Machines (discussed below) and Neural Networks which are not found in the typical statistical toolbox.

Tree-based methods can be designed to serve both regression and classification analysis. The foundation of all tree-based models is known as decision tree models, and two components construct a decision tree model. They are nodes and branches, whereas, at each node, one of the features (predictor variables) in the model is evaluated to divide the observations. In general, a decision tree model partitions the prediction space into disjoint regions and run a simple model on each region. Let $\mathbf{X} = (X_1, \ldots, X_j)$ be the set of j predictor variables, Y be an outcome and $\{R_1, \ldots, R_M\}$ be the partitions of the predictor space $\mathbf{X}$. A tree-based regression

model can be given as

$$f(X) = \sum_{m=1}^{M} c_m I\{X \in R_m\}, \tag{2.8}$$

where c_m is the predicted value of the response variable at the mth region R_m. For example, based on the minimum sum of squares criteria, c_m is simply the mean of Y in the region R_m. A greedy algorithm can be run to find a partition that minimises an appropriate objective function, such as the sum of squares.

Compared to other machine learning algorithms, decision tree models are considered simply manageable. Also, another advantage of decision tree models is that they can handle missing values (Friedman, 2017). However, in many practical cases, a single decision tree method is inaccurate, so it cannot produce a reliable prediction. Therefore, bagging, also known as bootstrap aggregation, is a technique employed to increase accuracy. It is a simple yet powerful idea that builds many decision tree models by randomly sampling with replacement, or bootstrapping, from the original dataset and taking the average of the estimated prediction function. The random forest technique (Breiman, 2001) is a powerful statistical learning approach that combines multiple decision tree models using a way similar to bagging. A regression or classification tree is built on a bootstrap sample using a randomly chosen subset of variables in random forest models. Let the number of bootstrap sample be B and T_b represent a tree based on the bth bootstrap sample, i.e., T_b follows the model in equation (2.8). In random forest methods, the

prediction at new data points x is made as

$$\frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Since a pair of trees in the random forest approach shares some data, the outcome of the pair of trees may be correlated. It is recognised that as the number of trees grow in the forest, the effect of the correlation on the precision of prediction decreases.

The Bayesian counterpart of random forest is the Bayesian additive regression trees (BART) (Chipman et al., 2010). While tree parameters are considered fixed in the random forest models technique, BART treats the parameters characterizing the ensemble of trees as random. A prior distribution is assigned to each parameter, and a backfitting MCMC is used. Regularization of the model through carefully chosen prior distribution is a distinctive feature of BART.

Another procedure of supervised machine learning methods that can also be utilised for classification and regression purposes is support vector machines (SVMs) (Boser et al., 1992). The idea of the SVMs algorithm is to find a hyperplane or line that is a function applied to separate the observations into classes. Maximising the margin distance of data points from the hyperplane increases the confidence that the data points are classified correctly. Support vectors are defined as the nearest points of the data to the hyperplane, and they are considered critical to determine the hyperplane. The SVMs technique works relatively well on small and clean datasets, and accuracy is one of its advantages. However, it becomes less efficient

when there are some noises in datasets (Ray, 2019).

There are supervised learning approaches that can be used only for classification analysis, such as logistic regression. On the other hand, simple linear regression, multiple linear regression and polynomial regression analysis are examples of algorithms employed for regression designs.

2.8 Discussion

The chapter began to present an introduction to the generalised linear models and discussed the logistic regression model considering that this thesis focuses on binary responses. Next presented was matched pairs analysis since this study design is used to compare two dependent sequences taken from the same person. Logistic regression analysis of matched data and the binomial exact test were discussed in this chapter within the context of comparing tumour and healthy cells to detect the impact of somatic mutations located in tumour cells. Three association genetic methods used for categorical data, including the chi-square test, CATT and Wilcoxon rank-sum test, were introduced as they are adopted in association analysis of genetic variants, discussed in the following chapter. Moreover, large sample tests, including the likelihood ratio test, the Wald test and the score test, were mentioned. The Bonferroni correction and GHC procedures used for adjusting multiple testing burden were presented in this chapter. Finally, the chapter ended by giving a short brief of Bayesian methods and machine learning techniques.

In the following chapter, some evaluative approaches for genetic association anal-

ysis are discussed. Furthermore, as this thesis focuses on somatic mutations, a detection analysis of somatic mutations is presented beside a newly released method for studying somatic mutations' effect.

Chapter 3

Association analysis of genetic mutations

Using advanced technologies (next-generation sequencing) assists in determining genetic mutations as complex diseases, including cancer, are triggered by genetic variations. In consequence, investigating the relationship between each of genetic variants and a disease trait outcome helps examine the degree of a variant's impact. Even though the GWAS successfully identified thousands of variants using single-variant association tests, it works only on common variants. Note also that GWAS findings have been used in fine-mapping, which involves rare variants, e.g., through target sequencing. Set-based association methods have been developed with the aim of dealing with low-frequency and rare variants.

Before applying genetic association approaches in cancer studies, it is pivotal to introduce some somatic mutation calling methods as we in this thesis are interested in this type of mutation. Somatic variant calling methods are proposed owing to

the fact that low-frequency mutations are not easy to identify.

In this chapter, we first discuss the association analysis of single-variant methods. A simulation study is conducted to evaluate and compare a selection of single association tests. Then, we explore a number of standard statistical techniques applied to test the relationship between low-frequency and rare mutations grouped within a genetic unit and a trait outcome. Next, a comparative simulation analysis of rare variant association tests is offered to appraise their performances in terms of type 1 error and power. Later on, we broach the ideology of finding somatic mutations and introduce various approaches adopted in the somatic mutation calling procedure. Lastly, a very recent association approach that has been constructed for testing the effect of somatic mutations on a cancer subtype outcome is included at the end of this chapter.

3.1 Association testing for genetic variants

Two primary forms of tests can be applied in genetic association analysis. The first type is to test the probable association of a single mutation with a disease outcome. This type of tests is called a single-variant association test. Another form of association tests is named set-based association tests. This kind of test is proposed for testing the relationship between an entire genetic set of mutations, for example, a gene or pathway, and a disease outcome.

3.1.1 Data description and model

Before discussing the two main categories of genetic variant association tests, we introduce a model used to illustrate the association testing analysis. Consider a study with n independent individuals and a set of j variants grouped within a gene of the ith individual indicated as $\mathbf{z}_i = (z_{i1}, \ldots, z_{ij}), i = 1, \ldots, n$. We assume an additive genetic model so that $z_{ix} = 0, 1$ or 2 where $x = 1, \ldots, j$ represents the number of genetic alleles for the xth variant of the ith individual. We suppose that the model does not include non-genetic covariates. The variable Y_i denotes a binary outcome of a disease. The outcome Y_i can be modelled by a generalised linear model as

$$\text{logit}P(Y_i = 1) = \beta_0 + \boldsymbol{\beta}\mathbf{z}_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where β_0 is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j)'$ are the regression coefficients of the genotypes. The association test is regarding $\boldsymbol{\beta}$, so the hypothesis test is $H_0 : \boldsymbol{\beta} = 0$ that is $\beta_1 = \beta_2 = \cdots = \beta_j = 0$.

3.1.2 Single-variant association tests

Every SNP marker in for example GWAS approaches is tested through a single-marker association marginal test. A linear model can be implemented for continuous outcomes and logistic regression for binary outcomes. Although thousands of disease-associated SNPs have been identified in the way of the GWAS, the

approach only works well for common variants, defined as alleles that have a frequency of more than 5% (Hindorff et al., 2009). Additional examples of methods that can be applied in the GWAS to evaluate the relationship between a single variant and disease outcome comprise the chi-square test and CATT.

We appraise in the following section the performances of some standard single-variant association tests that can be utilised in the additive genetic model to determine the relationship between a single genetic mutation and binary disease outcome.

3.1.2.1 *Simulation studies and results*

A simulation study was created in order to examine the performance of the chi-square test, CATT and GLM in terms of type I error and power. The assessment of the methods was composed of $1,000$ replications at significance level 0.05 with various rates of variant allele frequency. In terms of evaluating type I error, it was done by simulating the data under the null hypothesis (β=0). In terms of evaluating the power, we set $\beta = 0.4, 0.8, 1.2, 1.6, 2.0$.

A dataset of a sample size of $n = 400$ was conducted, and the genotype for a genetic variant of the ith individual z_i was simulated with a pre-selected variant allele frequency ($VAF = 0.1, 0.05, 0.01$). A dichotomous outcome Y_i was simulated by a Bernoulli distribution with probability of success p_i, and $\log(p_i/(1-p_i)) = -0.5 + \beta z_i$.

3.1.2.1.1 Type I error and power

Based on the investigation, the chi-square test, CATT and GLM control the type I error. In terms of power, all methods can obtain high rates of power at a genetic variant frequency VAF ≥ 0.05 and large effect size. In other words, the tests produce more than 99% and 93% powers when a genetic variant has a frequency of VAF =0.1 and VAF =0.05, respectively, for effect size $\beta \geq 1.2$ as shown in Figure 3.1. However, the performance of the single-variant association tests decreases for a low frequency (VAF =0.01) even when the effect size is very large ($\beta = 2$). For instance, the chi-square test and CATT approach achieve 71% power while the GLM obtains 43% power for VAF=0.01 and $\beta = 2$. For this sample size ($n = 400$), this finding can confirm that single-variant tests are powerful when a genetic variant is deemed common (e.g., its frequency is ≥ 0.05). However, their performance decreases for analysing low-frequency and rare variants (variants have a frequency of < 0.05).

3.1.2.2 Conclusion

In summation, single-association tests can be robust in genetic studies when a genetic variant has a standard frequency rate. Based on cancer mutation analysis, single variant association approaches can be employed to examine genetic associations of germline mutations as their recurrence is frequent. Since it is widely known that somatic mutations are considered low-frequency and rare, these developed methods might be inapplicable for the purpose of exposing somatic mutations

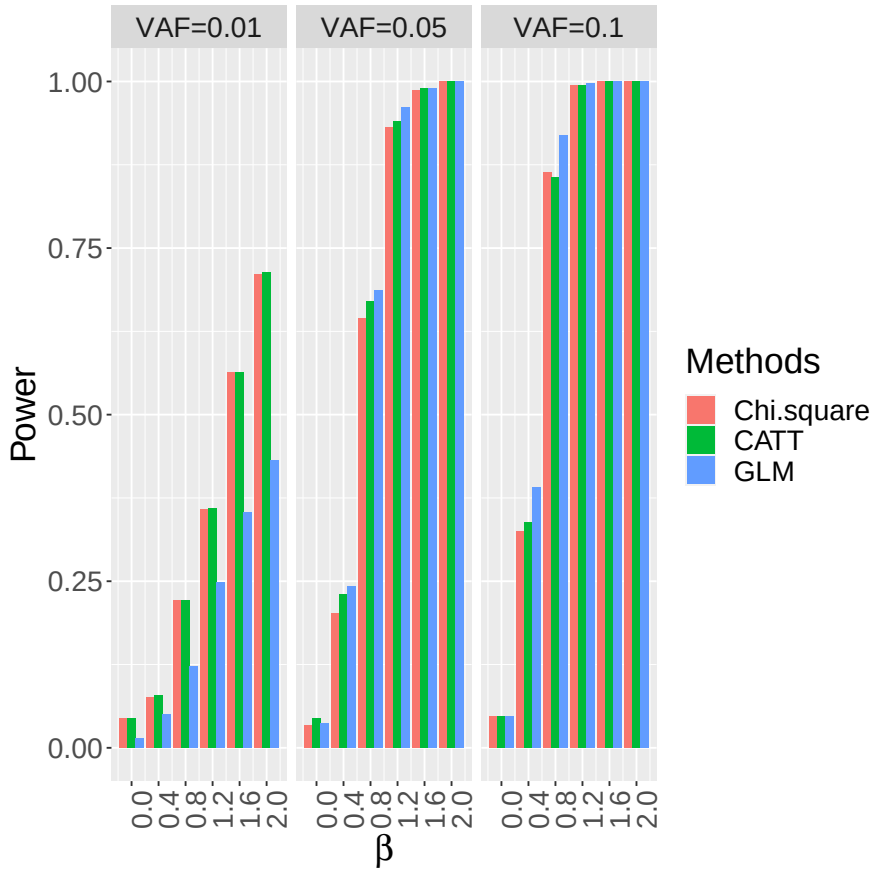

Figure 3.1: Comparison of type I error and power for single-mutation analysis of the chi-square test (red bars), Cochran-Armitage Trend Test (CATT) (green bars) and GLM (blue bars) with various rates of variant allele frequency (VAF) and effect size β. The comparison is based on a sample size of ($n = 400$).

association to a cancer subtype outcome. A new strategy has been suggested in order to study the impact of low-frequency mutations on disease outcomes. It is to aggregate the genetic effects of the variants within a genetic unit such as a gene to enhance genetic association signals and avoid penalties that may occur when multiple tests are made.

3.1.3 Set-based association tests

Set-based, also called region-based association tests, have become more desirable in genetic studies to inspect the effect of rare and low-frequency mutations. This is believed due to the fact that considering every single low-frequency mutation may produce a deficiency in statistical association approaches as there will be a dearth of variation in data (Sugasawa et al., 2017). Set-based association tests combine the complete genetic information from multiple mutations grouped in a genetic set such as a gene to investigate the relationship between this genetic set and trait outcomes. The most common gene-based association tests follow two classes of tests. The first type is burden tests, while the second type is variance-component tests. In addition, there are some set-based association tests that use a combination of burden tests and variance-component tests. Table 3.1 compiles a number of common association tests.

3.1.3.1 Burden tests

The purpose of this sort of test is to epitomise the whole genetic information from the j genetic markers within a target gene or any genomic unit for each

Set-based association methods		
Type	Description	Tests
Burden tests	collapse rare variants into genetic scores	CAST, CMC test, WSS, Mz test, VT test
Variance-component tests	test variance of genetic effects	SSU test, SKAT, C-alpha test
Combined tests	combine burden tests and variance-component tests	SKAT-O, Fisher method, MiST

Table 3.1: A summary of some set-based association tests.

individual into one genetic score in order to use this burden score for the association procedure. Burden tests include the cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007), combined and multivariate collapsing (CMC) test (Li and Leal, 2008), weighted-sum statistic (WSS) (Madsen and Browning, 2009), Morris and Zeggini (MZ) test (Morris and Zeggini, 2010), variable allele-frequency threshold (VT) test (Price et al., 2010).

3.1.3.1.1 *Cohort allelic sums test (CAST)*

The cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007) follows a collapsing strategy for rare genetic variants. It assumes that the risk of a disease can be risen by the attendance of any rare mutation in a genetic set. The genetic score of the ith individual in the CAST is given by

$$B_i = I_{\{\sum_{x=1}^{j} z_{ix>0}\}}, \tag{3.2}$$

where I is an indicator function that takes 0 or 1, and z_{ix} is the genotype of the ith individual at the xth variant that is formed of (0, 1, 2). Then, the chi-square

test can be applied in order to evaluate the association of the burden score and binary trait outcome.

3.1.3.1.2 Combined and multivariate collapsing (CMC) test

The combined and multivariate collapsing (CMC) test (Li and Leal, 2008) is a technique that is designed to analyse both rare and common variants. In the same manner as the CAST, the CMC test collapses genetic variants but in different categories of the variant allele frequency and appraises the combined effect of common and rare variants by utilising Hotelling's t-test. The stages of the CMC test can be formulated as follows:

- Collect genetic variants based on their allele frequency.
- Collapse each group by using the CAST method.
- Perform Hotelling's t-test.

The CMC test improves the power when a genetic set includes common variants. This means that when a genetic set does not contain common variants, the CMC method becomes similar to the CAST.

3.1.3.1.3 Weighted-sum statistic (WSS)

Different assumptions about the relationship between a genetic set and trait outcome can be produced in the weighted-sum statistic (WSS) method (Madsen and Browning, 2009). The WSS approach is a collapsing procedure that is proposed to give more weight to rare variants, and its summary genetic score of the ith

individual is given as

$$B_i = \sum_{x=1}^{j} w_j z_{ix}. \tag{3.3}$$

A suggested weight is given as $w_j = I_{\{VAF_x<\zeta\}}$, where VAF_x is the allele frequency of the xth variant and ζ is a pre-determined threshold for the allele frequency. Another weight is suggested as $w_j = 1/\sqrt{VAF_x(1-VAF_x)}$. In order to test the association, the Wilcoxon rank-sum test is applied in the WSS approach.

3.1.3.1.4 The limitations of burden tests

The burden methods assume that all of the genetic variants grouped in a gene or any genomic unit are causal and have the same direction and magnitude of effect. Consequently, this means that it might result in a loss of power when this assumption is violated. Another limitation is that in order to obtain sufficient power by using burden tests, the methods need large rates of variants that are causal. Some of the burden approaches such as the CAST, CMC test, and WSS have been constructed to deal with qualitative data, and they do not incorporate covariates as they use the collapsing model.

3.1.3.2 Variance-component tests

Instead of grouping variants, proposed techniques use a variance-component test to assess the association by evaluating the distribution of the effect of a set of rare variants. Approaches that utilise the type of variance-component test are the sum of square score (SSU) test (Pan, 2009), C-alpha test (Neale et al., 2011), sequence

kernel association test (SKAT) (Wu et al., 2011).

3.1.3.2.1 C-alpha test

The C-alpha test is a good and robust procedure for the presence of a mixture of biased and unbiased coins (Neyman and Scott, 1965; Zelterman and Chen, 1988). Neale et al. (2011) develops the C-alpha test to evaluate the association of a group of rare variants. Assuming that the rare variants are randomly distributed across the individuals, the probability of observing a particular variant m_1 times in the affected subjects out of m total is evaluated by the binomial distribution (m, p). On the assumption that the sample consisting of affected and unaffected individuals is in equilibrium, it can indicate that $p = 0.5$ and m_1 is $0, 1$ and 2 are expected with probabilities of 0.25, 0.5, and 0.25, respectively. It is natural to observe a higher proportion of $m_1 = 0$ or $m_1 = 2$ than expected if some variants are protective or harmful.

For the xth variant observed m_x times, it is assumed that m_{1x} follows the binomial distribution (m_x, p_x) under the null hypothesis $H_0 : p_x = p_0$ ($p_0 = \frac{1}{2}$ if the number of affected and unaffected individuals is equal, so rare variants are expected to be in either sample at random). The C-alpha test statistic T compares the variance of each observed count with its expected variance under the assumption of a binomial distribution.

$$T = \sum_{x=1}^{m} [(m_{1x} - m_x p_0)^2 - (m_x p_0 (1 - p_0))],$$

where m_{1x} is the number observations of the xth variant in the affected subjects

out of n individuals, and m_x is the number of copies of the xth variant.

In order to standardise the test statistic, c is required which is the variance of T and given by

$$c = \sum_{n=2}^{maxn} m(n) \sum_{u=0}^{n} [(u - np_0)^2 - np_0(1 - p_0)]^2 f(u \mid n, p_0),$$

where $m(n)$ is the number of variants with n copies, and $f(u \mid n, p_0)$ indicates the probability of observing u copies of the xth variant assuming the binomial model. The resulting test statistic is defined as $Z = T/\sqrt{c}$. We reject the null hypothesis when Z is larger than expected using a one-tailed standard normal distribution for reference (Neale et al., 2011).

3.1.3.2.2 *Sequence kernel association test (SKAT)*

The sequence kernel association test (SKAT) (Wu et al., 2011) is a regression approach that tests the effect of variants located within a genetic set on a trait. Genetic variants are possible to be common or rare variants, and SKAT allows the variants to have different direction and magnitude. Moreover, SKAT can deal with continuous and dichotomous traits and adjust the covariates, which could be age, gender or any environmental variable. The idea of SKAT is that it aggregates the entire genetic information from variants within a genetic unit such as a gene set and gives all variants the same weight.

Recall model 3.1, SKAT assumes that each β_x, $x = 1, \ldots, j$ follows an arbitrary distribution with mean 0 and variance $w_x\tau$ where w_x is a pre-specified weight of

the xth variant, and τ is a variance component. The null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = 0$ is equivalent to testing $H_0 : \tau = 0$.

The variance component score statistic for SKAT is given by

$$Q = (\mathbf{y} - \hat{\boldsymbol{\mu}})^T \mathbf{K} (\mathbf{y} - \hat{\boldsymbol{\mu}}),$$

where $\hat{\boldsymbol{\mu}}$ is the predicted mean of y under the null hypothesis H_0 that is in our model $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\beta_0)$ for the dichotomous outcomes, and K is the kernel matrix. There are several types of pre-specified kernels, including linear, weighted linear, IBS, weighted IBS and quadratic. The kernel matrix for the weighted linear kernel is $K = ZWWZ$ where Z is a genotypes matrix, and W is a diagonal weight matrix that contains the weights of the j variants. Good choices of weights can improve power. SKAT authors suggest using $\sqrt{w_x} = \text{Beta}(VAF_x; a_1, a_2)$ where VAF_x is the allele frequency of the xth variant, and it is recommended setting $a_1 = 1$ and $a_2 = 25$ as it enhances the weight of rare mutations while still giving proper non zero weights for low-frequency mutations with allele frequency 1%–5% (Wu et al., 2011). The Q statistics has a mixture of chi-square distributions under the null hypothesis that can be evaluated explicitly and used as a reference distribution to compute the p-value.

The sequence kernel association test (SKAT) becomes comparable to the C-alpha approach (Neale et al., 2011) in the event of investigating the association between an assemblage of genetic variants and a dichotomous outcome in the absence of covariates; therefore, it means that the C-alpha test is a particular case of the

SKAT.

The following section evaluates the performance of some selected approaches of burden tests and variance-component tests in terms of type I error and power.

3.1.3.3 Simulation studies and results

In this section, we made different setups of simulation studies to evaluate the abilities of the cohort allelic sums test (CAST), weighted-sum statistic (WSS) and C-alpha test. The performance assessment of the gene-based association methods was made in terms of type I error and power through $1,000$ replications. Similar to assessing the analysis of single-variant approaches, evaluating type I error was done by generating the data under the null hypothesis ($\boldsymbol{\beta} = 0$). In terms of testing the power, the genetic effect size $\boldsymbol{\beta}$ was set to be a nonzero value. Diverse situations were designed in order to test the power of the tests.

A dataset of a sample size of $n = 400$ was constructed, and we assume that a candidate gene contains 10 genetic mutations. For the ith individual, each mutation from the 10 genetic mutations $z_{ix}, x = 1, \ldots, 10$, was simulated with a pre-selected variant allele frequency ($VAF_x = 0.01, 0.008, 0.005$). A binary outcome Y_i was simulated by a Bernoulli distribution with probability of success p_i, and $\log(p_i/(1-p_i)) = -0.5 + \boldsymbol{\beta}\mathbf{z}_i$.

In the first scenario, it is supposed that all of the 10 mutations within a gene are causal and have the same effect magnitude. In the second scenario, we assume that some of the 10 genetic variants are not causal, and four cases are considered.

In the first case, 3 of the 10 mutations are expected to be natural and do not have an effect on a trait outcome. In the second and third cases, it is presumed that 5 and 7 mutations are not causal, respectively. In all of these three cases, the active mutations that are deemed to produce an impact on a trait outcome have the same effect size. Finally, in the fourth case, we assume that only 1 mutation within a gene is causal; namely, the remaining 9 mutations are harmless and do not have an impact on a disease trait.

3.1.3.3.1 *Type I error and power*

In all of the simulated frameworks, the type I error was protected by all of the gene-based methods. With relevance to tests power, as Figure 3.2 illustrates, when assuming that all mutations that are gathered within a gene have a contribution to a trait outcome, the burden tests, the cohort allelic sums test (CAST) and weighted-sum statistic (WSS), can perform productively (obtaining in excess of 90% power) when the effect size $\beta > 0.8$ for a sporadic genetic mutation frequency. In contrast, the C-alpha test requires an effect size $\beta > 1.2$ to get more than 90% power.

In the second scenario, when supposing that some mutations within a gene do not affect a trait outcome, the WSS procedure continues to perform better than the CAST and C-alpha methods as displayed in Figure 3.3. However, all the approaches carry out inadequately in this sample size (n=400) when most mutations of the 10 genetic variants do not influence a trait outcome. To put it another way, the WSS method's power does not surpass 68% for VAF=0.01, while the CAST

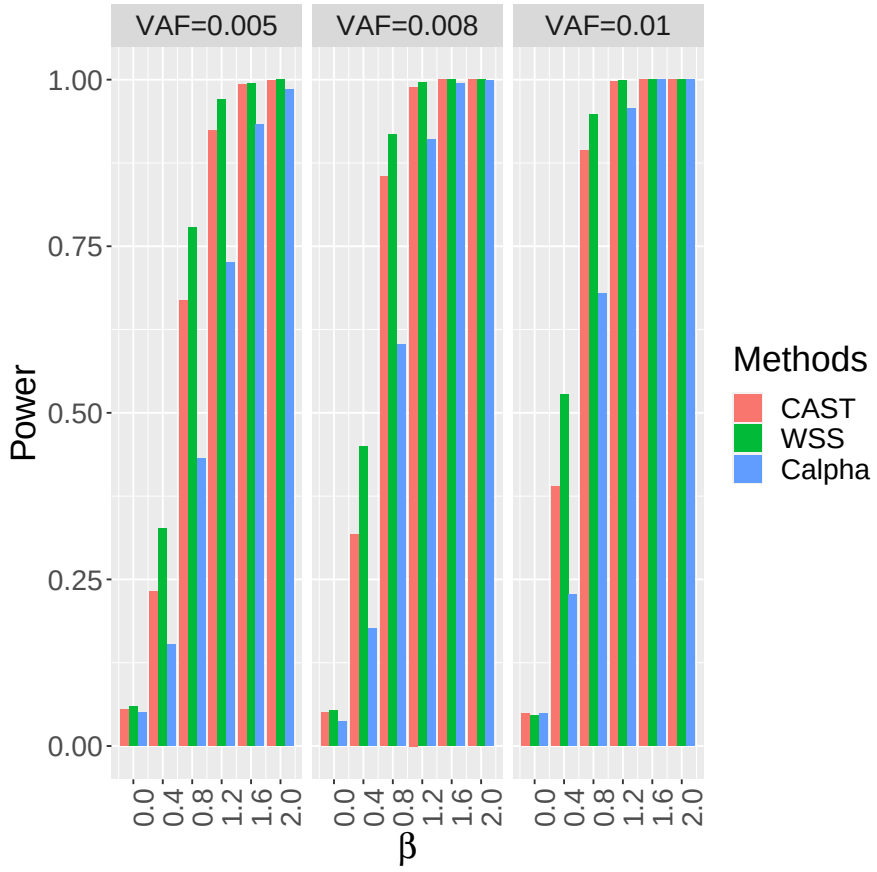

Figure 3.2: Comparison of type I error and power for gene-level mutation analysis of the cohort allelic sums test (CAST) (red bars), weighted-sum statistic (WSS) (green bars) and C-alpha test (blue bars) with various rates of variant allele frequency (VAF) and effect size β. In the comparison, the sample size is ($n = 400$), and the number of variants within a gene is 10. It is assumed that all of the 10 mutations are causal and have the same magnitude and direction of the association.

and C-alpha tests reach less than 53% and 27% power, respectively, when 7 mutations within the 10 variants are assumed to be harmless. Moreover, none of the procedures overtops 20% power when only 1 mutation is believed to be leading to a trait outcome, as demonstrated in Figure 3.4.

3.1.3.4 Conclusion

When a genetic region contains a number of causal variants that hold the same direction of the association, burden tests can excel variance-component tests. On the contrary, variance-component tests might be more potent than burden tests in the circumstance that a genetic region is composed of variants with different directions of the association (detrimental and protective variants) (Lee et al., 2014). Because both of the cases are possible to occur, several methods have been developed that combine burden tests and variance-component tests (Lee et al., 2014). Approaches that use a combination of burden tests and variance-component tests include the SKAT-O (Lee et al., 2012), Fisher method (Derkach et al., 2013) and MiST (Sun et al., 2013).

3.2 Genetic mutation analysis in cancer

In the light of the analysis of cancer genetic mutations, it is too important to recognise that the detection procedure of germline mutations or somatic mutations is a critical stage that needs to be studied and considered prior to testing the potential associations. As long as in the thesis, we are interested in a point somatic

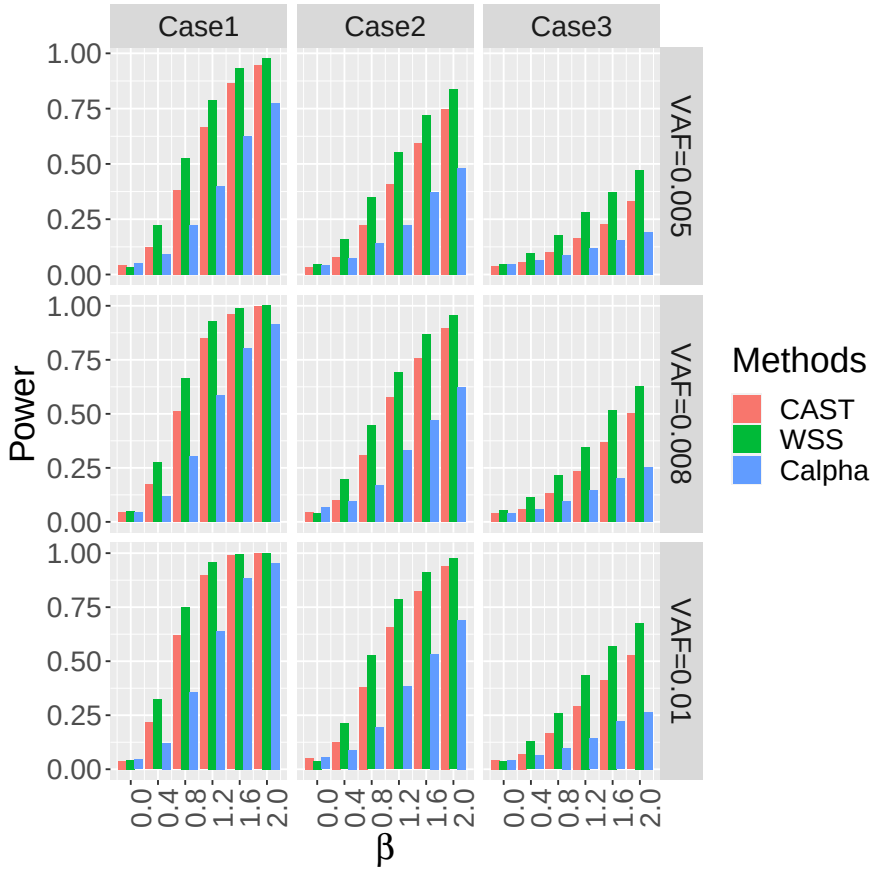

Figure 3.3: Comparison of type I error and power for gene-level mutation analysis of the cohort allelic sums test (CAST) (red bars), weighted-sum statistic (WSS) (green bars) and C-alpha test (blue bars) with various rates of variant allele frequency (VAF) and effect size β. In the comparison, the sample size is ($n = 400$), and the number of variants within a gene is 10. In case 1, it is assumed that 7 of the 10 mutations are causal, while in cases 2 and 3, 5 and 3 mutations are supposed to cause an effect, respectively. The effective variants are assumed to have the same magnitude and direction of the association.

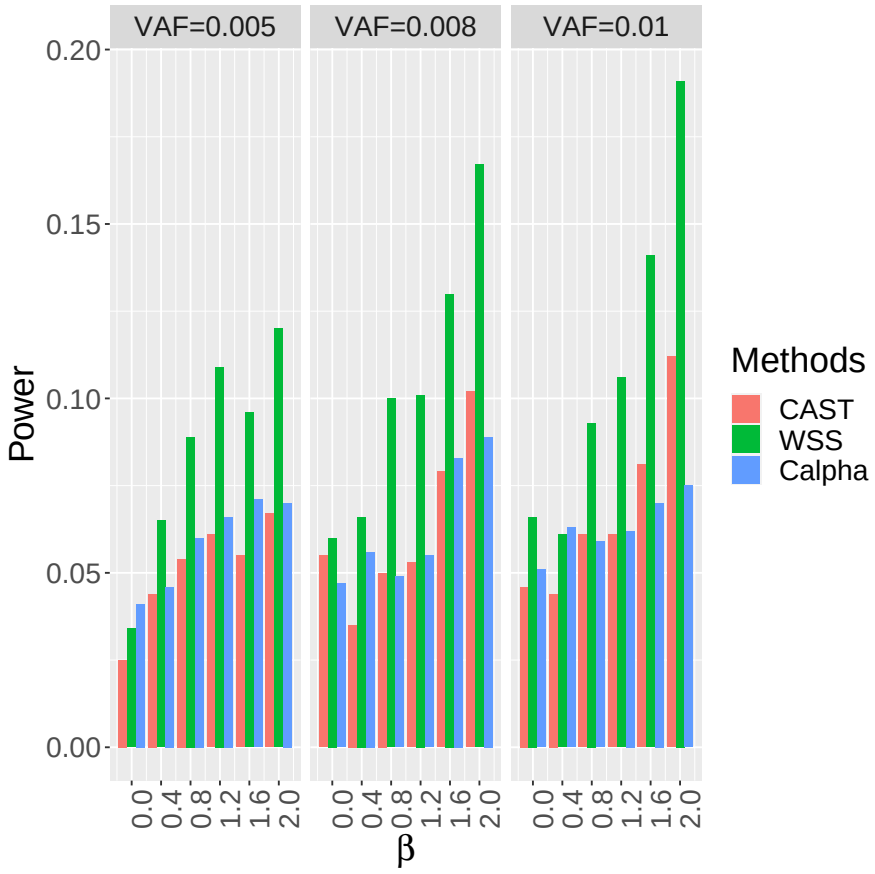

Figure 3.4: Comparison of type I error and power for gene-level mutation analysis of the cohort allelic sums test (CAST) (red bars), weighted-sum statistic (WSS) (green bars) and C-alpha test (blue bars) with various rates of variant allele frequency (VAF) and effect size β. In the comparison, the sample size is ($n = 400$), and the number of variants within a gene is 10. In this setup, it is assumed that only 1 mutation is causal.

mutation; this section presents an introduction to some common somatic mutation calling approaches. Then, it proceeds to introduce a recent technique that has been released in order to study the association between somatic mutations and a cancer subtype outcome.

3.2.1 Somatic mutation calling methods

Next-generation sequencing (NGS) technologies, also known as massively parallel sequencing, have provided essential and sophisticated help in medical and biological fields. The remarkable feature of the NGS technologies, unlike Sanger sequencing, is that millions of DNA fragments are simultaneously aligned so that the whole chain of the genome can be sequenced (Behjati and Tarpey, 2013). It is believed that a variant given enough coverage (read-depth) can be identified irrespective of the variant allele frequency (VAF) or the variant position. However, calling the variant with certainty is not a simple task due to possible errors of reads when the genomic sequences are scanned (Xu, 2018). Abundant bioinformatic methods have been released to call mutations. In respect to cancer data, calling somatic mutations is thought to be more challenging compared to germline mutations as somatic variants are considered low frequency and rare (Liu et al., 2018).

Several techniques have been conducted for identifying somatic mutations, such as heuristic methods, probabilistic methods and machine learning methods. Some somatic mutation calling methods are constructed only to call the type of single nucleotide variants (SNVs) of a somatic mutation, while other methods can detect SNVs and insertion/deletion polymorphisms (indels). Few methods can detect all

types of genetic variation, (i.e. they can call SNVs, indels and structural variations (SV)). Table 3.2 summarises some common somatic mutation calling methods and classifies their categories and variation types. Some other approaches are not mentioned in the table.

3.2.1.1 Heuristic methods

Normal and tumour cells are separately analysed to detect a somatic mutation in light of the heuristic calling approach. Then, the results are compared to each other. The heuristic methods are used by Koboldt et al. (2012); Hansen et al. (2013); Radenbaugh et al. (2014); Lai et al. (2016).

The VarSan2 approach (Koboldt et al., 2012) is a heuristic method that looks for a genotype with a variant with a minimum frequency level (0.20 by default) and is adjustable by a user. If genotypes do not match in normal and tumour cells, then one-tailed Fisher's exact test is used in order to examine for a significant difference between the two sets of cells. A genotype is called somatic if the normal sample is a homozygous reference in case of a significant difference. Otherwise, the genotype can be called loss-of-heterozygosity (LOH) if the normal sample is a heterozygous reference and called unknown if the normal sample is a homozygous variant (non-reference) and the tumour sample does not match.

The Shimmer approach (Hansen et al., 2013) is a method that follows VarScan2's idea. It applies the Fisher's exact test but conducts multiple testing correction to control the type 1 error. In addition to these two somatic mutation calling methods, the approaches of RADIA (Radenbaugh et al., 2014) and VarDict (Lai

et al., 2016) adopt the notion of heuristic methods.

Somatic mutation calling methods		
Somatic mutation callers	Type of variation	Category
JointSNVMix2	SNV	Probabilistic method
MutationSeq	SNV	Machine learning
MuTect	SNV	Probabilistic method
SAMtools	SNV, indel	Probabilistic method
RADIA	SNV	Heuristic method
Shimmer	SNV, indel	Heuristic method
SNooPer	SNV, indel	Machine learning
SomaticSeq	SNV	Machine learning
SomaticSniper	SNV	Probabilistic method
VarDict	SNV, indel, SV	Heuristic method
VarScan2	SNV, indel	Heuristic method

Table 3.2: A summary of some methods used for detecting somatic mutations.

3.2.1.2 *Probabilistic methods*

In this method, the tumour and normal cells are diploid, and the likelihood of joint genotypes are evaluated. Examples of approaches that use this type of method are SAMtools (Li, 2011), SomaticSniper (Larson et al., 2012), JointSNVMix2 (Roth et al., 2012) and MuTect (Cibulskis et al., 2013).

The SomaticSniper procedure (Larson et al., 2012) is based on a Bayesian method and calculates the posterior probability of a genotype across tumour and normal cells given the observed reads and calculate prior genotype likelihoods depending on some genetic information such as somatic mutation rate, population mutation rate and sequencing error rate (Roberts et al., 2013). SomaticSniper discovers a somatic mutation by constructing a somatic score for each genetic position that

there is no difference in the genotypes of tumour and healthy cells. The somatic score is given by

$$-10\mathbf{log}_{10}P(Z_T = Z_N \mid D_T, D_N),$$

where Z_T, Z_N are genotypes of tumour and normal cells, respectively, and D_T, D_N are read-depth values in tumour and normal cells, respectively. A higher somatic score means that the genetic position is more likely to have different genotypes in tumour and normal samples. Therefore, that position has a high possibility to be called somatic.

3.2.1.3 Machine learning methods

Some developed methods use machine learning techniques for detecting somatic mutations, such as mutationSeq (Ding et al., 2012), SomaticSeq (Fang et al., 2015) and SNooPer (Spinella et al., 2016). The MutationSeq method (Ding et al., 2012) utilises genotypes and other genetic features on every genetic position to train four classifiers. The trained classifiers are random forests, Bayesian adaptive regression tree, support vector machine and logistic regression. The learning methods were tested on naive datasets based on the features. A mutation could be somatic if validated, or it is considered non-somatic if it was found a wild-type or germline variant. All of these four machine learning methods were found to be reasonable compared to the subtraction method. The SNooPer approach (Spinella et al., 2016) is a technique that follows the machine learning methods, and it only uses a random forest classifier and claims it works well on low-coverage data (Xu, 2018).

3.2.2 A recent proposed method of somatic mutation association analysis

After introducing the detection analysis of genetic mutations in cancer and discussing a number of somatic mutation calling approaches as we are interested in somatic mutations rather than germline mutations, in this part, we present a recently developed technique to analyse the effect of somatic mutation on a trait outcome.

The burden tests and variance-component tests mentioned in Section 3.1 might not be suitable for analysing somatic mutations due to the false positive and false negative rates that the calls of somatic mutations regularly have (Liu et al., 2018). Therefore, a recent somatic mutation association approach, called Somatic mutation Association test with Measurements Errors (SAME) (Liu et al., 2018), has been released to study the effect of a single somatic mutation or gene-level somatic mutations on a continuous or binary outcome of cancer trait-related with considering the uncertainty of calling somatic mutations.

The SAME test uses read count data (read-depth and alternative reads) to model the potential errors of the somatic mutation calling procedure and applies the likelihood ratio test to investigate the relationship between somatic mutations and a cancer subtype outcome. The SAME test has been compared to the GLM that does not account for the uncertainty of somatic mutation calling, and it has been shown that the SAME test performs better than the GLM. It indicates that taking the uncertainty of somatic mutation calling can improve the association analysis

of somatic mutations.

3.3 Discussion

The association analysis of single genetic variants was firstly discussed in this chapter by introducing some approaches used in the genome-wide association study (GWAS). Also, a simulation study was shown in order to evaluate a selection of single-variant tests. Next presented was the theory of the set-based analysis of genetic variants as single-variant tests do not perform sufficiently in rare genetic variants. A number of region-based approaches, including burden tests and variant-component tests, was mentioned and evaluated via simulation studies. Due to the fact that this thesis is interested in cancer genetic mutations, specifically in somatic mutations, the identification procedure of somatic mutations was disclosed in the chapter. Furthermore, several approaches of calling somatic mutations, including heuristic methods, probabilistic methods and machine learning methods, were considered. In conclusion, the chapter ended by displaying an introduction to a new approach produced to test the effect of somatic mutations on cancer subtype outcomes.

In the next chapter, we propose a novel score test procedure based on using the GHC test in order to compare two genetic sequences taken from the same patient with the objective of detecting the effect of somatic mutations that tumour samples contain.

Chapter 4

Novel use of GHC in somatic mutation association analysis

Somatic mutation calling methods, introduced in Chapter Three, aim to discover a single somatic mutation by utilising statistical tests. Somatic mutations are found in tumour cells but not in healthy cells, and their significant role in developing cancer leads to investigating the influence of somatic mutations on cancer outcomes. The work motivation in this chapter has been risen by the 100,000 Genomes Project (100kGP) (Caulfield et al., 2017) since two samples received from an individual are matched, seeking to study the crucial impact of somatic mutations. This chapter proposes a score test procedure to detect an association of a set of somatic mutations assorted within a gene and cancer outcomes based on applying the GHC test. For the sake of examining the validity of our proposed method, it is evaluated through different scenarios of simulation studies, in terms of type I error and power, and compared to the performance of the binomial exact

test corrected by the Bonferroni correction.

4.1 Score test for matched pairs data

Consider a study where two genomes are taken from each person. One sample is sequenced from tumour cells, and the second sample is sequenced from normal (healthy) cells. Denote the ith patient's tumour cells by $y_{1i} = 1$ and the disease-free cells by $y_{2i} = 0$. Following the conditional probability approach used in matched pairs case-control designs, the likelihood function for the ith patient is given by

$$L_i = \frac{e^{\boldsymbol{\beta}\mathbf{z}_{1i}}}{e^{\boldsymbol{\beta}\mathbf{z}_{1i}} + e^{\boldsymbol{\beta}\mathbf{z}_{2i}}}, \tag{4.1}$$

and its log likelihood function is

$$l_i = \boldsymbol{\beta}\mathbf{z}_{1i} - \log(e^{\boldsymbol{\beta}\mathbf{z}_{1i}} + e^{\boldsymbol{\beta}\mathbf{z}_{2i}}), \tag{4.2}$$

where $\mathbf{z}_{1i} = (z_{1i1}, \ldots, z_{1ij})$ is a vector of the ith patient's genotypes for the j mutations within a gene of tumour cells, and $\mathbf{z}_{2i} = (z_{2i1}, \ldots, z_{2ij})$ is a vector of the ith patient's genotypes for the j mutations within a gene of normal cells. We suppose an additive genetic model and let $z_{1ix} = 0, 1$ or 2 and $z_{2ix} = 0, 1$ or 2 indicate the number of variant allele counts where $x = 1, \ldots, j$. Finally, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_j)'$ is a vector of the genotypes coefficients.

The score function for the ith individual for β_x is the first derivative of l_i with

respect to β_x, $x = 1, \ldots, j$, and it is given as

$$u_i = \frac{\partial l_i}{\partial \beta_x} = z_{1ix} - \frac{z_{1ix} e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + z_{2ix} e^{\boldsymbol{\beta} \mathbf{z}_{2i}}}{e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + e^{\boldsymbol{\beta} \mathbf{z}_{2i}}}. \tag{4.3}$$

Under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ that is $\beta_1 = \beta_2 = \cdots = \beta_j = 0$, the score function for the xth marker becomes

$$u_i = \frac{z_{1ix} - z_{2ix}}{2}. \tag{4.4}$$

The Fisher Information of $\beta_x, \beta_{x'}$ is given as

$$\begin{aligned} I_i = -E\left[\frac{\partial l_i}{\partial \beta_{x'} \partial \beta_x}\right] = \frac{-z_{1ix} z_{1ix'} e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + z_{2ix} z_{2ix'} e^{\boldsymbol{\beta} \mathbf{z}_{2i}}}{e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + e^{\boldsymbol{\beta} \mathbf{z}_{2i}}} \\ + \frac{(z_{1ix} e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + z_{2ix} e^{\boldsymbol{\beta} \mathbf{z}_{2i}}).(z_{1ix'} e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + z_{2ix'} e^{\boldsymbol{\beta} \mathbf{z}_{2i}})}{(e^{\boldsymbol{\beta} \mathbf{z}_{1i}} + e^{\boldsymbol{\beta} \mathbf{z}_{2i}})^2}. \end{aligned} \tag{4.5}$$

The Fisher Information under the null becomes

$$I_i = \frac{(z_{1ix} - z_{2ix})(z_{1ix'} - z_{2ix'})}{4}. \tag{4.6}$$

It follows that the score test statistic for the xth marker is

$$V = \frac{(\sum_{i=1}^{n} u_i)^2}{\sum_{i=1}^{n} I_i} = \frac{(\sum_{i=1}^{n} z_{1ix} - z_{2ix})^2}{\sum_{i=1}^{n} (z_{1ix} - z_{2ix})^2}. \tag{4.7}$$

The performance of the proposed score test corrected by the GHC test correction is evaluated below in terms of type I error and power and compared to the binomial exact test based on the Bonferroni correction.

4.2 Simulation studies and results

Recall from Chapter One in the 100,000 genomes project, tumour and normal sequences were compared in cancer patients to detect somatic mutations that are likely to be a cause of cancer. In this section, we made a variety of simulation configurations in order to assess the performance of our developed gene-based method for detecting the impact of somatic mutations grouped within a gene and compare it to the binomial exact test in terms of type I error and power. The evaluation procedure was produced $1,000$ replications at a significance level of 0.05. The score test statistic for association at each marker and the correlation matrix for the score test statistics are calculated to compute the GHC test. For comparison, the binomial exact tests corrected by the Bonferroni correction are computed.

Two sets of genotypes are simulated for each individual. One set is for tumour cells, and the other set is for normal cells. The rare variants are simulated with a pre-selected variant allele frequency ($VAF = 0.01, 0.008, 0.005$) and with the majority occurring in both tumour and healthy cells. In other words, these rare variants that present in both cells are not considered somatic mutations.

Evaluating the type I error was performed by assuming that the simulated sequences of a gene of tumour cells do not contain somatic mutations. In terms of evaluating the power, we set a number of somatic mutations in tumour sequences. Different scenarios are considered in order to assess the proposed approach's ability.

In the first case, a dataset of a sample size of $n = 400$ was generated. In this case, we supposed that there are 10, 7, 5, 2 or 0 mutations within a total of 50 rare variants that occur in tumour sequences, and they are deemed somatic mutations. In the second and third cases, we constructed simulation setups in the same way as in the first case, but we changed the length of genes to contain 100 and 150 rare variants, respectively. In specific, in the second case, it is assumed that tumour sequences include 10, 7, 5, 2 or 0 somatic mutations within a total of 100 rare variants, whereas within 150 rare variants in the third case.

4.2.1 Type 1 error and power

The proposed gene-based score test and binomial exact test protect the level of type I error in all of the various simulated cases. Regarding power, in the first case, when a gene contains 50 rare variants, the proposed score test performs much better than the binomial exact test at all different mutation rates, as displayed in Figure 4.1. The score test procedure works perfectly (obtaining more than 98% power) when a tumour gene includes more than 5 somatic mutations with a variant allele frequency (VAF=0.01). In comparison, the binomial exact test obtains only 85% power even when a gene has a chance to include 10 somatic mutations at a frequency rate of 0.01.

When a gene gets longer to include 100 and 150 rare variants in the second and third cases, respectively, the power drop rates are less than 1.5% and 5%, respectively, for our proposed method when 10 somatic mutations occur in a tumour sequence at a mutation frequency of 0.01. In contrast, the rates are 17.3% and

37%, respectively, for the binomial exact test as shown in Figures 4.2 and 4.3. This finding implies that the proposed test is robust and is not sensitive to the gene's length when there are 10 somatic mutations with frequency ≥ 0.01.

It is evident that at this sample size ($n = 400$), the performance of our developed method decreases when a gene includes less number of somatic mutations. However, the proposed test can obtain around 99% and 88% powers when 7 and 5 somatic mutations happen, respectively, in a tumour sample with 50 rare variants at VAF=0.01, as exhibited in Figure 4.1. Compared to the binomial exact test, it obtains only 77% and 60.2% powers for 7 and 5 somatic mutations, respectively, occur in a gene with 50 rare variants at the probability of 0.01.

When a rare variant gets extremely low frequency (VAF<0.01), the tests' performance decreases. However, it is possible to have more than 98% and 75% powers by using the proposed test when 10 somatic mutations occur in a gene with 50 and 100 rare variants, respectively, at VAF=0.008. In contrast, the binomial test performs inadequately and gets only 47% and 25% powers, respectively, for the rare mutation frequency. As Figure 4.1 illustrates, the proposed test does not perform well at a sample size of $n = 400$ for VAF=0.005 even when a gene contains 50 rare variants. Therefore, we double the sample size to be $n = 800$ in order to expand our proposed test's evaluation procedure.

At a sample size of $n = 800$, as Figure 4.4 exhibits, it is likely to have a high rate of power ($\geq 90\%$ power) by using the proposed test when there are 5 or more somatic mutations in a gene with 50 rare variants at an extremely low frequency (VAF=0.005). In contrast, the binomial exact test can give only 65.2 power in

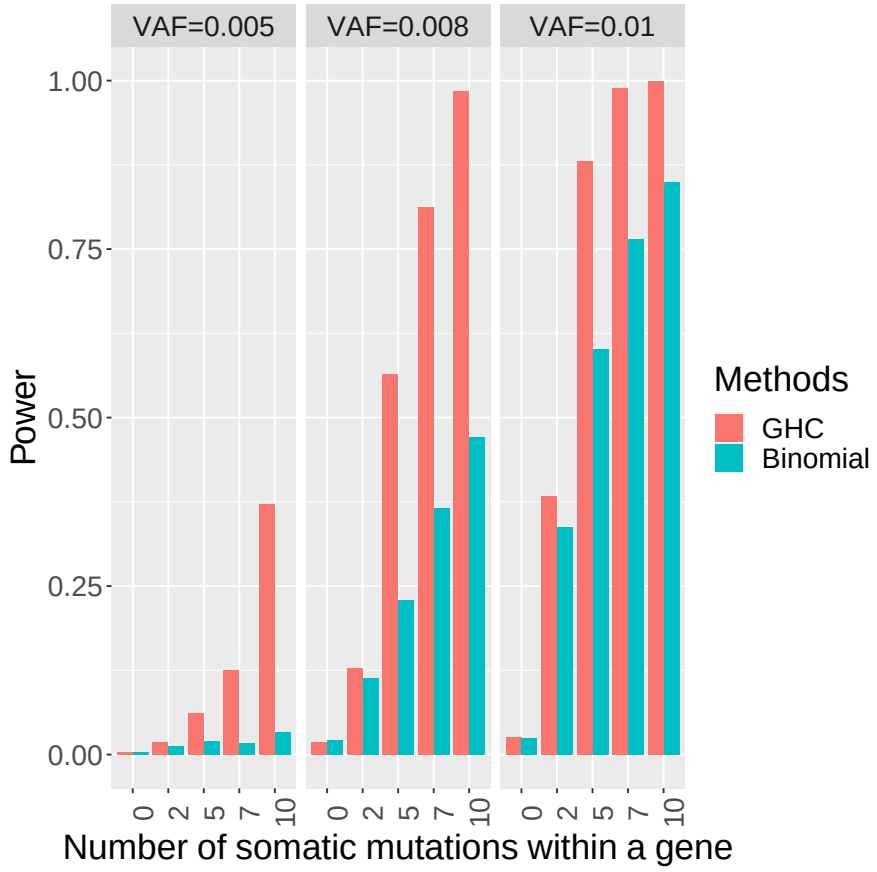

Figure 4.1: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 50 rare variants at a sample size of $n = 400$.

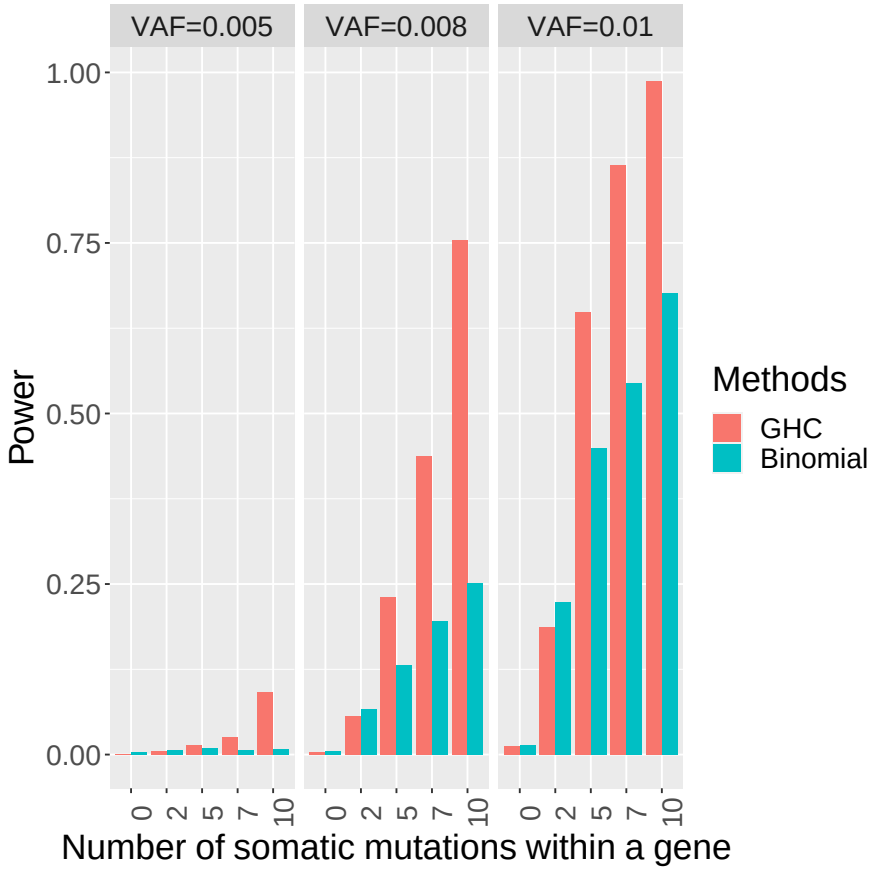

Figure 4.2: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 100 rare variants at a sample size of $n = 400$.

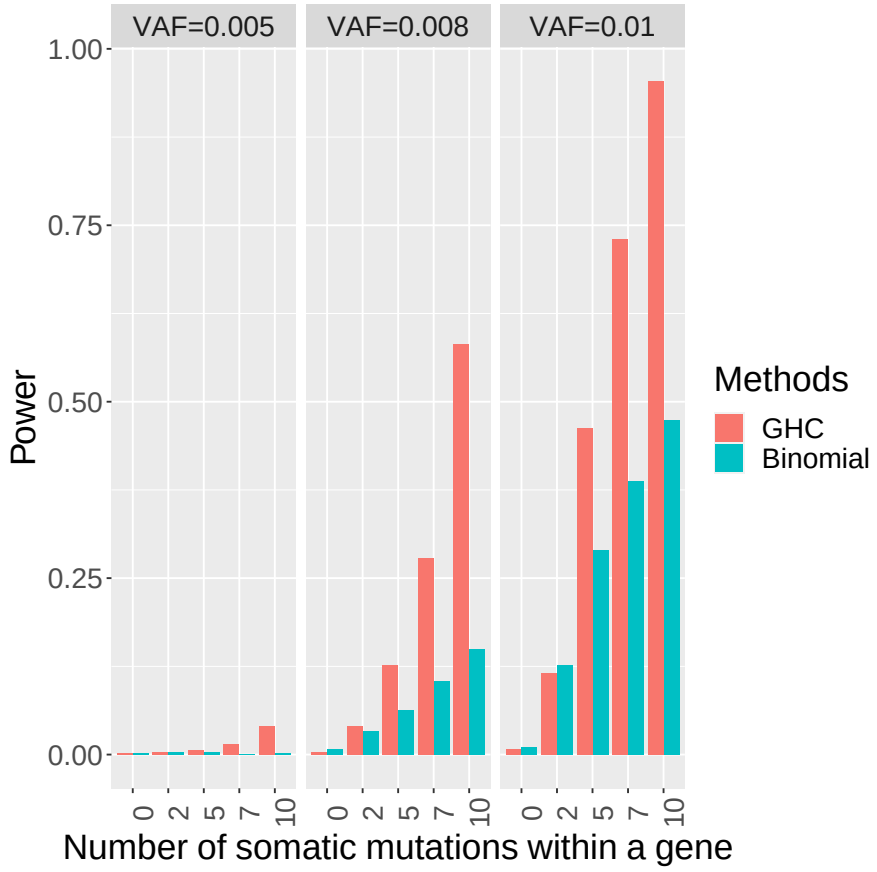

Figure 4.3: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 150 rare variants at a sample size of $n = 400$.

that case. The score test has robustness, as shown in Figures 4.5 and 4.6, since its rates of power drop by less than 1% and 5% when 10 somatic mutations occur in a gene with 100 and 150 rare variants, respectively, at VAF=0.005. However, the binomial exact test's power decreases by 19% and 41%, respectively. This can confirm that the binomial exact test is insufficient for the very low frequency (VAF=0.005) even after doubling the sample size.

4.3 Discussion

Matching tumour and normal sequences is an excellent way to understand the impact of somatic mutations. It can lead to identifying the potential association between a set of somatic mutations and a cancer trait. In this chapter, we proposed a novel use of the GHC test with the object of detecting the effect of an entire gene. Our developed method was evaluated in terms of type I error and power by comparing it to the exact binomial test adjusted by the Bonferroni correction through various simulated cases. The simulation results discovered that both of the tests controlled well type I error. Regarding power, the performance of our gene-based score test was better than the binomial exact test in the different scenarios.

A novel association approach is proposed to evaluate a single somatic mutation while considering somatic mutation calling errors in the next chapter.

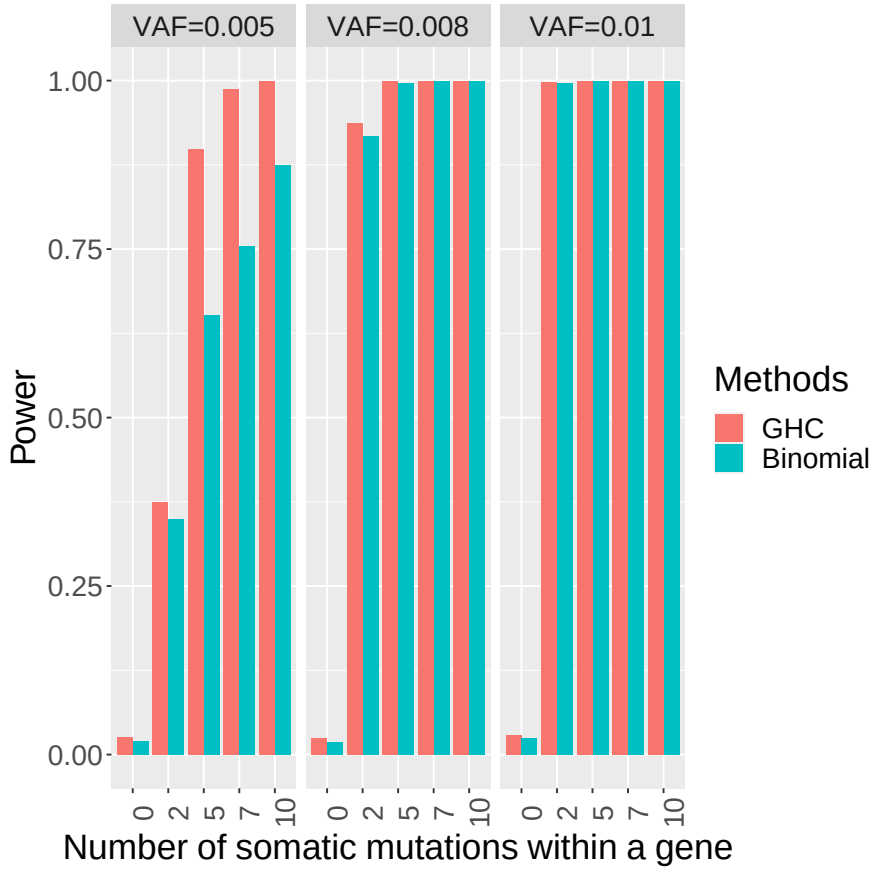

Figure 4.4: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 50 rare variants at a sample size of $n = 800$.

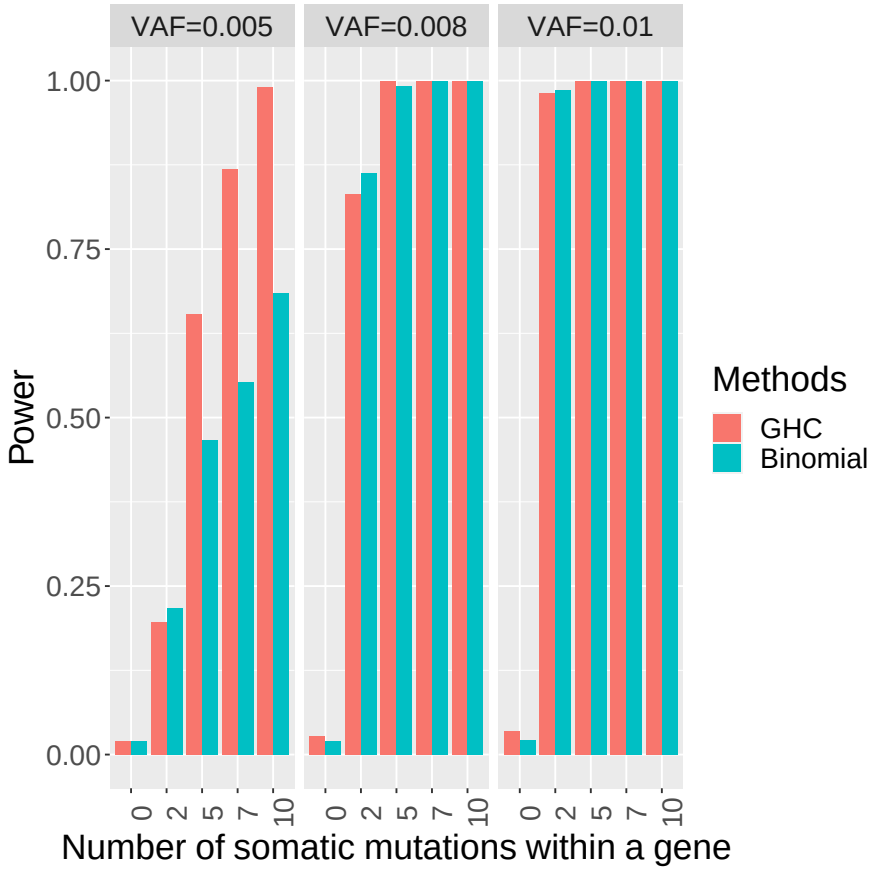

Figure 4.5: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 100 rare variants at a sample size of $n = 800$.

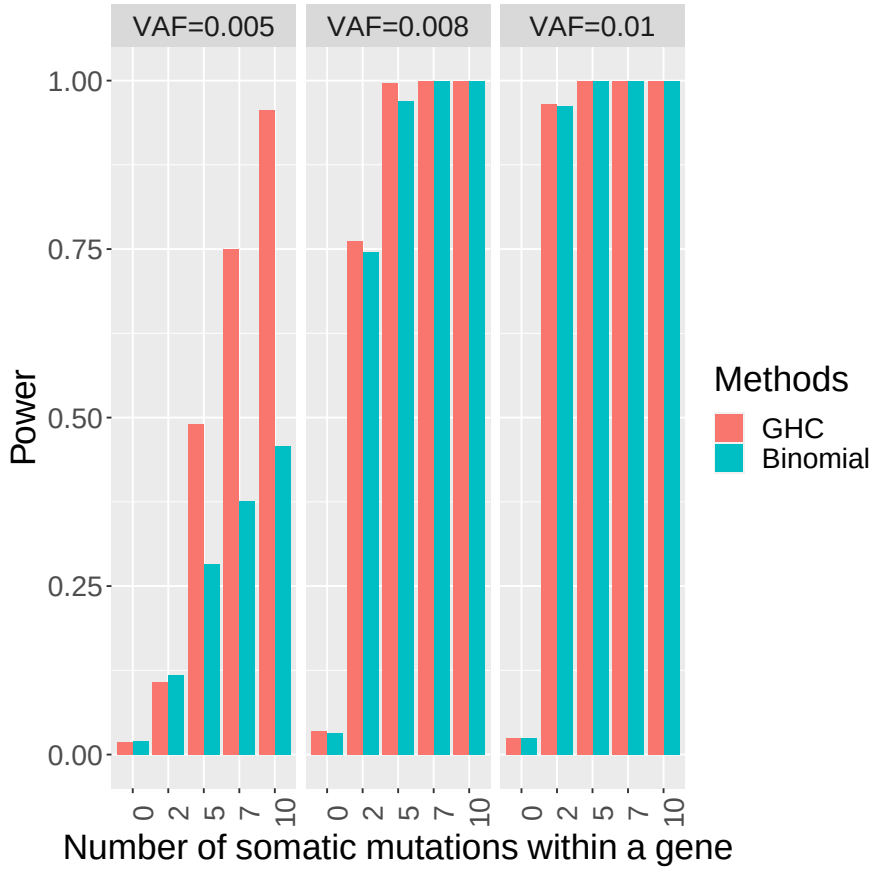

Figure 4.6: Comparison of type I error and power for the gene-level score test based on using the GHC test (red bars) and the binomial exact test (green bars) corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 150 rare variants at a sample size of $n = 800$.

Chapter 5

Association analysis of a single somatic mutation and cancer subtype outcome

Since somatic mutations, unlike germline mutations, are not easy to be confidently called as they are assumed to be low-frequency and rare mutations, there might be extremely low coverage reads in reading the whole genome sequencing data for somatic mutations. Therefore, it is essential to take the uncertainty of somatic mutation calling into consideration. In this chapter, we introduce a model for studying the relationship between a single somatic mutation and cancer subtype outcome that takes the uncertainty of somatic mutation calling into consideration. Estimating the parameters in this model is a considerable analytical and computational challenge. We construct a novel score test for the association between a single-mutation and cancer subtype outcome and evaluate its performance by com-

paring it to the recently proposed mSAME test, introduced in Chapter Two, and the commonly used GLM. Simulation results on type I error and power for a wide range of scenarios are presented. Lastly, we apply our developed single-mutation score test on multiple somatic mutations using multiple testing correction and compare it to the mSAME test and GLM. We then give an introduction to the extension of the single-mutation score test to a gene-based setting.

5.1 Introduction

Scrutinising the potential relationship between mutations and cancer outcomes can help develop a good understanding of cancer aetiology and provide scientists with an inspirational perspective of tumour cells growth. Moreover, it can assist in producing effective cancer treatments. Although it is believed that somatic mutations play the most significant role in the development of cancer (Kuijjer et al., 2018), few computational approaches have been proposed to inspect the association between somatic mutations and cancer outcomes. Liu et al. (2018) postulate that the reason for the lack of studies on somatic mutation association is that somatic mutation data are relatively new, arising from recent technological developments, and the focus has therefore been on methodology to detect these mutations reliably.

Some association approaches have been developed to study the relationship between germline mutations and cancer subtype outcomes and have considered the calling uncertainty, such as in (Lin and Zeng, 2006; Tzeng and Zhang, 2007). Still,

these approaches do not suit somatic mutations because they are not proposed for low-frequency mutations (Liu et al., 2018). Liu et al. (2018) have developed a test that accounts for the uncertainty of somatic mutation calling using the likelihood ratio test. While the approach of taking the uncertainty of somatic mutation calling into account is a good way in genetic association analysis, instead of the likelihood ratio test, a score test may be more attractive primarily as it only requires one optimisation for each hypothesis test, which is under the null model. Reducing the number of optimisations is particularly helpful in a genome-wide setting with a vast number of tests. A score test for a single-mutation setting is developed and evaluated below.

5.2 A score test

Consider a study with n independent individuals and, for the ith individual, let the actual somatic mutation status and the mutation call (observed mutation status) be denoted as S_i and O_i, respectively, $i = 1, 2, \ldots, n$. The actual status S_i can be equal to either 1 or 0, where 1 means this mutation is present in the ith individual, and 0 means it is not present. The mutation call O_i depends on the read-depth D_i. A mutation is called only if there is enough coverage, i.e. $D_i \geq D_0$, where D_0 is a selected threshold used in mutation calling methods. In this case O_i is 1 or 0 depending on whether or not the mutation was observed. If $D_i < D_0$ then O_i cannot be observed. The number of alternative reads is indicated by A_i. The outcome of cancer subtype for the ith individual is indicated by Y_i and all

covariates are indicated by the vector C_i. Finally let $\rho_0 = P(S_i = 0)$ denote the probability that the specific somatic mutation in the ith individual is not present, and let $\rho_1 = P(S_i = 1) = 1 - \rho_0$ denote the probability that the specific somatic mutation is present.

5.2.1 The likelihood function

As here S_i is not observed and ignoring the covariates C_i for the time being, we can write the probability for the ith individual, $P(Y_i, A_i, D_i, O_i, S_i) = P(S_i)P(Y_i, A_i, D_i, O_i \mid S_i)$, as the sum over all possible values, or marginal probability,

$$\sum_{x=0}^{1} P(S_i = x)P(Y_i, A_i, D_i, O_i, \mid S_i) = \sum_{x=0}^{1} \rho_x P(Y_i, A_i, D_i, O_i \mid S_i). \tag{5.1}$$

Written in this form we can see that the probability is a two-component mixture, with ρ_x being the mixing proportion. The right hand side of the above can be decomposed by a second application of the chain rule for probabilities to get

$$P(Y_i, A_i, D_i, O_i \mid S_i) = P(Y_i \mid S_i)P(O_i \mid Y_i, S_i)P(A_i, D_i \mid Y_i, O_i, S_i). \tag{5.2}$$

Now given S_i, we assume Y_i carries no additional information about O_i and further, has no more additional information about A_i and D_i. Taking this into account gives the complete marginal likelihood for the data as $L = \prod_{i=1}^{n} L_i$, where

$$L_i = \sum_{x=0}^{1} \rho_x f(y_i \mid S_i = x)f(o_i \mid S_i = x)f(a_i, d_i \mid O_i, S_i = x), \tag{5.3}$$

is the likelihood function for the ith individual and $f(y_i \mid S_i), f(o_i \mid S_i)$ and $f(a_i, d_i, \mid O_i, S_i)$ are probability functions characterising the conditional probabilities $P(Y_i \mid S_i), P(O_i \mid S_i)$ and $P(A_i, D_i \mid O_i, S_i)$.

The outcome Y_i given somatic mutation S_i can be modelled by a generalised linear model with mean $E(Y_i) = g^{-1}(\boldsymbol{\beta_0}\mathbf{C}_i^T + \beta S_i)$, for some canonical link function g. Here $\boldsymbol{\beta_0}$ and β are the regression coefficients and primary interest is inference regarding β. For a continuous outcome, $f(y_i \mid S_i)$ in equation (5.3) can be replaced by a normal density function and for a binary outcome by a Bernoulli density. The terms $f(o_i \mid S_i)$ and $f(a_i, d_i \mid O_i, S_i)$ capture the uncertainty in correctly calling S_i. The first term is modelled by Bernoulli distributions and the second by beta-binomial distributions. Further details of this approach will be provided in section 5.3.

5.2.2 The score and Fisher Information for binary outcomes

The score and Fisher Information for the parameter β are derived here. We assume that the other parameters in the model given by equation (5.3) do not affect the variance of the score statistic and thus we only require the first and second derivatives of the log-likelihood with respect to β. We also assume there are no covariates in the model. These assumption are for convenience and leads to a much simpler problem. For notational convenience in the derivation we write $k_{ix} = \rho_x f(a_i, d_i \mid O_i, S_i = x)$ and $f_x(y_i) = f(y_i \mid S_i = x)$. Then the score function

for β for the ith individual and assuming a binary outcome is given by

$$\begin{aligned} u_i &= \frac{\partial}{\partial\beta}\left[\log\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i)\right] \\ &= \frac{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x)\frac{\partial f_x(y_i)}{\partial\beta}}{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i)}, \end{aligned}$$

where $f_x(y_i) = p_{ix}^{y_i}(1-p_{ix})^{1-y_i}$ for logit $p_{ix} = \beta_0 + \beta x$ and

$$\frac{\partial f_x(y_i)}{\partial\beta} = \frac{\partial}{\partial p_{ix}}\left[p_{ix}^{y_i}(1-p_{ix})^{1-y_i}\right]\frac{\partial p_{ix}}{\partial\beta}. \tag{5.4}$$

Now as the partial derivatives

$$\frac{\partial}{\partial p_{ix}} p_{ix}^{y_i}(1-p_{ix})^{1-y_i} = \frac{f_x(y_i)}{p_{ix}(1-p_{ix})}(y_i - p_{ix}), \tag{5.5}$$

and

$$\frac{\partial p_{ix}}{\partial\beta} = x p_{ix}(1-p_{ix}), \tag{5.6}$$

the score function for the ith individual can be written

$$u_i = \frac{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) x f_x(y_i)}{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i)}(y_i - p_{ix}). \tag{5.7}$$

The score test is evaluated under the null hypothesis $H_0 : \beta = 0$. In this case,

$$f_x(y_i) = p_{ix}^{y_i}(1-p_{ix})^{1-y_i} = \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0}}\right)^{1-y_i} = \frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}, \tag{5.8}$$

and the score function given by equation (5.7) evaluates to

$$u_i = \frac{k_{i1}}{f(o_i \mid S_i = 0)k_{i0} + f(o_i \mid S_i = 1)k_{i1}} \left[y_i - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right] f(o_i \mid S_i = 1). \tag{5.9}$$

The second derivative of $\log L_i$ is

$$\frac{\partial u_i}{\partial \beta} = \frac{\partial}{\partial \beta} \left(\frac{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) x f_x(y_i)}{\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i)} (y_i - p_{ix}) \right). \tag{5.10}$$

Now

$$\frac{\partial}{\partial \beta} \left[f_x(y_i)(y_i - p_{ix}) \right] = \frac{\partial f_x(y_i)}{\partial \beta} (y_i - p_{ix}) - f_x(y_i) \frac{\partial p_{ix}}{\partial \beta},$$

which using equations (5.4)-(5.6) can be written as,

$$= x f_x(y_i) \left[(y_i - p_{ix})^2 - p_{ix}(1 - p_{ix}) \right].$$

Using the above in evaluating equation (5.10) gives the observed Fisher Information

$$\begin{aligned} I_i = \frac{1}{(\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i))^2} \Bigg[\Bigg(\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) \\ x f_x(y_i)(y_i - p_{ix}) \Bigg)^2 - \Bigg(\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) f_x(y_i) \Bigg) \\ \Bigg(\sum_{x=0}^{1} k_{ix} f(o_i \mid S_i = x) x^2 f_x(y_i) [(y_i - p_{ix})^2 - p_{ix}(1 - p_{ix})] \Bigg) \Bigg]. \end{aligned} \tag{5.11}$$

The observed Fisher Information evaluated under the null becomes

$$I_i = \frac{1}{(k_{i0}f(o_i \mid S_i = 0) + k_{i1}f(o_i \mid S_i = 1))^2}\Bigg[\Bigg(k_{i1}f(o_i \mid S_i = 1) \\ (y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}})\Bigg)^2 - \Bigg((k_{i0}f(o_i \mid S_i = 0) + k_{i1}f(o_i \mid S_i = 1))\Bigg) \\ \Bigg(k_{i1}f(o_i \mid S_i = 1)(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}})^2 - (\frac{e^{\beta_0}}{(1+e^{\beta_0})^2})\Bigg)\Bigg]. \tag{5.12}$$

It follows therefore that, as

$$E(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}})^2 = var(y_i) = \frac{e^{\beta_0}}{(1+e^{\beta_0})^2},$$

the expected Fisher Information under the null is given by

$$E(I_i) = \left(\frac{k_{i1}f(o_i \mid S_i = 1)}{(k_{i0}f(o_i \mid S_i = 0) + k_{i1}f(o_i \mid S_i = 1)}\right)^2 \frac{e^{\beta_0}}{(1+e^{\beta_0})^2}. \tag{5.13}$$

Finally the score test statistic

$$V = \frac{(\sum_{i=1}^{n} u_i)^2}{\sum_{i=1}^{n} I_i} \sim \chi_1^2. \tag{5.14}$$

The score test here is evaluated by comparing its performance, in terms of type I error and power, with the mSAME test that was developed in (Liu et al., 2018) and the GLM. Like the score test, the mSAME test accounts for the uncertainty in somatic mutation calling, but this uses the likelihood ratio procedure. On the other hand, the GLM ignores the uncertainty in observing the somatic mutation

and simply treats the observed somatic mutation O_i as the truth.

5.3 Parameter estimation

The mSAME test requires estimating the parameters in the likelihood $L = \prod_{i=1}^{n} L_i$, where L_i is given by equation (5.3), under both the null $H_0 : \beta = 0$ and alternative hypotheses, whereas the score test requires estimating the parameters only under the null. First notice that the conditional density $f(a_i, d_i \mid O_i, S_i)$ in L_i can be decomposed as $f(a_i \mid O_i, S_i, D_i) f(d_i \mid S_i)$. Assuming read depth D_i does not depend on S_i, the term $f(d_i \mid S_i)$ can be ignored in the estimation procedure. The term $f(a_i \mid O_i, S_i, D_i)$ is modelled using beta-binomial distributions that depends on the mutation call O_i and actual somatic mutation status S_i.

When there is enough coverage ($D_i \geq D_0$),

$$f(a_i \mid O_i, S_i, D_i) = \begin{cases} f(a_i \mid D_i, \pi_{00}, \varphi_{00}) \text{ if } O_i = 0, S_i = 0, \\ f(a_i \mid D_i, \pi_{01}, \varphi_{01}) \text{ if } O_i = 0, S_i = 1, \\ f(a_i \mid D_i, \pi_{10}, \varphi_{10}) \text{ if } O_i = 1, S_i = 0, \\ f(a_i \mid D_i, \pi_{11}, \varphi_{11}) \text{ if } O_i = 1, S_i = 1, \end{cases} \tag{5.15}$$

where $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ and $\varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}$ are mean and over-dispersion parameters for the beta binomial distributions, and the probability density function of

the beta-binomial distribution is given by

$$f(a_i|D_i,\pi,\varphi) = \frac{\Gamma(d_i+1)}{\Gamma(a_i+1)\Gamma(d_i-a_i+1)}\frac{\Gamma(a_i+\pi)\Gamma(d_i-a_i+\varphi)}{\Gamma(d_i+\pi+\varphi)}\frac{\Gamma(\pi+\varphi)}{\Gamma(\pi)\Gamma(\varphi)}. \tag{5.16}$$

On the other hand, when there is not enough coverage ($D_i < D_0$), the mutation call O_i is not observed and the conditional density $f(a_i \mid S_i, D_i)$ can be written as

$$f(a_i \mid S_i, D_i) = \begin{cases} f(a_i \mid D_i, \pi_0, \varphi_0) \text{ if } S_i = 0, \\ f(a_i \mid D_i, \pi_1, \varphi_1) \text{ if } S_i = 1, \end{cases} \tag{5.17}$$

where π_0, π_1 and φ_0, φ_1 are mean and over-dispersion parameters for the distributions, and the probability density function of the beta-binomial distribution is defined as equation (5.16).

The conditional density $f(o_i \mid S_i)$ can be modelled using Bernoulli distributions that depends on the somatic mutation status S_i. In particular,

$$f(o_i \mid S_i) = \begin{cases} f(o_i, 1-\gamma_0) \text{ if } S_i = 0, \\ f(o_i, \gamma_1) \text{ if } S_i = 1, \end{cases} \tag{5.18}$$

where γ_0 and γ_1 are the specificity (1- false positive rate) and the sensitivity (1- false negative rate) of somatic mutation calls. Based on evaluation of somatic mutation calling methods (Xu et al., 2014), suggested values for γ_0 and γ_1 are 0.98 and 0.9, respectively.

In summary, the parameters in the likelihood that now need to be estimated are

$$\rho_0, \beta_0, \beta, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}, \pi_0, \pi_1, \varphi_0, \varphi_1.$$

Now, a single mutation does not provide enough information to estimate the mean and over-dispersion parameters $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}$ in equation (5.15). Instead, they are estimated by pooling the data across all genes and samples in the study. In this evaluation these parameters are estimated using real data in (Liu et al., 2018). The remaining parameters are estimated using an EM algorithm.

5.3.1 EM algorithm for estimating the parameters

In a population, it is commonly seen that different groups or clusters of observations follow different distributions. To attain flexibility in such a situation, a mixture of distributions are fit to the data. In most cases, the groups or clusters of observations are not known or observed (latent). Therefore, we fit the data into a model that assumes latent variables and fits the sample to a finite mixture distribution. Depending on the nature of the latent variable, such models have different names in the literature, but in general, they are known as finite mixture models.

The Expectation Maximisation (EM) algorithm is an iterative process for maximum likelihood estimation of the model parameters that depend on unobserved variables. Each iteration of the EM algorithm consists of two steps; the expectation

(E) step and the maximisation (M) step. In the E-step, a function is created for the expectation of the log-likelihood evaluated at the current estimates of the parameters. In the M-step, the parameters are estimated by maximising the function from the E-step. An EM algorithm is used to estimate $\boldsymbol{\theta} = \{\rho_0, \beta_0, \beta, \pi_0, \pi_1, \varphi_0, \varphi_1\}$.

In the finite mixture model framework, the observed data Y_i, A_i, D_i, O_i for $i = 1, \ldots n$ are viewed as being incomplete as the vectors of component labels are missing. The data we are missing here are whether an observation has a true mutation or not ($S = 0$ or 1). The missing data can be modelled as a two dimensional latent class vector $\boldsymbol{s}_i$, $i = 1, \ldots, n$, defined as

$$s_{ix} = \begin{cases} 1 & \text{if the } i\text{th observation comes from S=x,} \\ 0 & \text{otherwise,} \end{cases}$$

for $x = 0, 1$. The component-label vectors $\boldsymbol{s}_{ix}$ are taken to be a realisation of the Bernoulli random vectors with density function

$$f(\boldsymbol{s}_{ix}) = \prod_{x=0}^{1} \rho_x^{s_{ix}} = \rho_0^{s_{i0}} \rho_1^{s_{i1}}.$$

The complete data likelihood is given by

$$L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{D}, \boldsymbol{O}, \boldsymbol{S}) = \prod_{i=1}^{n} \prod_{x=0}^{1} f(y_i, a_i, d_i, o_i|S_i = x)^{s_{ix}} \rho_x^{s_{ix}},$$

and the log-likelihood is given by

$$\ell(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{A},\boldsymbol{D},\boldsymbol{O},\boldsymbol{S}) = \sum_{i=1}^{n}\sum_{x=0}^{1}\left[s_{ix}\log f(y_i,a_i,d_i,o_i|S_i=x) + s_{ix}\log\rho_x\right].$$

Let $\boldsymbol{\theta}^{(t)}$ be the current estimate of the parameters. On the expectation step of the EM algorithm, we need $E[\boldsymbol{S}|\boldsymbol{Y},\boldsymbol{A},\boldsymbol{D},\boldsymbol{O},\boldsymbol{\theta}^{(t)}]$ which is calculated as

$$\eta_{ix} = E[S_{ix}|Y_i,A_i,D_i,O_i,\boldsymbol{\theta}^{(t)}] = \frac{\rho_x f(y_i,a_i,d_i,o_i|S_i=x,\boldsymbol{\theta}^{(t)})}{\sum_{x=0}^{1}\rho_x f(y_i,a_i,d_i,o_i|S_i=x,\boldsymbol{\theta}^{(t)})}.$$

Using the expected values η_{ix}, the expected log-likelihood in the E-step can be written as

$$\begin{aligned}
Q = E_{\boldsymbol{S}|Y,A,D,O;\boldsymbol{\theta}^t}\log L(\boldsymbol{\theta};Y,A,D,O,S) &= \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log\rho_x f(y_i,a_i,d_i,o_i|S_i=x,\boldsymbol{\theta}^{(t)}) \\
&= \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log\rho_x + \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log f(y_i|S_i=x) \\
&\quad + \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log f(a_i \mid S_i,D_i)I_{D_i<D_0} \\
&\quad + \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log f(a_i \mid O_i,S_i,D_i)I_{D_i\geq D_0}f(o_i \mid S_i).
\end{aligned} \tag{5.19}$$

Maximising equation (5.19) with respect to ρ_0 and recognising the second, third and fourth sums in equation (5.19) as constants gives

$$\rho_0^{(t+1)} = \arg\max_{\rho_0}\left[\sum_{i=1}^{n}\eta_{i0}^{(t)}\log\rho_0 + \sum_{i=1}^{n}\eta_{i1}^{(t)}\log\rho_1\right] = \frac{1}{n}\sum_{i=1}^{n}\eta_{i0}^{(t)}, \tag{5.20}$$

similarly

$$\rho_1^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\eta_{i1}^{(t)} = 1 - \rho_0^{(t+1)}. \tag{5.21}$$

Now, in order to estimate β_0 and β, we need to maximise Q with respect to β_0 and β and the other terms in equation (5.19) are constants

$$\begin{aligned} Q &\approx \sum_{i=1}^{n}\sum_{x=0}^{1}\eta_{ix}\log f(y_i \mid S_i = x) \\ &= \sum_{i=1}^{n}\eta_{i0}\log f(y_i \mid S_i = 0) + \sum_{i=1}^{n}\eta_{i1}\log f(y_i \mid S_i = 1). \end{aligned}$$

The above shows that Q is a sum of two weighted logistic regression likelihoods. Substituting for the density functions give,

$$\begin{aligned} Q \approx &\sum_{i=1}^{n}\eta_{i0}\left[y_i\log\left(p_{i0}\right) + \left(1-y_i\right)\log\left(1-p_{i0}\right)\right] \\ &+ \sum_{i=1}^{n}\eta_{i1}\left[y_i\log\left(p_{i1}\right) + \left(1-y_i\right)\log(1-p_{i1})\right]. \end{aligned}$$

Using the logit transformation to parametrising the probabilities in terms of β_0 and β, we get

$$\begin{aligned}
Q &\approx \sum_{i=1}^{n} \eta_{i0} \left[y_i \log\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) + (1-y_i)\log\left(1-\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\right] \\
&+ \sum_{i=1}^{n} \eta_{i1} \left[y_i \log\left(\frac{e^{\beta_0+\beta}}{1+e^{\beta_0+\beta}}\right) + (1-y_i)\log\left(1-\frac{e^{\beta_0+\beta}}{1+e^{\beta_0+\beta}}\right)\right] \\
&= \sum_{i=1}^{n} \eta_{i0} \left[y_i\beta_0 - y_i \log\left(1+e^{\beta_0}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\beta_0}}\right)\right] \\
&+ \sum_{i=1}^{n} \eta_{i1} \left[y_i(\beta_0+\beta) - y_i \log\left(1+e^{\beta_0+\beta}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\beta_0+\beta}}\right)\right] \\
&= \sum_{i=1}^{n} \eta_{i0} \left[y_i\beta_0 - \log\left(1+e^{\beta_0}\right)\right] + \sum_{i=1}^{n} \eta_{i1} \left[y_i\beta_0 + y_i\beta - \log\left(1+e^{\beta_0+\beta}\right)\right].
\end{aligned}$$

The estimates of β_0 and β are updated on the $(t+1)$th iteration of EM algorithm as

$$\begin{aligned}
(\beta_0^{(t+1)}, \beta^{(t+1)}) = \arg\max_{\beta_0,\beta} \Big[&\sum_{i=1}^{n} \eta_{i0} \left(y_i\beta_0 - \log(1+exp(\beta_0))\right) + \\
&\sum_{i=1}^{n} \eta_{i1} \left(y_i\beta_0 + y_i\beta - \log(1+exp(\beta_0+\beta))\right) \Big].
\end{aligned} \tag{5.22}$$

Finally, we need to estimate $\pi_0, \varphi_0, \pi_1, \varphi_1$, and note that only the third term of equation (5.19) depends on $\pi_0, \varphi_0, \pi_1, \varphi_1$. Therefore, the updated estimates are given as

$$(\pi_0^{t+1}, \varphi_0^{t+1}) = \arg\max_{\pi_0,\varphi_0} \sum_{i=1}^{n} \eta_{i0} \log f(a_i \mid D_i; \pi_0, \varphi_0) I_{D_i<D_0}, \tag{5.23}$$

$$(\pi_1^{t+1}, \varphi_1^{t+1}) = \arg\max_{\pi_1,\varphi_1} \sum_{i=1}^{n} \eta_{i1} \log f(a_i \mid D_i; \pi_1, \varphi_1) I_{D_i<D_0}. \tag{5.24}$$

All optimisations are done in R using the optim function. We can ignore the fourth term of equation (5.19) because it does not contain any unknown parameters. The iteration process of E-step and M-step will be terminated if $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||_\infty < \epsilon$ for some pre-set threshold ϵ.

5.4 Simulation studies and results

In this section, multiple simulation setups were composed in order to compare the performance of our developed score test to the mSAME test that was developed in (Liu et al., 2018) and the GLM in terms of type I error and power. The mSAME test accounts for the uncertainty of somatic mutation calling and performs the likelihood ratio test. In contrast, the GLM does not account for the uncertainty of somatic mutation calling, and it deals with the observed somatic mutation O_i as actual somatic mutation status.

The evaluation of the performance of tests was made $1,000$ replications at significance level 0.05 with different settings of somatic mutation probabilities ($\rho_1 = 0.02, 0.05, 0.1$). In terms of evaluating type I error, it was done by simulating the data under the null hypothesis (β=0). In terms of evaluating the power of test procedures, we set $\beta = 0.4, 0.8, 1.2, 1.6, 2.0$.

The main simulation setup

A dataset of a sample size of $n = 400$ was generated. The true somatic mutation status for the ith sample S_i was simulated by a Bernoulli distribution with prob-

ability of success ρ_1, and a binary outcome Y_i was simulated by a Bernoulli distribution with probability of success p_i, $\text{logit}[p(Y_i = 1)] = -0.5 + \beta S_i$. The observed mutation call O_i was simulated by a Bernoulli distribution with the sensitivity value (1- false negative rate) $\gamma_1 = 0.9$ and specificity value (1- false positive rate) $\gamma_0 = 0.98$. The sensitivity and specificity values are set according to suggestions in (Xu et al., 2014). The read-depth values D_i were simulated in two stages. First step, the mean read-depth for each mutation was simulated by a negative binomial distribution with mean $\mu = 113$ and over-dispersion 3.28. Second step, the read-depth for each mutation across samples was simulated by a negative binomial distribution with mean value that was simulated in the first step and over-dispersion 1.9. When there is a high coverage which means ($D_i \geq D_0 = 20$), the number of alternative reads A_i was simulated by a beta-binomial distribution with parameters $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.001, 0.002, 0.1179, 0.3207)$ and $(\varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}) = (0.0006, 0.3457, 0.0001, 0.1018)$. On the contrary, when there is a low coverage ($D_i < D_0 = 20$), in this case, the number of alternative reads A_i was simulated by a beta-binomial distribution with parameters $\pi_0 = 0.001, \varphi_0 = 0.001, \pi_1 = 0.146, \varphi_1 = 0.10$.

From the investigation using this simulated dataset when $\beta = 0$, it can be seen in Figure 5.1, all of our proposed score method, the mSAME and GLM procedures, control the type I error. In terms of power, our proposed score test has higher power than the mSAME test and GLM in most scenarios, significantly when the somatic mutation frequency, ρ_1 diminishes. To illustrate, for mutation frequency, $\rho_1 = 0.05$, our score test can have more than 80% power when the effect size,

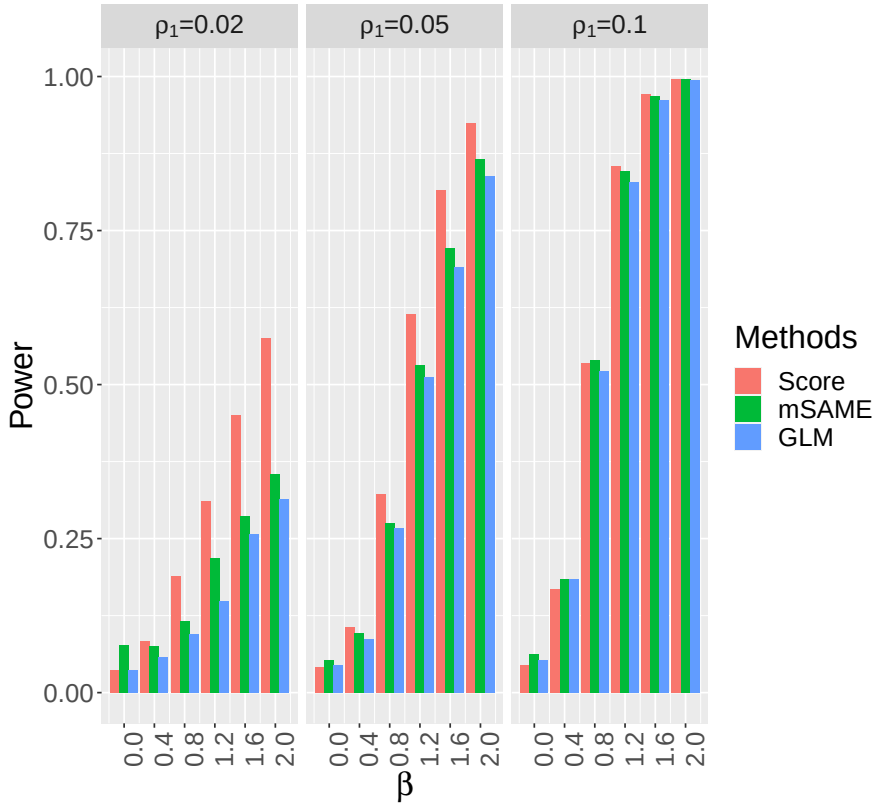

Figure 5.1: Comparison of type I error and power for single-mutation analysis of our developed score test (red bars), the mSAME test (green bars) and GLM (blue bars) with various rates of mutation frequency ρ_1 and effect size β. This setup is the main simulation model of the single-mutation analysis, which is constructed of a sample size of n=400.

$\beta \geq 1.6$ while the mSAME test and GLM necessitate $\beta = 2$ to surpass this rate of power. Additionally, for somatic mutation low-frequency, $\rho_1 = 0.02$ and $\beta = 2$, our proposed score method has the power of 0.575, but the mSAME test and GLM have the power of 0.355 and 0.314, respectively. The power rates of all the approaches are not adequate for low-frequency mutation and small effect size. Accordingly, we extended the evaluation procedure by increasing the sample size.

Sample size and error rates

Analogously to the main simulation setup, datasets of sample sizes of $n = 800, 1000$, 3000 and 5000 were generated. Evaluating our proposed method, the mSAME test and GLM based on these datasets with increased sample sizes, all of the methods control the type I error. In relation to power, our proposed score test outperforms the mSAME test and GLM for low-frequency mutation, as Figure 5.2 presents. By doubling the sample size ($n = 800$), our score test can gain more than 80% power for low-frequency mutation, $\rho_1 = 0.02$ and effect size, $\beta = 2$ while the mSAME and GLM methods can not reach 60%. Our proposed method's performance continues to operate better when the sample size is increased to be $n = 1000$. Case in point, for low-frequency mutation, $\rho_1 = 0.02$, our method can attain 80% power with $\beta = 1.6$ and more than 92% power when $\beta = 2$. However, the mSAME test and GLM have the powers of 58% and 55%, respectively, for effect size, $\beta = 1.6$, and both have less than 71% power when $\beta = 2$. In a state of $n = 3000$, our developed method requires an effect size of $\beta \geqslant 1.2$ to get above 95% power for low-frequency mutation, $\rho_1 = 0.02$, whereas the mSAME test and GLM demand an effect size of

$\beta \geqslant 1.6$ to possess this rate of power. Even though all of the methods may perform similarly with the large sample size ($n = 5000$), our score test acts better when the frequency mutation lowers, and the effect size becomes small. For mutation frequency, $\rho_1 = 0.05$, our score test tops 65% power at an effect size of $\beta = 0.4$, but the mSAME test and GLM have below 62% and 52% power, respectively. Moreover, for low-frequency mutation, $\rho_1 = 0.02$, our proposed method has the power of 0.864 when $\beta = 0.8$ whilst the mSAME test and GLM have 0.699 and 0.589 power, respectively.

This finding can verify that for the various sample sizes investigated, despite the fact that single-mutation association tests might not be robust to detect the association when genetic mutations are low-frequency and rare, our proposed score test performed better than the mSAME test and GLM for low-frequency somatic mutations ($\rho_1 = 0.02$).

Low read-depth and error rates

In the previous simulation parts, it is apparent that accounting for the error in somatic mutation calling process can improve the performance of the approaches as our developed score test and the mSAME test, which consider the somatic mutation calling uncertainty, have higher power than the GLM that does not. The GLM is less powerful even with relatively high read-depth (average read-depth of a somatic mutation was 113). To elongate the methods testing framework and compare our proposed score test to the mSAME test and GLM, we generated a dataset of a sample size of $n = 400$ identically to the main simulation setup; how-

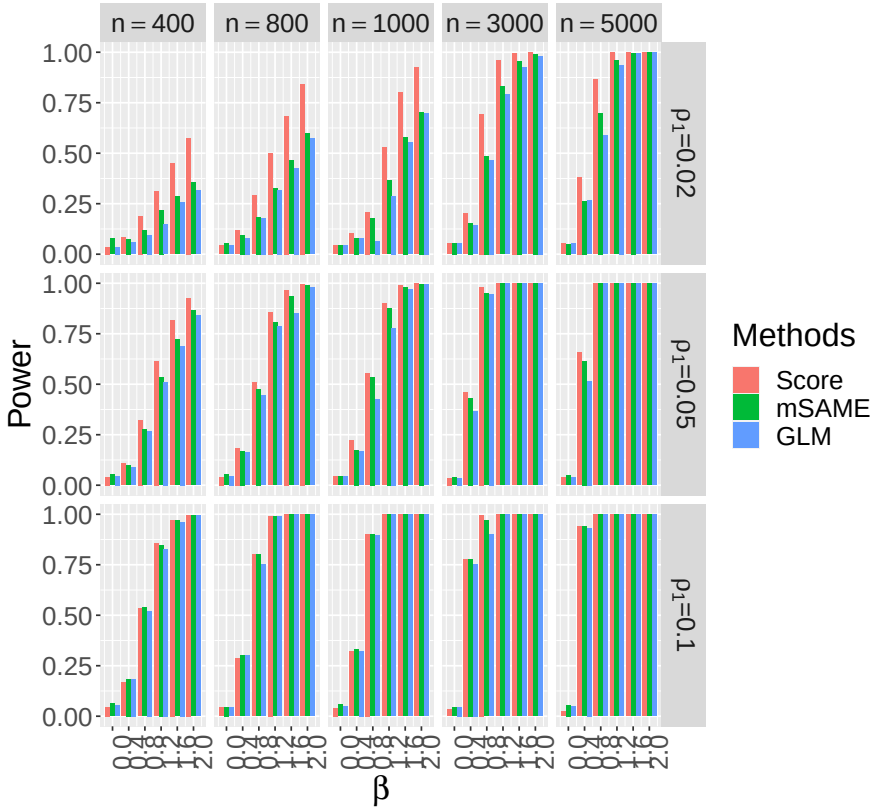

Figure 5.2: Comparison of type I error and power for single-mutation analysis of our developed score test (red bars), the mSAME test (green bars) and GLM (blue bars) with various rates of mutation frequency ρ_1 and effect size β. The comparison is based on the sample sizes (n).

ever, the read-depth of a somatic mutation was simulated by a negative binomial distribution with mean 40 and over-dispersion 1.9.

Applying this setup, the type I error is well-controlled by all methods, but in terms of power, decreasing the read-depth values affects the GLM approach much more than our score method and the mSAME procedure. This is because the GLM does not consider the error of the somatic mutation calling procedure, and the observed call O_i needs high coverage to be called. For instance, as displayed in Figure 5.3, for somatic mutation frequency, $\rho_1 = 0.1$, and effect size, $\beta = 1.2$, the power of GLM is above 82% power for high read-depth data, but it does not reach 72% power when the read-depth reduces. Furthermore, the GLM loses around 12% power when the somatic mutation rate, $\rho_1 = 0.05$ and effect size, $\beta = 1.6$ and $\beta = 2$.

Our developed score method obtains the highest power even though reducing the mutations read-depth has an effect on its power. A second reduction that might influence the methods' achievement is diminishing the mutation calling accuracy.

Somatic mutation calling accuracy and error rates

To examine the association methods' performance under different situations, we set two datasets of a sample size of $n = 400$ comparatively to the main simulation setup. However, in these cases, we decreased the mutation calling accuracy. In the first case, we lower only the specificity value, i.e., the observed mutation O_i was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and

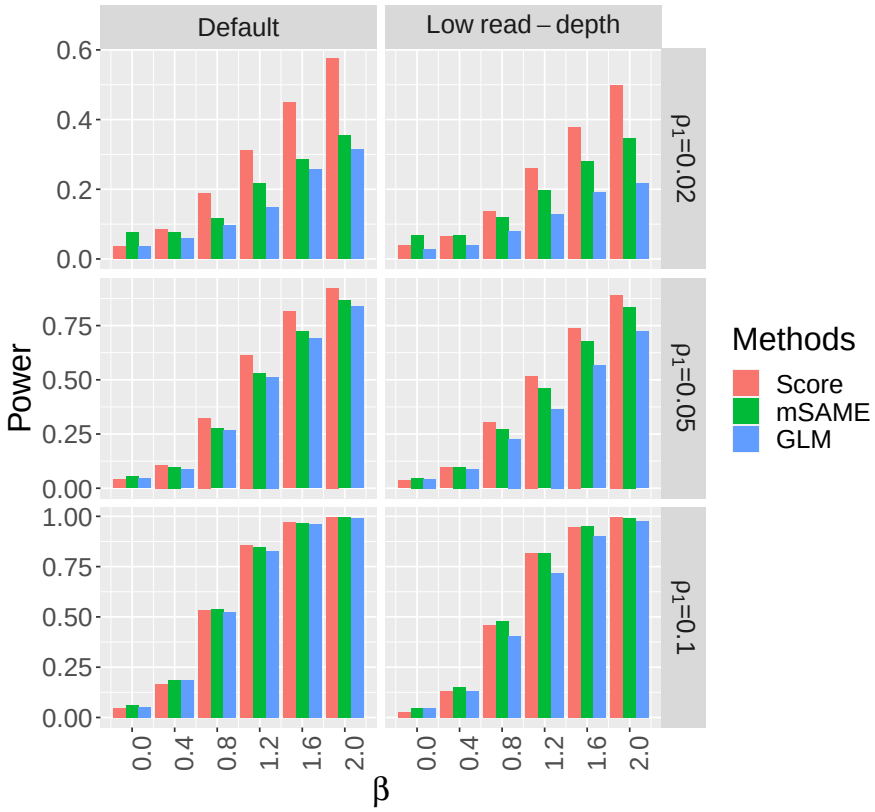

Figure 5.3: Comparison of type I error and power for single-mutation analysis of our developed score test (red bars), the mSAME test (green bars) and GLM (blue bars) with various rates of mutation frequency ρ_1 and effect size β. The comparison is based on the read-depth values of somatic mutations. In the default case (the main simulation setup), the mean of the somatic mutation read-depth was simulated by a negative binomial distribution with mean $\mu = 113$ and over-dispersion 3.28, whereas, in the low-read depth case, the mean was set 40.

specificity value $\gamma_0 = 0.95$. In the second case, both the sensitivity and specificity values are lessened so that the observed mutation O_i was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.85$ and specificity value $\gamma_0 = 0.95$.

Based on these simulation setups, it can be evident in Figure 5.4 that the methods preserve the type I error. In reference to power, all of our method, the mSAME test and GLM, show robustness in their performances for the different cases. Our developed score test is still the most effective approach and acts better than the mSAME test and GLM in all scenarios. For example, when the false rate of somatic mutation calling increases, our developed score's power exceeds 91% level of power for mutation frequency, $\rho_1 = 0.05$ and effect size, $\beta = 2$. By contrast, the mSAME test and GLM have less than 85% and 83% powers, respectively. For low-frequency mutation, $\rho_1 = 0.02$, our developed method has higher power in the case of the somatic mutation calling is defective than the mSAME test and GLM when the somatic mutation calling technique is accurate. It is achievable by using our developed score test to have beyond 55% power for inexact mutation calling procedure. On the other side, the mSAME test and GLM have less than 36% and 32% power even with a high certainty level of mutation calling.

Summary

Several simulation arrangements were constructed with the intention of analysing the performance of our proposed score method and compare it to a compatible test (mSAME) that accounts for mutation calling uncertainty. In addition to opposing our proposed approach to the mSAME test, it is a good idea to correspond to

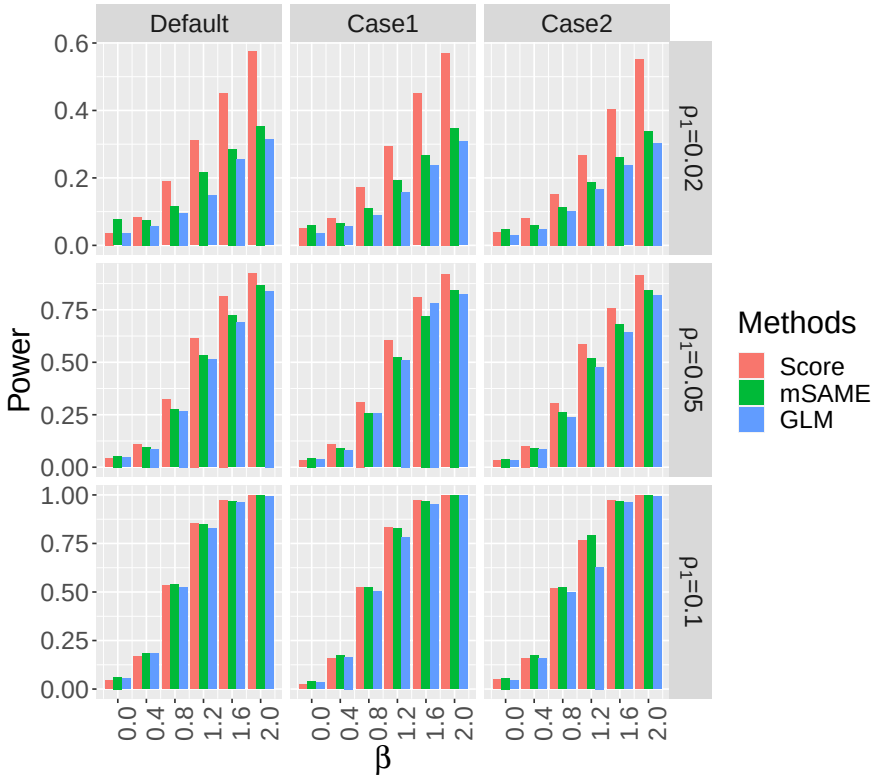

Figure 5.4: Comparison of type I error and power for single-mutation analysis of our developed score test (red bars), the mSAME test (green bars) and GLM (blue bars) with various rates of mutation frequency ρ_1 and effect size β. The comparison is based on the somatic mutation calling accuracy. In the default case (the main simulation setup), the sensitivity and specificity values are set, $\gamma_1 = 0.9$ $\gamma_0 = 0.98$, respectively. In case1, the sensitivity value is remaining as in the default setting, $\gamma_1 = 0.9$,but the specificity value is decreased to be, $\gamma_0 = 0.95$. In case2, both values are decreased so that the sensitivity value, $\gamma_1 = 0.85$, and specificity value, $\gamma_0 = 0.95$.

the commonly used GLM approach, which does not care about somatic mutation calling, and it treats the observed mutation O_i as true mutation.

According to the simulation results, considering the somatic mutation calling uncertainty can advance the association testing methods even if the read-depth is high, and the mutation calling method is accurate. The techniques' ability was assessed through varying sample sizes. It was demonstrated that when the mutation frequency $\rho_1 = 0.1$, our score test and the mSAME test have relatively similar performance, and they have better work than the GLM. When the mutation frequency becomes low (0.05 - 0.02), this affects the mSAME test and GLM performance much more than our developed score. We finally set datasets at low read-depth and less exact mutation calling to evaluate the methods under different circumstances. Our proposed score method has better performance than the mSAME test and GLM in both of the cases.

5.5 Evaluating the association of multiple somatic mutations using the single analysis of association tests

It is believed that some disease outcomes are linked or induced by multiple genetic markers rather than a single marker. In order to evaluate our proposed single test and compare its performance to the mSAME and GLM methods in light of this study framework, we generated a dataset of a sample size of $n = 400$.

We assume that 10 somatic mutation markers are expected to provoke a cancer subtype outcome. The dataset was made similar to the main simulation setup, so the actual somatic mutation status S_i, mutation calls O_i, read-depth D_i and alternative reads A_i of the ith individual were simulated in the same way of the main simulation setting. It is supposed that all of the 10 somatic mutations have the same effect size. We applied our single score test, adopted the Bonferroni procedure for multiple testing correction, and compared its performance to the mSAME and GLM methods corrected by the Bonferroni correction.

In this simulation frame, as shown in Figure 5.5, it is within the bounds of possibility to obtain more than 30% power by using our combined single score tests for somatic mutation frequency, $\rho_1 = 0.1$ when $\beta \geq 1.6$. In comparison, the combined mSAME tests need $\beta = 2$ to produce 30% power. The GLM does not reach 30% even with large effect size, $\beta = 2$. The methods perform insufficiently when the mutation frequency declines. Our combined score tests' power is 0.298, and the mSAME tests and GLM get powers of 0.297 and 0.0.268, respectively, for mutation frequency, $\rho_1 = 0.05$. All of the approaches produce less than 10% power for $\rho_1 = 0.02$.

By doubling the sample size to be $n = 800$, Figure 5.6 displays that all of the approaches extend their power. The increase rate of power of using our proposed method is 56%, and it is 57% by using the mSAME test and GLM when the somatic mutation frequency, $\rho_1 = 0.1$ and effect size, $\beta = 2$. It was not accomplishable at a sample size of $n = 400$ to approach a level of 10% power for low-frequency mutation, $\rho_1 = 0.02$ regardless of the effect size β; however, by increasing the

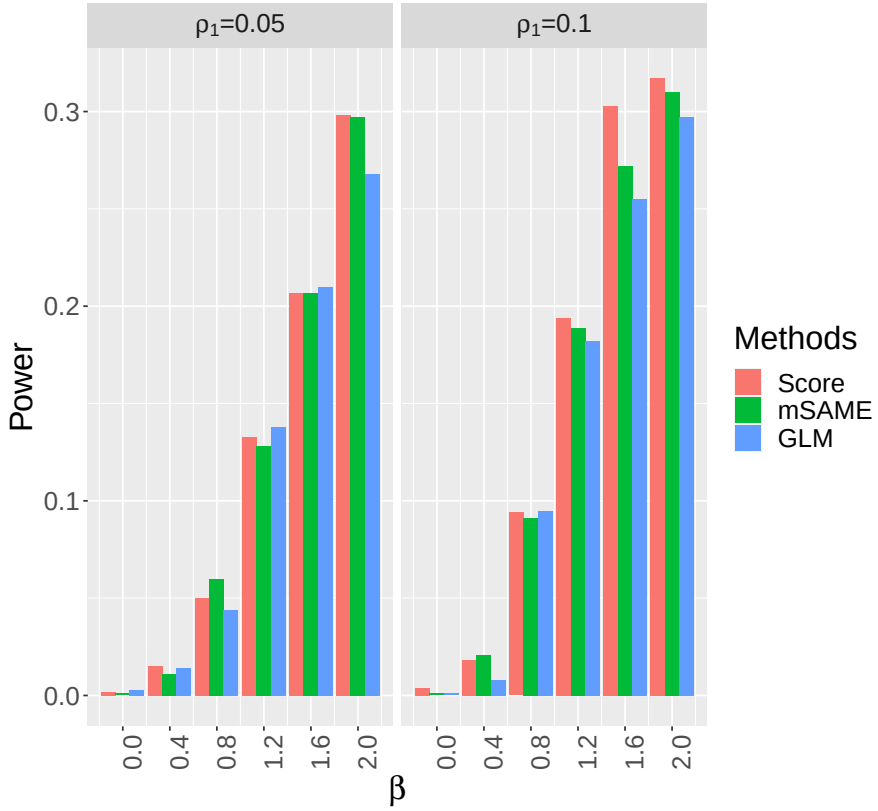

Figure 5.5: Comparison of type I error and power for our developed single score tests (red bars), the mSAME tests (green bars) and GLM (blue bars) corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400 and number of mutation=10.

sample size to be $n = 800$, all of the approaches can develop and attain 40% power.

In addition to expanding the sample size, detracting the targeted mutations can result in satisfactoriness in the methods' performance. We set a dataset of a sample size of $n = 400$ and reduce the number of mutations to be 5. What can be clearly seen in Figure 5.7 is the growth of the ability in all of the approaches. To be specific, for somatic mutation frequency, $\rho_1 = 0.1$ and $\beta = 1.6$, the powers of our combined score tests and the mSAME tests rise from 30% and 27%, respectively, to reach the level of 70%. Also, the GLM's power increase by 42% when the number of mutations shrinks. For somatic mutation frequency, $\rho_1 = 0.05$ and $\beta = 2$, using our score method and the mSAME test can lead to 60% power while the GLM obtains 0.462 power. However, it is still not likely to receive more than 20% power when the somatic mutation frequency is low, $\rho_1 = 0.02$.

In summary, considering each of the single mutations to assess the association of multiple somatic mutation markers and a cancer subtype outcome does not lead to good results even with high-frequency mutation, $\rho_1 = 0.1$ with a sample size of $n = 400$. In spite of the fact that doubling the sample size or reducing the number of targeted mutations, which are thought to be associated with a cancer trait-related outcome, can improve the performance of our combined score tests, the mSAME and GLM methods, the approaches are still less effective for low-frequency mutation and small effect size. These limitations and challenges motivate developing a set-based association approach to deal with a whole genetic set instead of dealing with every genetic marker.

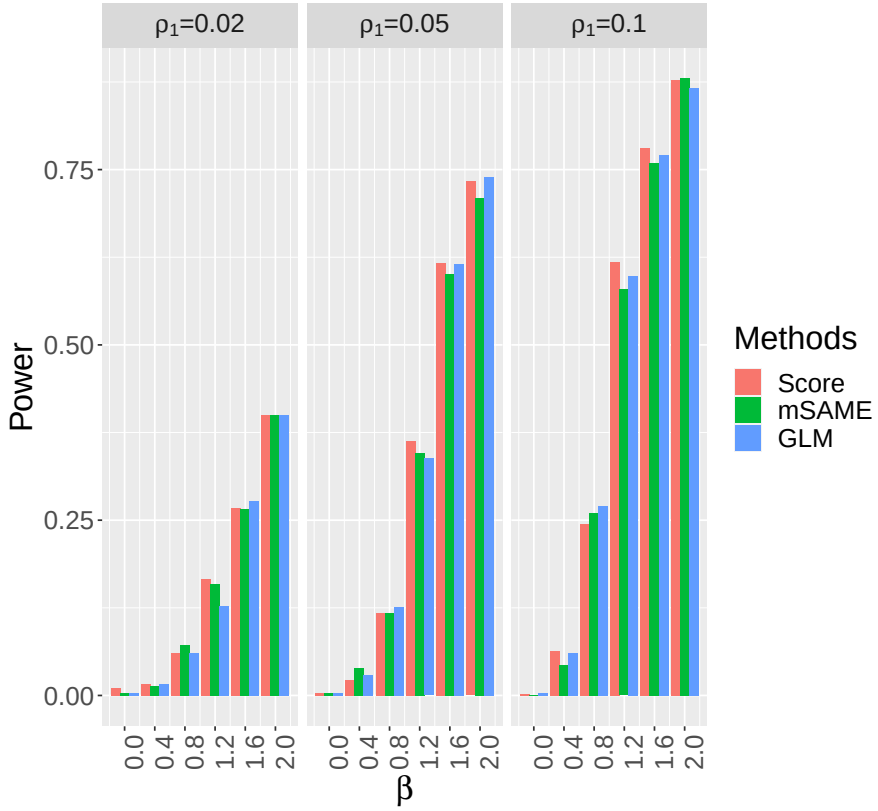

Figure 5.6: Comparison of type I error and power for our developed single score tests (red bars), the mSAME tests (green bars) and GLM (blue bars) corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=800 and number of mutation=10.

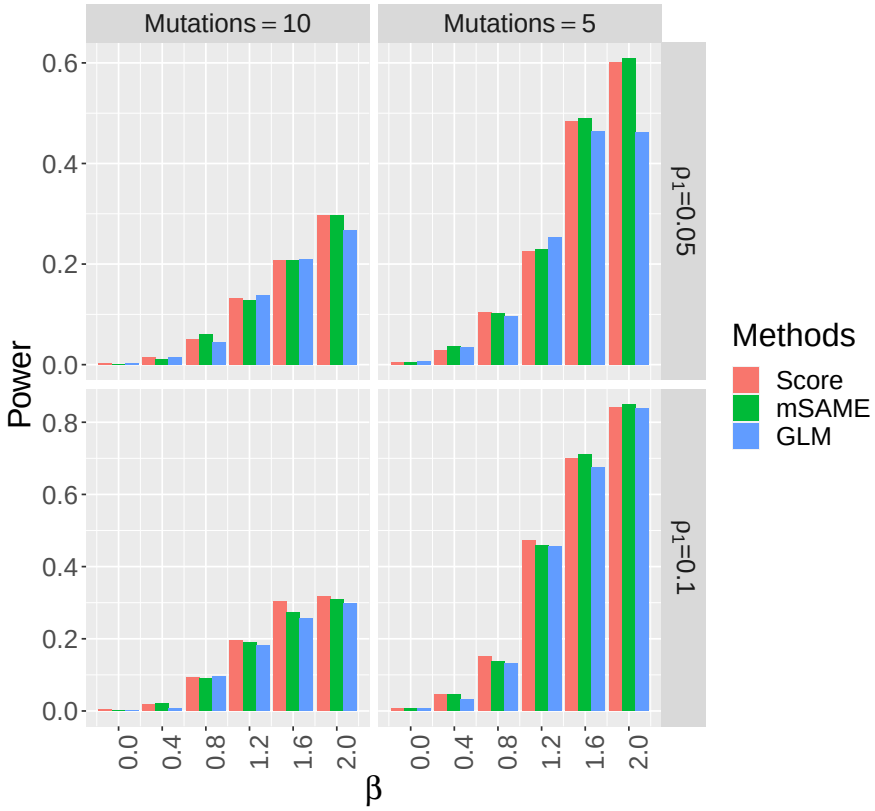

Figure 5.7: Comparison of type I error and power for our developed single score tests (red bars), the mSAME tests (green bars) and GLM (blue bars) corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400 and number of mutations=10 and 5.

5.6 Extension of a single-mutation method to a gene-based setting

Reviewing the association of a single mutation and a trait outcome by utilising single-variant association tests can be efficient for working with common variants denoted by variant allele frequency (VAF $> 5\%$). In addition to this, single-variant association tests are possibly powerful for low-frequency variants when the sample size is large enough. However, they become less powerful when mutations frequencies are rare (Madsen and Browning, 2009).

Since somatic mutations are considered low-frequency and rare, it might be more appropriate to consider a gene-based association analysis rather than studying a single somatic mutation. Grouping mutations into sets (such as a gene) combines the effects of multiple mutations to increase power and reduce the number of tests performed in the whole -genome study.

5.7 Discussion

Understanding the association between somatic mutations and cancer subtype outcomes is an essential procedure of cancer treatment. In view of the fact that the frequency of somatic mutations is sparse, they are arduous to be indubitably identified. On account of this, taking the uncertainty of the somatic mutation calling procedure into consideration can produce computational power. In this chapter, we

developed a score test for a single-mutation framework to test the relationship between a single somatic mutation and cancer subtype outcome. The developed score method was compared to the mSAME test, which uses the likelihood ratio test, and the GLM through simulation studies under different scenarios. The simulation results revealed that all of the approaches preserved the type I error, and in terms of power, our developed score test performed better than the mSAME test and GLM. At the end of this chapter, the ability of our proposed single-mutation score method was evaluated based on the idea of exploring the association of multiple somatic mutations and a cancer subtype outcome using multiple testing correction and compared to the performance of the mSAME test and GLM. It was disclosed that applying the single association tests is unproductive when the sample size is not large enough, and it is inappropriate for low-frequency mutations. Instead, it is suggested to scrutinise the association of multiple somatic mutations within a complete genetic unit such as a gene.

Chapter 6

Association analysis of gene-based somatic mutations and a cancer subtype outcome

Considering somatic mutations are assumed to be low-frequency and rare mutations, examining a single-mutation association might be infeasible. Therefore, appraising the whole genetic unit, such as a gene, becomes more desirable. In this chapter, we introduce a model for investigating the relationship between gene-based somatic mutations and a cancer subtype outcome that takes the uncertainty of somatic mutation calling into consideration. We evaluate our gene-based score test by comparing its performance to the gSAME test and GLM. Simulation studies on type I error and power for a wide range of scenarios are provided.

6.1 Introduction

In view of the fact that a substantial portion of heritability of a large number of diseases is now presumed to be due to low-frequency and rare genetic variants, several approaches have been proposed in an effort to study and analyse the low-frequency and rare mutations seeking to detect the association between an entire genetic set such as genes, gene networks or pathways and disease outcomes. Association analysis testing approaches for rare and low-frequency variants, such as burden tests and variance-component tests, can be applied to identify the effect of somatic mutations on cancer subtype outcomes. However, due to the challenge and difficulty in calling the somatic mutations confidently, it is vital to consider the uncertainty of somatic mutation calling.

None of the set-based association methods considers the uncertainty of the mutation calling; subsequently, this makes them inappropriate to investigate somatic mutations. In contrast, the gSAME test is an extended test from the single-mutation test mSAME (Liu et al., 2018) and accounts for the error in the somatic mutation calling process. We extend our proposed score test of a single-mutation to gene-based form analysis. Our gene-based score test collects the genetic information of somatic mutations within a gene to examine their effect on a cancer subtype outcome considering the somatic mutation calling uncertainty. We compare our gene-based score test performance to the gSAME test and GLM for the evaluation. A gene-based score test is developed and evaluated below.

6.2 A gene-based score test

Assume that there are j mutation markers in a gene, and for the sake of notational clarification, we use ^{m} and ^{g} to refer to the settings of mutation-level and gene-level data. The actual somatic mutation status and the mutation calls for the ith individual are denoted by $S_i^m = (S_{i1}^m, \ldots, S_{ij}^m)$ and $O_i^m = (O_{i1}^m, \ldots, O_{ij}^m)$, respectively, $i = 1, 2, \ldots, n$. Also, the read-depth and the number of alternative reads are denoted by $D_i^m = (D_{i1}^m, \ldots, D_{ij}^m)$ and $A_i^m = (A_{i1}^m, \ldots, A_{ij}^m)$, respectively. The gene-level mutation is denoted by S_i^g and can be equal to 1 if there is at least one mutation within the gene or 0 if there is no mutation in that gene.

$$S_i^g = \begin{cases} 1 & \text{if any } S_{ix}^m = 1, \\ 0 & \text{if all } S_{ix}^m = 0. \end{cases}$$

The outcome of cancer subtype and all covariates are indicated as in the single somatic mutation analysis by Y_i and the vector C_i, respectively. Finally let $\rho_0^g = P(S_i^g = 0)$ denote the probability that the targeted gene in the ith individual does not have any somatic mutation, and let $\rho_1^g = P(S_i^g = 1) = 1 - \rho_0^g$ denote the probability that the targeted gene contains at least one somatic mutation.

6.2.1 The likelihood function

As the actual somatic mutation status for a gene S_i^g is not observed and ignoring the covariates C_i for the time being, the probability for the ith individual

$P(Y_i, A_i^m, D_i^m, O_i^m, S_i^g)$ can be written as the sum over all possible values,

$$\sum_{x=0}^{1} P(S_i^g = x)P(Y_i, A_i^m, D_i^m, O_i^m \mid S_i^g) = \sum_{x=0}^{1} \rho_x^g P(Y_i, A_i^m, D_i^m, O_i^m \mid S_i^g). \quad (6.1)$$

The probability $P(Y_i, A_i^m, D_i^m, O_i^m \mid S_i^g)$ can be decomposed as in the single somatic mutation analysis

$$P(Y_i, A_i^m, D_i^m, O_i^m \mid S_i^g) = P(Y_i \mid S_i^g)P(O_i^m \mid Y_i, S_i^g)P(A_i^m, D_i^m \mid Y_i, O_i^m, S_i^g), \quad (6.2)$$

and recall that given S_i^g, we assume Y_i carries no additional information about O_i^m and has no more additional information about A_i^m and D_i^m. In the gene-based setting, the outcome Y_i is modelled as a function of S_i^g, so the likelihood function for the ith individual can be given by

$$L_i = \sum_{x=0}^{1} \rho_x^g f(y_i \mid S_i^g = x)f(o_i^m \mid S_i^g = x)f(a_i^m, d_i^m \mid O_i^m, S_i^g = x). \quad (6.3)$$

Because the data of read-depth D_i^m, alternative reads A_i^m and observed calls O_i^m are obtained for each mutation, their distributions can be modelled given S_i^m as in the single-mutation analysis, and we need to model S_i^m conditional on S_i^g. Therefore, the equation (6.3) can be written as

$$L_i = \sum_{x=0}^{1} \rho_x^g f(y_i \mid S_i^g = x)P(S_i^m \mid S_i^g = x)f(o_i^m \mid S_i^m = x)f(a_i^m, d_i^m \mid O_i^m, S_i^m = x), \quad (6.4)$$

where $f(y_i \mid S_i^g), f(o_i^m \mid S_i^m)$ and $f(a_i^m, d_i^m \mid O_i^m, S_i^m = x)$ are probability functions characterising the conditional probabilities $P(Y_i \mid S_i^g), P(O_i^m \mid S_i^m)$ and $P(A_i^m, D_i^m \mid O_i^m, S_i^m)$.

It is obvious that when $S_i^g = 0$, $S_{ix}^m = 0$ for all j mutations within the gene. However, when $S_i^g = 1$, this indicates that S_i^m has $2^j - 1$ potential values. This procedure means that it is computationally hard when j is large. In genetics, it is believed that to call a somatic mutation for a genetic marker, its alternative read should be larger than 0, i.e., $A_{ix}^m > 0$, which is reasonable. When $A_{ix}^m = 0$, S_{ix}^m is set to be 0 directly, and this assumption helps to reduce the computational difficulty of the possible combinations. Consequently, the number of combinations can be minimised to $2^{j'} - 1$ where j' is the number of mutations with $A_{ix}^m > 0$. For a specific combination, the tth combination, in the ith individual, the true mutation status are $s_i^t = s_{i1}^t, \ldots, s_{ij}^t$ where $t = 1, \ldots, 2^{j'} - 1$, we have

$$P(S_i^m = s_i^t \mid S_i^g = 1) = \frac{P(S_i^m = s_i^t, S_i^g = 1)}{\sum_{l=1}^{2^{j'}-1} P(S_i^m = s_i^l)} = \frac{\delta_{it}}{\sum_{l=1}^{2^{j'}-1} \delta_{il}}, \tag{6.5}$$

where

$$\delta_{it} = P(S_i^m = s_i^t) = \prod_{x=1}^{j} w_x^{s_{ix}^t} (1 - w_x)^{1 - s_{ix}^t}, \tag{6.6}$$

and $w_x = P(S_{ix}^m = 1)$ and can be estimated through the observed frequency of the xth mutation across all samples, or in an external reference population.

Given the above models, the likelihood function for the ith individual in equation (6.4) can be re-written as a summation of two likelihood functions L_{i0} and

L_{i1} when $S_i^g = 0$ and $S_i^g = 1$, respectively,

$$L_i = L_{i0} + L_{i1}, \tag{6.7}$$

where

$$L_{i0} = \rho_0^g f(y_i \mid S_i^g = 0) \prod_{x=1}^{j} f(o_{ix}^m \mid S_{ix}^m = 0) f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m = 0), \tag{6.8}$$

and

$$L_{i1} = \rho_1^g f(y_i \mid S_i^g = 1) \sum_{t=1}^{2^{j'}-1} \delta_{it}^* \prod_{x=1}^{j} f(o_{ix}^m \mid S_{ix}^m = s_{ix}^t) f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m = s_{ix}^t), \tag{6.9}$$

and $\delta_{it}^* = P(S_i^m = s_i^t \mid S_i^g = 1)$. The outcome Y_i given the somatic mutation status of the entire gene S_i^g can be modelled by a generalised linear model as in the single somatic analysis with mean $E(Y_i) = g^{-1}(\boldsymbol{\beta_0}\mathbf{C}_i^T + \beta S_i^g)$, for some canonical link function g. Here $\boldsymbol{\beta_0}$ and β are the regression coefficients, and primary interest is inference regarding the gene-based effect size parameter β. For a continuous outcome, $f(y_i \mid S_i^g)$ can be replaced by a normal density function, and for a binary outcome by a Bernoulli density.

6.2.2 The score and Fisher Information for binary outcomes

Recalling that we assume the other parameters in the model given by equation (6.7) do not affect the variance of the score statistic and thus we only require the first and second derivatives of the log-likelihood of L_i with respect to β. We also assume there are no covariates in the model. For notational convenience in the derivation we write $f_0(y_i) = f(y_i \mid S_i^g = 0)$ and in similar way $f_1(y_i) = f(y_i \mid S_i^g = 1)$. Also, we write

$$\Delta_{i0} = \rho_0^g \prod_{x=1}^{j} f(o_{ix}^m \mid S_{ix}^m = 0) f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m = 0),$$

and

$$\Delta_{i1} = \rho_1^g \sum_{t=1}^{2^{j'}-1} \delta_{it}^* \prod_{x=1}^{j} f(o_{ix}^m \mid S_{ix}^m = s_{ix}^t) f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m = s_{ix}^t).$$

Then the score function for β for the ith individual and assuming a binary outcome is given by

$$\begin{aligned} u_i = \frac{\partial \log L_i}{\partial \beta} &= \frac{\frac{\partial}{\partial \beta}(L_{i1})}{L_{i0} + L_{i1}} \\ &= \frac{\frac{\partial}{\partial \beta}\Delta_{i1} f_1(y_i)}{\Delta_{i0} f_0(y_i) + \Delta_{i1} f_1(y_i)}. \end{aligned}$$

The function $f_1(y_i) = p_{i1}^{y_i}(1-p_{i1})^{1-y_i}$ for logit $p_{i1} = \beta_0 + \beta$, so the partial derivative can be written

$$\frac{\partial}{\partial \beta}\Delta_{i1} f_1(y_i) = \Delta_{i1} f_1(y_i)(y_i - p_{i1}), \tag{6.10}$$

as

$$\frac{\partial f_1(y_i)}{\partial \beta} = f_1(y_i)(y_i - p_{i1}). \tag{6.11}$$

The score function for the ith individual can be written

$$u_i = \frac{\Delta_{i1} f_1(y_i)(y_i - p_{i1})}{\Delta_{i0} f_0(y_i) + \Delta_{i1} f_1(y_i)}. \tag{6.12}$$

The score test is evaluated under the null hypothesis $H_0 : \beta = 0$. In this case,

$$f_1(y_i) = p_{i1}^{y_i}(1 - p_{i1})^{1-y_i} = \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0}}\right)^{1-y_i} = \frac{e^{\beta_0 y_i}}{1 + e^{\beta_0}}, \tag{6.13}$$

and the score function given by equation (6.12) evaluates to

$$\begin{aligned} u_i &= \frac{\Delta_{i1}\left(\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}\right)\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\Delta_{i0}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}} + \Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}} \\ &= \frac{\Delta_{i1}\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\Delta_{i0} + \Delta_{i1}}. \end{aligned} \tag{6.14}$$

The second derivative of $\log L_i$ is

$$\frac{\partial u_i}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\frac{\Delta_{i1} f_1(y_i)(y_i - p_{i1})}{\Delta_{i0} f_0(y_i) + \Delta_{i1} f_1(y_i)}\right), \tag{6.15}$$

and this gives the observed Fisher Information

$$I_i = \frac{1}{\left(\Delta_{i0} f_0(y_i) + \Delta_{i1} f_1(y_i)\right)^2} \Bigg[\left(\Delta_{i1} f_1(y_i)(y_i - p_{i1})\right)^2 - \left(\Delta_{i1} f_1(y_i)\left[(y_i - p_{i1})^2 - p_{i1}(1-p_{i1})\right]\right)\left(\Delta_{i0} f_0(y_i) + \Delta_{i1} f_1(y_i)\right)\Bigg]. \tag{6.16}$$

The observed Fisher Information evaluated under the null becomes

$$I_i = \frac{1}{\left(\Delta_{i0}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}} + \Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}\right)^2} \Bigg[\left(\Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\right)^2 - \left(\Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}\left[\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^2 - \frac{e^{\beta_0}}{(1+e^{\beta_0})^2}\right]\right)\left(\Delta_{i0}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}} + \Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}\right)\Bigg]. \tag{6.17}$$

The expected Fisher Information under the null is given by

$$\begin{aligned} E(I_i) &= \left(\frac{\Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}}{\Delta_{i0}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}} + \Delta_{i1}\frac{e^{\beta_0 y_i}}{1+e^{\beta_0}}}\right)^2 \frac{e^{\beta_0}}{(1+e^{\beta_0})^2} \\ &= \left(\frac{\Delta_{i1}}{\Delta_{i0}+\Delta_{i1}}\right)^2 \frac{e^{\beta_0}}{(1+e^{\beta_0})^2}. \end{aligned} \tag{6.18}$$

Finally, the score test statistic

$$V = \frac{(\sum_{i=1}^n u_i)^2}{\sum_{i=1}^n I_i} \sim \chi_1^2. \tag{6.19}$$

The performance of the gene-based score test is appraised and compared, in terms of type I error and power, to the gSAME test that was developed in (Liu et al., 2018) and the GLM. The gene-based score test and gSAME test account for the somatic mutation calling error. On the other side, the GLM ignores the uncertainty in observing the somatic mutations.

6.3 Parameter estimation

Since the read-depth D_i^m, alternative reads A_i^m and observed calls O_i^m are calculated for each single somatic mutation, the conditional density of the ith individual for the xth mutation that is given by equation (6.4), $f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m)$, can be modelled as in the same method as in the single somatic mutation analysis. The conditional density $f(a_{ix}^m, d_{ix}^m \mid O_{ix}^m, S_{ix}^m)$ can be written as $f(a_{ix}^m \mid O_{ix}^m, S_{ix}^m, D_{ix}^m) f(d_{ix}^m \mid S_{ix}^m)$, and the term $f(d_{ix}^m \mid S_{ix}^m)$ can be ignored in the estimation as D_{ix}^m does not depend on S_{ix}^m. When there is enough coverage for the xth somatic mutation ($D_{ix}^m \geq D_0$), the term $f(a_{ix}^m \mid O_{ix}^m, S_{ix}^m, D_{ix}^m)$ is modelled by using beta-binomial distributions as given by equation (5.15), and the parameters $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \varphi_{00}, \varphi_{01}, \varphi_{10}$ and φ_{11} can be pre-estimated in a similar manner to the single-mutation analysis. On the other hand, When there is not enough coverage for the xth somatic mutation ($D_{ix}^m \geq D_0$), the mutation call of the xth mutation O_{ix} is not observed, and the term $f(a_{ix}^m \mid S_{ix}^m, D_{ix}^m)$ is modelled by using equation (5.17). In the gene-level analysis, there are plenty of low-coverage data, so to reduce the complexity of the model, the parameters π_0, π_1, φ_0 and φ_1 are

estimated by pooling together all the low read-depth data of all individuals and mutations within the gene set and fitting a mixture of two beta-binomial distributions. The conditional density $f(o_{ix}^m \mid S_{ix}^m)$ can be modelled by using Bernoulli distributions as in equation (5.18), and the parameters γ_0 and γ_1 are pre-determined as in the single analysis.

Now, the parameters ρ_0^g, β_0 and β can be estimated using the EM algorithm. Let $\boldsymbol{\theta}^{(t)}$ be the current estimate of the parameters, the density of S_i^g conditional on observed data is given

$$\eta_{i0} = \frac{\rho_0^g f(y_i, a_i^m, d_i^m, o_i^m \mid S_i^g = 0, \boldsymbol{\theta}^{(t)})}{\sum_{x=0}^{1} \rho_x^g f(y_i, a_i^m, d_i^m, o_i^m \mid S_i^g = x, \boldsymbol{\theta}^{(t)})}. \tag{6.20}$$

The parameters ρ_0^g, β_0, β are updated using functions (5.20)-(5.22) in the single-mutation association analysis. Under the null hypothesis, we can estimate β_0 using the maximum likelihood estimator under the null model and use the EM algorithm in order to estimate ρ_0^g.

6.4 Simulation studies and results

In this section, various simulation setups were designed to examine our developed gene-based score test performance and compare it to the gSAME test that was developed in (Liu et al., 2018) and the GLM in terms of type I error and power. The proposed gene-based score and gSAME tests account for the uncertainty of calling the somatic mutations within a gene, whereas the GLM does not consider the uncertainty and deals with the observed mutation as actual status.

The evaluation of the performance of tests was made $1,000$ replications at significance level 0.05 with different settings of gene-based somatic mutation frequencies ρ_1^g. In the single somatic analysis, the mutation-level frequencies ρ_1 were considered to be 0.02, 0.05 or 0.1. The gene-based mutation frequencies are usually higher than mutation-level frequencies, so we set $\rho_1^g = 0.05, 0.1$ or 0.15. In terms of evaluating type I error, it was done by simulating the data under the null hypothesis (β=0). In terms of evaluating the power of the tests, we set $\beta = 0.4, 0.8, 1.2, 1.6, 2.0$.

The study aim in the gene-based association analysis is to test the relationship between a disease subtype outcome Y_i and a gene-based mutation S_i^g. The frequency of gene-based mutation is indicated as $P(S_i^g = 1) = \rho_1^g$. The true mutation statuses for a single mutation in the ith individual S_{ix}^m, $x = 1, \ldots, j$ were generated independently by a Bernoulli distribution with probability of success $P(S_{ix}^m = 1) = 1 - (1 - \rho_1^g)^{1/j}$. By collapsing the single-mutation data S_{ix}^m, we can obtain the gene-based mutation S_i^g. The variable S_i^g is set to be 1 if a gene has at least one mutation, and it is set 0 if the gene does not contain any mutation. The binary outcome Y_i was simulated from a Bernoulli distribution, $\text{logit}[p(Y_i = 1)] = -0.5 + \beta S_i^g$.

The main simulation setup

Our main simulation framework was generated with a sample size of $n = 400$, and it is assumed that there are 10 somatic mutations within a gene, i.e., $j = 10$. For each of the somatic mutations, the xth mutation, the calculation of its read-

depth was done equivalently to the single-mutation analysis. The mean read-depth of the xth mutation D_{ix}^m was simulated by a negative binomial distribution with mean $\mu = 113$ and over-dispersion 3.28. Then, the read-depth of that xth mutation was simulated by a negative binomial distribution with the simulated mean and over-dispersion 1.9. Also, the alternative reads number A_{ix}^m and the observed mutation call O_{ix}^m were simulated by a beta-binomial distribution and Bernoulli distribution, respectively, as they were produced in the single-mutation analysis. If the xth mutation receives enough red-depth which means ($D_{ix}^m \geq D_0 = 20$), its alternative reads number, A_{ix}^m was simulated by a beta-binomial distribution with parameters $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.001, 0.002, 0.1179, 0.3207)$ and $(\varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}) = (0.0006, 0.3457, 0.0001, 0.1018)$, and O_{ix}^m was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and specificity value $\gamma_0 = 0.98$. In contrast, if the read-depth of the xth mutation is low, i.e., ($D_{ix}^m < D_0 = 20$), A_{ix}^m was simulated by a beta-binomial distribution with parameters $\pi_0 = 0.001, \varphi_0 = 0.001, \pi_1 = 0.146, \varphi_1 = 0.10$, and in this case, O_{ix}^m is not observed.

In this simulated dataset framework, all our proposed gene-based score method, the gSAME test and GLM, control the type I error. On the question of power, as sown in Figure 6.1, our developed gene-based score test has greater power than the gSAME test and GLM in all of the scenarios. Using our developed gene-based score leads to more than 78% power for gene-level mutation frequency, $\rho_1^g = 0.15$ and effect size, $\beta = 1.2$. In opposition, the gSAME test and GLM do not reach 64%. When $\beta = 1.6$, our method has 0.893 power while the power of the gSAME test and GLM are 0.821 and 0.822, respectively. For gene-based mutation, $\rho_1^g = 0.1$, our

proposed method's power exceeds 81% and 94% when the effect size, $\beta = 1.6$ and 2. Contrastingly, the gSAME test and GLM have smaller than 76% for $\beta = 2$. Even though all of the methods' power is not satisfactory for the low-frequency gene-level mutation in this sample size ($n = 400$), our method obtains much higher power than the gSAME test and GLM. For example, for low-frequency cases, $\rho_1^g = 0.05$ when $\beta = 2$, our score test has 0.681 power, but the gSAME test and GLM have 0.325 and 0.328, respectively. A simple way to increase power is to enlarge the sample size.

Sample size and error rates

In a similar fashion to the main simulation setup, datasets of sample sizes of $n = 800, 1000, 3000$ and 5000 were constructed. As Figure 6.2 illustrates, the type I error is preserved by all of the methods. In terms of power, when the base sample size is doubled ($n = 800$), the power of our proposed gene-based score method increases to reach 0.82 and 0.958 levels of power for low-frequency gene-based mutation, $\rho_1^g = 0.05$ and effect sizes, $\beta = 1.6$ and $\beta = 2$, respectively. By contrast, the gSAME test and GLM are still much lower than a satisfactory level, and their power rates are 0.424 and 0.425, respectively, for $\beta = 1.6$, and 0.555 and 0.559, respectively, for $\beta = 2$. The gSAME test and GLM are not sufficient for obtaining over 65% power for a low-frequency situation even at a sample size of $n = 1000$. At the same time, it is acquirable to pass this level of power (65%) by using our developed gene-level method even at a smaller sample size (n=400).

Testing the methods' performance at high sample size, $n = 3000$, 77% power

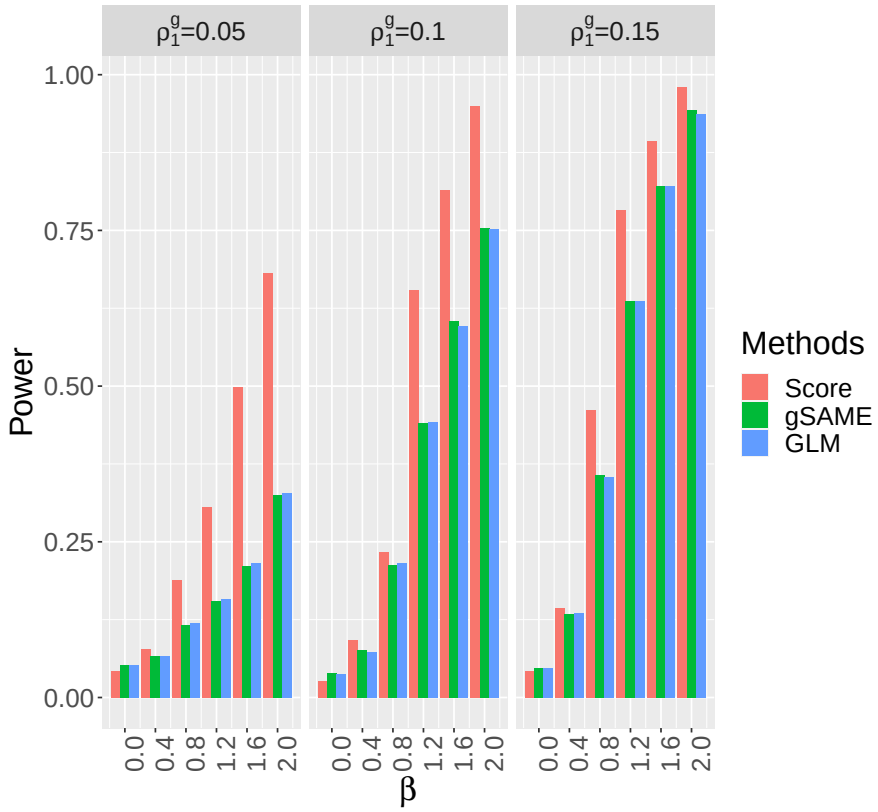

Figure 6.1: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. This setup is the main simulation model of the gene-based analysis, which is constructed of a sample size of $n = 400$ and number of mutations $j = 10$.

is accomplished by applying our method and not more than 62% utilising the gSAME test and GLM approach when the gene-level somatic mutation, $\rho_1^g = 0.15$ and $\beta = 0.4$. For low-frequency gene-level mutation, $\rho_1^g = 0.05$, our gene-based score test can gain more than 90% power when $\beta \geq 0.8$, whereas the gSAME test and GLM require $\beta = 2$ to amount to this rate of power. The gSAME test and GLM need higher β than our developed gene-based score method requires even with a larger sample size of $n = 5000$. Specifically, for low-frequency of gene-level mutation, $\rho_1^g = 0.05$, the gSAME test and GLM obtain 92% power when $\beta = 1.2$ while our gene-based score test gains 97% power with $\beta = 0.8$.

In conclusion, we evaluated the approaches with different sample sizes, and the difference in the power of the methods is more significant by using our developed score test. Our method has better performance than the gSAME test and GLM at all of the sample sizes. In the following simulation setup, we changed the number of mutations within a gene (j) to analyse the methods at varying possible scenarios.

Number of somatic mutations and error rates

In human genetics, somatic mutations are considered low-frequency variants; therefore, to expand the framework of testing the performance of our developed gene-based score test and compare it to the gSAME test, we set a dataset of a sample size of $n = 400$ with a reduced number of mutations ($j = 5$) per gene. For each of the single somatic mutations, the read-depth values, alternative reads, actual somatic mutation status and observed somatic mutation calls were simulated in

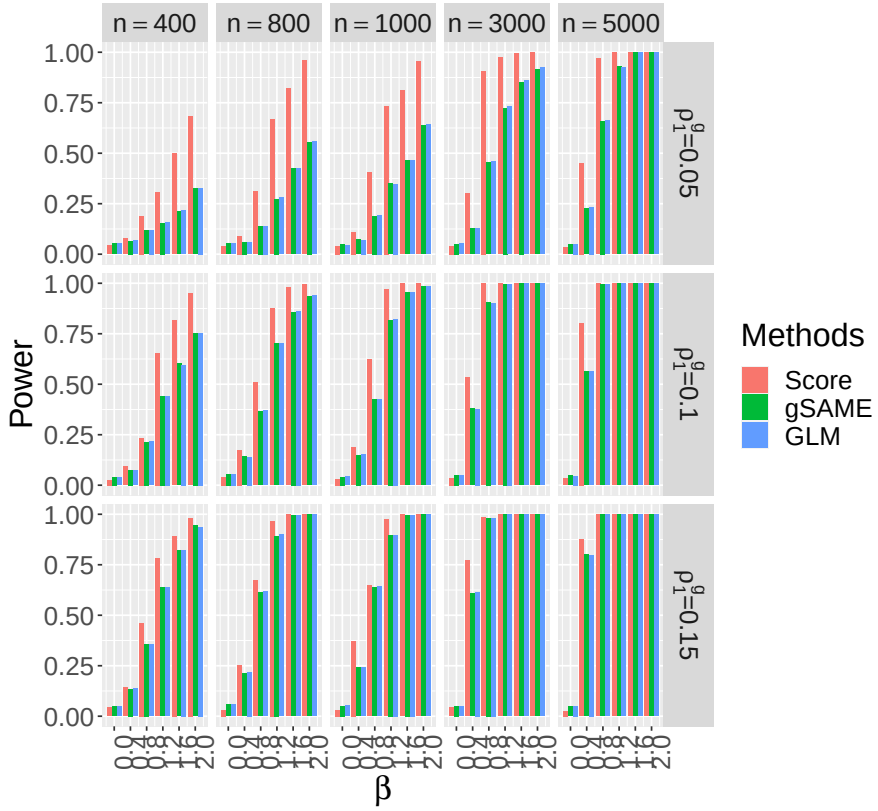

Figure 6.2: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. The comparison is based on the sample sizes (n).

the same prior process.

Figure 6.3 exhibits that the type I error is well-controlled by all of the procedures. With regard to power, shortening the number of variants within a gene makes an impact and difference on the gSAME test and GLM performance more considerable than it does on our developed gene-level score method. It can indicate that the gSAME test and GLM are more sensitive to increasing the number of somatic mutations within a genetic set. To give examples, for gene-level mutation, $\rho_1^g = 0.15$ and effect size, $\beta = 0.8$, the powers of gSAME test and GLM increase by 16.8% and 17.4%, respectively, whilst our gene-based method gain an increase of 6.8%. When $\beta = 1.2$, our method's increase power is 4.8%, but it is 18.6% and 18.2% for the gSAME test and GLM, respectively.

However, the gSAME test and GLM still have less power than our proposed method when the frequency of gene-level mutation decreases. For instance, even though the gSAME and GLM approaches obtain increases of 15% and 14.4% power, respectively, for $\rho_1^g = 0.05$ and effect size $\beta = 2$, they are less potent than our score test as their level of power is smaller than 0.48 while our method has 0.746 power.

Data quality and error rates

Recalling that, our developed gene-based method and the gSAME test account for the uncertainty of somatic mutation calling and deal with the true somatic mutation status as an unobservable variable. They use the read-depth and alternative

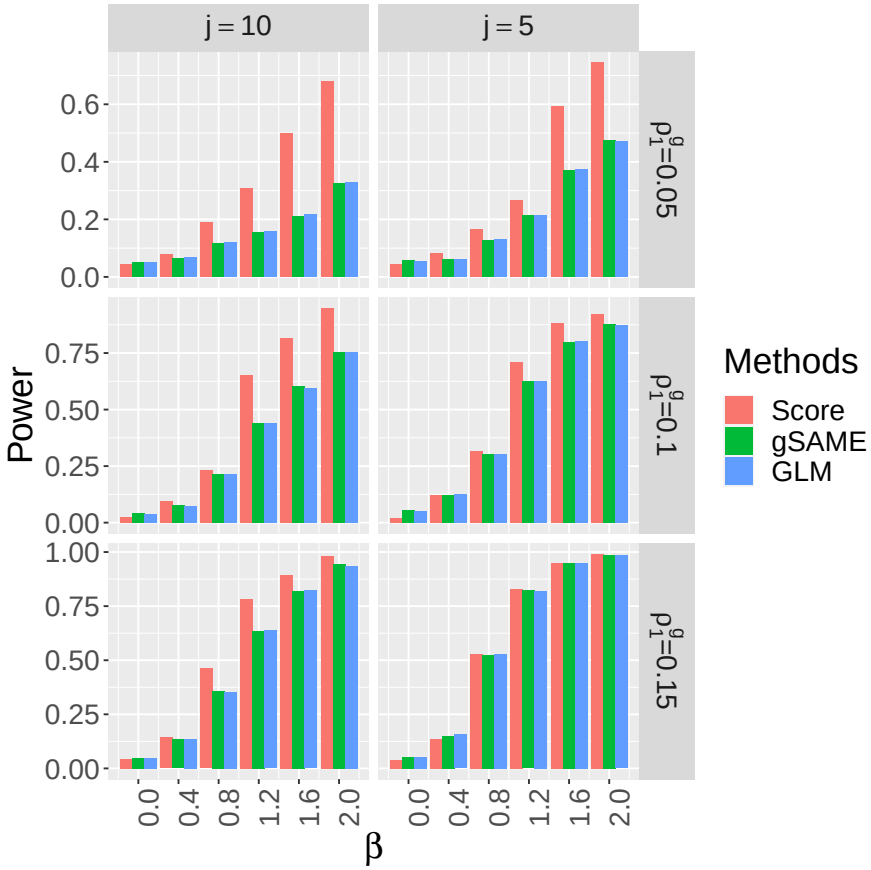

Figure 6.3: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. The comparison is based on the number of somatic mutations within a gene (j).

reads data to raise the probability of mutation calls. In the somatic mutation association analysis, the typical read-depth of a mutation for the whole genome sequencing can be 20x to 40x (Liu et al., 2018), so in our main simulation setup, the adopted threshold of read-depth (D_0) was chosen to be 20x. When a somatic mutation, the xth mutation, is covered, there are two potential cases. Firstly, if the read-depth of the xth mutation is less than ($D_0 = 20$), the observed mutation call O_{ix}^m is not observed. Otherwise, O_{ix}^m can be observed and modelled by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and specificity value $\gamma_0 = 0.98$.

In this simulation part, we aim to assess the performance of our proposed gene-level method and match it to the gSAME test and GLM in different levels of data quality. Datasets of a sample size of $n = 400$ and the number of mutations $j = 10$ were created similar to the main simulation setup with changing the read-depth threshold D_0 of a somatic mutation to be $10, 30$ and 40.

Figure 6.4 reveals that the procedures preserve the type I error in all of the conditions. Pertaining to power, our proposed score method performs much better based on these simulation setups than the gSAME test and GLM for high-quality data (when $D0 > 20$). Namely, for gene-based mutation frequency, $\rho_1^g = 0.15$ and effect size, $\beta = 1.2$, the power of our proposed method is under 70% for low-quality data ($D_0 = 10$), and it increases to become above 85% when $D_0 = 30$ and $D_0 = 40$. Conversely, the gSAME test and GLM obtain not more than 67% for all the scenarios. With the same gene-level mutation frequency when $\beta = 1.6$, our developed score increases by 10% and gets 92% power for high-quality samples while using the gSAME test and GLM not reach 80%. For a low-frequency muta-

tion and effect size, $\beta = 2$, our proposed method's power escalates from 0.557 with low-quality data to become 0.746 by using the high-quality data. By contrast, the gSAME test and GLM powers decline from 0.310 and 0.314 to be 0.295 and 0.296, respectively. This finding can signify that improving the quality of the data can help our score test perform better.

Since observing and modelling the observed somatic mutation call O_i^m depends on the selected read-depth threshold D_0, in the next simulation part, we tested the procedure's ability based on the model parameters of the observed mutation call.

Low somatic mutation calling accuracy and error rates

Another critical scenario of examining the implementation of the methods that should be studied is to reduce the accuracy of the somatic mutation calling procedure (e.g., changing the sensitivity value γ_1 and specificity value γ_0). Two cases were constructed. In the first scenario, we made a dataset of a sample size of $n = 400$ and the number of mutations $j = 10$ in an identical process to the primary simulation setting. However, in this case, we reduced only the specificity value, so the observed call of the xth mutation O_{ix}^m was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and specificity value $\gamma_0 = 0.95$. In the other scenario, we changed both the sensitivity and specificity values so that the observed mutation call of the xth mutation O_{ix}^m, in this case, was simulated by a Bernoulli distribution with $\gamma_1 = 0.85$ and $\gamma_0 = 0.95$.

From the investigation based on these two cases, all of our proposed score test,

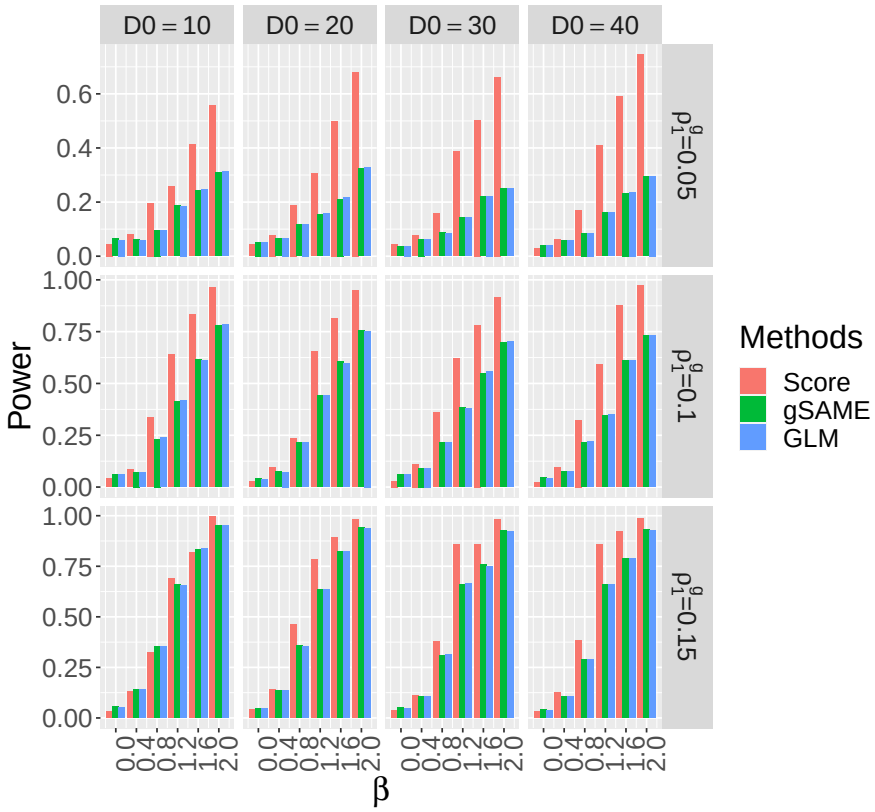

Figure 6.4: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. The comparison is based on the data quality, so the methods' performance is evaluated under different read-depth thresholds (D_0).

the gSAME test and GLM, control the type I error. In the matter of power, it is manifest as Figure 6.5 demonstrates that the impact of reducing the somatic mutation calling accuracy (by decreasing only the specificity value or both the sensitivity and specificity values) is considerably more significant on the gSAME test and GLM than our developed method. Our proposed gene-based score test is much more robust comparing to the gSAME test and GLM. By way of illustration, when the gene-level mutation frequency, $\rho_1^g = 0.15$ and effect size, $\beta = 2$, the powers of the gSAME test and GLM slump from 0.943 and 0.937 to become 0.675 and 0.677, respectively, when only the specificity value is decreased, and their abilities continue declining to become 0.595 and 0.592, respectively, when both the sensitivity and specificity values are reduced. On the other hand, our developed gene-based test obtains more than 95% power even after dropping the specificity value or both the sensitivity and specificity values. Reducing the somatic mutation calling accuracy obstructs the gSAME test and GLM to excel the 40% level of power for $\rho_1^g = 0.1$ and $\beta = 2$ while it is still likely to have more than 85% power by employing our developed gene-based method even if the somatic mutation calling is less precise. For the low-frequency gene-level mutation, $\rho_1^g = 0.05$, the powers of the gSAME test and GLM dip to be under 15%, whereas our developed method still holds more than 61% when $\beta = 2$.

In contrast to this simulation setup, we dilated our proposed method's evaluation procedure by increasing the somatic mutation calling accuracy in the next simulation part.

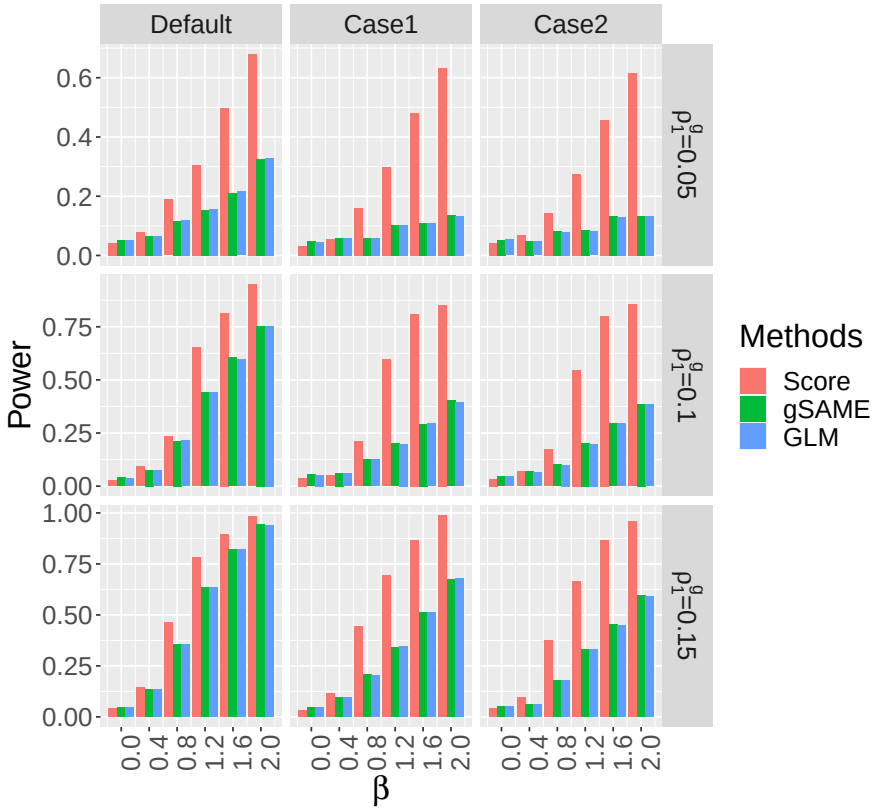

Figure 6.5: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. The comparison is based on the somatic mutation calling accuracy. In the default case (the main simulation setup), the sensitivity and specificity values are set, $\gamma_1 = 0.9$ $\gamma_0 = 0.98$, respectively. In case1, the sensitivity value is remaining as in the default setting, $\gamma_1 = 0.9$,but the specificity value is decreased to be, $\gamma_0 = 0.95$. In case2, both values are decreased so that the sensitivity value, $\gamma_1 = 0.85$, and specificity value, $\gamma_0 = 0.95$

High somatic mutation calling accuracy and error rates

Analogously to the previous cases, but in this setup, the quality of somatic mutation calling procedure is elevated. A dataset of a sample size of $n = 400$ and the number of mutations $j = 10$ was constructed in the same process as the main simulation. In this case, we increased the sensitivity value and specificity value, so the observed call of the xth mutation O_{ix}^{m} was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.95$ and specificity value $\gamma_0 = 0.99$.

Employing this arrangement, increasing the mutation calling method lifts all of the methods' performance, as explained in Figure 6.6. For somatic gene-level mutations, $\rho_1^g = 0.15$ and effect size, $\beta = 1.6$, it is possible to produce 0.97 power by using our gene-based score test, the gSAME test or GLM when the somatic mutation calling is more accurate. However, when the frequency of gene-based somatic mutations is low ($\rho_1^g = 0.05$), the gSAME test and GLM still have low power. Specifically, for $\rho_1^g = 0.05$ and $\beta = 2$, the powers of gSAME and GLM are 0.504 and 0.508, respectively, while our proposed method's power is 0.791. It can imply that the gSAME test and GLM do not perform sufficiently for low-frequency mutation even if the somatic mutation calling is relatively perfect.

Summary

Various simulation studies were constructed in order to evaluate our proposed gene-level score test's type I error and power and compare its performance to the gSAME test and GLM. In all of the scenarios, all of the procedures control

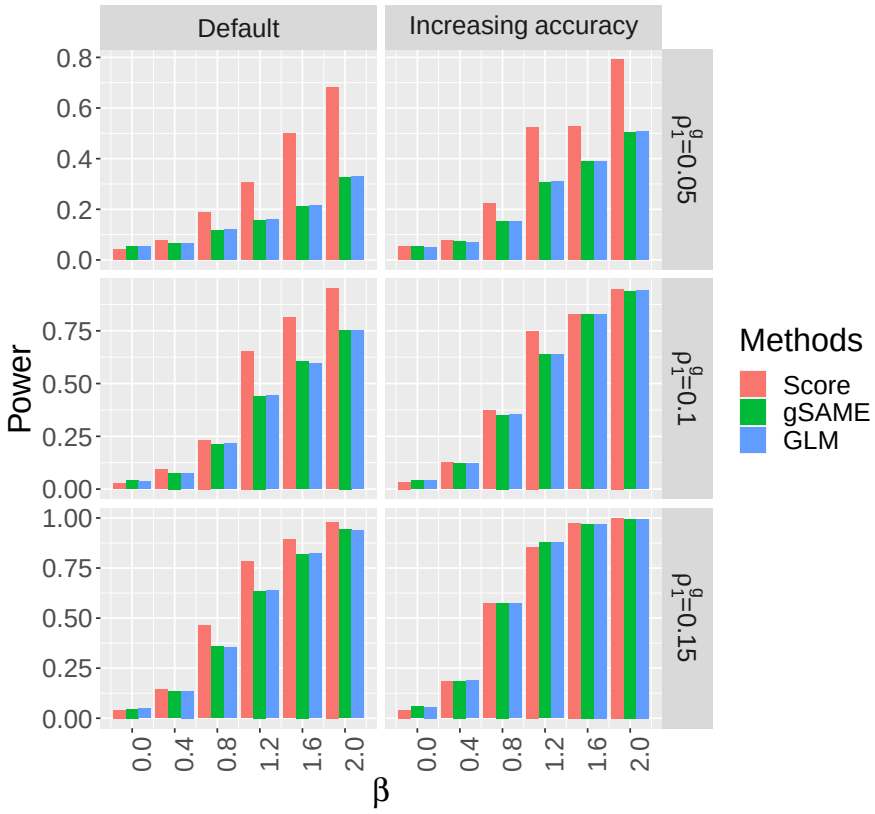

Figure 6.6: Comparison of type I error and power for gene-based mutation analysis of our developed score test (red bars), the gSAME test (green bars) and GLM (blue bars) with various rates of gene-based mutation frequency ρ_1^g and effect size β. The comparison is based on the somatic mutation calling accuracy. In the default case (the main simulation setup), the sensitivity and specificity values are set, $\gamma_1 = 0.9$ $\gamma_0 = 0.98$, respectively. In the increasing accuracy case, both the sensitivity and specificity values are increased so that the sensitivity value, $\gamma_1 = 0.95$, and specificity value, $\gamma_0 = 0.99$

the type I error. Concerning power, our developed gene-level score test has higher power than the gSAME test and GLM. In the gene-level analysis; unlike the single-variant analysis which the mSAME test's power is approximating to the power of our single developed method for the high-frequency somatic mutation ($\rho_1 = 0.1$), our gene-based score method acts much better than the gSAME test for the gene-level high-frequency mutation ($\rho_1^g = 0.15$).

Decreasing the number of somatic mutations within a gene can improve the performance of our gene-based method, the gSAME test and GLM. It might be due to the possibility of committing measurement errors of mutation-level is less when the number of mutations lessens. Nevertheless, the impact size of raising the number of mutations within a gene is more distinguished on the gSAME test and GLM than our gene-level score test. To give an example, when the gene-level mutations frequency, $\rho_1^g = 0.1$ and effect size, $\beta = 1.6$, our score test has 81% power for the case of 10 mutations within a gene are studied, i.e., $j = 10.$, and the power increases to be 88% for $j = 5$. On the other side, the power levels of the gSAME test and GLM approach grow from 60% to 80% for the cases of $j = 10$ and $j = 5$, respectively.

Our gene-based score method works much better than the gSAME test and GLM with high data quality. Moreover, one significant feature of our gene-based score method is that it is more robust than the gSAME test and GLM as it has sturdy performance even when the somatic mutation calling is less accurate. In the single-variant analysis, the mSAME and GLM were solid when the mutation calling's reliability reduces; however, in the gene-level analysis, the performance of the

gSAME test and GLM weaken with lower accurate mutation calling procedure.

6.5 Discussion

Studying a set of somatic mutations within a gene can help realise a comprehensive and better understanding of somatic mutations' critical role in a cancer consequence as they are deemed low-frequency and rare mutations. A considerable number of set-based association techniques have been proposed to discover the association of a group of genetic mutations that are located within a genetic construct in character with a gene, pathway or any genetic set. However, these set-based methods do not account for the variant calling. In this chapter, a novel approach was established using a score test to detect the association of gene-based somatic mutations and a cancer subtype outcome considering the somatic mutation calling method. Our developed gene-level model was then appraised by comparing its performance, in terms of type I error and power, with the gSAME test, which considers the error of somatic mutation calling procedure, and the GLM. The GLM deals with the observed somatic mutation O_i as actual status, and it does not take the somatic mutation calling into consideration. The simulation results determined that all of the approaches control the type I error. In terms of power, our developed gene-based test has better performance than the gSAME test and GLM under diverse situations.

Chapter 7

Discussion and future work

This final chapter gives a discussion of the outcomes achieved and the conclusions of the thesis. Moreover, some insights into alluring future work based on the results in the thesis are introduced in this chapter. The research has focused on statistical methodologies for identifying genomic regions linked to disease outcomes. As cancer is often deemed a severe and fatal disease, and somatic mutations play the most significant role in cancer development, novel methods for investigating and testing the association between somatic mutations and cancer outcomes have been developed in this thesis.

In Chapter Three, we presented several standard association analysis methods of set-based rare variants and made a preliminary observation through simulation studies. An additional investigation that can be added to improve the numerical comparison of the approaches is to study the impact of multiple rare mutations within a gene with different effect magnitudes and various directions of the association. This adjustment in the simulations can help variance-component procedures

expel collapsing techniques in detecting the impact of gene-level rare mutations on a disease trait outcome.

In Chapter Four, since exciting bioinformatics tools have been proposed in order to identify a single somatic mutation, we developed an approach based on adopting the GHC test to compare tumour and normal cells and examine the effect of somatic mutations grouped within a gene. This designed method was assessed by comparing its performance through simulation studies for different scenarios in terms of type 1 error and power to the binomial exact test corrected by the Bonferroni correction. The results of using our gene-based score test were promising; therefore, a further evaluation procedure can be produced by simulating the sequences of genetic variants being in linkage disequilibrium (LD). This modification can help the GHC test perform better as mutations will be correlated within a gene by an LD structure.

Concerning the association analysis of somatic mutations with accounting for the uncertainty of the mutation calling process, as mentioned, standard association methods of rare genetic variants do not account for calling errors for somatic mutations and are restrained in their abilities to analyse the functional effect of somatic mutations. A new somatic mutation association test with measurement errors (SAME) addresses this issue through the likelihood ratio test. It has demonstrated that considering the uncertainty in somatic mutation calling increases the power of an association. We developed and evaluated a score procedure that models actual somatic mutation as an unobservable variable and uses read-depth to increase the mutation calls. The score test is computationally efficient as only op-

timisation under the null model is required for each genetic variant. Additionally, the risk of non-convergence of optimisation routines is reduced. These computational advantages are particularly beneficial in genomewide settings.

In Chapter Five, the score test was developed to examine the association between a single somatic mutation and binary outcome of cancer subtype. Chapter Six extended the method to study the relationship between a group of somatic mutations within a gene and binary trait of cancer subtype. The developed score test was evaluated by comparing its performance in terms of type I error and power to the SAME test and the GLM. Throughout a variety of genetic cases of simulation studies, our proposed score procedure controlled type I error and performed better than the SAME test and GLM. Therefore, one of the future interests is to apply our single and gene-based score approaches in different real cancer genomic datasets in order to compare them to existing methods.

The promising results that we obtained in Chapters Five and Six motivate us to produce additional work to extend our proposed score test procedure to evaluate the effect of somatic mutations on continuous or multinomial cancer subtype outcomes. In the case of continuous outcomes, the variables will be modelled to follow a normal distribution. For multinomial outcomes, we can use a multinomial regression framework. In both of the cases, we need to derive the score test statistic and evaluate it through simulation studies in order to check type I error and investigate the power of the score test.

Appendix A

Results of GHC on somatic mutation association analysis

This appendix contains the results of our developed score test based on applying the generalised higher criticism (GHC) test and the binomial exact test corrected by the Bonferroni correction for comparing tumour and healthy sequences.

Number of somatic mutations= 0 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.026	0.024
0.008	0.018	0.022
0.005	0.004	0.003
Number of somatic mutations= 2 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.384	0.338
0.008	0.128	0.114
0.005	0.019	0.013
Number of somatic mutations= 5 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.881	0.602
0.008	0.564	0.229
0.005	0.062	0.02
Number of somatic mutations= 7 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.989	0.765
0.008	0.813	0.365
0.005	0.126	0.017
Number of somatic mutations= 10 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	0.85
0.008	0.984	0.471
0.005	0.371	0.034

Table A.1: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 50 rare variants at a sample size of $n = 400$.

Number of somatic mutations= 0 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.012	0.014
0.008	0.003	0.005
0.005	0.001	0.003
Number of somatic mutations= 2 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.187	0.223
0.008	0.057	0.067
0.005	0.005	0.007
Number of somatic mutations= 5 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.648	0.45
0.008	0.231	0.131
0.005	0.014	0.01
Number of somatic mutations= 7 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.864	0.545
0.008	0.437	0.195
0.005	0.026	0.006
Number of somatic mutations= 10 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.988	0.677
0.008	0.754	0.251
0.005	0.091	0.008

Table A.2: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 100 rare variants at a sample size of $n = 400$.

Number of somatic mutations= 0 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.008	0.011
0.008	0.004	0.008
0.005	0.002	0.002
Number of somatic mutations= 2 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.115	0.127
0.008	0.04	0.034
0.005	0.004	0.003
Number of somatic mutations= 5 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.462	0.29
0.008	0.127	0.063
0.005	0.007	0.003
Number of somatic mutations= 7 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.731	0.387
0.008	0.278	0.104
0.005	0.015	0.001
Number of somatic mutations= 10 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.955	0.474
0.008	0.581	0.15
0.005	0.041	0.002

Table A.3: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 150 rare variants at a sample size of $n = 400$.

Number of somatic mutations= 0 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.029	0.025
0.008	0.025	0.019
0.005	0.026	0.02
Number of somatic mutations= 2 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.998	0.997
0.008	0.937	0.918
0.005	0.374	0.35
Number of somatic mutations= 5 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	0.997
0.005	0.898	0.652
Number of somatic mutations= 7 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	0.987	0.755
Number of somatic mutations= 10 within 50 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	1	0.874

Table A.4: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 50 rare variants at a sample size of $n = 800$.

Number of somatic mutations= 0 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.035	0.021
0.008	0.027	0.02
0.005	0.02	0.02
Number of somatic mutations= 2 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.981	0.986
0.008	0.831	0.863
0.005	0.196	0.218
Number of somatic mutations= 5 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	992
0.005	0.653	0.466
Number of somatic mutations= 7 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	0.868	0.552
Number of somatic mutations= 10 within 100 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	0.991	0.684

Table A.5: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 100 rare variants at a sample size of $n = 800$.

Number of somatic mutations= 0 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.024	0.025
0.008	0.035	0.032
0.005	0.018	0.02
Number of somatic mutations= 2 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	0.965	0.962
0.008	0.762	0.745
0.005	0.108	0.118
Number of somatic mutations= 5 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	0.999
0.008	0.997	0.97
0.005	0.491	0.282
Number of somatic mutations= 7 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	0.75	0.376
Number of somatic mutations= 10 within 150 rare variants		
Variant allele frequency (VAF)	The GHC test	The binomial exact test
0.01	1	1
0.008	1	1
0.005	0.957	0.458

Table A.6: Type I error and power for the gene-level score test based on using the GHC test and the binomial exact test corrected by the Bonferroni procedure with various rates of the variant allele frequency (VAF) and different numbers of somatic mutations occurring in a gene of tumour cells that includes 150 rare variants at a sample size of $n = 800$.

Appendix B

Results of somatic mutation association analysis

This appendix includes the results of our developed score test procedures, the Somatic mutation Association test with Measurement Errors (SAME) test and the generalised linear model (GLM). We have explored the performance of the methods over a range of effect sizes ranging from 0.4, which is considered a small effect, to 2, which is typically deemed to be large.

B.1 Association analysis of a single somatic mutation

This section contains the results of our single score test, the mSAME test and GLM for testing the relationship between a single somatic mutation and cancer

subtype outcome.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.036	0.084	0.189	0.311	0.45	0.575
mSAME	0.077	0.075	0.116	0.218	0.286	0.355
GLM	0.036	0.058	0.095	0.148	0.257	0.314
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.107	0.323	0.615	0.816	0.924
mSAME	0.053	0.097	0.275	0.532	0.722	0.866
GLM	0.045	0.087	0.267	0.512	0.69	0.839
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.045	0.168	0.534	0.855	0.972	0.996
mSAME	0.062	0.185	0.539	0.847	0.968	0.995
GLM	0.053	0.185	0.522	0.828	0.962	0.994

Table B.1: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.044	0.117	0.291	0.502	0.685	0.843
mSAME	0.054	0.095	0.185	0.324	0.463	0.599
GLM	0.042	0.081	0.179	0.318	0.427	0.572
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.04	0.185	0.508	0.855	0.967	0.996
mSAME	0.053	0.169	0.476	0.806	0.933	0.988
GLM	0.046	0.161	0.443	0.785	0.849	0.98
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.042	0.287	0.8	0.992	1	1
mSAME	0.045	0.301	0.8	0.989	1	1
GLM	0.044	0.3	0.752	0.99	1	1

Table B.2: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=800.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.042	0.103	0.208	0.527	0.8	0.925
mSAME	0.045	0.08	0.178	0.366	0.58	0.704
GLM	0.044	0.077	0.065	0.289	0.553	0.7
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.042	0.221	0.554	0.899	0.99	1
mSAME	0.044	0.175	0.536	0.878	0.982	0.997
GLM	0.044	0.168	0.423	0.778	0.969	0.995
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.04	0.321	0.903	1	1	1
mSAME	0.057	0.33	0.9	1	1	1
GLM	0.047	0.323	0.895	1	1	1

Table B.3: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=1000.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.052	0.202	0.694	0.959	0.997	1
mSAME	0.056	0.155	0.486	0.829	0.956	0.991
GLM	0.054	0.141	0.467	0.79	0.923	0.982
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.036	0.46	0.981	1	1	1
mSAME	0.037	0.432	0.948	1	1	1
GLM	0.036	0.366	0.947	1	1	1
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.032	0.778	0.996	1	1	1
mSAME	0.044	0.778	0.969	1	1	1
GLM	0.045	0.753	0.9	1	1	1

Table B.4: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=3000.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.052	0.383	0.864	0.998	1	1
mSAME	0.051	0.263	0.699	0.96	0.997	1
GLM	0.054	0.265	0.589	0.937	0.994	1
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.038	0.658	1	1	1	1
mSAME	0.047	0.616	1	1	1	1
GLM	0.037	0.516	1	1	1	1
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.024	0.94	1	1	1	1
mSAME	0.055	0.94	1	1	1	1
GLM	0.048	0.932	1	1	1	1

Table B.5: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=5000.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.038	0.065	0.135	0.26	0.378	0.497
mSAME	0.068	0.068	0.119	0.197	0.279	0.346
GLM	0.027	0.038	0.078	0.129	0.191	0.216
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.036	0.096	0.302	0.517	0.738	0.892
mSAME	0.048	0.098	0.27	0.461	0.677	0.836
GLM	0.042	0.089	0.226	0.365	0.565	0.723
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.026	0.13	0.458	0.817	0.949	0.996
mSAME	0.048	0.15	0.48	0.82	0.95	0.994
GLM	0.045	0.13	0.407	0.718	0.901	0.975

Table B.6: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400 with low read-depth. In this detaset, the somatic mutation read-depth was simulated by a negative binomial distribution with mean 40 and over-dispersion 1.9.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.051	0.081	0.173	0.294	0.45	0.571
mSAME	0.061	0.066	0.111	0.194	0.268	0.348
GLM	0.036	0.056	0.091	0.159	0.238	0.308
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.033	0.109	0.311	0.603	0.808	0.917
mSAME	0.044	0.09	0.255	0.522	0.72	0.844
GLM	0.036	0.079	0.256	0.51	0.783	0.824
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.024	0.156	0.525	0.834	0.973	0.999
mSAME	0.037	0.173	0.522	0.829	0.965	0.997
GLM	0.033	0.165	0.504	0.781	0.951	0.995

Table B.7: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400. In this dataset, the sensitivity value γ_1 is set as in the default setting ($\gamma_1 = 0.9$), but the specificity value γ_0 is decreased. Consequently, the observed mutation O_i was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and specificity value $\gamma_0 = 0.95$.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.04	0.082	0.153	0.269	0.404	0.552
mSAME	0.049	0.061	0.114	0.188	0.262	0.34
GLM	0.029	0.049	0.101	0.167	0.237	0.303
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.031	0.102	0.303	0.586	0.757	0.914
mSAME	0.04	0.089	0.26	0.517	0.68	0.843
GLM	0.032	0.087	0.236	0.476	0.642	0.817
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.049	0.158	0.516	0.765	0.971	0.995
mSAME	0.057	0.172	0.523	0.79	0.967	0.995
GLM	0.044	0.157	0.499	0.629	0.962	0.992

Table B.8: Type I error and power for single-mutation analysis of our developed score test, the mSAME test and GLM with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400. In this dataset, both the sensitivity value γ_1 and the specificity value γ_0 are decreased. Consequently, the observed mutation O_i was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.85$ and specificity value $\gamma_0 = 0.95$.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.004	0.01	0.012	0.026	0.036	0.082
mSAME	0.001	0.003	0.014	0.018	0.027	0.071
GLM	0.001	0.004	0.011	0.017	0.039	0.076
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.002	0.015	0.05	0.133	0.207	0.298
mSAME	0.001	0.011	0.06	0.128	0.207	0.297
GLM	0.003	0.014	0.044	0.138	0.21	0.268
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.004	0.018	0.094	0.194	0.303	0.317
mSAME	0.001	0.021	0.091	0.189	0.272	0.31
GLM	0.001	0.008	0.095	0.182	0.255	0.297

Table B.9: Type I error and power for our developed single score tests, the mSAME tests and GLM corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400 and number of mutation=10.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.011	0.017	0.06	0.166	0.268	0.4
mSAME	0.004	0.013	0.072	0.159	0.266	0.4
GLM	0.003	0.017	0.061	0.127	0.278	0.4
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.004	0.022	0.117	0.363	0.617	0.733
mSAME	0.003	0.039	0.118	0.346	0.601	0.709
GLM	0.004	0.029	0.126	0.338	0.615	0.74
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.002	0.063	0.244	0.618	0.781	0.878
mSAME	0.001	0.043	0.26	0.579	0.76	0.88
GLM	0.003	0.06	0.27	0.598	0.771	0.866

Table B.10: Type I error and power for our developed single score tests, the mSAME tests and GLM corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=800 and number of mutation=10.

Type I error and power when $\rho_1 = 0.02$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.005	0.018	0.047	0.094	0.157	0.185
mSAME	0.003	0.012	0.042	0.1	0.147	0.19
GLM	0.001	0.017	0.04	0.102	0.158	0.19
Type I error and power when $\rho_1 = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.005	0.028	0.105	0.225	0.485	0.601
mSAME	0.004	0.036	0.102	0.229	0.49	0.61
GLM	0.007	0.034	0.096	0.253	0.465	0.462
Type I error and power when $\rho_1 = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.006	0.046	0.15	0.473	0.7	0.841
mSAME	0.008	0.047	0.137	0.458	0.71	0.85
GLM	0.006	0.032	0.132	0.456	0.676	0.839

Table B.11: Type I error and power for our developed single score tests, the mSAME tests and GLM corrected by the Bonferroni correction with various rates of mutation frequency ρ_1 and effect size β of a sample size of n=400 and number of mutation=5.

B.2 Association analysis of gene-based somatic mutations

This section contains the results of our gene-based score test, the gSAME test and GLM for testing the relationship between somatic mutations that are grouped within a gene and a cancer subtype outcome.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.043	0.078	0.189	0.306	0.499	0.681
gSAME	0.052	0.066	0.116	0.155	0.211	0.325
GLM	0.052	0.067	0.119	0.158	0.216	0.328
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.026	0.093	0.233	0.654	0.815	0.949
gSAME	0.04	0.076	0.213	0.441	0.604	0.754
GLM	0.037	0.073	0.216	0.442	0.596	0.753
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.042	0.143	0.462	0.782	0.893	0.981
gSAME	0.047	0.134	0.357	0.636	0.821	0.943
GLM	0.048	0.136	0.354	0.637	0.822	0.937

Table B.12: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.037	0.091	0.311	0.667	0.82	0.958
gSAME	0.052	0.061	0.139	0.273	0.424	0.555
GLM	0.052	0.057	0.139	0.282	0.425	0.559
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.174	0.508	0.875	0.98	0.993
gSAME	0.055	0.142	0.366	0.702	0.858	0.933
GLM	0.056	0.139	0.371	0.703	0.863	0.94
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.031	0.251	0.671	0.966	0.999	1
gSAME	0.057	0.211	0.612	0.893	0.993	1
GLM	0.059	0.219	0.617	0.899	0.993	1

Table B.13: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 800$ and number of mutations $j = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.11	0.405	0.734	0.811	0.954
gSAME	0.047	0.074	0.19	0.349	0.463	0.636
GLM	0.044	0.071	0.195	0.347	0.465	0.641
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.031	0.188	0.622	0.97	1	1
gSAME	0.041	0.15	0.425	0.819	0.954	0.985
GLM	0.042	0.151	0.423	0.822	0.953	0.985
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.03	0.373	0.648	0.973	0.999	1
gSAME	0.05	0.242	0.638	0.898	0.997	1
GLM	0.052	0.242	0.641	0.898	0.997	1

Table B.14: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 1000$ and number of mutations $j = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.301	0.907	0.975	0.995	0.998
gSAME	0.051	0.127	0.455	0.723	0.852	0.914
GLM	0.053	0.126	0.458	0.73	0.861	0.925
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.035	0.532	0.998	1	1	1
gSAME	0.051	0.379	0.908	0.997	1	1
GLM	0.049	0.378	0.9	0.997	1	1
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.043	0.771	0.984	1	1	1
gSAME	0.05	0.609	0.981	1	1	1
GLM	0.048	0.614	0.982	1	1	1

Table B.15: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 3000$ and number of mutations $j = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.033	0.452	0.97	1	1	1
gSAME	0.047	0.226	0.66	0.928	1	1
GLM	0.051	0.233	0.663	0.925	1	1
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.034	0.804	1	1	1	1
gSAME	0.047	0.563	0.993	1	1	1
GLM	0.046	0.564	0.993	1	1	1
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.026	0.876	0.999	1	1	1
gSAME	0.049	0.802	1	1	1	1
GLM	0.05	0.799	1	1	1	1

Table B.16: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 5000$ and number of mutations $j = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.043	0.081	0.164	0.266	0.592	0.746
gSAME	0.059	0.06	0.127	0.215	0.369	0.475
GLM	0.053	0.06	0.129	0.215	0.372	0.472
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.021	0.121	0.318	0.708	0.882	0.924
gSAME	0.054	0.121	0.304	0.625	0.799	0.878
GLM	0.05	0.124	0.304	0.624	0.801	0.875
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.037	0.135	0.53	0.83	0.95	0.992
gSAME	0.053	0.151	0.525	0.822	0.949	0.985
GLM	0.052	0.157	0.528	0.819	0.95	0.984

Table B.17: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 5$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.045	0.081	0.197	0.258	0.412	0.557
gSAME	0.065	0.061	0.097	0.187	0.243	0.31
GLM	0.06	0.058	0.096	0.185	0.246	0.314
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.083	0.337	0.64	0.834	0.965
gSAME	0.059	0.073	0.232	0.414	0.616	0.78
GLM	0.06	0.07	0.238	0.418	0.61	0.786
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.032	0.133	0.322	0.69	0.819	0.998
gSAME	0.057	0.142	0.352	0.662	0.833	0.954
GLM	0.053	0.14	0.355	0.656	0.84	0.954

Table B.18: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, the read-depth threshold $D_0 = 10$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.043	0.077	0.157	0.388	0.503	0.663
gSAME	0.035	0.063	0.088	0.142	0.221	0.251
GLM	0.037	0.062	0.085	0.142	0.223	0.251
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.029	0.111	0.358	0.622	0.782	0.916
gSAME	0.061	0.092	0.214	0.384	0.55	0.697
GLM	0.059	0.092	0.213	0.381	0.556	0.701
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.037	0.113	0.377	0.857	0.858	0.98
gSAME	0.051	0.108	0.311	0.662	0.759	0.926
GLM	0.047	0.104	0.312	0.663	0.749	0.922

Table B.19: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, the read-depth threshold $D_0 = 30$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.028	0.064	0.17	0.409	0.59	0.746
gSAME	0.042	0.059	0.086	0.161	0.233	0.295
GLM	0.041	0.057	0.086	0.161	0.237	0.296
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.024	0.095	0.324	0.591	0.876	0.974
gSAME	0.046	0.076	0.215	0.347	0.609	0.732
GLM	0.041	0.075	0.218	0.35	0.611	0.733
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.032	0.128	0.381	0.856	0.921	0.988
gSAME	0.041	0.106	0.287	0.662	0.789	0.93
GLM	0.038	0.104	0.288	0.662	0.79	0.929

Table B.20: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, the read-depth threshold $D_0 = 40$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.033	0.056	0.161	0.298	0.48	0.634
gSAME	0.048	0.059	0.058	0.104	0.109	0.137
GLM	0.044	0.058	0.058	0.104	0.111	0.134
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.038	0.049	0.211	0.597	0.81	0.852
gSAME	0.055	0.062	0.125	0.202	0.293	0.402
GLM	0.053	0.059	0.125	0.198	0.294	0.394
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.03	0.115	0.443	0.696	0.864	0.989
gSAME	0.049	0.097	0.209	0.341	0.513	0.675
GLM	0.047	0.094	0.205	0.345	0.51	0.677

Table B.21: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, the sensitivity value γ_1 is set as in the default setting ($\gamma_1 = 0.9$), but the specificity value γ_0 is decreased. Consequently, the observed mutation of the xth mutation O_{ix} was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.9$ and specificity value $\gamma_0 = 0.95$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.07	0.144	0.274	0.457	0.617
gSAME	0.053	0.05	0.082	0.085	0.134	0.132
GLM	0.054	0.047	0.079	0.081	0.131	0.132
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.034	0.069	0.172	0.547	0.8	0.855
gSAME	0.047	0.069	0.101	0.203	0.295	0.386
GLM	0.046	0.065	0.099	0.199	0.295	0.383
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.042	0.094	0.376	0.665	0.865	0.96
gSAME	0.053	0.063	0.18	0.332	0.452	0.595
GLM	0.053	0.062	0.177	0.331	0.448	0.592

Table B.22: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, both the sensitivity value γ_1 and the specificity value γ_0 are decreased. Consequently, the observed mutation of the xth mutation O_{ix} was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.85$ and specificity value $\gamma_0 = 0.95$.

Type I error and power when $\rho_1^g = 0.05$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.052	0.078	0.222	0.523	0.529	0.791
gSAME	0.054	0.074	0.151	0.308	0.39	0.504
GLM	0.051	0.071	0.153	0.31	0.391	0.508
Type I error and power when $\rho_1^g = 0.1$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.034	0.126	0.372	0.749	0.827	0.948
gSAME	0.039	0.123	0.351	0.637	0.828	0.938
GLM	0.039	0.121	0.353	0.637	0.829	0.94
Type I error and power when $\rho_1^g = 0.15$						
Test	$\beta = 0$	$\beta = 0.4$	$\beta = 0.8$	$\beta = 1.2$	$\beta = 1.6$	$\beta = 2$
Score test	0.041	0.184	0.573	0.854	0.975	1
gSAME	0.059	0.186	0.576	0.877	0.969	0.995
GLM	0.055	0.189	0.573	0.877	0.97	0.995

Table B.23: Type I error and power for gene-based mutation analysis of our developed score test, the gSAME test and GLM with various rates of gene-based mutation frequency ρ_1^g and effect size β of a sample size of $n = 400$ and number of mutations $j = 10$. In this dataset, both the sensitivity value γ_1 and the specificity value γ_0 are increased. Consequently, the observed mutation of the xth mutation O_{ix} was simulated by a Bernoulli distribution with the sensitivity value $\gamma_1 = 0.95$ and specificity value $\gamma_0 = 0.99$.

Bibliography

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Barnett, I., Mukherjee, R., and Lin, X. (2017). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, 112(517):64–76.

Barnett, I. J. and Lin, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika*, 101(4):964–970.

Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6):236–238.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breslow, N., Day, N., Halvorsen, K., Prentice, R., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American*

Journal of Epidemiology, 108(4):299–307.

Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., Hubbard, T., Jostins, L., Maltby, N., Mahon-Pearson, J., et al. (2017). The national genomics research and healthcare knowledgebase. *figshare*.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219.

Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology*, 37(1):110–121.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M. A., Condon, A., et al. (2012). Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*, 28(2):167–175.

Donoho, D., Jin, J., et al. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.

Fang, L. T., Afshar, P. T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J. C., Gibeling, G., Barr, S., Asadi, N. B., Gerstein, M. B., et al. (2015). An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome*

biology, 16(1):1–13.

Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):361–369.

Hall, P., Jin, J., et al. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

Hansen, N. F., Gartner, J. J., Mei, L., Samuels, Y., and Mullikin, J. C. (2013). Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics*, 29(12):1498–1503.

He, Q., He, Q., Liu, X., Wei, Y., Shen, S., Hu, X., Li, Q., Peng, X., Wang, L., and Yu, L. (2014). Genome-wide prediction of cancer driver genes based on snp and cancer snv data. *American journal of cancer research*, 4(4):394.

He, Q., Liu, Y., Peters, U., and Hsu, L. (2018). Multivariate association analysis with somatic mutation data. *Biometrics*, 74(1):176–184.

Helena Mangs, A. and Morris, B. J. (2007). The human pseudoautosomal region (par): origin, function and future. *Current genomics*, 8(2):129–136.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Pro-*

ceedings of the National Academy of Sciences, 106(23):9362–9367.

Kaler, A. S. and Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC genomics*, 20(1):1–8.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576.

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *British journal of cancer*, 118(11):1492–1501.

Lai, E., Fan, S.-T., Lo, C.-M., Chu, K.-M., Liu, C.-L., and Wong, J. (1995). Hepatic resection for hepatocellular carcinoma. an audit of 343 patients. *Annals of surgery*, 221(3):291.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C., and Dry, J. R. (2016). Vardict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research*, 44(11):e108–e108.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association

analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

Lin, D. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, 101(473):89–104.

Liu, M., Liu, Y., Wu, M. C., Hsu, L., and He, Q. (2021). A method for subtype analysis with somatic mutations. *Bioinformatics*.

Liu, Y., He, Q., and Sun, W. (2018). Association analysis using somatic mutations. *PLoS genetics*, 14(11):e1007746.

Luzzatto, L. (2011). Erratum to: Somatic mutations in cancer development. *Environmental health*, 10(1):1–8.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Mittelhammer, R. C., Miller, D. T., Judge, G. G., and Miller, D. J. (2000). *Econometric foundations pack with CD-ROM*. Cambridge University Press.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56.

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2):188–193.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Neyman, J. and Scott, E. (1965). On the use of c (alpha) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, 41(1):477–497.

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(6):497–507.

Parsa, N. (2012). Environmental factors inducing human cancers. *Iranian journal of public health*, 41(11):1.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838.

Radenbaugh, A. J., Ma, S., Ewing, A., Stuart, J. M., Collisson, E. A., Zhu, J., and Haussler, D. (2014). Radia: Rna and dna integrated analysis for somatic mutation detection. *PloS one*, 9(11):e111516.

Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE.

Richards, J. E. and Hawley, R. S. (2011). *The human genome*. Academic Press.

Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., Glonek, G., and Adelson, D. L. (2013). A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–2230.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., et al. (2012). Jointsnvmix: a probabilistic

model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913.

Spinella, J.-F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., Healy, J., and Sinnett, D. (2016). Snooper: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC genomics*, 17(1):1–11.

Sugasawa, S., Noma, H., Otani, T., Nishino, J., and Matsui, S. (2017). An efficient and flexible test for rare variant effects. *European Journal of Human Genetics*, 25(6):752–757.

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, 37(4):334–344.

Taylor, C. M. and Frankl, F. E. K. (2012). Developing a strategy for the management of rare diseases.

Tzeng, J.-Y. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, 81(5):927–938.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24.

Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., and Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics*, 15(1):1–10.

Zelterman, D. and Chen, C.-F. (1988). Homogeneity tests against central-mixture alternatives. *Journal of the American Statistical Association*, 83(401):179–182.

Ziegler, A., Konig, I. R., and Pahlke, F. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons.