

ULRR

Statistical methods for the detection of single nucleotide polymorphisms (SNPs) using new generation genome sequencers

Item Type	Thesis
Authors	Sheikhi, Ali
Download date	2026-05-11 05:24:46
Item License	https://creativecommons.org/licenses/by-nc-sa/1.0/
Link to Item	https://hdl.handle.net/10344/5996



Statistical Methods for the Detection of Single
Nucleotide Polymorphisms (SNPs) Using New
Generation Genome Sequencers

Ali Sheikhi (B.Sc., M.Sc.)

(A Thesis Submitted for the Award of Ph.D.)

Centre of Biostatistics
Department of Mathematics and Statistics
University of Limerick
Ireland

Supervisors:

Dr. David Ramsey
Professor Gilbert MacKeznie

February 2014

Declaration

I hereby declare that the work described in this thesis is entirely my own work except where otherwise stated. No part of this thesis has already been submitted for any degree at this or any other university.

Dedication

I lovingly dedicate this thesis to my mother, father and my wife who supported me each step of the way.

Acknowledgments

This research project would not have been possible without the support of many people.

I would like to express my gratitude to my supervisors, Dr. David Ramsey and Professor Gilbert MacKenzie who were abundantly helpful and offered invaluable assistance, support and guidance, without whose knowledge and assistance this study would not have been successful.

This research would not have been possible without the financial assistance of Science Foundation Ireland, BIO-SI project under grant 07MI012. I express my gratitude to this organization.

Special thanks also to all my friends in the Department of Mathematics and Statistics, especially Dr. Joseph Lynch and Mr. Kevin Burke for sharing the literature and invaluable assistance. I owe a deep gratitude to Dr. Mark Burke, Head of the Department of Mathematics and Statistics, who kindly provided the opportunity of being a lecturer at the department while doing the research. Not forgetting my best Iranian friends who have always been there, helping me to deal with living away from home.

I would like to express my deepest gratitude to my entire family for their understanding and endless love, through the duration of my studies:

I have been blessed with the greatest parents in the world, my mother and father, the best gifts of my life, who have been the best encourages I could ever had. I do not think words can say how much I am grateful to them.

My brother and sister, Farshid and Mahshid, with whom I have been sharing a beautiful friendship throughout our lives.

My lovely niece, Melisa, a little princess with a kind and caring heart. Although she is still a child, her kindness and love have been always with me during these years.

Finally, and most crucially, I would like to thank my loving and dearest friend. I would like to express my deepest gratitude to my lovely wife, Zohreh, who is not just my wonderful wife, but also the greatest friend I could ever had. There are no words that can express my love and gratitude to her. Without her endless love, understanding and patience, it would have been impossible for me to accomplish this long journey.

Abstract

A single nucleotide polymorphism, or SNP, is a site of the genome where variation occurs within a population. Almost all SNPs have only two alleles (variants).

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. The importance of SNPs also comes from their ability to influence disease risk, drug efficacy and side-effects. Thus, there should be confidence in the methods for the detection of SNPs that a SNP found is not a sequencing error.

In this work, a statistical method based on a likelihood ratio test has been considered to detect these SNPs. The efficiency of this test has been compared with a threshold test using simulations. Results of the analysis of real genome sequence data are also presented.

Contents

1	Introduction	7
1.1	Genome sequencing	7
1.1.1	Importance	10
1.1.2	Error types in DNA sequencing	10
1.1.3	How genome sequencers work	11
1.2	Pooling	12
1.2.1	Advantages and disadvantages	13
1.3	Gene tagging	14
2	Review of Multiple Comparisons	15
2.1	Bonferroni correction	16
2.2	Benjamini-Hochberg procedure	17
2.2.1	An algorithm for the Benjamini-Hochberg procedure	19
2.3	FDR control by modelling the distribution of the p -value by a mixture of distributions	20
2.3.1	Calculation of the q -value	21
2.4	A pseudo-Bayesian approach	22
3	Description of the Problem	24
3.1	The data	24
3.2	The model	26
4	Methods	28
4.1	The idea of a fixed threshold rule	28
4.2	The likelihood ratio test	29
4.2.1	The likelihood ratio test statistic	31
4.2.2	Adaptation of the test: four possible variants at a site	31

4.2.3	A brief note on the overdispersion of the data	32
4.2.4	Bayesian score	33
4.2.5	Adaptation of the likelihood ratio test to pooled data	34
4.3	Estimation of the power for the threshold test when the read rate and error rate are fixed	36
4.4	A fixed significance level based on the assumed density of SNPs	37
5	Simulations	39
5.1	Introduction	39
5.2	Calling SNPs with unpooled or tagged DNA	39
5.3	Calling SNPs with pooled, untagged DNA	57
5.4	Estimation of the power for the threshold test when the read rate and error rate are fixed: calculations	81
5.5	A fixed significance level based on the assumed density of SNPs: calculations	92
5.6	Overview of the results	96
6	Analysis of Real Genome Sequence Data	100
6.1	Sites inferred to be SNPs	100
6.2	Genotyping of the SNPs	101
6.2.1	How genotyping is done	102
7	Conclusion	104
7.1	Introduction	104
7.2	Maximum likelihood test versus the threshold test	105
7.2.1	Tagged (or unpooled) DNA	106
7.2.2	Untagged (pooled) DNA	106
7.2.3	Other findings	107
7.3	Benjamini-Hochberg procedure: advantages & disadvantages	107
7.4	Possible problems and further investigation	109
	Appendices	112
A	Bayesian Score & Genotyping	113
A.1	SNP Genotyping and the Bayesian Scores	114
A.2	An Example of Genotyping	117
A.2.1	Illustrating how the likelihoods are calculated	124

B R Programmes	128
B.1 programme 1: SNP Detection from the Real Data . . .	129
B.1.1 A brief note about the programme	129
B.2 programme 2: Genotyping of the SNPs	134
B.3 programme 3: SNP Detection (Tagged Data - Poisson Distribution)	143
B.4 programme 4: SNP Detection (Tagged Data - Nega- tive Binomial Distribution)	148
B.5 programme 5: SNP Detection (Untagged Data - Pois- son Distribution)	153
B.6 programme 6: SNP Detection (Untagged Data - Neg- ative Binomial Distribution)	158
Bibliography	163

List of Tables

2.1	Benjamini-Hochberg Procedure	17
3.1	The Structure of the Data	24
3.2	The Quality Scores	25
5.1	Power and FDR of Tests for Tagged DNA Samples by Read Rate and Error Rate (Number of Reads Has a Poisson Distribution)	42
5.2	Power and FDR of Tests for Tagged DNA Samples by Read Rate and Error Rate (Number of Reads Has an NB Distribution)	43
5.3	Regression Analysis to Assess the Effect of Read Rate and Error Rate on the Power (Number of Reads Has a Poisson Distribution)	45
5.4	Regression Analysis to Assess the Effect of Read Rate and Error Rate on the Power (Number of Reads Has an NB Distribution)	48
5.5	Power and FDR of Tests for Tagged DNA Samples (Number of Reads Has a Poisson Distribution with Fixed Read Rate and Empirical Error Rate)	54
5.6	Power and FDR of Tests for Tagged DNA Samples (Number of Reads Has an NB Distribution with Fixed Read Rate and Empirical Error Rate)	54
5.7	Power and FDR of Tests for Tagged DNA Samples (Empirical Read Rate Used - Mean Empirical Error Rate = 0.0008)	56
5.8	Power and FDR of Tests for Tagged DNA Samples (Empirical Read Rate and Empirical Error Rate Used)	56
5.9	Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has a Poisson Distribution - Error Rate = 0.01)	59

5.10	Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has an NB Distribution - Error Rate = 0.01)	60
5.11	Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has a Poisson Distribution)	63
5.12	Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has an NB Distribution)	67
5.13	Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has a Poisson Distribution - Mean Empirical Error Rate = 0.0008)	70
5.14	Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has an NB Distribution - Mean Empirical Error Rate = 0.0008)	71
5.15	Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has a Poisson Distribution)	73
5.16	Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has an NB Distribution)	77
5.17	Power and FDR of Tests for Untagged DNA Sample (Empirical Read Rate Used - Mean Empirical Error Rate = 0.0008)	80
5.18	Power and FDR of Tests for Untagged DNA Samples (Empirical Read Rate and Empirical Error Rate Used)	80
5.19	Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate	83
5.20	Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate	90
5.21	Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate	91
5.22	Estimated Power of the Threshold Test Using a Fixed Significance Level	93
5.23	FDR of Various Parameters of the Sequencer	95
6.1	Positions of the Sites Inferred to be SNPs	101
A.1	Details of the SNPs	114
A.2	Details of the Site 55447	117
A.3	Details of the Site 59754	118
A.4	Likelihoods and Genotyping of Individuals at Site 59754119	

List of Figures

1.1	DNA Molecule 1 Differs from DNA Molecule 2 at a Single Base-Pair Location (an AG Polymorphism) . .	8
1.2	Genome Sequencing	12
5.1	3D Figure of Regression Analysis	46
5.2	3D Figure of Regression Analysis	49
5.3	Histogram of the Mean Number of Reads from a Site from Real Data	51
5.4	Histogram of Quality Scores from Real Data	52
5.5	3D Figure of Regression Analysis	64
5.6	3D Figure of Regression Analysis	68
5.7	3D Figure of Regression Analysis	74
5.8	3D Figure of Regression Analysis	78

Chapter 1

Introduction

1.1 Genome sequencing

The genome of every organism, including humans, contains all of the biological information needed to build and maintain a living example of that organism. The biological information contained in a genome is encoded in its deoxyribonucleic acid (DNA) and is divided into discrete units called genes. Most living organisms have the same sort of genetic material, DNA, in their cells. DNA contains two strands wrapped around each other in a double helix, and these strands are held in place by nucleotides or bases. Nucleotides are organic molecules that form the basic building blocks of DNA. There are four nucleotides, denoted by A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). In a helix C is always paired with G and A with T. Hence, if the nucleotides on one side of the helix (say the left) are known, then the whole sequence is known. Also, higher organisms are diploid, i.e., their genome is made up of chromosome pairs which have the helix structure (Mange, E. J. and Mange, A. P. Basic Human Genetics, 1990).

It should be noted that each chromosome is essentially a sequence of nucleotides. The number of nucleotides on each chromosome depends on the size of the chromosome. For example, human chromosome 1, the largest, is estimated to have 247,249,719 base pairs, i.e., the DNA sequence in the chromosome is about 247 million base pairs long. Since DNA is double stranded, that equates to about 494 million nucleotides (Ensembl Genome Browser (www.ensembl.org)).

Base pairs are building blocks of the DNA double helix. So, for example one can have an AT base pair (one base from the top strand, one base from the bottom).

The genome of a diploid animal can be thought of as a sequence of pairs of these nucleotides. Diploid cells contain two complete sets of chromosomes, whereas haploid cells, i.e., eggs and sperm, have half the number of chromosomes as diploid, i.e., a haploid cell contains only one complete set of chromosomes.

A **single nucleotide polymorphism** or **SNP** (pronounced snip) is a site at which variation is observed within a population (Figure 1.1).

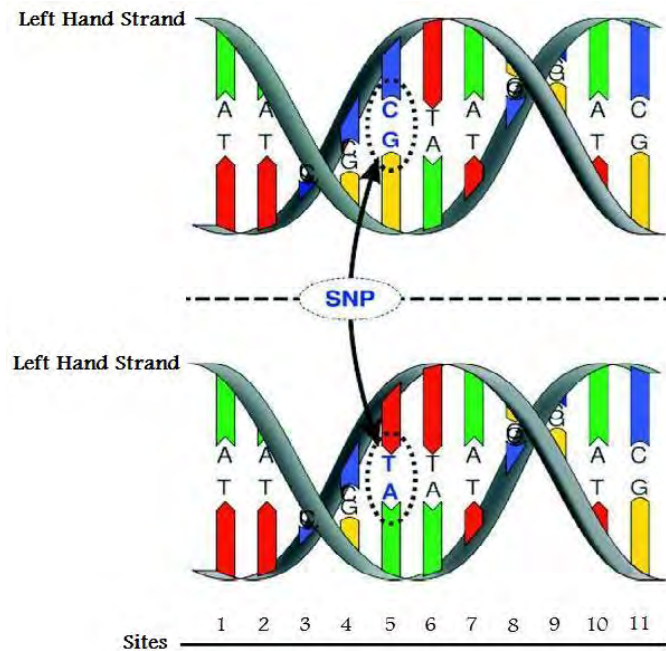


Figure 1.1: DNA Molecule 1 Differs from DNA Molecule 2 at a Single Base-Pair Location (an AG Polymorphism)

The genotype of an individual at a site (or locus) can be defined by the two nucleotides on the left hand sites of the two helices in a pair at the corresponding position. For example, the genotype at the first site in the above picture is AA, but the genotype at the

fifth site, i.e., at the SNP, is AG (base pairs are always written in alphabetical order).

At a SNP just two of the four nucleotides appear within a population. For example, at the SNP in the picture above, A and G appeared. These variants of the genes are called alleles, the most common (rare) is termed the major allele (minor allele, respectively).

Genome sequencing includes methods and technologies that are used for determining the order of nucleotides along the sequence. A genome can be thought of as a sequence where each site can be occupied by one of four nucleotides. The whole genome cannot be read at once, so instead, the DNA must be broken into small pieces. The small pieces will be read, and then the whole sequence inferred using sequence alignment procedures (not considered here).

In general, there are two approaches to the task of cutting up the genome and then deriving the whole sequence (Human Genome Project (http://www.ornl.gov/sci/techresources/Human_Genome)).

One strategy, known as the clone-by-clone approach, involves first breaking the genome up into relatively large chunks, called clones. Scientists use genome mapping techniques to figure out where in the genome each clone belongs. Next they cut each clone into smaller, overlapping pieces of the right size for sequencing. Finally, they sequence the pieces and use the overlaps to reconstruct the sequence of the whole clone.

The other strategy, called the whole-genome shotgun method, involves breaking the genome up into small pieces, sequencing the pieces, and reassembling the pieces into the full genome sequence.

The genome sequencer only reads the left hand side of the strand. Once the nucleotide on the left hand side is read, the nucleotide on the right hand side is known (if the left is A, the right is T and if the left is C, the right is G and vice versa). So, for example, the beginning of the sequence defining the first DNA molecule in Figure 1 (upper panel) is (AAGGAT) (it starts with AA on the left hand strand at the top, and continues with GGAT at the bottom in the

middle).

1.1.1 Importance

Approximately 99.9% of the human genome is identical in all individuals. On average, however, at one in every 500 to 1000 base pairs there is variation within the population. These randomly occurring changes are passed from generation to generation and account for a high proportion of the DNA differences between us. It is estimated that between three and six million such variations lie hidden in our genome. When the frequency of the minor allele in a species - or to be more precise, of a particular population, i.e., an ethnic group - is at least 1%, it is referred to as a SNP. These variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. However, their greatest importance in biomedical research is for comparing regions of the genome in genome-wide association studies.

In order to carry out appropriate statistical analyses, the frequency of the minor allele should not be very small.

1.1.2 Error types in DNA sequencing

There are some factors which can cause errors during the determination of a DNA sequence. Basically, there are three types of errors: insertions, deletions and mismatches.

In brief, insertions are wrongly called bases at places where there are none, deletions are bases that were not called in a sequence and mismatches represent wrongly called bases (calling is the process of assigning nucleotides to loci which is done by the genome sequencer). These types of errors can be reduced by using improved chemistry (Lario et al. (1997); Rosenblum et al. (1997)), by applying image processing algorithms (Sanders et al. (1991)) or by using different base calling algorithms (Berno (1996)).

Having a viable numerical estimate of the base quality was a major advance achieved by Ewing et al. (1998) and Ewing and

Green (1998), who presented an improved base caller that also gives estimates of the probability of error for the called bases. The probability of making an error while calling a nucleotide, denoted by p , describes the probability with which the base caller has produced a wrong base call, where a value of 1 represents a certain wrong call. The quality of a base call is assigned to be,

$$q = -10\log_{10}(p),$$

where q is the quality and p the probability of an error. For instance, a quality of 40 corresponds to an error probability of 0.0001. It can be clearly seen that the higher the quality score, the lower the error probability.

1.1.3 How genome sequencers work

Here, how a genome sequencer works will be briefly described. DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases - adenine, guanine, cytosine, and thymine - in a strand of DNA. This is a complex nucleotide-sequencing technique, which in general includes three identifiable steps:

1. Polymerase Chain Reaction (PCR).
2. Sequencing Reaction.
3. Gel Electrophoresis & Computer Processing.

Chromosomes, which range in size from 50 million to 250 million bases, must first be broken into much shorter pieces (PCR step). This was described earlier in this chapter. Each short piece is used as a template to generate a set of fragments that differ in length from each other by a single base that will be identified in a later step (template preparation and sequencing reaction steps). The fragments in a set are separated by gel electrophoresis (separation step). New fluorescent dyes allow separation of all four fragments in a single lane on the gel.

Cassettes with a specific number of lanes each containing a DNA sample (or more if the DNA is pooled, see Section 1.2) will be in-

serted into the genome sequencer.

The final base at the end of each fragment is identified (base-calling step). This process then recreates the original sequence of As, Ts, Cs, and Gs for each short piece generated in the first step. The fluorescently labelled fragments that migrate through the gel, are passed through a laser beam at the bottom of the gel. The laser excites the fluorescent molecule, which sends out light of a distinct colour. That light is collected and focused by lenses into a spectrograph. Based on the wavelength, the spectrograph separates the light across a CCD camera (charge coupled device). Each base has its own colour, so the sequencer can detect the order of the bases in the sequenced gene (Sanger, Nicklen & Coulson (1977)).

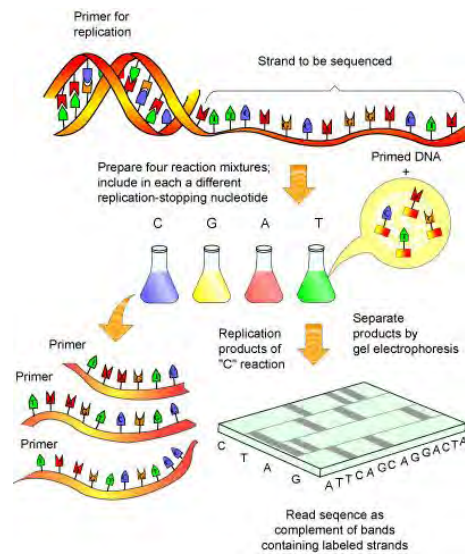


Figure 1.2: Genome Sequencing

1.2 Pooling

Compared to the separate sequencing of individuals, pooling is a cost effective sequencing strategy, where DNA material from more than one individual is placed in a single lane of the sequencer. This has been shown to be a practical way to reduce the cost of large-scale association studies to identify susceptibility loci for common

diseases (Sham et al. (2002)). Large pools increase the chance of capturing rare alleles, but make it more difficult to avoid false positives due to sequencing errors.

Ramsey & Futschik (2012) extended the results of Futschik and Schlötterer (2010) on the sequencing of anonymous, independently (because DNA samples are from different groups of individuals) pooled DNA samples. They assumed that the number of reads of a site from a lane has a Poisson distribution and the probability of an error in any given read is a known constant. A conservative way of choosing the probability of error is to take the maximum of the estimates of the probabilities of error. Another is to take the average of these estimates in some way. Futschik and Schlötterer (2010) considered the probability of detecting a SNP using a threshold rule, i.e., an allele is inferred to exist if the number of times it is read in a lane is greater than some chosen number. They also considered the probability of wrongly inferring that a site is a SNP.

It should be noted that the tests of Ramsey & Futschik (2012) are adapted to experiments where DNA pooling is used but not tagging (see Section 1.3). Their model will be considered in more detail in Chapter 4.

1.2.1 Advantages and disadvantages

Sequencing a large pool of individuals simultaneously keeps the number of redundant DNA reads low (redundancy means having multiple reads of the same nucleotide), and thus provides an economic alternative to the sequencing of individual genomes. Pooling DNA can be a cost-effective procedure since samples can be processed relatively cheaply. Pooling can be particularly important if the goal of analyses is just to detect SNPs, i.e only the presence or absence of an allele is important, not who has it.

On the other hand, more care has to be taken to establish appropriate control of sequencing errors. Another disadvantage of pooling is that one cannot infer genotypes of individuals, since one cannot infer which pair of chromosomes come from a particular individual.

1.3 Gene tagging

Even when pooling, one might like to know which individual a read comes from. In order to do this, gene tagging can be used, i.e., if pooling is used, then after segmentation, DNA from individuals is tagged before pooling. A tag is a short unique string of nucleotides which is added to each segment. Using this method, genes of interest can be located on the chromosome. Tagged SNPs can be used to discover or further investigate the genes which might be responsible for certain diseases (International HapMap Consortium (2003)).

Gene tagging is only practical for large organisms, which are by nature diploid. The data obtained from such sequencing can be thought of as using a pool size of 2 when sequencing haploid individuals. This is due to the fact that one cannot directly observe from which chromosome in a pair a read comes from.

Chapter 2

Review of Multiple Comparisons

Statistical analysis of a dataset typically involves testing many hypotheses. For any particular test, one may assign a pre-set probability α of a type I error (i.e., a false positive, rejecting the null hypothesis when in fact it is true). The problem is that using a value of $\alpha = 0.05$ means that roughly one out of every twenty such tests will show a false positive (rejecting the null hypothesis when in fact it is true). Thus, if the experiment involves performing 100 tests, one expects the null hypothesis to be rejected 5 times if a value of $\alpha = 0.05$ for each has been used, even if the null hypothesis is always true.

Suppose that the null hypothesis is always true. In general, if one performs m hypothesis tests, then α is the probability of making an error (for a single test). Hence, the probability of not making an error (for a single hypothesis) is $1 - \alpha$. Suppose that the test statistics are independent. The probability of not making an error in m tests is thus $(1 - \alpha)^m$, i.e., the probability of making at least one error in m tests is $1 - (1 - \alpha)^m$. So, if $\alpha = 0.05$ and one wants to test 100 tests, then the probability of making at least one error is $1 - (1 - 0.05)^{100} \cong 0.99$. This is the problem of multiple comparisons, in that one would like to control the false positive rate not just for any single test but also for the entire collection (or family) of tests that makes up our experiment.

Experience indicates that less than 1% of the sites are SNPs. Due to this, if one tests at a fixed significance level of 5%, the majority of the sites inferred to be SNPs will be false positives.

In order to reduce this danger, the tests should not be conducted on a per comparison basis. Classical Multiple Comparison Procedures (MCPs), aim at controlling the probability of committing even a single type I error within the tested family of hypotheses referred to as Family-Wise Error-Rate (FWER) (Benjamini and Hochberg (2000)). Thus FWER controlling procedures answer both concerns that, according to Cox (1965), should lead a researcher to control for the multiplicity effect: selection bias, and simultaneous correctness. For example, FWER control has been recommended by Tukey (1977) for the specific problem of subgroup analysis in clinical trials.

The main problem with such classical MCPs, which hinders their application in applied research, is that since they control the FWER, they tend to have substantially less power than the per comparison procedures of the same levels. Benjamini and Hochberg (1995) introduced the False Discovery Rate (FDR) - the expected ratio of erroneous rejections to the number of rejected hypotheses - as an appropriate error rate to control in many problems where the control of the family-wise error-rate is not of concern.

In this chapter, the Bonferroni correction as one of the classical methods for controlling the type I error, will be reviewed and then the Benjamini-Hochberg procedure will be explained. Another method to control the FDR introduced by Storey and Tibshirani (2003) will be also introduced.

2.1 Bonferroni correction

The Bonferroni correction (Miller, R. G. (1981)) is a statistical adjustment for the multiple comparisons problem. It was developed and introduced by Italian mathematician Carlo Emilio Bonferroni (1892-1960). This correction is based on the idea that if one is testing m independent correct null hypotheses on a set of data, then the probability of not making any type I (false positive) errors in m independent tests, each of significance level α , is $(1 - \alpha)^m$. Hence,

the probability of at least one false positive (the family-wise error-rate, FWER, the probability that at least one true null hypothesis is rejected) is $1 - (1 - \alpha)^m$. One way of maintaining the family-wise error-rate is to test each individual hypothesis at a statistical significance level of $\frac{1}{m}$ times what it would be if only one hypothesis were tested. Thus, if the desired significance level for the whole family of tests should be at most α , then the Bonferroni correction would be to test each of the individual tests at a significance level of $\frac{\alpha}{m}$. It should be noted that this procedure controls the FWER to be $\leq \alpha$ even when the test statistics are correlated.

While helpful when used correctly, concerns have been expressed about possible misuse and misunderstanding of the Bonferroni correction. For example, suppose one has 5 independent tests and the p -values are all between 0.01 and 0.05. Using the Bonferroni correction, one would not reject any of the H_0 . Given all the H_0 are true, then the p -values have a uniform distribution on $[0, 1]$ (if the test statistic has the appropriate distribution). Hence, this conclusion seems highly unreasonable. The Bonferroni correction is very conservative.

2.2 Benjamini-Hochberg procedure

Suppose one tests m null hypotheses of which m_0 are true (m_0 is an unknown constant). The results of these tests can be presented in the form of a contingency table (Benjamini-Hochberg (1995)):

Table 2.1: Benjamini-Hochberg Procedure

	H_0 not rejected	H_0 rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	$m_1 = m - m_0$
	$m - R$	R	m

V is the number of false positives, S is the number of true positives, and R is the total number of hypotheses called significant. Also, m_0 is the number of true null features in the study, and $m_1 = m - m_0$ is the number of true alternative features (in this context, features can be thought of as SNPs). These quantities can be used to form an overall error measure for any given

p -value cut-off. Regardless of whether the p -value threshold is fixed or data-dependent, the quantities V , S , and R are random variables, therefore, it is common statistical practice to write the overall error measure in terms of an expected value, which is denoted by $E[\cdot]$.

The Family-Wise Error-Rate (FWER) is defined to be the probability that at least one true null hypothesis is rejected, i.e., $P(V \geq 1)$. If the false positive rate is the error measure used, then a simple p -value threshold is used. By assuming $m_0 = m$, an upper bound on the FWER is obtained.

A p -value threshold of 0.05, for example, guarantees only that the expected number of false positives is $E[V] \leq 0.05m$. For genome-wide studies this number is much too large, and thus control of the false positive rate is too liberal. However, control of the family-wise error-rate is much too conservative for many of the genome-wide studies currently being performed, especially where many of the alternative hypotheses are expected to be true.

It is therefore useful to find an error measure in between these, specifically, one that provides a sensible balance between the number of false positive features, V , and the number of true positive features, S . This balance can be achieved efficiently by considering the ratio,

$$Q = \frac{\text{no. of false positive features}}{\text{no. of significant features}} = \frac{V}{V + S} = \frac{V}{R},$$

which can be stated in words as the proportion of false positive features among all of those called significant. If the number of rejections, R , is zero, Q is defined to be zero. FDR is defined to be the expected value of this quantity,

$$FDR = E\left[\frac{V}{V + S}\right] = E\left[\frac{V}{R}\right].$$

It can be shown that if all the null hypotheses are true, then $FDR = FWER$, therefore control of the FDR implies control of the FWER in the weak sense. Also when $m_0 < m$, it can be shown that the FDR is smaller than or equal to FWER.

The term false positive ratio usually refers to the proportion of tests in which the null hypothesis is falsely rejected. Using the terminology suggested here, this is simply $\frac{V}{m_0}$. Since V is a random variable, m_0 is a constant and $V < m_0$, the false positive ratio is also a random variable ranging between 0 and 1.

The False Positive Rate (FPR) refers to the expected value of the false positive ratio, expressed by $E(\frac{V}{m_0})$. Note that $E(\frac{V}{m_0}) \geq E(\frac{V}{m})$.

To use the Benjamini-Hochberg (B-H) procedure, first the p -values are sorted in increasing order from the smallest to the largest. The ordered p -values are denoted by $p_{(1)}, p_{(2)}, \dots, p_{(m)}$. Now suppose k is the largest value of i such that $p_{(i)} \leq \frac{i\alpha}{m}$. Then the k null hypotheses corresponding to the p -values $p_{(1)}, p_{(2)}, \dots, p_{(k)}$ will be rejected. This controls the FDR to be $\leq \alpha$ when the test statistics are independent.

It should be noted that using the FDR approach in practice, dependent test statistics are met more often than independent ones. This issue is addressed by using the multiple endpoints example in a paper by Benjamini and Yekutieli (2001). Benjamini, Hochberg and Kling (1997) showed that the same procedure as described above, can control the FDR for equally positively correlated normally distributed (possibly Studentized) test statistics. Benjamini and Yekutieli (2001) proved that the procedure controls the FDR in families with positively dependent test statistics.

2.2.1 An algorithm for the Benjamini-Hochberg procedure

Since ordering a large vector of p -values is numerically intensive and it is expected that less than 1% of the null hypotheses should be rejected, the following algorithm for carrying out the procedure has been adopted:

1. Set the significance level to be α .
2. Do not reject any null hypothesis for which the p -value is greater than the current significance level. These hypotheses are removed from the set of hypotheses to be tested. If no hypotheses are re-

moved, then reject the null hypotheses that remain.

3. Suppose l hypotheses remain in the set to be tested. Set the significance level to be $\frac{\alpha l}{m}$ and return to 2.

Suppose k hypotheses at any step are removed where l hypotheses are left at the end. All the p -values, $p_{(l+1)}, p_{(l+2)}, \dots, p_{(l+k)}$ are greater than $\frac{(l+k)\alpha}{m}$. This holds for all the steps, so all the null hypothesis retained under this procedure, are retained under the B-H procedure.

Also, if $l > 0$ at the end of the procedure, then $p_{(l)} < \frac{\alpha l}{m}$ and thus the null hypotheses rejected under this procedure are rejected under the B-H procedure.

It is important at each stage to remember which p -value corresponds to which test that is still to be tested.

2.3 FDR control by modelling the distribution of the p -value by a mixture of distributions

D. Storey and J. Tibshirani (2003) proposed an approach which was claimed to avoid a flood of false positive results. This approach is based on a measure of statistical significance, called the q -value, which gives each feature its own individual measure of significance.

The difference between the p -value and the q -value is that the p -value is a measure of significance in terms of the false positive rate, whereas the q -value is a measure of significance in terms of the FDR.

Consider Table 1.1. The FDR can also be written in terms of the well known specificity, $(m_0 - V)/m_0$, and sensitivity, S/m_1 :

$$FDR = E \left[\frac{m_0(1 - \text{specificity})}{m_0(1 - \text{specificity}) + m_1 \text{sensitivity}} \right].$$

Clearly, the FDR is a useful measure of the overall accuracy of a set of significant features of many genome-wide studies. But one would also like a measure of significance that can be attached to each individual feature.

The q -value is a measure designed to reflect this level of significance. Suppose that the features are listed in order of their evidence against the null hypothesis. It is practical to arrange the features in this way because calling one feature significant means that any other feature with more evidence against the null should also be called significant. Hence, as under the B-H procedure, the features will be listed from smallest to largest p -value. If a threshold value is chosen, features are called significant when they correspond to p -values smaller than this threshold. The q -value for a particular feature is the expected proportion of false positives incurred when calling that feature (and all those corresponding to smaller p -values) significant, while features corresponding to larger p -values are considered to be non-significant. Therefore, calculating the q -values for each feature and thresholding them at q -value level α produces a set of significant features such that a proportion α of them are expected to be false positives.

2.3.1 Calculation of the q -value

The definition $FDR = E(\frac{V}{R})$ is somewhat loose. It will almost always be the case that $R = 0$ with positive probability, which implies that $E(\frac{V}{R})$ is undefined. The quantity $E(\frac{V}{R} | R > 0) \times P(R > 0)$ was proposed as a solution to this problem, which is the result of setting $\frac{V}{R} = 0$ whenever $R = 0$ as in the definition given above.

This quantity is technically called the FDR in the statistics literature. In this case one wants to place a measure of significance on each feature, which is done under the assumption that the feature is called significant. Thus, the inclusion of $P(R > 0)$ is somewhat awkward. An alternative quantity, called the pFDR, was recently proposed, which is simply defined as $pFDR = E(\frac{V}{R} | R > 0)$. The q -value corresponding to a feature is most technically defined as the minimum pFDR at which a feature can be called significant. Because m is large in genome-wide studies, $P(R > 0) \approx 1$ and $FDR \approx pFDR \approx \frac{E(V)}{E(R)}$, so the distinction is not crucial here. Suppose that each feature's statistic probabilistically follows a random mixture of a null distribution and an alternative distribution. Then under a fixed significance rule, the pFDR can be written as $P(\text{feature}$

i is truly null | feature i is significant), for any $i = 1, \dots, m$. Similarly, the false positive rate can be written as $P(\text{feature } i \text{ is significant} \mid \text{feature } i \text{ is truly null})$, for any $i = 1, 2, \dots, m$. Notice the similarity between the pFDR and false positive rate: the arguments have simply been swapped in the conditional probabilities. This connection is the motivation for calling our proposed quantity the q -value.

Indeed, the p -value of a feature is technically defined to be the minimum possible false positive rate when calling that feature significant. Likewise, the q -value is based on the minimum possible pFDR.

2.4 A pseudo-Bayesian approach

Bayesian inference is a method of inference in which Bayes' rule is used to update the estimate of the probability that a particular hypothesis is true as additional evidence is learned. This theorem is named after Reverend Thomas Bayes (1702-1761), and is also referred to as Bayes' law or Bayes' rule (Bayes and Price 1763). Bayes' theorem expresses the conditional probability, or posterior probability, of an event A given B is observed in terms of the prior probability of A , prior probability of B , and the conditional probability of B given A .

Genome sequencing of human DNA by the HapMap and the Seattle SNP projects indicated that around 1 in 300 sites showed variation. This can be used as a conservative estimate of the proportion of sites which are SNPs on the basis that the larger the sample, the more SNPs are expected to be found. Let $P(H_0)$ and $P(H_A)$ denote the probabilities that H_0 and H_A are true, respectively, where H_0 is the hypothesis that a site is not a SNP and H_A is the hypothesis that a site is a SNP. Thus,

$$P(H_0) = \frac{299}{300} = 1 - \pi; \quad P(H_A) = \frac{1}{300} = \pi.$$

One can get a posterior measure of the probability, W , of H_A given the data, X (in this context, the reads), using Bayes' law.

Define,

$$W = \frac{\pi L(X|H_A)}{\pi L(X|H_A) + (1 - \pi)L(X|H_0)},$$

where $L(X|H_0)$ is the likelihood of the data under H_0 , i.e., the site is not a SNP, and $L(X|H_A)$ is the maximum likelihood of data under H_A , i.e., the site is a SNP. Note that here the maximum likelihood has been calculated with respect to the frequency γ of the minor allele. This will be considered in more detail when the maximum likelihood test is defined in Chapter 4.

Hence, W is by definition between 0 and 1. It should be noted that since $L(X|H_A)$ is calculated by maximising the likelihood function under the alternative hypothesis, which is a composite hypothesis (because the parameter of the distribution, γ , is not specified) this can be thought of as a positively biased estimate of the posterior probability of a site being a SNP given the prior probabilities. A simple rule for stating which sites are SNPs is as follows:

Accept that a site is a SNP if and only if $W > w^*$.

It is necessary to choose an appropriate value of w^* . Using Bayesian decision theory, this value depends on the relative costs of errors in inference. Errors of type I occur when a site that is not a SNP is determined to be a SNP. Errors of type II occur when a site that is a SNP is determined not to be a SNP. In the case where the costs of these errors are equal, then from the argument above, some w^* should be chosen, such that $w^* \geq 0.5$. It should be noted that costs for wrongly inferring a SNP are high compared to wrongly inferring no variation. So in this case, a higher threshold should be used.

Chapter 3

Description of the Problem

3.1 The data

The data used in this study can be represented as two matrices. The rows of each matrix correspond to sites (loci) and the columns to individuals. Each entry in the first matrix gives the set of reads for a given individual at a given site (Table 3.1). For example, the first read for individual 1 at site 1 is A, the second read is T and so on. The main issue here is to determine whether a read is from a major allele, a minor allele or an error, i.e., it is necessary to distinguish reads from individuals with a rare allele from sequencing errors. For example, for individual 1 in Table 3.1, the following two possibilities are the most likely: at site 1 the Ts could be errors, or the genotype of the individual might be AT.

Table 3.1: The Structure of the Data

Sites	Individual 1	Individual 2	...	Individual n
Site 1	ATAATAAAAA	AATAATAAAAA	...	AATAAAAAAAA
Site 2	AAAAAAA	AAAAAAA	...	AA
Site 3	CCCCCG	CCC	...	CCCCCC
...
Site n_s	GGGGG	GGGGCA	...	GGGGT

It should be noted that the different reads for an individual at the same site are not treated as repeated measurements. This is because these reads are obtained from different cells of that individual

(read 1 from one cell, read 2 from another cell and so on). These cells are taken from an individual at random.

In general, there are two reasons why multiple reads are obtained from an individual at a site. As noted previously, the genome sequencer makes errors reading a site, so multiple reads are needed to say whether the nucleotide is in fact there or it is an error. But more importantly, for diploid individuals one wants to observe both alleles. An allele is chosen each time at random, due to the physical process of chopping up the DNA. Hence, multiple reads are needed at each site to observe both of the alleles.

For each set of reads in Table 3.1, there is a corresponding set of quality scores (Table 3.2). As described before, when a read is made there is always a possibility of an error. These quality scores correspond to estimates of the probability that a read is not in fact the nucleotide which is at that site (the probability of reading a base incorrectly). The quality of a base call is defined to be $q = -10\log_{10}(p)$ where q is the quality and p the error probability. As noted previously, the higher the quality score, the lower the probability of error.

Table 3.2: The Quality Scores

Sites	Individual 1	...	Individual n
Site 1	28,30,38,38,40,28,25,29,30,33	...	25,30,33,40,42,38,30,25,25,30
Site 2	35,40,30,28,25,26,24	...	38,40
Site 3	40,38,35,28,29,30	...	35,40,35,40,38,25,25
...
Site n_s	28,26,30,40,38,25	...	28,40,38,35,25,25

It should be noted that there is no information on the probability of a specific type of an error (i.e., A is read as G). A probability of an error only gives the probability with which a base is called wrongly.

The data are obtained from the Trinity Sequencing Laboratory in Dublin. The dataset consists of 160 individuals, for which there are 320 columns (160 for the DNA sequence and 160 for the quality scores). 60,000 sites were considered for each individual, therefore the dataset is a matrix with 60,000 rows and 320 columns.

3.2 The model

Assume that one knows from which individual (assumed to be diploid) any read comes from (i.e., either DNA is not pooled or it is tagged). Initially a particular site under a model where there are just two possible alleles is considered. Suppose there are n individuals. As the number of reads varies between different individuals, let m_i be the realisation of the number of reads for individual i , and $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$ be the vector of reads for individual i . A read should be understood as the inferred type of the nucleotide chosen (A, C, T or G). As said before, the genome sequencer can read a base (nucleotide) incorrectly. An error is made when the inferred type does not correspond to the actual type. Let $\hat{\mathbf{p}}_i = (\hat{p}_{i,1}, \hat{p}_{i,2}, \dots, \hat{p}_{i,m_i})$ be the vector of estimates of the probability of error. In practice, these estimates are of order 0.01 or less. It should be noted that these estimates of the probability of error are externally calculated (using the software installed in the genome sequencer). It is assumed that the number of reads for a site has a Poisson distribution. It should be noted that the likelihood test defined below is based on the conditional likelihood given the number of reads for each individual. Thus, the number of reads takes the form of an ancillary statistic (Lehmann, 1986, Chapter 10, Section 2).

The major allele at a site is assumed to be the allele with the largest total number of reads. When the minor allele is relatively rare, the major allele will be correctly determined with probability close to 1. On the other hand, when the minor allele is relatively common, any sensible test will detect it with a probability close to 1. Hence, for practical purposes the possibility of calling the wrong allele as the major allele is ignored.

Other definitions are also possible. For example, one could define the score of an allele at a site to be the number of individuals for which that allele has the greatest number of reads. The allele with the largest score is then defined to be the major allele. In practice, one can use either of the above definitions for a major allele. Any other sensible definition of the major allele would lead to inferring the same major allele unless the minor allele frequency was close to

50%, and hence the minor allele is easily detectable.

First consider a simple model with the following assumptions. If a site is not a SNP then all errors at a site result in a read of the same (incorrect) nucleotide, the prospective minor allele. If the site is a SNP, then all errors in reading the major allele result in a read of the minor allele and all errors in reading the minor allele result in a read of the major allele. Hence, just two alleles can be read: the major allele and the (prospective) minor allele.

To simplify the notation, let \mathbf{X} be the $n \times M$ matrix of reads for a particular site, where M is the maximum number of reads for an individual. If $j \leq m_i$ then $X_{i,j}$ is the j -th read for individual i , otherwise $X_{i,j} = 0$ which indicates no read.

The matrix $\mathbf{P}_{n \times M}$ of estimates of the probability of error can be defined in a similar manner (note all the $\hat{p}_{i,j}$ are greater than 0).

Chapter 4

Methods

4.1 The idea of a fixed threshold rule

Futschik & Schlötterer (2010) considered the probability of detecting a SNP using a threshold rule, i.e. an allele is inferred to exist if the number of times it is read in a lane is greater than some chosen number. Ramsey & Futschik (2012) extended the results of that paper on the sequencing of anonymous, independently pooled DNA samples. They presented a test based on the threshold rule, to detect the SNPs while controlling the probability of wrongly inferring that a site is a SNP.

Suppose that there is a given number of available lanes and there are good estimates of the mean number of reads per lane and the error rate. Consider a single site. Let λ and ϵ be the read and error rates respectively (λ and ϵ are fixed, when the error rate is variable one can take some form of average error rate). The test statistic $K = \max_{1 \leq i \leq n} K_i$ can be defined, where K_i is the number of reads for the prospective minor allele from individual i and n is the number of individuals. Let k be the realisation of K . Then,

$$P(K \geq k) = 1 - P(K < k) = 1 - \prod_{i=1}^n P(K_i < k),$$

from the independence of the number of reads for each individual.

By assumption, $K_i \sim \text{Poisson}(\lambda\epsilon)$, therefore,

$$P(K \geq k) = 1 - P(K_i < k)^n.$$

Let γ denote the relative frequency of the minor allele. First the following hypothesis for each site is tested:

H_0 : The site is not a SNP, i.e., $\gamma = 0$,

H_A : The site is a SNP, i.e., $\gamma > 0$.

$P(K \geq k)$ gives the p -value for testing whether a site is a SNP. Since this test is carried out for a large number of sites, a multiple testing procedure will be used. If the Bonferroni correction is used, when the read rate is independent of the site H_0 will be rejected if and only if $k \geq k^*$ where k^* is the smallest k such that $P(K \geq k) \leq \frac{\alpha}{n_s}$ where α is the required significance level and n_s is the number of sites. Thus it would be a fixed threshold rule, whereas if the B-H procedure is used the threshold would be a random variable.

Note that λ is unknown and depends on the site and is thus estimated from the data for each site. The effect of the variability of the read rate will be investigated by the aid of simulations.

4.2 The likelihood ratio test

Consider the simplified model in which it is assumed that only two alleles can be read at a site. Again, initially consider a single site. As noted previously, the one allele with the largest number of reads is defined to be the major allele. The other is the prospective minor allele.

Let $I_{i,j} = 1$ if the j -th read from individual i indicates the major allele and $I_{i,j} = 0$ otherwise. Now $L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, 0)$ can be defined to be the likelihood of the sequence of reads for individual i under the null hypothesis, $\gamma = 0$, by

$$L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, \gamma = 0) = \prod_{j:I_{i,j}=1} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0} \hat{p}_{i,j},$$

where $1 \leq j \leq m_i$, \mathbf{x}_i is the vector of reads and $\hat{\mathbf{p}}_i$ the vector of the estimates of the probability of error for individual i .

Under H_0 a read is an error if and only if it does not indicate the major allele. Therefore under H_0 the likelihood for the whole sample is,

$$L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0) = \prod_{i=1}^n L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, \gamma = 0) = \prod_{i=1}^n \left(\prod_{j:I_{i,j}=1} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0} \hat{p}_{i,j} \right).$$

If there are only reads of one nucleotide (variant) at a site, obviously H_0 cannot be rejected.

Assume that there is a minor allele of relative frequency γ , where $\gamma > 0$. In this case, the probability of a read from an individual indicating a minor allele depends on the genotype of that individual (i.e., on the number of minor alleles that individual i has at the site considered, denoted by A_i). $A_i \sim Bin(2, \gamma)$. Assume that the genotype is given by MM (homozygote with 2 copies of the major allele, $A_i=0$), Mm (heterozygote with 1 copy of the minor allele and 1 copy of major allele, $A_i=1$) or mm (homozygote with 2 copies of the minor allele, $A_i=2$). For example, given the genotype is Mm, the minor allele is sampled with probability 0.5. The probability that the minor allele is sampled and it is read correctly is thus $0.5(1 - \hat{p}_{i,j})$, and the probability that the major allele is sampled and it is read wrongly (as the minor allele) is $0.5\hat{p}_{i,j}$. The probability that a read indicates the minor allele when the genotype is Mm is the sum of these two probabilities, i.e., 0.5.

Let $f_j(a_i)$ be the probability that the j -th read indicates the prospective minor allele given that it comes from individual i with a_i minor alleles in their genotype. Thus $f_j(0) = \hat{p}_{i,j}$, $f_j(1) = 0.5$ and $f_j(2) = 1 - \hat{p}_{i,j}$. For example, note that $f_j(0)$ is the probability that the j -th read indicates the prospective minor allele given that it comes from an individual with 0 minor allele in its genotype. Obviously if there is no copy of the minor allele in the genotype of that individual then the probability of observing a minor allele is equal to the probability of error.

Using the law of total probability, the likelihood of the reads from individual i given the minor allele frequency is given by,

$$L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, \gamma) = \sum_{a_i=0}^2 \left[\binom{2}{a_i} \gamma^{a_i} (1-\gamma)^{2-a_i} \prod_{j:I_{i,j}=1} [1-f_j(a_i)] \prod_{j:I_{i,j}=0} f_j(a_i) \right].$$

Multiplying the likelihoods for each individual,

$$L(\mathbf{X}; \hat{\mathbf{P}}, \gamma) = \prod_{i=1}^n \left\{ \sum_{a_i=0}^2 \left[\binom{2}{a_i} \gamma^{a_i} (1-\gamma)^{2-a_i} \prod_{j:I_{i,j}=1} [1-f_j(a_i)] \prod_{j:I_{i,j}=0} f_j(a_i) \right] \right\}.$$

Let,

$$S = \frac{\max_{0 \leq \gamma \leq 0.5} L(\mathbf{X}; \hat{\mathbf{P}}, \gamma)}{L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)}.$$

Since this maximisation is carried out numerically, it can be assumed that $\gamma \in \{0, \frac{1}{2n}, \frac{2}{2n}, \dots, \frac{n}{2n}\}$. The numerator gives the expected total number of copies of the minor allele in the sample (out of $2n$, given γ). Suppose the maximum is achieved at $\hat{\gamma} = \frac{l}{2n}$. The maximum likelihood estimate of the total number of minor alleles in the n genotypes is l . Now consider a method of testing whether a particular site is a SNP.

4.2.1 The likelihood ratio test statistic

The likelihood ratio statistic is $T = -2 \ln S$. Using standard asymptotic theory, this statistic will have approximately a chi-square distribution with one degree of freedom. A p -value for this test can thus be calculated under this assumption. However, the minor allele frequency under H_0 is at the boundary of the parameter space and so this approximation may well not be appropriate. This will be investigated in the section on results from simulations.

4.2.2 Adaptation of the test: four possible variants at a site

Consider the case in which all four possible alleles (variants) may occur. In this case, calculation of the likelihood function under H_0

is straightforward because under H_0 (no minor allele) any variant other than the major allele is assumed to be an error. Therefore, once the major allele - the allele with the largest number of reads - is detected, likelihood function under H_0 can be calculated using the procedure introduced in section 4.2.

When k non-major alleles are read at a site, then k likelihood ratio statistics may be calculated (one for each non-major allele). For example, if the major allele is A and the current prospective minor allele is G, then any reads of C or T are assumed to be errors. The likelihood function under the hypothesis that the minor allele is G can be calculated using an approach analogous to the one given above. The prospective minor allele is the one for which the greatest likelihood ratio is obtained and the test based on this ratio.

4.2.3 A brief note on the overdispersion of the data

As stated previously, it is assumed that the number of reads from a lane has a Poisson distribution. When models based on that distribution are fitted, the problem of overdispersion could be faced. When data come from the Poisson distribution, the variance is equal to the expected value. Overdispersion occurs when the variance exceeds the expected value.

By looking at the real data for individual sites, it can be easily seen that the variance of the number of reads is larger than the mean. The mean and the variance of the number of reads per site is approximately 38.6 and 357.9, respectively. Obviously the variance of the number of reads is larger than the mean (approximately 9 times).

The Poisson distribution has one free parameter and does not allow for the variance to be adjusted independently of the mean. An alternative model with additional free parameters may provide a better fit. In this case, a Poisson mixture model, which can lead to the negative binomial distribution (McCullagh P. and Nelder J. 1989), can be used instead, where the mean of the Poisson distribution can itself be thought of as a random variable.

Therefore, the number of reads have been allowed to come from either the Poisson distribution or the negative binomial distribution in the simulations.

4.2.4 Bayesian score

The Bayesian score for a site is an estimate of the posterior probability that the site is a SNP and is defined as follows (see Section 2.4),

$$W = \frac{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma})}{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma}) + (1 - \pi)L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)}.$$

Based on the Bayesian scores of sites, the FDR (false discovery rate) and FNR (false negative rate) can be estimated. Consider a large number of sites, n_s . Let S' be the set of sites inferred to be SNPs. Then the empirical estimate of the FDR is,

$$FDR = \frac{\sum_{k \in S'} (1 - W_k)}{|S'|},$$

where W_k is the Bayesian score for locus k and $|S'|$ is the number of elements in the set S' . Similarly, the empirical estimate of the false negative rate can be calculated as follows,

$$FNR = \frac{\sum_{k \in S'^c} W_k}{n_s - |S'|}.$$

The dataset used for this study had 60,000 sites and 171 sites were inferred to be SNPs. The Bayes scores were calculated for each site, and the FDR and FNR were calculated to be as follows (based on the set of sites inferred to be SNPs using the B-H procedure):

$$FDR = 0.00635,$$

$$FNR = 0.00021.$$

It is assumed that there is a false negative when a minor allele appears in the sample but is not detected. In practice, it is possible that minor alleles of low frequency do not appear in a sample.

If one replaces $\hat{\gamma}$ with $\frac{1}{2n}$ in the formula for the calculation of the Bayesian score, a lower bound, W_2 , can be defined for the posterior probability, i.e., by assuming that the alternative hypothesis is that there is one copy of the minor allele in the sample. Hence,

$$W_2 = \frac{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \frac{1}{2n})}{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \frac{1}{2n}) + (1 - \pi)L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)}.$$

These lower bounds are calculated for the Dublin data.

See Appendix A.1 for a full list of sites inferred to be SNPs (from the real data) with the corresponding Bayesian Scores and lower bounds of the Bayesian Scores.

See Appendix A.2 for an example of how the Bayesian score is calculated for a SNP.

4.2.5 Adaptation of the likelihood ratio test to pooled data

The likelihood ratio test can be adapted to pooled data, where an estimate of the error probability is given for each read.

Suppose there are k lanes. In this case, the estimate of the error probability for the j -th read from lane i is denoted by $\hat{p}_{i,j}$ (here i represents the number of the lane). Let $I_{i,j} = 1$ if the j -th read from lane i indicates the major allele and $I_{i,j} = 0$ if it is the prospective minor allele. Define $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,m_i})$ to be the vector of reads for lane i , where m_i is the number of reads in lane i . The likelihood of the reads from lane i assuming the null hypothesis is true is given by,

$$L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, \gamma = 0) = \prod_{j:I_{i,j}=1}^{m_i} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0}^{m_i} \hat{p}_{i,j}.$$

Under H_0 , there is no minor allele present in the lane, so any read of the prospective minor allele is an error.

Now, the likelihood under H_0 for the whole sample can be obtained by multiplying the likelihoods over the k lanes,

$$L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0) = \prod_{i=1}^k \left(\prod_{j:I_{i,j}=1}^{m_i} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0}^{m_i} \hat{p}_{i,j} \right).$$

Under H_A , $\gamma > 0$ (γ is the relative frequency of the minor allele). Define A_i to be the number of copies of the minor allele in lane i , so $\mathbf{A} = (A_1, A_2, \dots, A_k)$, is the vector of the number of copies of the minor allele in lanes 1 to k . Obviously, for a lane containing the DNA of l diploid individuals, $A_i \sim \text{Bin}(2l, \gamma)$.

Suppose the number of times the minor allele appears in the genotypes of the l individuals in lane i is, $a_i \in \{0, 1, 2, \dots, 2l\}$ (a_i is the realization of A_i , which is not observed). Also, if an individual is a homozygote with two copies of the minor allele, then that is counted as two appearances of the minor allele). There are pools of l diploid individuals, so the maximum a_i can take is $2l$. If one further assumes that individuals in a pool all contribute a large and equal amount of DNA, then reads are obtained by binomial sampling from the pool.

Thus, the likelihood of the reads in a lane given γ can be obtained. Suppose $f_j(a_i)$ is the probability that the j -th read in a lane indicates the prospective minor allele given that it comes from lane i with a total of a_i minor alleles in the genotypes of the individuals in that lane. Thus,

$$f_j(a_i) = \frac{a_i(1 - \hat{p}_{i,j})}{2l} + \frac{\hat{p}_{i,j}(2l - a_i)}{2l}.$$

(Note that similar to tagged data, $f_j(0) = \hat{p}_{i,j}$, ..., $f_j(2l) = 1 - \hat{p}_{i,j}$).

The likelihood of the reads from lane i under H_A can be found by conditioning on the number of individuals in the lane with the minor allele, and then using the law of total probability,

$$L_i(\mathbf{x}_i; \hat{\mathbf{p}}_i, \gamma) = \sum_{a_i=0}^{2l} \left[\binom{2l}{a_i} \gamma^{a_i} (1 - \gamma)^{2l - a_i} \prod_{j:I_{i,j}=1} [1 - f_j(a_i)] \prod_{j:I_{i,j}=0} f_j(a_i) \right].$$

Multiplying the likelihoods for each lane,

$$L(\mathbf{X}; \hat{\mathbf{P}}, \gamma) = \prod_{i=1}^k \left\{ \sum_{a_i=0}^{2l} \left[\binom{2l}{a_i} \gamma^{a_i} (1-\gamma)^{2l-a_i} \prod_{j:I_{i,j}=1} [1-f_j(a_i)] \prod_{j:I_{i,j}=0} f_j(a_i) \right] \right\}.$$

Using a method analogous to the case where the DNA is tagged or is not pooled, the likelihood ratio is given by $T = -2 \ln S$, where,

$$S = \frac{L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma})}{L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)}.$$

4.3 Estimation of the power for the threshold test when the read rate and error rate are fixed

Here a method to estimate the power of the threshold test analytically is presented when there are reads from individuals. Suppose there are n_s sites of which k are SNPs. Also suppose the pool size is denoted by m and λ is the mean number of reads from an individual. It is assumed that the number of reads from an individual has a Poisson distribution. Suppose \hat{p} is our estimate of the error rate.

Let $pval_{(i)}$ be the i -th smallest p -value from the n_s sites. Also let $X_{(i)}$ be the number of reads of the minor allele associated with $pval_{(i)}$, i.e., $X_{(i)}$ is the i -th largest number of reads of the minor allele from a lane.

It is assumed that at each SNP there is one copy of the minor allele present and that these k SNPs correspond to the k smallest p -values. Also, suppose that none of the non-SNPs are inferred to be SNPs. These assumptions mean that a lower bound on the probability of detecting a minor allele which is actually present in a sample can be obtained.

Let X be the number of reads of a prospective minor allele for an individual. Let L be the number of SNPs detected using the test. First $P(L = i), i = 1, \dots, k$ is calculated. The power of the test is calculated using,

$$Power = \frac{E(L)}{k} = \frac{\sum_{l=1}^k lP(L=l)}{k},$$

i.e., the expected proportion of the minor alleles that are detected.

Results of the calculations are presented in Section 5.3.

4.4 A fixed significance level based on the assumed density of SNPs

Genome-wide association studies (GWAS) have so far yielded a large number of associations between SNPs and complex diseases (McCarthy, M. I., Abecasis, G. R., Cardon, L. R. et al. 2008). In many studies, a lenient cut-off value of $P < 10^{-5}$ is used, which has been also used by the National Human Genome Research Institute for the identification and archiving of putative associations. However, not all of these associations are genuine (Ioannidis, J. P. 2007). Some of these associations have exceedingly low p -values and can be considered definitively true, whereas many others have modest p -values and only a small fraction of them are likely to be true.

The CARDIoGRAMplusC4D Consortium (2012, consisted of more than 160 authors) used a fixed threshold method to detect SNPs in large-scale association analysis to identify new risk loci for coronary artery disease. They used a fixed significance level called the genome-wide significance level (5×10^{-8}). Using this fixed significance level, any p -value $\leq 5 \times 10^{-8}$ is considered to be significant, and hence the corresponding locus a risk factor.

This approach is based on the expectation that one in a million loci is a risk factor for cancer. Suppose one expects to reject a proportion q of the null hypotheses. Using the B-H procedure, one would expect that a null hypothesis will be rejected iff the p -value is $\leq q\alpha$. Hence, one can approximate the B-H procedure by using a fixed significance level of $q\alpha$.

The power of such a test has been calculated with an assumed density of SNPs to detect a single copy of the minor allele using the

threshold test. This is simply the probability that the number of reads of the minor allele from an individual with one copy of the minor allele exceeds the appropriate threshold. From this a bound on the FDR has been obtained by using such an approach.

Results of the calculations are presented in Section 5.4.

Chapter 5

Simulations

5.1 Introduction

Two different groups of simulations have been presented. The first simulates data when DNA is either unpooled or tagged. The second simulates data when DNA is pooled, but not tagged. The likelihood ratio test (L-R test) and the threshold test have been applied to these data in order to allow us to compare the results from these two methods.

It should be noted that these simulations reflect the sample obtained from Dublin and the parameters of the genome sequencer used. It is not intended to be a full investigation into the properties of such tests.

The simulations have been done using R DEVELOPMENT CORE TEAM. (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.

5.2 Calling SNPs with unpooled or tagged DNA

As described before, when the DNA is not pooled or it is tagged, one knows from which individual any read comes from. In this case, one may assume that one individual is placed in each lane of a sequencer, which gives a random number of reads of the nucleotide at

a particular site.

Simulations have been carried out to investigate the power of the test, defined as the proportion of the actual SNPs found and the false discovery rate, FDR. The genotype of 160 individuals with 500 sites for each has been simulated. The minor allele frequency has been set to 0.01 for the first 5 sites and 0 for the rest. Therefore, the first 5 sites are SNPs and the expected number of minor alleles for each SNP is $(0.01 \times 160 \times 2) = 3.2$.

It should be noted that it is possible that such a minor allele does not appear in the sample. As explained before, the total number of the minor alleles, A , has a Binomial distribution with parameters 320 and 0.01, i.e., $A \sim Bin(320, 0.01)$. Note that if $A \sim Bin(320, 0.01)$, then A has approximately a Poisson distribution with parameter 3.2. So,

$$P(A = 0) = \frac{e^{-3.2} \times 3.2^0}{0!} = e^{-3.2} = 0.040762203.$$

So it implies that if the power is around 96%, then the test is essentially discovering a minor allele whenever it appears in the sample.

Also, since the likelihood function for the whole set of the reads from a site is very close to zero, for numerical reasons the logarithm of the likelihood function has been calculated. Firstly, the likelihood of the reads from each single individual is calculated. Secondly, the logarithms of these likelihoods are calculated and summed together to obtain the value of the logarithm of the likelihood function for the whole set of reads.

Four different error rates, including the mean empirical error rate which is calculated using the real data (Dublin data), and three different read rates for generating the data have been considered (simulations were also carried out using the empirical distribution of the error rate). Two different distributions (Poisson and negative binomial (NB)) have been also used for the number of reads.

The process has been repeated 50 times, with the error rates,

read rates and the two distributions, thus the total number of SNPs after 50 times running the simulations is $(5 \times 50) = 250$.

Table 5.1: Power and FDR of Tests for Tagged DNA Samples by Read Rate and Error Rate (Number of Reads Has a Poisson Distribution)

		Error Rate			
		0.01	0.005	0.001	0.0008
Read Rate		Likelihood Ratio Test			
5	Real SNPs Detected	219 (87.6%)	211 (84.4%)	209 (83.6%)	218 (87.2%)
	FDR	0 (0.0%)	1 (0.472%)	3 (1.415%)	1 (0.457%)
10	Real SNPs Detected	235 (94.0%)	235 (94.0%)	236 (94.4%)	238 (95.2%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.418%)
20	Real SNPs Detected	239 (95.6%)	231 (92.4%)	234 (93.6%)	242 (96.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Read Rate		Threshold Test			
5	Real SNPs Detected	152 (60.8%)	158 (63.2%)	187 (74.8%)	197 (78.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
10	Real SNPs Detected	204 (81.6%)	222 (88.8%)	235 (94.0%)	237 (94.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.420%)
20	Real SNPs Detected	239 (95.6%)	231 (92.4%)	234 (93.6%)	242 (96.8%)
	FDR	0 (0.0%)	0 (0.0%)	2 (0.847%)	7 (2.811%)

Table 5.2: Power and FDR of Tests for Tagged DNA Samples by Read Rate and Error Rate (Number of Reads Has an NB Distribution)

		Error Rate			
		0.01	0.005	0.001	0.0008
Read Rate		Likelihood Ratio Test			
5	Real SNPs Detected	177 (70.8%)	182 (72.8%)	191 (76.4%)	202 (80.8%)
	FDR	0 (0.0%)	0 (0.0%)	1 (0.521%)	1 (0.493%)
10	Real SNPs Detected	213 (85.2%)	224 (89.6%)	223 (89.2%)	226 (90.4%)
	FDR	2 (0.930%)	1 (0.444%)	2 (0.889%)	1 (0.441%)
20	Real SNPs Detected	242 (96.8%)	234 (93.6%)	235 (94.0%)	239 (95.6%)
	FDR	1 (0.412%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Read Rate		Threshold Test			
5	Real SNPs Detected	140 (56.0%)	126 (50.4%)	162 (64.8%)	170 (68.0%)
	FDR	0 (0.0%)	0 (0.0%)	3 (1.818%)	1 (0.585%)
10	Real SNPs Detected	171 (68.4%)	203 (81.2%)	215 (86.0%)	224 (89.6%)
	FDR	0 (0.0%)	0 (0.0%)	4 (1.826%)	1 (0.444%)
20	Real SNPs Detected	231 (92.4%)	225 (90.0%)	234 (93.6%)	239 (95.6%)
	FDR	0 (0.0%)	0 (0.0%)	9 (3.704%)	5 (2.049%)

Table 5.1 shows the results of the simulations for the tagged DNA samples when the number of reads comes from a Poisson distribution with four different error rates and three different read rates.

Table 5.2 shows the results of the simulations for the tagged DNA samples when the number of reads comes from a negative binomial distribution with four different error rates and three different read rates.

It can be seen from Table 5.1 that at a low read rate and a high error rate (reads per individual = 5, error rate = 0.01) the L-R test gives better results than the threshold test (87.6% v.s. 60.8%). However, by increasing the read rate per individual and decreasing the error rate (reads per individual = 20, error rate = 0.0008) the number of SNPs detected by the L-R test and the threshold test get close to each other (96.8% v.s. 96.8%). Although the FDR for the threshold test in this case is higher in comparison with the L-R test (2.811% v.s. 0.0%).

It seems that the read rate is a more important factor than the error rate in determining the relative effectiveness of the two tests. To further investigate the effect of the read rate and the error rate on the power of the test, regression analysis has been carried out. A quadratic surface model has been fitted with the power of the test as the dependent variable and the read rate and the error rate and the interaction term as explanatory variables. It should be noted that there is no constant term in the regression model, since when the read rate and the error rate are zero, the power of the test is by definition zero (if there are no reads, one cannot state the presence of a minor allele).

It should be noted that the power of the test is presented as a percentage in all regression results.

The results of the regression analysis are presented in Table 5.3.

Table 5.3: Regression Analysis to Assess the Effect of Read Rate and Error Rate on the Power (Number of Reads Has a Poisson Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate	3.505	5.229×10^{-1}	6.703	0.000535***
(Read Rate) ²	-1.147×10^{-1}	1.987×10^{-2}	-5.770	0.001183**
Error Rate	-8.428×10^2	6.105×10^2	-1.380	0.216668
(Error Rate) ²	9.469×10^4	5.207×10^4	1.818	0.118874
Read Rate \times Error Rate	-1.076×10^1	1.70×10^1	-0.546	0.604607
Adjusted $R^2 = 0.8792$				
Threshold Test				
Read Rate	7.152	6.944×10^{-1}	10.299	4.89×10^{-5} ***
(Read Rate) ²	-2.400×10^1	2.639×10^2	-9.094	9.93×10^{-5} ***
Error Rate	-4.208×10^3	8.108×10^2	-5.190	0.00203**
(Error Rate) ²	1.558×10^5	6.915×10^4	2.253	0.06518
Read Rate \times Error Rate	1.258×10^2	2.616×10^1	4.810	0.00297**
Adjusted $R^2 = 0.9724$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a Poisson distribution, the read rate and the $(\text{read rate})^2$ have a significant contribution in determining the power of both tests whereas the error rate and the interaction term only contribute significantly to the power of the threshold test. The results show that by increasing the read rate, the power of both tests increases, and by decreasing the error rate, the power of the threshold test increases. Also, the coefficient of the read rate in the regression analysis shows that the impact of the read rate on the power is higher for the threshold test.

Naturally as the read rate increases the power must increase. This condition cannot be enforced using such a regression model, but here the qualitative, rather than quantitative, behaviour of the power is of interest. The fact that the coefficient of $(\text{read rate})^2$ is negative here indicates that the power plateaus as the read rate increases.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.3 are presented in Figure 5.1.

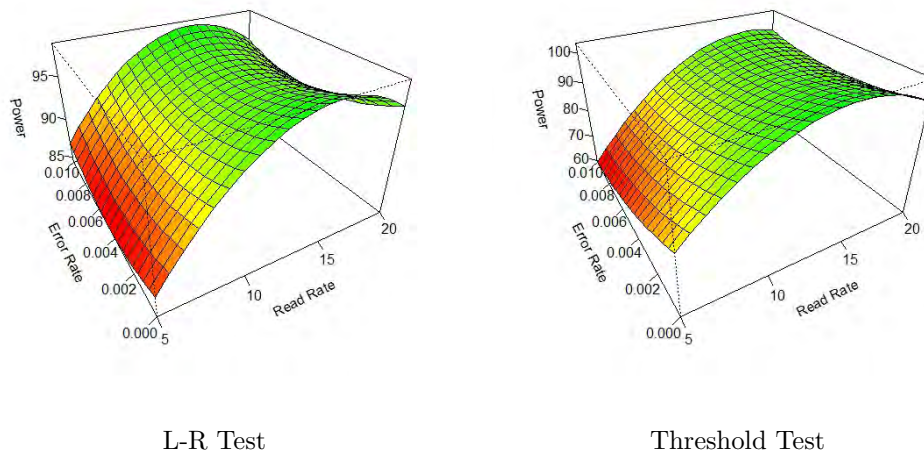


Figure 5.1: 3D Figure of Regression Analysis

Figure 5.1 clearly shows how the power of the tests are affected by the read rate and the error rate. It can be seen that as read

rate increases the power of both tests increases as well. Increasing the error rate, however, does not have a great impact on the power of the L-R test, whereas it does on the power of the threshold test. The power of the threshold test decreases as the error rate increases.

It should be noted that results given in Table 5.1 correspond to the case in which the assumptions of the threshold test are satisfied (i.e., the threshold test works at the highest level of efficiency compared to the L-R test).

Table 5.2 also shows that results for the negative binomial distribution follow the same pattern. At a low read rate and a high error rate (reads per individual = 5, error rate = 0.01) the L-R test gives better results than the threshold test (70.8% v.s. 56.0%). However, by increasing the read rate per individual and decreasing the error rate (reads per individual = 20, error rate = 0.0008) the number of SNPs detected by the L-R test and the threshold test get closer to each other (95.6% v.s. 95.6%). Although the FDR for the threshold test in this case is higher in comparison with the L-R test (2.049% v.s. 0.0%).

Also here, the read rate seems to be a more important factor than the error rate in determining the relative effectiveness of the two tests. To further investigate the effect of the read rate and the error rate on the power of the test, regression analysis has been carried out. As before, a quadratic surface model has been fitted with the power of the test as the dependent variable and the read rate and the error rate and the interaction term as explanatory variables.

Table 5.4: Regression Analysis to Assess the Effect of Read Rate and Error Rate on the Power (Number of Reads Has an NB Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate	4.420	5.777×10^{-1}	7.651	0.000260***
(Read Rate) ²	-1.360×10^{-1}	2.195×10^{-2}	-6.195	0.000815***
Error Rate	-1.732×10^3	6.745×10^2	-2.567	0.042478*
(Error Rate) ²	4.722×10^4	5.753×10^4	0.821	0.443154
Read Rate \times Error Rate	7.149×10^1	2.176×10^1	3.285	0.016717*
Adjusted $R^2 = 0.962$				
Threshold Test				
Read Rate	7.088	1.481	4.785	0.00305**
(Read Rate) ²	-2.093×10^{-1}	5.630×10^{-2}	-3.718	0.00987**
Error Rate	-4.487×10^3	1.730×10^3	-2.595	0.04096*
(Error Rate) ²	2.143×10^5	1.475×10^5	1.453	0.19643
Read Rate \times Error Rate	8.374×10^1	5.580×10^1	1.501	0.18413
Adjusted $R^2 = 0.9178$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a negative binomial distribution, the read rate, the $(\text{read rate})^2$ and the error rate have a significant contribution in determining the power of both tests whereas the interaction term only contributes significantly to the power of the L-R test. The results show that by increasing the read rate or decreasing the error rate, the power of both tests increases. Also, based on the value of the coefficients in the regression analysis, the impact of those variables which are significant for both tests, is higher for the threshold test.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.4 are presented in Figure 5.2.

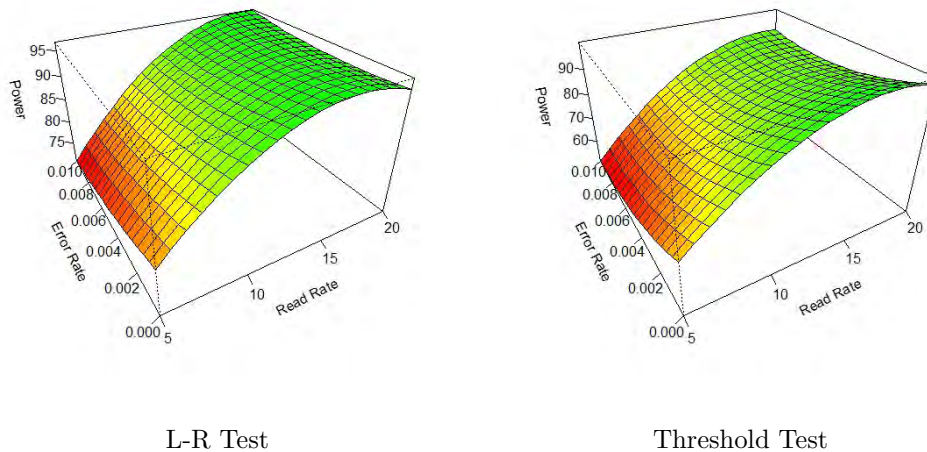


Figure 5.2: 3D Figure of Regression Analysis

Figure 5.2 clearly shows how the power of the tests are affected by the read rate and the error rate. It can be seen that as the read rate increases the power of both tests increases as well. Also, increasing the error rate decreases the power of the tests.

In conclusion, given such a distribution of the number of reads according to site, it is expected that the likelihood ratio test works slightly better than the threshold test.

Since the average read rate and estimates of the probability of error vary, these factors will now be investigated.

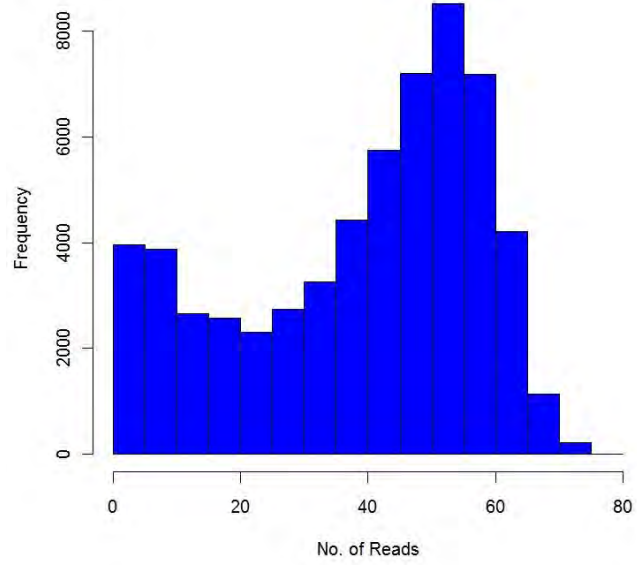


Figure 5.3: Histogram of the Mean Number of Reads from a Site from Real Data

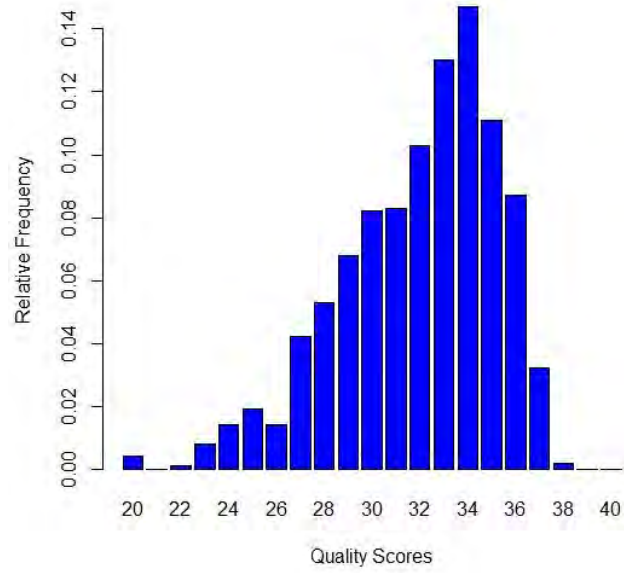


Figure 5.4: Histogram of Quality Scores from Real Data

Figures 5.1 and 5.2 show the histogram of the mean number of the reads and the quality scores from real data, respectively. It can be seen that in many cases the average number of reads per individual is around 40-50, which means that either test will find a real minor allele with probability 1. In quite a lot of cases the mean number of reads is small, which means that neither method works well, but the likelihood ratio method works better.

Tables 5.5 and 5.6 show the results of simulations for tagged data when the number of reads comes from Poisson and negative binomial distributions, respectively, with three different fixed read rates and the probability of an error is generated from the empirical distribution of error probabilities.

Table 5.5: Power and FDR of Tests for Tagged DNA Samples (Number of Reads Has a Poisson Distribution with Fixed Read Rate and Empirical Error Rate)

Distribution		Read Rate		
		5	10	20
L-R test	Real SNPs Detected	222 (88.8%)	236 (94.4%)	238 (95.2%)
	FDR	3 (1.333%)	0 (0.0%)	4 (1.653%)
Threshold Test	Real SNPs Detected	199 (79.8%)	235 (94.0%)	237 (94.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 5.6: Power and FDR of Tests for Tagged DNA Samples (Number of Reads Has an NB Distribution with Fixed Read Rate and Empirical Error Rate)

Distribution		Read Rate		
		5	10	20
L-R test	Real SNPs Detected	199 (79.6%)	238 (95.2%)	240 (96.0%)
	FDR	1 (0.500%)	2 (0.833%)	0 (0.0%)
Threshold Test	Real SNPs Detected	177 (70.8%)	228 (91.2%)	240 (96.0%)
	FDR	0 (0.0%)	1 (0.437%)	8 (3.226%)

It can be seen from Table 5.5 that assuming the number of reads follows a Poisson distribution and using the empirical error rate based on the real data, at a low read rate (read rate per individual = 5) the L-R test works better than the threshold test (88.8% v.s. 79.8%).

However, by increasing the read rate per individual the difference between the proportion of SNPs detected by the two tests gets smaller (read rate per individual = 20, 95.2% v.s. 94.8%).

Table 5.6 shows that assuming the number of reads follows a negative binomial distribution, at a low read rate (read rate per individual = 5) the L-R test gives better results than the threshold test (79.6% v.s. 70.8%).

However, by increasing the read rate per individual the difference between the proportion of SNPs detected by the two tests gets smaller (read rate per individual = 20, 96.0% v.s. 96.0%).

Tables 5.7 shows the results of simulations for tagged DNA samples with the read rate coming from the empirical distribution and mean empirical error rate (0.0008).

Table 5.7: Power and FDR of Tests for Tagged DNA Samples (Empirical Read Rate Used - Mean Empirical Error Rate = 0.0008)

		L-R test	Threshold Test
Distribution			
Poisson	Real SNPs Detected	236 (94.4%)	236 (94.4%)
	FDR	0 (0.0%)	2 (0.840%)
NB	Real SNPs Detected	230 (92.0%)	227 (90.8%)
	FDR	1 (0.433%)	6 (2.575%)

Table 5.8 shows the results of simulations for tagged DNA samples with the mean read rate and error rate both coming from their empirical distributions.

Table 5.8: Power and FDR of Tests for Tagged DNA Samples (Empirical Read Rate and Empirical Error Rate Used)

		L-R test	Threshold Test
Distribution			
Poisson	Real SNPs Detected	239 (95.6%)	238 (95.2%)
	FDR	0 (0.0%)	4 (1.653%)
NB	Real SNPs Detected	226 (90.4%)	210 (84.0%)
	FDR	1 (0.441%)	2 (0.943%)

Table 5.7 shows that assuming the number of reads comes from a Poisson distribution, the L-R test and the threshold test detect the same proportion of SNPs (94.4% v.s. 94.4%).

Assuming that the number of reads follows a negative binomial distribution does not change the proportion of SNPs detected by the two tests dramatically, but on the other hand the FDR for the threshold test is approximately 3 times greater than the FDR for the L-R test (2.575% v.s. 0.840%).

Table 5.8 shows that assuming the number of reads comes from a Poisson distribution, the L-R test and the threshold test, again, approximately detect the same proportion of SNPs (95.6% v.s. 95.2%), although the FDR for the threshold test is 1.653% and for the L-R test is 0.0%.

Assuming that the number of reads follows a negative binomial distribution lowers the proportion of SNPs detected by the L-R test and the threshold test (90.4% v.s. 84.0%). This decrease is more marked for the threshold test.

5.3 Calling SNPs with pooled, untagged DNA

As described before, compared to the separate sequencing of individuals, pooling is a cost effective sequencing strategy, where DNA material from more than one individual is placed in a single lane of the sequencer. Large pools increase the chance of capturing rare alleles, but make it more difficult to avoid false positives due to sequencing errors. In this case, one does not know from what individual a read comes from.

Initially, three different read rates per individual and three different pool sizes have been considered. Obviously, the read rate per lane in this case is the read rate per individual times the pool size. For example, a read rate of 100 corresponds to 20 individuals in a pool with read rate per individual = 5 (or vice versa) or 10 individuals with read rate per individual = 10.

Eight lanes are used in total to generate the data. It should be

noted that the total number of individuals is the number of lanes times the pool size.

Table 5.9: Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has a Poisson Distribution - Error Rate = 0.01)

		Pool Size		
		5	10	20
Read Rate Per Individual		L-R test		
5	Real SNPs Detected	78 (31.2%)	137 (54.8%)	181 (72.4%)
	FDR	2 (2.500%)	1 (0.725%)	1 (0.549%)
10	Real SNPs Detected	126 (50.4%)	176 (70.4%)	220 (88.0%)
	FDR	0 (0.0%)	1 (0.565%)	1 (0.452%)
20	Real SNPs Detected	140 (56.0%)	194 (77.6%)	242 (96.8%)
	FDR	0 (0.0%)	1 (0.513%)	0 (0.0%)
Read Rate Per Individual		Threshold Test		
5	Real SNPs Detected	26 (10.4%)	20 (8.0%)	20 (8.0%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)
10	Real SNPs Detected	62 (24.8%)	65 (26.0%)	60 (24.0%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)
20	Real SNPs Detected	119 (47.6%)	133 (53.2%)	132 (52.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 5.10: Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has an NB Distribution - Error Rate = 0.01)

		Pool Size		
		5	10	20
Read Rate Per Individual		L-R test		
5	Real SNPs Detected	74 (29.6%)	129 (51.6%)	187 (74.8%)
	FDR	2 (2.632%)	1 (0.769%)	2 (1.058%)
10	Real SNPs Detected	110 (44.0%)	177 (70.8%)	222 (88.8%)
	FDR	0 (0.0%)	1 (0.561%)	0 (0.0%)
20	Real SNPs Detected	136 (54.4%)	192 (76.8%)	238 (95.2%)
	FDR	0 (0.0%)	1 (0.518%)	0 (0.0%)
Read Rate Per Individual		Threshold Test		
5	Real SNPs Detected	26 (10.4%)	23 (9.2%)	26 (10.4%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)
10	Real SNPs Detected	46 (18.4%)	70 (28.0%)	72 (28.8%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)
20	Real SNPs Detected	102 (40.8%)	129 (51.6%)	119 (47.6%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 5.9 shows the results of the simulations for the untagged DNA samples when the number of reads comes from a Poisson distribution with three different pool sizes and three different read rates per individual, and the error rate is 0.01.

Table 5.10 shows the results of the simulations for the untagged DNA samples when the number of reads comes from a negative binomial distribution with three different pool sizes and three different read rates per individual, and the error rate is 0.01.

It can be seen from Table 5.9 that regardless of the pool size and the read rate, the L-R test always gives better results than the threshold test, and the difference between the proportions of SNPs detected by the tests is sometimes very large.

At a low pool size and read rate per individual (pool size = 5, read rate per individual = 5, read rate = 25), the L-R test detects 31.2% of SNPs whereas the threshold test only detects 10.4% of SNPs. Also, the FDR for the L-R test in this case is 2.500% whereas FDR for the threshold test is 0.0%. As the pool size increases (leaving the read rate per individual fixed), the power of the L-R test clearly increases, but the power of the threshold test does not change much.

If the pool size and the read rate per individual increase at the same time (pool size = 20, read rate per individual = 20, read rate = 400), the L-R test detects 96.8% of SNPs, i.e., all the minor alleles that actually appear in the sample, but the threshold test still does not perform very well (52.8% of SNPs).

Table 5.9 also shows that for the threshold test, the read rate per individual seems a more important factor than the pool size, since the proportion of SNPs detected by the two tests increases more rapidly by increasing the read rate per individual at a fixed pool size, than by increasing the pool size at a fixed read rate per individual.

However, increasing the pool size at a fixed read rate per individual is more useful when the L-R test is used.

To further investigate the effect of the read rate per individual and the pool size on the power of the test, regression analysis has been carried out. A quadratic surface model has been fitted with the power of the test as the dependent variable, and the read rate per individual and the pool size (and the interaction term) as explanatory variables.

The results of the regression analysis are presented in Table 5.11.

Table 5.11: Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has a Poisson Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate per Individual	5.965714	0.510675	11.682	0.001348**
(Read Rate per Individual) ²	-0.176000	0.018989	-9.268	0.002658**
Pool Size	6.845714	0.510675	13.405	0.000897***
(Pool Size) ²	-0.168889	0.018989	-8.894	0.002998**
Read Rate per Individual \times Pool Size	0.002939	0.011302	0.260	0.811670
Adjusted $R^2 = 0.9968$				
Threshold Test				
Read Rate per Individual	3.48000	0.71181	4.889	0.0164*
(Read Rate per Individual) ²	-0.04000	0.02647	-1.511	0.2279
Pool Size	0.32000	0.71181	0.450	0.6835
(Pool Size) ²	-0.02489	0.02647	-0.940	0.4164
Read Rate per Individual \times Pool Size	0.02971	0.01575	1.886	0.1557
Adjusted $R^2 = 0.9958$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a Poisson distribution, the read rate per individual, the $(\text{read rate per individual})^2$, the pool size and the $(\text{pool size})^2$ have a significant contribution in determining the power of the L-R test, whereas only the read rate per individual contributes to the power of the threshold test. The results show that increasing the read rate per individual or the pool size, will result in increasing the power of the L-R test and increasing the read rate per individual will result in increasing the power of the threshold test.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.11 are presented in Figure 5.5.

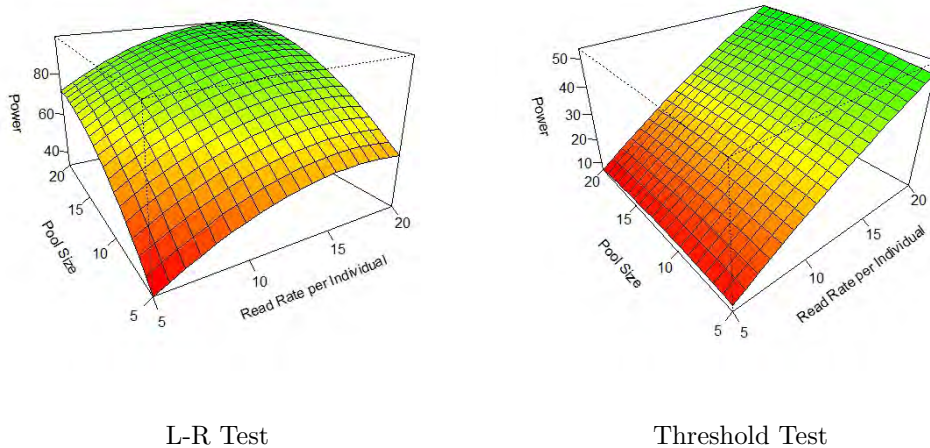


Figure 5.5: 3D Figure of Regression Analysis

Figure 5.5 clearly shows how the power of the tests is affected by the read rate per individual and the pool size. It can be seen that by increasing the read rate per individual, the power of both tests increases. However, increasing the pool size does not have an impact on the power of the threshold test, but it does increase the power of the L-R test.

One important practical aspect which should be noted here is the choice of the pool size when the read rate per lane is fixed, i.e., if

the read rate is 100 a position on the diagonal from bottom left to top right can be chosen.

Table 5.9 shows that using the L-R test, at a fixed read rate per lane (read rate per lane = 100), it is optimal to have a large pool size (under the assumption that there are at least several reads per individual). At a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), only 140 (56.0%) of SNPs are detected. By increasing the pool size and decreasing the read rate per individual to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), 181 (72.4%) of SNPs are detected. This suggests that a large pool might increase the chance of capturing SNPs when using the L-R test.

However, using the threshold test at a fixed read rate per lane gives different results. In this case, at a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), 119 (47.6%) of SNPs are detected. If the pool size increases and the read rate per individual decreases in order to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), only 20 (8.0%) of SNPs are detected. These results suggest that for a given read rate per lane the optimal pool size is smaller when using the threshold rule rather than the L-R test.

Table 5.10 shows the results when it is assumed that the number of reads come from a negative binomial distribution. The power is slightly reduced compared to the case where the number of reads comes from the Poisson distribution. In all cases, regardless of the pool size and read rate per individual, the L-R test performs better than the threshold test.

At a low pool size and read rate per individual (pool size = 5, read rate per individual = 5, read rate = 25), the L-R test detects 29.6% of SNPs whereas the threshold test only detects 10.4% of SNPs. Although in this case, the FDR for the L-R test is 2.632% whereas the FDR for the threshold test is 0.0%. As the pool size increases, the power of the L-R test increases, but the power of the

threshold test remains fairly constant, and there is thus a large difference between the proportions of SNPs detected by the tests.

If the pool size and the read rate per individual increase at the same time (pool size = 20, read rate per individual = 20, read rate = 400), the L-R test detects 95.2% of SNPs, but the threshold test still does not perform very well (47.6% of SNPs).

Again Table 5.10 shows that for the threshold test, the read rate per individual seems a more important factor than the pool size, since the proportion of SNPs detected by the two tests increases more rapidly by increasing the read rate per individual at a fixed pool size, than by increasing the pool size at a fixed read rate per individual.

However, increasing the pool size at a fixed read rate per individual is more useful when the L-R test is used.

To further investigate the effect of the read rate per individual and the pool size on the power of the test, regression analysis has been carried out. A quadratic surface model has been fitted with the power of the test as the dependent variable, and the read rate per individual and the pool size (and the interaction term) as explanatory variables.

The results of the regression analysis are presented in Table 5.12.

Table 5.12: Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has an NB Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate per Individual	5.83714	0.63182	9.239	0.00268**
(Read Rate per Individual) ²	-0.16089	0.02349	-6.848	0.00637**
Pool Size	7.75714	0.63182	12.277	0.00116**
(Pool Size) ²	-0.18400	0.02349	-7.832	0.00433**
Read Rate per Individual \times Pool Size	-0.02147	0.01398	-1.535	0.22226
Adjusted $R^2 = 0.9943$				
Threshold Test				
Read Rate per Individual	3.70286	1.59082	2.328	0.102
(Read Rate per Individual) ²	-0.05689	0.05915	-0.962	0.407
Pool Size	2.46286	1.59082	1.548	0.219
(Pool Size) ²	-0.08978	0.05915	-1.518	0.226
Read Rate per Individual \times Pool Size	0.01404	0.03521	0.399	0.717
Adjusted $R^2 = 0.9377$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a negative binomial distribution, the read rate per individual, the $(\text{read rate per individual})^2$, the pool size and the $(\text{pool size})^2$ have a significant contribution in determining the power of the L-R test. By increasing either the read rate per individual or the pool size, the power of the test increases. However, the results show that none of these variables contribute significantly to the power of the threshold test.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.12 are presented in Figure 5.6.

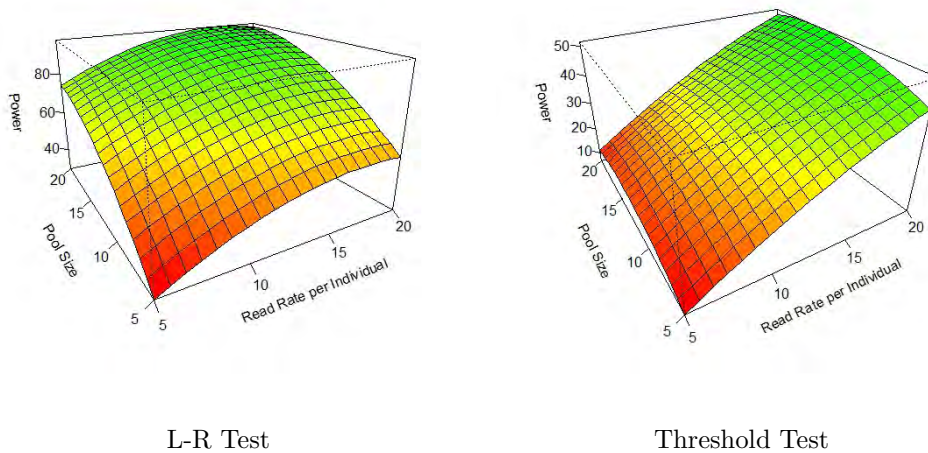


Figure 5.6: 3D Figure of Regression Analysis

Figure 5.6 clearly shows how power of the tests is affected by the read rate per individual and the pool size. It can be seen by increasing the read rate per individual or the pool size, the power of the L-R test increases. However, these two factors do not have any impact on the power of the threshold test.

Also here, the choice of the pool size when the read rate per lane is fixed, (i.e., if the read rate is 100 a position on the diagonal from bottom left to top right can be chosen) is important.

Table 5.10 shows that using the L-R test, at a fixed read rate

per lane (read rate per lane = 100), it is again optimal to have a large pool size (under the assumption that there are at least several reads per individual). At a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), only 136 (54.4%) of SNPs are detected. By increasing the pool size and decreasing the read rate per individual to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), 187 (74.8%) of SNPs are detected. This suggests that a large pool might increase the chance of capturing SNPs.

However, using the threshold test at a fixed read rate per lane gives different results. In this case, at a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), 102 (40.8%) of SNPs are detected. If the pool size increases and the read rate per individual decreases in order to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), only 26 (10.4%) of SNPs have been detected. These results suggest that for a given read rate per lane the optimal pool size is smaller when using the threshold rule rather than the L-R test.

Table 5.13: Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has a Poisson Distribution - Mean Empirical Error Rate = 0.0008)

		Pool Size		
		5	10	20
Read Rate Per Individual		L-R test		
5	Real SNPs Detected	83 (33.2%)	126 (50.4%)	212 (84.8%)
	FDR	1 (1.190%)	1 (0.787%)	1 (0.469%)
10	Real SNPs Detected	106 (42.4%)	185 (74.0%)	231 (92.4%)
	FDR	0 (0.0%)	2 (1.700%)	1 (0.431%)
20	Real SNPs Detected	150 (60.0%)	196 (78.4%)	240 (96.0%)
	FDR	0 (0.0%)	0 (0.0%)	1 (0.415%)
Read Rate Per Individual		Threshold Test		
5	Real SNPs Detected	80 (32.0%)	123 (49.2%)	171 (68.4%)
	FDR	1 (1.235%)	2 (1.600%)	1 (0.581%)
10	Real SNPs Detected	105 (42.0%)	177 (70.8%)	229 (91.6%)
	FDR	0 (0.0%)	1 (0.562%)	7 (2.966%)
20	Real SNPs Detected	150 (60.0%)	197 (78.8%)	240 (96.0%)
	FDR	0 (0.0%)	7 (3.431%)	6 (2.439%)

Table 5.14: Power and FDR of Tests for Untagged DNA Samples (Number of Reads Has an NB Distribution - Mean Empirical Error Rate = 0.0008)

		Pool Size		
		5	10	20
Read Rate Per Individual		L-R test		
5	Real SNPs Detected	88 (35.2%)	126 (50.4%)	193 (77.2%)
	FDR	2 (2.222%)	4 (3.077%)	0 (0.0%)
10	Real SNPs Detected	127 (50.8%)	190 (76.0%)	229 (91.6%)
	FDR	0 (0.0%)	0 (0.0%)	0 (0.0%)
20	Real SNPs Detected	136 (54.4%)	192 (76.8%)	238 (95.2%)
	FDR	0 (0.0%)	1 (0.518%)	0 (0.0%)
Read Rate Per Individual		Threshold Test		
5	Real SNPs Detected	80 (32.0%)	113 (45.2%)	161 (64.4%)
	FDR	1 (1.235%)	5 (4.237%)	0 (0.0%)
10	Real SNPs Detected	127 (50.8%)	181 (72.4%)	232 (92.8%)
	FDR	1 (0.781%)	1 (0.549%)	4 (1.695%)
20	Real SNPs Detected	121 (48.4%)	196 (78.4%)	242 (96.8%)
	FDR	1 (0.817%)	9 (4.390%)	5 (2.024%)

Table 5.13 shows the results of the simulations for untagged DNA samples when the number of reads comes from a Poisson distribution with three different pool sizes and three different read rates per individual, and the error rate is set to the mean empirical error rate, 0.0008.

Table 5.14 shows the results of the simulations for untagged DNA samples when the number of reads comes from a negative binomial distribution with three different pool sizes and three different read rates per individual, and the error rate is set to mean empirical error rate, 0.0008.

It can be seen from Table 5.13 that when the error rate is decreased, both the L-R test and the threshold test perform at almost the same level, and the difference between the proportions of SNPs detected by the tests is relatively small.

At a low pool size and read rate per individual (pool size = 5, read rate per individual = 5, read rate = 25), the L-R test detects 33.2% of SNPs whereas the threshold test detects 32.0% of SNPs. As the pool size increases, the power of the L-R test and the threshold test increases, with the L-R test being slightly more powerful. If the pool size and the read rate per individual increase at the same time (pool size = 20, read rate per individual = 20, read rate = 400), the L-R test and the threshold test detect the same proportion of SNPs (96.0%).

However, by increasing the pool size and the read rate per individual, the FDR for the threshold test becomes larger than the FDR for the L-R test.

To further investigate the effect of the read rate per individual and the pool size on the power of the test, regression analysis has been carried out. A quadratic surface model has been fitted with the power of the test as the dependent variable, and the read rate per individual and the pool size (and the interaction term) as explanatory variables.

The results of the regression analysis are presented in Table 5.15.

Table 5.15: Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has a Poisson Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate per Individual	5.42286	1.90279	2.850	0.0651
(Read Rate per Individual) ²	-0.12267	0.07076	-1.734	0.1814
Pool Size	7.50286	1.90279	3.943	0.0291*
(Pool Size) ²	-0.14222	0.07076	-2.010	0.1380
Read Rate per Individual × Pool Size	-0.07624	0.04211	-1.811	0.1679
Adjusted $R^2 = 0.9523$				
Threshold Test				
Read Rate per Individual	6.46571	1.80932	3.574	0.0375*
(Read Rate per Individual) ²	-0.17600	0.06728	-2.616	0.0793
Pool Size	6.90571	1.80932	3.817	0.0316*
(Pool Size) ²	-0.16089	0.06728	-2.391	0.0966
Read Rate per Individual × Pool Size	-0.01478	0.04004	-0.369	0.7366
Adjusted $R^2 = 0.954$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a Poisson distribution, only the pool size has a significant contribution in determining the power of the L-R test. By increasing the pool size the power of the L-R test increases. The results also show that both pool size and read rate per individual contribute to the power of the threshold test. By increasing either of them the power of the test increases.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.15 are presented in Figure 5.7.

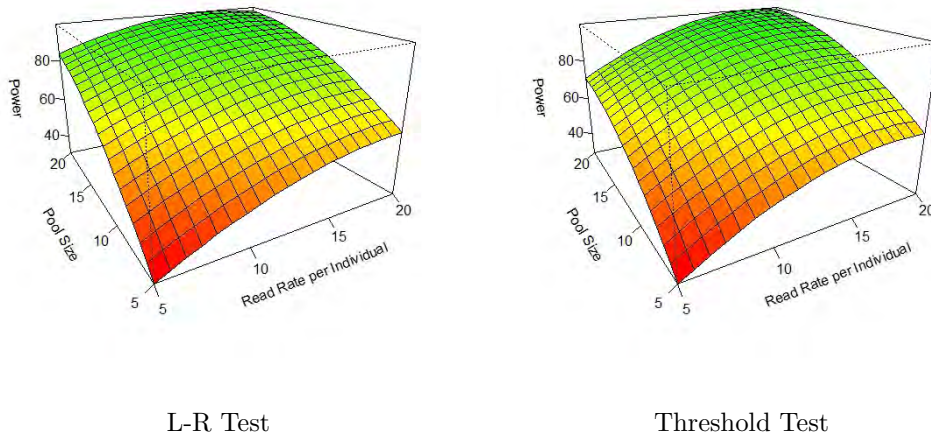


Figure 5.7: 3D Figure of Regression Analysis

Figure 5.7 clearly shows how the power of the tests is affected by the read rate per individual and the pool size. Increasing the read rate per individual or the pool size increases the power of the threshold test. However, although increasing the pool size will result in increasing the power of the L-R test, increasing the read rate per individual does not have any impact on the power of this test.

One important practical aspect which should be noted here is the choice of the pool size when the read rate per lane is fixed, i.e., if the read rate is 100 a position on the diagonal from bottom left to top right can be chosen.

Table 5.13 shows that using the L-R test, at a fixed read rate per lane (read rate per lane = 100), it is optimal to have a large pool size (under the assumption that there are at least several reads per individual). At a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), only 150 (60.0%) of SNPs are detected. By increasing the pool size and decreasing the read rate per individual to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), 212 (84.8%) of SNPs are detected. This suggests that a large pool might increase the chance of capturing SNPs.

If the threshold test is used at a fixed read rate per lane, it seems that the pool size does not affect the power of the test. In this case, at a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), 150 (60.0%) of SNPs are detected. If the pool size increases and the read rate per individual decreases in order to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), 171 (68.4%) of SNPs are detected. These results suggest that using the threshold test and when the error rate is low, increasing the pool size and decreasing the read rate per individual (or vice versa) have approximately the same effect on the power of the test, which implies that the pool size does not affect the power much. However, the maximum power is gained when the pool size is 10.

Table 5.14 shows similar results to Table 5.13.

It can be seen that when the error rate is decreased, both the L-R test and the threshold test perform at almost the same level, and the difference between the proportions of SNPs detected by the tests is relatively small.

At a low pool size and read rate per individual (pool size = 5, read rate per individual = 5, read rate = 25), the L-R test detects 35.2% of SNPs whereas the threshold test detects 32.0% of SNPs. As the pool size increases, both the L-R test and the threshold test perform better. When the read rate per individual is reasonably high (≥ 10), then there is little difference between the power of the two methods.

If the pool size and the read rate per individual increase at the same time (pool size = 20, read rate per individual = 20, read rate = 400), the L-R test detects 95.2% of SNPs, and the threshold test detects 96.8% of SNPs.

In this case, the FDR for the threshold test is 2.024% whereas the FDR for the L-R test is 0.0%.

To further investigate the effect of the read rate per individual and the pool size on the power of the test, regression analysis has been carried out. A quadratic surface model has been fitted with the power of the test as the dependent variable, and the read rate per individual and the pool size (and the interaction term) as explanatory variables.

The results of the regression analysis are presented in Table 5.16.

Table 5.16: Regression Analysis to Assess the Effect of Read Rate per Individual and Pool Size on the Power (Number of Reads Has an NB Distribution)

Coefficients	Estimate	Standard Error	t value	p-value
L-R Test				
Read Rate per Individual	7.26571	1.40168	5.184	0.0139*
(Read Rate per Individual) ²	-0.22933	0.05212	-4.400	0.0218*
Pool Size	6.46571	1.40168	4.613	0.0192*
(Pool Size) ²	-0.14400	0.05212	-2.763	0.0700
Read Rate per Individual \times Pool Size	-0.01020	0.03102	-0.329	0.7638
Adjusted $R^2 = 0.9691$				
Threshold Test				
Read Rate per Individual	9.00857	1.54341	5.837	0.0100*
(Read Rate per Individual) ²	-0.31378	0.05739	-5.467	0.0120*
Pool Size	6.04857	1.54341	3.919	0.0295*
(Pool Size) ²	-0.15911	0.05739	-2.772	0.0694
Read Rate per Individual \times Pool Size	0.05641	0.03416	1.651	0.1972
Adjusted $R^2 = 0.968$				

By looking at p -values in the regression analysis, it is clear that assuming the number of reads follows a negative binomial distribution, the read rate per individual, the (read rate per individual)² and the error rate have a significant contribution in determining the power of the L-R and threshold tests. By increasing the read rate per individual or the pool size, the power of both tests increases.

The 3D plots of the surface model fitted to the data using the regression model presented in Table 5.16 are presented in Figure 5.8.

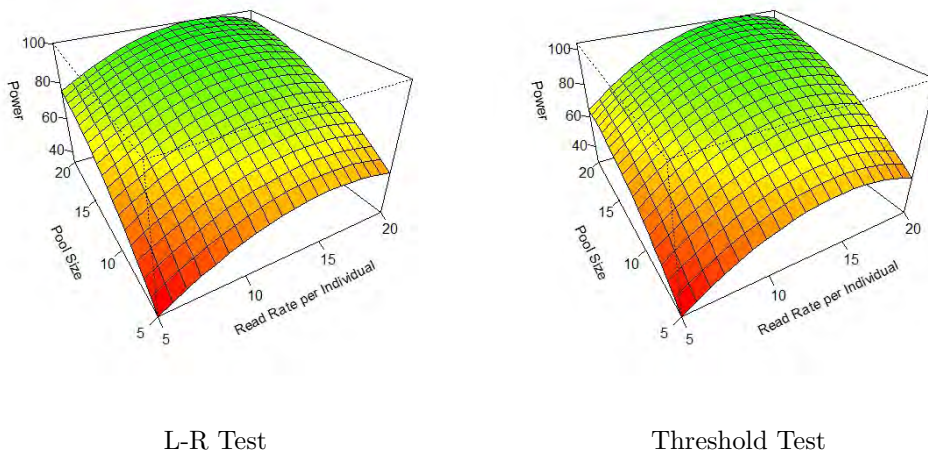


Figure 5.8: 3D Figure of Regression Analysis

Figure 5.8 clearly shows how the power of the tests is affected by the read rate per individual and the pool size. It can be seen that by increasing either of them, the power of the L-R test and threshold test increases.

Table 5.14 shows that using the L-R test, at a fixed read rate per lane (read rate per lane = 100), it is optimal to have a large pool size (under the assumption that there are at least several reads per individual). At a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), only 136 (54.4%) of SNPs are detected. By increasing the pool size and decreasing the read rate per individual to maintain the fixed read rate per lane (pool size = 20, read rate per individual

= 5, read rate = 100), 193 (77.2%) of SNPs are detected. This suggests that a large pool might increase the chance of capturing SNPs.

Using the threshold test at a fixed read rate per lane gives similar results. In this case, at a low pool size and a high read rate per individual (pool size = 5, read rate per individual = 20, read rate = 100), 121 (48.4%) of SNPs are detected. If the pool size increases and the read rate per individual decreases in order to maintain the fixed read rate per lane (pool size = 20, read rate per individual = 5, read rate = 100), 161 (64.4%) of SNPs are detected. Again here, the maximum power is gained when the pool size is 10.

It should be noted that the results from Tables 5.9, 5.10, 5.13 and 5.14 agree with the findings of Ramsey and Futschik (2012). They found out that the maximum power is associated with intermediate pool sizes when using the threshold test, and with higher pool sizes when using the L-R test.

Tables 5.17 and 5.18 show the results of simulations for untagged DNA samples when in both cases the mean read rate comes from the empirical distribution and the error rate firstly is the mean empirical error rate (0.0008) and secondly comes from the empirical distribution.

Table 5.17: Power and FDR of Tests for Untagged DNA Sample (Empirical Read Rate Used - Mean Empirical Error Rate = 0.0008)

Distribution		L-R test	Threshold Test
Poisson	Real SNPs Detected	234 (94.4%)	233 (93.2%)
	FDR	0 (0.0%)	2 (0.851%)
NB	Real SNPs Detected	229 (91.6%)	226 (90.4%)
	FDR	1 (0.435%)	6 (2.586%)

Table 5.18: Power and FDR of Tests for Untagged DNA Samples (Empirical Read Rate and Empirical Error Rate Used)

Distribution		L-R test	Threshold Test
Poisson	Real SNPs Detected	236 (94.4%)	235 (94.0%)
	FDR	0 (0.0%)	1 (0.424%)
NB	Real SNPs Detected	233 (93.2%)	232 (92.8%)
	FDR	0 (0.0%)	1 (0.429%)

Table 5.17 shows that assuming the number of reads comes from a Poisson distribution, the L-R test and the threshold test detect approximately the same proportion of SNPs (94.4% v.s. 93.2%).

Assuming the number of reads follows a negative binomial distribution, the proportions of SNPs detected by the two tests are still close to each other (91.6% v.s. 90.4%), but on the other hand the FDR for the threshold test is 2.586% while for the L-R test it is 0.435%.

Table 5.18 shows that assuming the number of reads comes from a Poisson distribution, the L-R test and the threshold test detect approximately the same proportion of SNPs (94.4% v.s. 94.0%).

Assuming the number of reads follows a negative binomial distribution, the proportions of SNPs detected by the L-R test and the threshold test still remain close (93.2% v.s. 92.8%).

5.4 Estimation of the power for the threshold test when the read rate and error rate are fixed: calculations

In this section, the same assumptions as in Section 4.3 are used for the calculation of the power of the threshold test when the read rate and error rate are fixed. Initially consider two cases where first $n_s = 1000$ and $k = 2$, and then $n_s = 2500$ and $k = 2$. First, consider the case where $n_s = 1000$ and $k = 2$. First $P(L = 2)$ is calculated, which is the probability that both SNPs are detected. Note that when the number of reads from an individual comes from a Poisson distribution with parameter λ , then $X \sim Poisson(\frac{\lambda}{2})$.

Using the B-H procedure, if both SNPs are to be detected and the significance level of the test is initially set to be 0.05 (i.e., $\alpha = 0.05$), then the new threshold should be $\frac{0.05 \times 2}{1000} = 0.0001$. This means that the largest p -value associated with either of the SNPs should be less than 0.0001 for both SNPs to be detected,

$$P(pval_{(2)} \leq 0.0001) = P(X_{(2)} > q_{1-0.0001}),$$

where q_p is the p -th quantile from the Poisson distribution with parameter $\lambda\hat{p}$. If $X_{(2)} > q_{1-p}$ then the maximum numbers of reads of minor allele from these two sites (SNPs) must be $> q_{1-p}$.

Since the number of reads from the lanes are independent of each other,

$$P(L = 2) = P(X > q_{1-0.0001})^2,$$

where X is the number of reads of the minor allele from a lane.

Similarly, $P(L = 1)$ can be calculated. First, a new threshold based on the B-H procedure should be calculated as only one of the SNPs is to be detected. In this case, the new threshold is $\frac{0.05 \times 1}{1000} = 0.00005$. The SNP which is detected must be associated with a p -value below 0.00005 and the SNP which is not detected with a p -value greater than 0.0001. Since the SNP which is not detected could be either of the two SNPs, there are two possibilities. Thus,

$$P(L = 1) = 2P(X > q_{1-0.00005})P(X \leq q_{1-0.0001}).$$

Using an analogous procedure, the above probabilities can be calculated for $n_s = 2500$, by calculating the new thresholds using the B-H procedure.

These calculations have been carried out using three different read rates and four different error rates to investigate the effect of these factors on the power of this test (Table 5.19).

Table 5.19: Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate

Error Rate	Read Rate		
	5	10	20
$n_s = 1000 \text{ \& } k = 2$			
0.01	0.4561869	0.7349741	0.9894684
0.005	0.4561869	0.8753480	0.9896639
0.001	0.7127025	0.9595723	0.9972306
0.0008	0.7127025	0.9595723	0.9972306
$n_s = 2500 \text{ \& } k = 2$			
0.01	0.3399398	0.7349741	0.9707473
0.005	0.4561869	0.8578501	0.9896639
0.001	0.7127025	0.8753480	0.9972306
0.0008	0.7127025	0.9561673	0.9972306

It can be seen from Table 5.19 that the power of the test is highly affected by the read rate. It can be clearly seen that the power is increasing in the read rate, regardless of the error rate.

It can be also seen that the power is affected by the error rate: as the error rate decreases, the power of the test increases.

Another result from Table 5.19 is that if the number of SNPs is fixed, increasing the number of sites does not have a large effect on the power of the test. The only difference in the powers is when the error rate is high (0.01). This is a desirable property when the proportion of sites which are SNPs is unknown.

Now consider the case where the number of sites has been decreased and the number of SNPs has been increased, say $n_s = 250$ and $k = 5$. Calculations in this case are more complicated. First $P(L = 5)$, the probability that all SNPs are detected will be calculated.

Using the B-H procedure, if all five SNPs are to be detected and the significance level of the test is initially set to be 0.05 (i.e., $\alpha = 0.05$), then the new threshold should be $\frac{0.05 \times 5}{250} = 0.001$. This means that the largest p -value associated with any of the SNPs should be less than 0.001 for all five SNPs to be detected,

$$P(pval_{(5)} \leq 0.001) = P(X_{(5)} > q_{1-0.001}).$$

$$\text{So, } P(L = 5) = (P(X_{(5)} > q_{1-0.001}))^5.$$

Similarly, $P(L = 4)$ can be calculated. Again, first a new threshold based on the B-H procedure should be calculated as four out of five SNPs are to be detected. In this case, the new threshold is $\frac{0.05 \times 4}{250} = 0.0008$. The four SNPs which are detected must be associated with the p -values below 0.0008 and the SNP which is not detected with a p -value greater than 0.001. Since the SNP which is not detected could be any of the five SNPs, there are five possibilities. Hence,

$$P(L = 4) = 5P(X > q_{1-0.0008})^4 P(X \leq q_{1-0.001}).$$

To calculate $P(L = 3)$,

$$pval_{(3)} < 0.0006,$$

$$pval_{(4)} > 0.0008,$$

and,

$$pval_{(5)} > 0.001.$$

Note that 0.0006 is calculated using the B-H method, i.e., $\frac{0.05 \times 3}{250} = 0.0006$.

Consider the two following cases,

1. Three p -values below 0.0008 and two above 0.001,
2. Three p -values below 0.0006, one between 0.0008 and 0.001 and one above 0.001.

Now the probabilities of these cases can be calculated as follows,

$$1. \binom{5}{2} P(X > q_{1-0.0006})^3 P(X \leq q_{1-0.001})^2.$$

$$2. 2 \binom{5}{2} P(X > q_{1-0.0006})^3 P(X \leq q_{1-0.001}) P(q_{1-0.0008} < X \leq q_{1-0.001}).$$

Therefore,

$$P(L = 3) = \binom{5}{2} P(X > q_{1-0.0006})^3 P(X \leq q_{1-0.001}) \times \\ \left(P(X \leq q_{1-0.001}) + 2P(q_{1-0.0008} < X \leq q_{1-0.001}) \right).$$

To calculate $P(L = 2)$, a new threshold now should be calculated. So, $\frac{0.05 \times 2}{250} = 0.0004$. If only two out of five SNPs are to be detected, then,

$$pval_{(2)} < 0.0004,$$

$$pval_{(3)} > 0.0006,$$

$$pval_{(4)} > 0.0008,$$

and,

$$pval_{(5)} > 0.001.$$

Consider the four following cases,

1. Two p -values below 0.0004 and three above 0.001,
2. Two p -values below 0.0004, two above 0.001 and one between 0.0006 and 0.001,
3. Two p -values below 0.0004, one above 0.001 and two between 0.0008 and 0.001,
4. Two p -values below 0.0004, one above 0.001, one between 0.0008 and 0.001 and one between 0.0006 and 0.0008.

Now the probabilities of each case can be calculated as follows,

$$1. \binom{5}{2} P(X > q_{1-0.0004})^2 P(X \leq q_{1-0.001})^3.$$

$$2. 3 \binom{5}{2} P(X > q_{1-0.0004})^2 P(X \leq q_{1-0.001})^2 P(q_{1-0.0006} < X \leq q_{1-0.001}).$$

$$3. 3 \binom{5}{2} P(X > q_{1-0.0004})^2 P(X \leq q_{1-0.001}) P(q_{1-0.0008} < X \leq q_{1-0.001})^2.$$

$$4. 3 \times 2 \binom{5}{2} P(X > q_{1-0.0004})^2 P(X \leq q_{1-0.001}) P(q_{1-0.0008} < X \leq q_{1-0.001}) P(q_{1-0.0006} < X \leq q_{1-0.0008}).$$

Hence,

$$\begin{aligned}
P(L = 2) &= \binom{5}{2} P(X > q_{1-0.0004})^2 P(X \leq q_{1-0.001}) \times \\
&\left(P(X \leq q_{1-0.001})^2 + 3P(X \leq q_{1-0.001})P(q_{1-0.0006} < X \leq q_{1-0.001}) + \right. \\
&\left. 3P(q_{1-0.0006} < X \leq q_{1-0.001})^2 + 6P(q_{1-0.0008} < X \leq q_{1-0.001})P(q_{1-0.0006} < X \leq q_{1-0.0008}) \right).
\end{aligned}$$

Finally, to calculate $P(L = 1)$, the new threshold is $\frac{0.05 \times 1}{250} = 0.0002$. If only 1 out of 5 SNPs is detected, then,

$$pval_{(1)} < 0.0002,$$

$$pval_{(2)} > 0.0004,$$

$$pval_{(3)} > 0.0006,$$

$$pval_{(4)} > 0.0008,$$

and,

$$pval_{(5)} > 0.001.$$

Consider the following eight cases,

1. One p -value below 0.0002 and four above 0.001,
2. One p -value below 0.0002, three above 0.001 and one between 0.0004 and 0.001,
3. One p -value below 0.0002, two above 0.001, one between 0.0004 and 0.0006 and one between 0.0006 and 0.001,
4. One p -value below 0.0002, two above 0.001 and two between 0.0006 and 0.001.
5. One p -value below 0.0002, one above 0.001, one between 0.0004 and 0.0006, one between 0.0006 and 0.0008 and one between 0.0008

and 0.001,

6. One p -value below 0.0002, one above 0.001, one between 0.0004 and 0.0006 and two between 0.0008 and 0.001,

7. One p -value below 0.0002, one above 0.001, two between 0.0006 and 0.0008 and one between 0.0008 and 0.001,

8. One p -value below 0.0002, one above 0.001, one between 0.0006 and 0.0008 and two between 0.0008 and 0.001 and 1.

Now the probabilities of each case can be calculated as follows,

$$1. 5P(X > q_{1-0.0002})P(X \leq q_{1-0.001})^4.$$

$$2. 5 \times 4P(X > q_{1-0.0002})P(X \leq q_{1-0.001})^3P(q_{1-0.0004} < X \leq q_{1-0.001}).$$

$$3. 5 \times 4 \times 3P(X > q_{1-0.0002})P(X \leq q_{1-0.001})^2P(q_{1-0.0004} < X \leq q_{1-0.0006})P(q_{1-0.0006} < X \leq q_{1-0.001}).$$

$$4. 5 \binom{4}{2} P(X > q_{1-0.0002})P(X \leq q_{1-0.001})^2P(q_{1-0.0006} < X \leq q_{1-0.001})^2$$

$$5. 5 \times 4 \times 3 \times 2P(X > q_{1-0.0002})P(X \leq q_{1-0.001})P(q_{1-0.0004} < X \leq q_{1-0.0006})P(q_{1-0.0006} < X \leq q_{1-0.0008})P(q_{1-0.0006} < X \leq q_{1-0.001}).$$

$$6. 5 \times 4 \times 3P(X > q_{1-0.0002})P(X \leq q_{1-0.001})P(q_{1-0.0008} < X \leq q_{1-0.001})^2P(q_{1-0.0004} < X \leq q_{1-0.0006}).$$

$$7. 5 \times 4 \times 3P(X > q_{1-0.0002})P(X \leq q_{1-0.001})P(q_{1-0.0006} < X \leq q_{1-0.0008})^2P(q_{1-0.0008} < X \leq q_{1-0.001}).$$

$$8. 5 \times 4 \times 3P(X > q_{1-0.0002})P(X \leq q_{1-0.001})P(q_{1-0.0006} < X \leq q_{1-0.0008})P(q_{1-0.0008} < X \leq q_{1-0.001})^2.$$

Now, $P(L = 1)$ can be calculated by summing the eight probabilities given above.

These calculations have been carried out using the same three

different read rates and four different error rates to investigate the effect of these factors on the power of this test (Table 5.20).

Table 5.20: Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate

Error Rate	Read Rate		
	5	10	20
0.01	0.6643702	0.9842505	0.9998932
0.005	0.8389083	0.9842505	0.9999923
0.001	0.9126046	0.9983630	0.9999998
0.0008	0.9126046	0.9983630	0.9999998

These powers are similar to those obtained in the simulations.

It can be seen from Table 5.20 that the power is increasing in the number of reads and decreasing in the error rate. It seems that, however, the read rate is a more important factor than the error rate in determining the effectiveness of the test.

To further investigate the effect of the density of SNPs, consider the following two cases,

1. $n_s = 1000$ and $k = 5$.
2. $n_s = 2000$ and $k = 5$.

Note that the density of SNPs in the above cases is less than when $n_s = 250$. Calculations of the power of the test for the above cases can be done by finding the appropriate thresholds using the B-H procedure. These calculations have been carried out using three different read rates and four different error rates. Table 5.21 shows the results for the above cases ($n_s = 1000$ & $n_s = 2000$) along with the previous results ($n_s = 250$). Again, the density of SNPs has very little influence on the power of the test, which is a desirable property.

Table 5.21: Estimated Power of the Threshold Test with Fixed Read Rate and Error Rate

Error Rate	Read Rate		
	5	10	20
$n_s = 250 \ \& \ k = 5$			
0.01	0.6643702	0.9842505	0.9998932
0.005	0.8389083	0.9842505	0.9999923
0.001	0.9126046	0.9983630	0.9999998
0.0008	0.9126046	0.9983630	0.9999998
$n_s = 1000 \ \& \ k = 5$			
0.01	0.6643702	0.9412681	0.9996976
0.005	0.6643702	0.9842505	0.9999715
0.001	0.9126046	0.9983630	0.9999986
0.0008	0.9126046	0.9983630	0.9999986
$n_s = 2000 \ \& \ k = 5$			
0.01	0.6643702	0.9261353	0.9996975
0.005	0.6643702	0.9842505	0.9998932
0.001	0.9126046	0.9949580	0.9999923
0.0008	0.9126046	0.9949580	0.9999923

5.5 A fixed significance level based on the assumed density of SNPs: calculations

Recall the assumptions of the threshold test from Section 4.3. Similarly, the power of the threshold test using a fixed significance level can be estimated. Consider the three cases where there are 5 SNPs and $n_s = 250$, $n_s = 1000$ and $n_s = 2000$, respectively. The actual SNP density in these cases is 0.02, 0.005 and 0.0025, respectively.

As described before, the appropriate fixed significance level can be obtained by multiplying the assumed SNP density, q , by the required FDR, which initially is assumed to be 0.05. Suppose one expects that a proportion 0.003 of sites are SNPs (i.e., $q = 0.003$). Therefore the fixed significance level should be $0.003 \times 0.05 = 0.00015$.

Using a method analogous to the one used in Section 5.3, the power of the threshold test to detect a single copy of the minor allele can be calculated.

Recall that it was assumed that only a single individual has one copy of the minor allele at each SNP. So the probability that the number of reads of the minor allele exceeds the appropriate threshold (i.e., the power of the test) can be obtained by,

$$P(X > q_{1-0.00015}) = P(X > q_{0.99985}),$$

where 0.00015 is the fixed significance level, X has a Poisson distribution with parameter $\frac{\lambda}{2}$ and q_p is the p -th quantile of a Poisson distribution with parameter $\lambda\epsilon$. Define β^* to be the power of the test.

Four different error rates and three different read rates have been considered (explained in Section 5.3).

Table 5.22 shows the power of the test to detect a single copy of a minor allele.

Table 5.22: Estimated Power of the Threshold Test Using a Fixed Significance Level

	Read Rate		
	5	10	20
Error Rate			
0.01	0.4561869	0.7349741	0.9896639
0.005	0.4561869	0.8753480	0.9896639
0.001	0.7127025	0.9595723	0.9972306
0.0008	0.7127025	0.9595723	0.9995006

By comparing Tables 5.20 and 5.22, it can be clearly seen that using a fixed significance level in the threshold test decreases the power of the test, especially when the read rate is low and the error rate is high. Here, again by increasing the read rate and decreasing the error rate the power of the test increases.

The following parameters are defined to approximate the FDR for this procedure in each of the cases,

1. m : the number of SNPs.
2. q : the assumed density of SNPs.
3. δ : the actual density of SNPs.
4. α : the FDR initially assumed.

Using the notation introduced in Table 1.1, V is the number of false detections, S is the number of correct detections and R is the total discoveries (obviously, $R = S + V$).

Also, for large m , the FDR can be approximated by $\frac{E(V)}{E(R)}$.

$E(V)$ is first calculated, which is the expected number of false positives (or false detections). Then, $E(S)$ is calculated, which is the expected number of true positives (or correct detections). Adding $E(V)$ to $E(S)$, $E(R)$ can be obtained, which is the expected number

of total discoveries.

$E(V)$ can be calculated by multiplying the number of non-SNPs in each case by the significance level. Hence,

$$E(V) = m(1 - \delta)\alpha q.$$

$E(S)$ can be calculated by multiplying the number of SNPs in each case by the power. Hence,

$$E(S) = m\delta\beta^*.$$

Note that $E(V)$ is an upper bound on the expected number of false rejections (since the test statistic is discrete, the probability of wrongly inferring that a non-SNP is a SNP will be $\leq q\alpha$). This bound is independent of the parameters of the sequencer. On the other hand, the power of the test, and hence $E(S)$, depend on the parameters of the sequencer.

By adding the expected number of false detections to the expected number of true detections, the expected number of total discoveries, $E(R)$, can be obtained.

Now, the FDR can be approximated by $\frac{E(V)}{E(R)}$,

$$\begin{aligned} FDR &\cong \frac{E(V)}{E(R)} = \frac{m(1 - \delta)\alpha q}{m(1 - \delta)\alpha q + m\delta\beta^*} \\ &= \frac{(1 - \delta)\alpha q}{(1 - \delta)\alpha q + \delta\beta^*}. \end{aligned}$$

This expression is independent of m (the only requirement is that m is large). Estimates of the FDR for various parameters of the sequencer are presented in Table 5.23.

Table 5.23: FDR of Various Parameters of the Sequencer

Error Rate	Read Rate		
	5	10	20
$\delta = 0.02$			
0.01	0.01585634	0.009901336	0.007372013
0.005	0.01585634	0.008326744	0.007372013
0.001	0.01020759	0.007601438	0.007316486
0.0008	0.01020759	0.007601438	0.007299991
$\delta = 0.005$			
0.01	0.06141509	0.03902858	0.02927866
0.005	0.06141509	0.03297621	0.02927866
0.001	0.04019918	0.03016912	0.02906296
0.0008	0.04019918	0.03016912	0.02899887
$\delta = 0.0025$			
0.01	0.11598008	0.07529968	0.05702640
0.005	0.11598008	0.06399714	0.05702640
0.001	0.07747046	0.05870972	0.05661820
0.0008	0.07747046	0.05870972	0.05649688

It can be seen from Table 5.23 that the FDR is higher for lower SNP density. Note that the assumed SNP density is 0.003, so when the actual SNP density is lower than expected (i.e., $\delta = 0.0025$) then the FDR can be higher than 0.05. Note however, this method uses an upper bound on $E(V)$, and hence tends to overestimate the FDR. Even when the SNP density is higher than expected (i.e., $\delta = 0.005$ and $\delta = 0.02$), the FDR might be higher than 0.05. This is the case where the power of a single test is low (i.e. the read rate is low and the error rate is high).

In conclusion, even though the parameters of genome sequencers are improving, this test is only recommended when one has a good prior estimate of the SNP density.

5.6 Overview of the results

The results of the simulations show that when DNA is not pooled or it is tagged, the L-R test is more efficient than the threshold test, especially when the read rate is small and the error rate is high. When the empirical read rate and error rate are used, the results from the tests are similar. This can be explained by looking at the histogram of the mean number of reads (Figure 5.1). As explained before, in many cases the average number of reads per individual is around 40-50, therefore either test will find a real minor allele with probability 1. On the other hand, in a lot of cases the mean number of reads is small. In this case neither method works well, but the likelihood ratio method works better.

When DNA is pooled and not tagged, the L-R test works considerably better than the threshold test, especially when the read rate and the pool size are both small and the error rate is high. The L-R test is considerably more efficient compared to the threshold test, even with a large read rate and pool size and high error rate. This difference is even larger when number of reads comes from the negative binomial distribution. When the empirical read rate and empirical error rate for the Dublin data are used, however, the powers of the two tests are similar. It should also be noted that because the mean read rate from real data is so high, larger pools would be more efficient.

Pooling is a very important factor in determining the power of the test. Simulations showed that when using the L-R test at a fixed read rate per lane, regardless of the distribution of the number of reads and the error rate, it is optimal to have a large pool size (under the assumption that there are at least several reads per individual). This suggests that when using the L-R test given a fixed number of lanes, large pool sizes increase the chance of capturing SNPs.

When the threshold test is used, it seems that the results are dependent on the error rate. When the error rate is high (0.01), regardless of the distribution of the number of reads, for a given read rate per lane the optimal pool size is smaller than when using the L-R test. On the other hand, when the error rate is low (0.0008), regardless of the distribution of the number of reads per lane, using the threshold test the pool size for which the power is maximized increases. Greater power, however, can be achieved using the likelihood ratio test with a similar (or slightly greater) pool size. These findings agree with findings of Ramsey and Futschik (2012). They found that the maximum power is associated with intermediate pool sizes when using the threshold test, and with higher pool sizes when using the L-R test.

It should also be noted that using a fixed significance level in the threshold test decreases the power of the test, especially when the error rate is high and the read rate is low. In addition, it was shown that when the assumed density of SNPs is higher than the actual density, the FDR increases. This suggests that the fixed significance level approach should be only used when a good prior estimate of the SNP density is provided.

Another point which should be noted here is the effect of overdispersion. Changing the distribution of the number of reads from Poisson to negative binomial (which takes overdispersion into account) affects the threshold test more, because that test explicitly assumes that the number of reads from a lane comes from a Poisson distribution. When the variance of the number of reads increases, the probability that the number of reads from an individual with the minor allele exceeds the critical value falls (since the critical value

for the test must be smaller than the expected number of reads from an individual). This results in a fall in power (see Ramsey & Futschik (2012)). Similarly, the FDR increases as the threshold must be greater than the expected number of errors. The L-R test is less affected by the distribution of the number of reads, since the likelihood ratio statistic is calculated conditional on the number of reads from each lane. Thus there is no explicit assumption regarding the distribution of the number of reads from each lane. When the variance of the number of reads increases, the probability of a small number of reads from a lane containing an individual with the minor allele increases, which will decrease the power of the test. Since this test does not employ a fixed threshold rule, however, it is more flexible with regard to variation in the number of reads from individual lanes.

Among the multiple comparison procedures introduced here, the B-H procedure is more reasonable to use than the Bonferroni or Storey-Tibshirani approach. The main problem with the Bonferroni correction is that it becomes very conservative as the number of comparisons increases. In GWAS, a large number of sites (in practice, millions) are examined and this makes the Bonferroni approach unsuitable for these studies. Although the power of a type I error is decreased using this approach, the power of the test will also be significantly decreased leading to a large number of false negatives.

The approach presented by Storey and Tibshirani, has two main problems. The first is that this approach estimates the distribution of the p -value under the alternative hypothesis (i.e., the site is a SNP). This hypothesis is rarely true in practice, and hence the estimation of the distribution of the p -value under this assumption is not accurate. The second problem is that this approach assumes that under the null hypothesis, the distribution of the test statistic is uniform. This assumption is untrue, both for the threshold test, where the test statistic is a discrete variable, and for the L-R test.

The B-H procedure has the advantage of being very simple and effective in practice. In addition, it has been shown that this method has a low FDR and high power in comparison with the other MCPs.

All the programmes (R codes) used for the simulations can be found in Appendix B.

Chapter 6

Analysis of Real Genome Sequence Data

In this chapter, the results of the analysis of the real genome sequence data obtained from the Trinity genome sequencing laboratory in Dublin have been presented.

6.1 Sites inferred to be SNPs

The sample consisted of 160 individuals. 60,000 sites in total were examined using a genome sequencer. Therefore, the data are in the form of a matrix with 60,000 rows (representing the sites) and 320 columns (160 for the reads and 160 for the quality scores). This dataset was loaded into R as a matrix with 60,000 rows and 320 columns. There were 2 columns for each individual, the first containing the reads and the second the quality scores.

These data have been analysed by reading one site at a time into R. First, the major allele was identified. Then, based on the major allele and using the quality scores, the likelihoods under each of the three alternative hypotheses corresponding to the other three nucleotides being the minor allele were calculated. The likelihood ratio was then defined to be the maximum of these three likelihoods divided by the likelihood under the null hypothesis.

Based on these results out of 60,000 sites, 171 sites are inferred to be SNPs. This means that the proportion of detected SNPs in

this sample is 0.00285 or 0.285%. The list of the positions of the SNPs is below:

Table 6.1: Positions of the Sites Inferred to be SNPs

621	890	1172	1378	1453	1458	1509	1524
1794	1868	2170	2495	3566	3992	4089	4233
4917	5968	6046	6086	6603	6766	7428	8107
8227	8230	8919	9015	9964	10101	10138	10769
10810	10927	11543	11776	12422	13226	13404	13679
14032	14649	14693	14883	15185	15251	16438	16534
17192	17258	18621	18835	19271	19318	19459	20168
21077	21403	21705	23400	24367	24662	25065	25392
25993	26275	26522	27159	28024	28402	28499	29850
30112	30293	30341	30867	30910	31003	31015	31160
31221	31358	31621	32258	32298	32713	32812	33354
33454	33533	33647	34086	35069	35123	35969	35997
36312	36407	36548	36759	36969	37034	37102	37118
37323	37629	37727	37826	38244	38361	38463	38571
38719	39560	40046	40185	40507	40715	42392	42805
43162	43228	43709	43800	44059	44119	44677	45136
45289	45556	46088	46122	46204	47313	47814	47945
48107	48913	50908	51146	51267	51481	51563	52754
53112	53343	53901	54397	54447	54448	54449	54452
54777	54862	55442	55447	55827	56443	56543	56824
57062	57158	57228	57631	58244	58250	58275	58946
59552	59631	59754					

6.2 Genotyping of the SNPs

Genotyping is the process of determining the genetic make-up (genotype) of an individual by examining the individual's DNA sequence. It reveals the alleles an individual has inherited from their parents. SNP genotyping first involves identifying SNPs that are common DNA variants present across the genome. Genotyping of SNPs has become extremely important to researchers working to understand and treat disease.

As described before, at almost all SNPs there are two alleles present. Therefore, once a SNP is identified one can genotype the SNP using the two alleles present there.

This process has been done for the detected SNPs from the real data. Overall, 171 SNPs were detected out of 60,000 sites. At all of those SNPs, it was reasonable to assume from the data that there were two alleles present (the minor and the major allele).

6.2.1 How genotyping is done

Genotyping begins by calculating the maximum likelihood estimator of the minor allele frequency, $\hat{\gamma}$, at each SNP.

Consider a SNP. Suppose the major and the minor allele are denoted by M and m , respectively. There are 3 possible genotypes at the site, say, MM , Mm and mm . Using the *MLE* of the minor allele frequency, the probabilities of obtaining two copies of the major allele (MM), one copy of the major allele and one copy of the minor allele (Mm) and two copies of the minor allele (mm) are presented below:

$$P(MM) = (1 - \hat{\gamma})^2,$$

$$P(Mm) = 2\hat{\gamma}(1 - \hat{\gamma}),$$

$$P(mm) = \hat{\gamma}^2.$$

These are treated as the prior probabilities of the genotypes.

Using the above equations, Bayesian scores (posterior probabilities) can be calculated for each genotype as follows:

Firstly, the likelihood of the data for an individual given the genotype is calculated. Suppose r_i is the total number of reads for the i -th individual. If the genotype of an individual is Mm , then the probability of a read being M (or m) is approximately 0.5. Hence,

$$L_i(\mathbf{x}_i|Mm) \cong (0.5)^k \prod_{1 \leq j \leq r_i: x_{i,j} \notin \{M,m\}} \hat{p}_{i,j},$$

where k is the total number of reads of the two alleles inferred to be present and the product is over all the reads for an individual which do not indicate either the major or minor allele (i.e., inferred errors).

If there are no alleles but the major allele, any reads of anything other than the major allele are errors. Hence,

$$L_i(\mathbf{x}_i|MM) = \prod_{1 \leq j \leq r_i: x_{i,j}=M} (1 - \hat{p}_{i,j}) \prod_{1 \leq j \leq r_i: x_{i,j} \neq M} \hat{p}_{i,j}.$$

Similarly, if there are no alleles but the prospective minor allele, any reads of anything other than the minor allele are errors. Hence,

$$L_i(\mathbf{x}_i|mm) = \prod_{1 \leq j \leq r_i: x_{i,j}=m} (1 - \hat{p}_{i,j}) \prod_{1 \leq j \leq r_i: x_{i,j} \neq m} \hat{p}_{i,j}.$$

Now, the likelihood of the data for an individual, $L_i(\mathbf{x}_i)$, can be calculated using:

$$L_i(\mathbf{x}_i) = L_i(\mathbf{x}_i|MM)P(MM) + L_i(\mathbf{x}_i|Mm)P(Mm) + L_i(\mathbf{x}_i|mm)P(mm).$$

Using the likelihood function of the data, the Bayesian score for each of the genotypes at a SNP can be calculated. Let $P_i(G|\mathbf{x}_i)$ denote the estimate of the probability that individual i has genotype G given the reads for that individual. Then,

$$P_i(mm|\mathbf{x}_i) = \frac{L_i(\mathbf{x}_i|mm)P(mm)}{L_i(\mathbf{x}_i)}.$$

$$P_i(Mm|\mathbf{x}_i) = \frac{L_i(\mathbf{x}_i|Mm)P(Mm)}{L_i(\mathbf{x}_i)}.$$

$$P(MM|\mathbf{x}_i) = 1 - P(mm|\mathbf{x}_i) - P(Mm|\mathbf{x}_i).$$

The genotype of an individual is inferred to be the one that is most likely given the data for that individual.

See Appendix A.1 for a full list of the sites inferred to be SNPs, with the corresponding major allele, genotype, Bayesian score and minor allele frequency.

See Appendix A.2 for an example on how to genotype the data.

Chapter 7

Conclusion

7.1 Introduction

The aim of this research was to develop and apply a new statistical method for the detection of SNPs.

SNP detection technologies are used to scan for new polymorphisms or to determine the allele(s) of a known polymorphism in target sequences. The number of SNP genotyping methods has exploded in recent years and many robust methods are currently available. The demand for SNP genotyping is great, however, and no one method is able to meet the needs of all studies using SNPs. Despite the considerable gains over the last decade, new approaches must be developed to lower the cost and increase the speed of SNP detection.

Kwok, P. Y. and Chen, X. (2003) reviewed different technologies for global and targeted SNP discovery. They concluded that despite advances in the field, none of the assays are ideally suited for all applications.

Xu, J. Y., Xu, G. B. and Chen, S. L. (2009) presented a new method for SNP discovery in which the DNA fragments containing SNPs could be isolated efficiently from background DNA. Their findings showed that the method they presented was cost-effective and applicable to essentially any organism.

Bansal, V. (2010) presented a statistical method for the detection

of variants from next-generation resequencing of DNA pools that is able to identify both rare and common variants. The results showed that this method can detect 80-85% of SNPs using individual sequencing while achieving a low false discovery rate (3-5%).

Muralidharan, O. et al. (2011) introduced an empirical Bayes method that learns the error properties of sequencing data by pooling information across samples and positions. The method uses mixture models to extend Efron's empirical null ideas (Efron, B. (2004)) to sequencing data in a computationally and statistically efficient way. By borrowing information across samples and positions, they showed that the model is able to detect SNPs with fewer false discoveries than existing methods, without sacrificing power.

Overall, it is agreed that accurate detection of SNPs is important in determining the genetic factors behind certain diseases, as well as the estimation of mutation rates. In population genetics, a local excess of SNPs where the minor allele is rare is a typical characteristic of genomic regions that have undergone recent positive selection. Also, SNP databases such as dbSNP (<http://www.ncbi.nlm.nih.gov/snp>) are collecting submissions of newly identified SNPs for many different organisms. Since there should be confidence in the SNPs submitted, sufficient statistical evidence is desirable, suggesting that a SNP found is not merely a sequencing error.

7.2 Maximum likelihood test versus the threshold test

A new method based on a maximum likelihood test was presented. In general, two types of data were generated for the simulations: data from tagged (or unpooled) DNA and untagged (pooled) DNA. Simulations have been done to assess the power of this new test and the FDR and to compare the efficiency with the threshold test.

7.2.1 Tagged (or unpooled) DNA

The results of the simulations showed that assuming the read rate follows a Poisson distribution (or negative binomial distribution), the L-R test generally gives better results (detects more SNPs) than the threshold test, especially when the read rate is small and the error rate is high. As the read rate gets larger and the error rate gets smaller, however, the proportion of SNPs detected by the two tests gets closer to each other (eventually both tests detect any minor allele that appears in the sample). It can be clearly seen that when using the empirical read rate and error rate, the proportion of SNPs detected by the two tests are very close.

Another interesting result from the simulations is that assuming the read rate follows a negative binomial distribution (which takes into account the overdispersion of the data), the proportion of SNPs detected by the two tests is smaller compared with the corresponding cases from Poisson distribution. It can be seen that this fall in power is greater for the threshold test. As explained before, the threshold test assumes that the distribution of the number of reads from a lane is Poisson, therefore changing this distribution to negative binomial affects this test more than the L-R test, in which the likelihood ratio statistic is calculated conditional on the number of reads from each lane.

7.2.2 Untagged (pooled) DNA

Simulations showed that assuming the read rate follows a Poisson distribution (or negative binomial distribution), the L-R test is generally more efficient than the threshold test, especially when the error rate is high. It can be clearly seen that with a high error rate, even when the read rate and the pool size are large, the threshold test fails to detect the same proportion of SNPs as the L-R test. When the error rate is decreased, however, the two tests detect almost the same proportion of SNPs.

Another result from the simulations is that pooling is proven to be a very important factor in determining the power of the test. Simulations showed that when using the L-R test at a fixed read rate per lane, regardless of the distribution of the number of reads

and the error rate, it is optimal to have a large pool size (under the assumption that there are at least several reads per individual). This might suggest that when using the L-R test, large pool sizes increase the chance of capturing more SNPs.

On the other hand, simulations showed that when the threshold test is used, for a given read rate per lane the optimal pool size is smaller rather than the L-R test. Ramsey and Futschik (2012) also showed that the maximum power is associated with intermediate pool sizes when using the threshold test, and with larger pool sizes when using the L-R test.

7.2.3 Other findings

Another advantage of the L-R test approach compared with the threshold test is that one can obtain estimates of,

1. the posterior probability of a minor allele being present,
2. an estimate of the frequency of the minor allele (which is needed for genotyping).

7.3 Benjamini-Hochberg procedure: advantages & disadvantages

The B-H procedure is a simple and effective approach to SNP detection. It has been shown that this method is a simple step-wise procedure that controls the FDR and have substantially better power than the traditional family-wise error-rate controlling methods (Benjamini, Y. and Hochberg, Y. (2000)).

The Bonferroni method has several problems. The most serious one is that the probability of type I errors cannot decrease (the whole point of Bonferroni adjustments) without inflating the probability of type II errors (the probability of accepting the null hypothesis when the alternative is true). This means that by using the Bonferroni correction, when the type I error rate is decreased, the power of the test will be significantly decreased at the same time. On the other hand, when testing a large number of hypotheses (in

this case, considering a large number of loci), the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

The approach D. Storey and J. Tibshirani (2003) have proposed is not appropriate for SNP detection, as it estimates the distribution of the p -value under the alternative hypothesis. As explained before, only a small proportion of sites are SNPs, hence one expects that the alternative hypothesis is rarely true. In such a case, estimation of the distribution of the p -value under the alternative hypothesis will not be accurate.

Also, this procedure assumes that the distribution of the test statistic under the null hypothesis is uniform. This assumption is not true either for the threshold test (where the test statistic is discrete) or the L-R test.

The CARDIoGRAMplusC4D Consortium (2012, consisted of more than 160 authors), Schunkert H, et. al. (2011), Clarke R, et. al. (2009), Wang F, et. al. (2011) and Soranzo N, et. al. (2009) used a different approach to identify new risk loci for coronary artery disease. This method is based on this assumption that one in a million loci is a risk factor for cancer. Using this, they obtained a fixed significance level of 5×10^{-8} .

This approach has the advantage of being very simple in practice, and can be adapted to the expected frequency of SNPs. It was shown that using a fixed significance level in the threshold test decreases the power of the test (compared to the B-H procedure). Also, if the assumed density of SNPs is higher than the actual density, then FDR increases. Therefore, it was suggested that this approach should be only used when a good prior estimate of the density of SNPs is available.

Having said that, it should be noted that the B-H procedure also has some disadvantages. Under H_0 , the distribution of the p -value is not a Uniform distribution on $[0, 1]$, since the minor allele frequency under H_0 is at the boundary of the parameter space. This means that both the density of SNPs and the number of sites investigated can affect the power of the test. The other problem with this approach is that when trying to use the FDR approach in

practice, dependent test statistics are encountered more often than independent ones. A simulation study by Benjamini, Hochberg and Kling (1997) showed that the same procedure controls the FDR for equally positively correlated normally distributed (possibly Studentized) test statistics.

Benjamini and Yekutieli (2001) proved that the procedure controls the FDR in families with positively dependent test statistics. In other cases of dependency, they proved that the procedure can still be easily modified to control the FDR, although the resulting procedure is more conservative.

7.4 Possible problems and further investigation

By looking at Figure 3.1, it can be clearly seen that there are some sites where there is very little information (not enough reads for individuals). These sites with little information might cause problems in the analysis, as one cannot infer whether they are SNPs or not. These sites should be further investigated.

The effect of the sample size is another issue which should be further investigated. One of the obvious limitations of genome-wide studies is the high cost and significant effort required to genotype hundreds of thousands of SNPs per individual (Hirschhorn, J. N. and Daly, M. J. 2005). Because of this high cost, there is pressure to limit the sample size, with a consequent reduction in power. However, because variants that contribute to complex traits are likely to have modest effects, large sample sizes are crucial.

Lindquist, K., J., et al. (2013), addressed the crucial question of whether future GWAS can detect new SNP associations and explain additional heritability given the new availability of larger GWAS SNP arrays, imputation, and reduced genotyping costs. They showed that increasing the sample size has a much larger impact than increasing coverage on the potential of future GWAS to detect additional SNP-disease associations and heritability. Meta-analysis of genome-wide association data and large-scale replication could be also used to reduce the costs of large sample sizes.

Another problem which should be further investigated is the rel-

ative frequency of the minor allele (RFMA). In the simulations, it was assumed that the relative frequency of SNPs is 1%. This is the minimum RFMA in GWAS, where a site is considered to be a SNP if the RFMA is at least 1%. Hence, using this relative frequency a lower bound on the power of the test and an upper bound on the FDR can be obtained. In practice, the RFMA is usually greater than 1%, and this increases the number of total discoveries. So if the power of the test is high, this will result in a lower FDR.

It should be noted that overdispersion may also affect the sampling of the minor allele. When pooling, different amount of DNA material from different individuals is placed on a single lane of the genome sequencer. In such a case, the binomial distribution used for the modelling of the number of copies of the minor allele in the genotype of the individual may not be appropriate. Although assuming a binomially distributed number of reads from an individual is fairly reasonable, but it does not mean that it is always true. To take overdispersion into account, one can use the beta-binomial distribution in which the probability of success at each trial is not fixed but random and follows the beta distribution. It is frequently used in Bayesian statistics, empirical Bayes methods and classical statistics as an overdispersed binomial distribution.

It has been mentioned previously that sequencing error probabilities are unknown and the genome sequencer provides the estimates of these probabilities. To improve these estimates, recalibration methods/software such as The Genome Analysis Toolkit (GATK) are available.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, et al (2011) presented a single framework and the associated tools capable of discovering high-quality variation and genotyping individual samples using diverse sequencing machines and experimental designs. They concluded that the inaccuracy and covariation patterns differ strikingly between sequencing technologies, which, if uncorrected, can propagate into downstream analyses. Accurately recalibrated base quality scores eliminate these sequencer-specific biases and enable integration of data generated from multiple systems.

But even after recalibration, these estimates still contain considerable noise. This might affect both the L-R test and the threshold

test. Both tests will become more conservative if these estimates of the probability of error are higher than what they are in practice, and they will become more liberal if these estimates of the probability of error are lower than what they are in practice.

Appendices

Appendix A

Bayesian Score & Genotyping

A.1 SNP Genotyping and the Bayesian Scores

Table A.1: Details of the SNPs

SNP Position	Major Allele	Alleles Present	Bayesian Score	Lower Bound	Minor Allele Frequency
621	C	CT	1	1	0.034375
890	G	AG	1	1	0.109375
1172	C	CT	1	1	0.215625
1378	G	AG	0.99976004	0.99974399	0.009375
1453	C	CT	1	1	0.025000
1458	C	CT	1	1	0.253125
1509	G	AG	0.99960662	0.99958032	0.006250
1524	C	CT	0.96533286	0.96309956	0.012500
1794	C	CT	1	1	0.015625
1868	G	AG	1	1	0.334375
2170	G	GT	1	1	0.006250
2495	G	AG	0.99953206	0.99950078	0.006250
3566	G	AG	1	1	0.003125
3992	C	CT	0.9998614	0.99985213	0.025000
4089	G	AG	0.99020966	0.98956163	0.240625
4233	G	AG	1	1	0.171875
4917	G	AG	1	1	0.500000
5968	G	AG	1	1	0.006250
6046	A	AG	1	1	0.003125
6086	G	GT	1	1	0.006250
6603	G	GT	1	1	0.018750
6766	A	AG	1	1	0.003125
7428	A	AC	1	1	0.003125
8107	G	AG	0.94428143	0.94428143	0.003125
8227	C	CT	1	1	0.003125
8230	C	AC	1	1	0.006250
8919	A	AG	1	1	0.184375
9015	C	CT	1	1	0.006250
9964	C	CT	1	1	0.003125
1010	A	AG	1	1	0.003125
1013	T	CT	1	1	0.003125
1076	T	CT	1	1	0.003125
10810	T	CT	1	1	0.006250
10927	A	AG	1	1	0.025000
11543	T	CT	1	1	0.018750
11776	C	CT	1	1	0.028125
12422	C	AC	1	1	0.012500
13226	A	AT	1	1	0.006250
13404	G	AG	1	1	0.003125
13679	A	AC	1	1	0.184375
14032	A	AC	1	1	0.021875
14649	A	AG	1	1	0.065625
14693	C	CT	1	1	0.500000
14883	C	AC	1	1	0.500000
15185	G	AG	1	1	0.100000
15251	G	AG	1	1	0.012500
16438	G	AG	1	1	0.003125
16534	G	AG	1	1	0.003125
17192	A	AG	1	1	0.190625
17258	T	CT	1	1	0.021875
18621	T	CT	1	1	0.003125
18835	G	AG	1	1	0.003125
19271	A	AC	1	1	0.006250
19318	T	AT	1	1	0.003125
19459	A	AT	1	1	0.003125
20168	T	CT	1	1	0.003125
21077	G	AG	1	1	0.003125
21403	C	CT	1	1	0.006250
21705	A	AG	1	1	0.003125

23400	G	AG	1	1	0.275000
24367	G	CG	1	1	0.500000
24662	C	AC	1	1	0.021875
25065	C	CT	1	1	0.003125
25392	C	AC	1	1	0.009375
25993	T	GT	1	1	0.128125
26275	A	AC	1	1	0.096875
26522	T	GT	1	1	0.015625
27159	C	CT	1	1	0.293750
28024	C	CT	0.86573042	0.85802429	0.012500
28402	A	AC	1	1	0.003125
28499	C	AC	1	1	0.409375
29850	G	AG	1	1	0.006250
30112	G	AG	1	1	0.500000
30293	C	CT	1	1	0.356250
30341	C	CT	1	1	0.375000
30867	C	CT	1	1	0.003125
30910	C	CT	1	1	0.003125
31003	T	CT	1	1	0.500000
31015	G	AG	1	1	0.500000
31160	G	AG	0.94526449	0.94526449	0.003125
31221	C	CT	1	1	0.500000
31358	G	AG	0.93839607	0.93454513	0.006250
31621	C	CT	0.99999999	0.99999999	0.003125
32258	T	CT	0.99947347	0.99947347	0.003125
32298	C	CT	1	1	0.500000
32713	C	CT	1	1	0.500000
32812	T	CT	1	1	0.284375
33354	G	AG	1	1	0.103125
33454	C	AC	1	1	0.009375
33533	A	AC	1	1	0.140625
33647	A	AG	1	1	0.453125
34086	A	AT	1	1	0.003125
35069	T	GT	1	1	0.500000
35123	G	AG	1	1	0.006250
35969	C	CT	1	1	0.500000
35997	G	AG	1	1	0.500000
36312	G	AG	1	1	0.371875
36407	G	AG	1	1	0.234375
36548	G	CG	1	1	0.500000
36759	C	CT	1	1	0.003125
36969	G	AG	1	1	0.003125
37034	G	AG	1	1	0.500000
37102	G	GT	1	1	0.081250
37118	C	CT	1	1	0.106250
37323	C	CG	1	1	0.034375
37629	G	AG	1	1	0.003125
37727	C	CT	1	1	0.500000
37826	C	CT	1	1	0.003125
38244	G	CG	1	1	0.003125
38361	G	GT	1	1	0.003125
38463	C	CG	1	1	0.021875
38571	C	CT	1	1	0.003125
38719	G	CG	1	1	0.003125
39560	G	AG	1	1	0.437500
40046	G	AG	1	1	0.003125
40185	C	CT	1	1	0.012500
40507	G	AG	1	1	0.034375
40715	T	CT	1	1	0.003125
42392	T	GT	1	1	0.003125
42805	G	AG	1	1	0.009375
43162	C	CT	1	1	0.003125
43228	C	CT	1	1	0.006250
43709	A	AG	1	1	0.003125
43800	C	CT	0.9604849	0.95795288	0.028125
44059	G	AG	1	1	0.025000
44119	C	CT	1	1	0.003125
44677	G	AG	1	1	0.003125

45136	T	GT	1	1	0.015625
45289	G	AG	1	1	0.500000
45556	T	AT	1	1	0.006250
46088	G	CG	0.81522869	0.81522869	0.003125
46122	C	CT	1	1	0.200000
46204	T	CT	1	1	0.003125
47313	G	AG	1	1	0.021875
47814	C	CT	1	1	0.003125
47945	A	AG	1	1	0.040625
48107	C	CT	1	1	0.003125
48913	C	CT	1	1	0.003125
50908	C	CT	1	1	0.021875
51146	C	AC	1	1	0.003125
51267	G	AG	1	1	0.009375
51481	T	CT	1	1	0.500000
51563	G	AG	1	1	0.046875
52754	G	AG	1	1	0.175000
53112	G	AG	0.99999996	0.99999996	0.003125
53343	C	CT	1	1	0.021875
53901	C	CT	1	1	0.146875
54397	G	AG	1	1	0.003125
54447	G	AG	0.99999965	0.99999963	0.090625
54448	C	CT	0.91986453	0.91496014	0.078125
54449	A	AG	0.99997375	0.99997199	0.112500
54452	A	AG	0.88765732	0.88103671	0.012500
54777	G	AG	1	1	0.015625
54862	C	CT	1	1	0.018750
55442	G	AG	0.97005494	0.96811579	0.006250
55447	A	AG	0.86032422	0.85236073	0.006250
55827	C	CG	1	1	0.003125
56443	G	AG	1	1	0.003125
56543	G	AG	1	1	0.003125
56824	A	AC	1	1	0.015625
57062	C	CT	1	1	0.009375
57158	T	CT	1	1	0.006250
57228	G	CG	1	1	0.062500
57631	G	AG	1	1	0.006250
58244	G	CG	1	1	0.500000
58250	C	CT	1	1	0.059375
58275	T	CT	1	1	0.500000
58946	T	CT	1	1	0.003125
59552	G	AG	1	1	0.003125
59631	A	AG	1	1	0.500000
59754	C	CT	1	1	0.465625

Two things should be noted here:

1. The bounds are tight (i.e., the upper and lower bounds are very similar). Also, the posterior probabilities of a site being a SNP are all greater than 0.8 for the sites inferred to be SNPs.
2. The minor allele frequency 0.5 appears more often than expected. This is due to the fact that this frequency was defined to be ≤ 0.5 and some mistakes may have been made in inferring which is the major allele when the minor allele frequency is close to 0.5. This is not of major concern, since the goal of the analysis was to locate SNPs.

A.2 An Example of Genotyping

Here, an example from the real data on how to calculate the Bayesian score and how to infer about the genotype of an individual is presented.

Consider site 55447. This site is inferred to be a SNP.

Table A.2: Details of the Site 55447

No. of Site	Major Allele	Minor Allele	Minor Allele Frequency
55447	A	G	0.006250

The Bayesian score of a SNP can be calculated by,

$$W = \frac{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma})}{\pi L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma}) + (1 - \pi)L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)}.$$

As explained in section 2.4, π can be assumed to be $\frac{1}{300} \cong 0.0033$. $L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = \hat{\gamma})$ for this site is approximately 7.11×10^{-66} . Also, $L(\mathbf{X}; \hat{\mathbf{P}}, \gamma = 0)$ is approximately 3.86×10^{-69} . Hence,

$$W = \frac{0.0033 \times 7.11 \times 10^{-66}}{(0.0033 \times 7.11 \times 10^{-66}) + (1 - 0.0033 \times 3.86 \times 10^{-69})} = \frac{6.138906}{7.135572} = 0.8603242.$$

(Programme in appendix B.1).

Consider the site 59754. This site is also inferred to be a SNP.

Consider the whole set of 160 individuals at this site. As before, suppose M and m represent the major and the minor alleles, respectively. The genotype of an individual is inferred to be the one that is most likely given the data for that individual. So, for example, if $P_i(MM|\mathbf{x}_i)$ is higher in comparison with $P_i(mm|\mathbf{x}_i)$ and $P_i(Mm|\mathbf{x}_i)$, then the inferred genotype is MM , which means that

Table A.3: Details of the Site 59754

No. of Site	Major Allele	Minor Allele	Minor Allele Frequency
59754	C	T	0.465625

the individual is assumed to have 2 copies of the major allele.

For the site considered, the major and the minor alleles are C and T , respectively. Therefore, the genotype of each individual at this site is CC (2 copies of the major allele), or TT (2 copies of the minor allele) or CT (1 copy of the major allele and 1 copy of the minor allele).

The sequence of reads for these individuals at this site, along with the corresponding likelihoods and the inferred genotype are presented in Table A.3. (See programme in appendix B.2). Calculation of the likelihoods was previously explained in subsection 6.2.1.

Table A.4: Likelihoods and Genotyping of Individuals at Site 59754

Individual	Sequence of Reads	$P(MM x_i)$	$P(mm x_i)$	$P(Mm x_i)$	Inferred Genotype
1	TTTCCCGCCCTTCTCTT	$< 10^{-8}$	$< 10^{-8}$	1	CT
2	CCCCCGCCCGCCCGCCCGCC	0.9999917	$< 10^{-8}$	8.3×10^{-7}	CC
3	TCCTTCG	$< 10^{-8}$	$< 10^{-8}$	1	CT
4	CCTTTC	$< 10^{-8}$	$< 10^{-8}$	1	CT
5	TCCCTTCTCCC	$< 10^{-8}$	$< 10^{-8}$	1	CT
6	TCCCTTCTTC	$< 10^{-8}$	$< 10^{-8}$	1	CT
7	TTC	5.62×10^{-6}	$< 10^{-8}$	1	CT
8	CTTTTT	$< 10^{-8}$	0.02147039	0.97852399	CT
9	TTTTTTCTTTTT	$< 10^{-8}$	0.94611342	0.05388658	TT
10	CCCC	$< 10^{-8}$	0.99955376	0.00044624	TT
11	CCCTT	0.87113431	$< 10^{-8}$	0.12886569	CC
12	CCCC	2.9×10^{-7}	$< 10^{-8}$	0.9999971	CT
13	CCCCCCCCCCCC	0.90140554	$< 10^{-8}$	0.09859446	CT
14	CTCTCTCCC	0.9992162	$< 10^{-8}$	0.0007838	CC
15	TT	$< 10^{-8}$	$< 10^{-8}$	1	CT
16	TTGCTCG	4×10^{-8}	0.63525511	0.36474485	TT
17	CCCTTTTCCGCTGCT	$< 10^{-8}$	$< 10^{-8}$	1	CT
18	CCCCCCCC	$< 10^{-8}$	$< 10^{-8}$	1	CT
19	TTCTTT	0.99829922	$< 10^{-8}$	0.00170078	CC
20	TC	$< 10^{-8}$	0.00139171	0.99860829	CT
21	TTTTTTTTTTTT	0.00720412	0.00017244	0.99262344	CT
22	CCCCCGCCCGCC	$< 10^{-8}$	0.99587711	0.00412289	TT
23	No Reads	0.9994195	$< 10^{-8}$	0.0005805	CC
24	TTTTCTTTTCTCTCTTT	N/A	N/A	N/A	N/A
25	TCCTTTT	$< 10^{-8}$	$< 10^{-8}$	1	CT
26	TCTTTT	$< 10^{-8}$	8.8×10^{-7}	0.9999912	CT
27	CCTTCTCCCC	$< 10^{-8}$	0.70522539	0.29477461	TT
28	CCCCCGCCCGCC	3.234×10^{-5}	$< 10^{-8}$	0.9996766	CT
29	TTTTTTTTTTT	0.99978509	$< 10^{-8}$	0.00021491	CC
30	CCC	$< 10^{-8}$	0.99887304	0.00112696	TT
31	TTTTTT	0.82108438	$< 10^{-8}$	0.17891562	CC
32	TTTTTTTTTTTTT	$< 10^{-8}$	0.96523169	0.03476831	TT
33	CGCCCGCC	$< 10^{-8}$	0.99589203	0.00410797	TT
34	CCCCCCCCCCCC	0.99085468	$< 10^{-8}$	0.00914532	CC
		0.996852	$< 10^{-8}$	0.003148	CC

35	C	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
36	CT	0.00037132	< 10 ⁻⁸	< 10 ⁻⁸	0.99962868	CT
37	CTCCCTCCCTTT	0.99489359	< 10 ⁻⁸	< 10 ⁻⁸	0.00510641	CC
38	CCCCCCCC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
39	TCTCCTCC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
40	CCCTTCT	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
41	TCCTCCTC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
42	CCC	0.82108058	< 10 ⁻⁸	< 10 ⁻⁸	0.17891942	CC
43	TTTTTTTTTTTT	< 10 ⁻⁸	0.99587897	< 10 ⁻⁸	0.00412103	TT
44	TTCTCTC	4 × 10 ⁻⁸	5 × 10 ⁻⁸	< 10 ⁻⁸	0.99999991	CT
45	CCCCCCCC	0.99539453	< 10 ⁻⁸	< 10 ⁻⁸	0.00460547	CC
46	CCCCC	0.94819785	< 10 ⁻⁸	< 10 ⁻⁸	0.05180215	CC
47	TTTTTTTTTTTTTTTT	< 10 ⁻⁸	0.99991261	< 10 ⁻⁸	8.739 × 10 ⁻⁵	TT
48	TTCCCTTTC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
49	CTTTTTTC	< 10 ⁻⁸	8.843 × 10 ⁻⁵	< 10 ⁻⁸	0.99991157	CT
50	CCCCCTC	0.01809787	< 10 ⁻⁸	< 10 ⁻⁸	0.98190213	CT
51	CCCCC	0.87135609	< 10 ⁻⁸	< 10 ⁻⁸	0.12864391	CC
52	TTTC	< 10 ⁻⁸	0.00055111	< 10 ⁻⁸	0.99944889	CT
53	TCTCT	< 10 ⁻⁸	2.78 × 10 ⁻⁶	< 10 ⁻⁸	0.99999722	CT
54	TTCTCTT	< 10 ⁻⁸	1.298 × 10 ⁻⁵	< 10 ⁻⁸	0.99998702	CT
55	TCTTTCCGCCCTTTTT	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
56	CCCCCC	0.96431844	< 10 ⁻⁸	< 10 ⁻⁸	0.03568156	CC
57	No Reads	N/A	N/A	N/A	N/A	N/A
58	CTTTCCTTCTCT	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
59	CTC	0.00212856	4.39 × 10 ⁻⁶	< 10 ⁻⁸	0.99786705	CT
60	TTTT	4 × 10 ⁻⁸	0.56243559	0.43756437	0.43756437	TT
61	TTTTT	< 10 ⁻⁸	0.58335643	0.41664357	0.41664357	TT
62	CCC	0.62875736	1 × 10 ⁻⁸	0.37124263	0.37124263	CC
63	TTT	1 × 10 ⁻⁷	0.56244084	0.43755906	0.43755906	TT
64	TTTTTTTT	< 10 ⁻⁸	0.9977354	0.0022646	0.0022646	TT
65	CTTCTTTC	< 10 ⁻⁸	2.802 × 10 ⁻⁵	0.99997198	0.99997198	TT
66	CCC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
67	CCCCCCC	0.82099408	< 10 ⁻⁸	< 10 ⁻⁸	0.17900592	CC
68	TTTTTTTTTTTTTTTT	0.98164961	< 10 ⁻⁸	< 10 ⁻⁸	0.01835039	CC
69	CCCTT	< 10 ⁻⁸	0.9999912	8.8 × 10 ⁻⁷	8.8 × 10 ⁻⁷	TT
70	CCCTTCTTCCCTCCTCCTCC	1.21 × 10 ⁻⁶	0.02695597	0.97304281	0.97304281	CT
71	CGGGCCCCCCCC	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
72	CCNCTTCCCTTCCCTC	0.99980445	< 10 ⁻⁸	< 10 ⁻⁸	0.00019555	CC
73	TTTTTTTTTTTTTTTTTTTT	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
			0.99999949	5.1 × 10 ⁻⁷	5.1 × 10 ⁻⁷	TT

74	TTTTTTTTTTTTTT	< 10 ⁻⁸	0.99961788	0.00038212	TT
75	CTTCTCTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
76	CCCCCICCC	3.42 × 10 ⁻⁶	< 10 ⁻⁸	0.99999658	CT
77	TTTTTTTTTT	< 10 ⁻⁸	0.99390738	0.00609262	TT
78	TCTTCCCTCCCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
79	CTCTCCCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
80	TTTTTCTTTTGT	< 10 ⁻⁸	0.9935027	0.0064973	TT
81	CCCC	0.94832691	< 10 ⁻⁸	0.05167309	CC
82	CCCCC	0.97345085	< 10 ⁻⁸	0.02654915	CC
83	CCCCCCCCCCCC	0.99882714	< 10 ⁻⁸	0.00117286	CC
84	TTTTTT	< 10 ⁻⁸	0.78958766	0.21041234	TT
85	TTTTTTTTTTTTTTTT	< 10 ⁻⁸	0.99997606	2.394 × 10 ⁻⁵	TT
86	CCCCCCCC	0.99658668	< 10 ⁻⁸	0.00341332	CC
87	TTTTTTTTTTTTTT	< 10 ⁻⁸	0.99985935	0.00014065	TT
88	CCCC	0.94827819	< 10 ⁻⁸	0.05172181	CC
89	GTTCCTTCTCTTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
90	CCCCCCCC	0.99914587	< 10 ⁻⁸	0.00085413	CC
91	CCCCCCCCCCCC	0.99999312	< 10 ⁻⁸	6.88 × 10 ⁻⁶	CC
92	CCCCC	0.931003	< 10 ⁻⁸	0.068997	CC
93	TCTTTTC	< 10 ⁻⁸	1.6 × 10 ⁻⁷	0.99999984	CT
94	CCCCCCTCCCTCC	2.357 × 10 ⁻⁵	< 10 ⁻⁸	0.99997643	CT
95	ACTCTCCCTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
96	TTTCTTCCCTTCTCCTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
97	CCCTCTT	< 10 ⁻⁸	1.76 × 10 ⁻⁶	0.99999824	CT
98	CTTCTCTCCTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
99	CCCCCCCCCCCC	0.99884537	< 10 ⁻⁸	0.00115463	CC
100	CCCCC	0.96425899	< 10 ⁻⁸	0.03574101	CC
101	TTTTCTTTTT	< 10 ⁻⁸	1.3 × 10 ⁻⁶	0.9999987	CT
102	TTCC	2.7 × 10 ⁻⁶	4.4 × 10 ⁻⁷	0.99999686	CT
103	CCCCCCCCCCCC	0.99883796	< 10 ⁻⁸	0.00116204	CC
104	TTTTTTTTTTTTTT	< 10 ⁻⁸	0.99923616	0.00076384	TT
105	CCCTCCTTCC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
106	TTTTCCCTTTTTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
107	TCTTCCCTTCCCTTTTTCC	< 10 ⁻⁸	0.9955231	0.0044769	TT
108	TTTTTTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
109	CCCCTTTTTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
110	CTTCTCTCTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
111	CTTCTCTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
112	TTCCCT	< 10 ⁻⁸	2.2 × 10 ⁻⁷	0.99999978	CT

113	TTTTTTTTTTTT	< 10 ⁻⁸	0.99943842	0.00056158	TT
114	CTTCCTTTT	< 10 ⁻⁸	1.1 × 10 ⁻⁷	0.9999989	CT
115	TTTTCTCCTTTTTCTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
116	CTCTTTTCTTTTCACTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
117	CTCTCCCTCCTTTTTTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
118	TCCTCTTTCCCTTCTCTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
119	CTCTCTCCTCCCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
120	CTCTTTCTTTTTCCCTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
121	CCCCCCCCCCCC	0.9997083	< 10 ⁻⁸	0.0002917	CC
122	CCCCCCCCCCCC	0.99577492	< 10 ⁻⁸	0.00422508	CC
123	CCCCCCCCCCCC	0.9992757	< 10 ⁻⁸	7.243 × 10 ⁻⁵	CC
124	CCCCCCCCCCCC	0.9998555	< 10 ⁻⁸	0.00014445	CC
125	CTTCTTTCTCCTTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
126	TCCTCTTTTTTCTTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
127	CCCCCCCC	0.9937694	< 10 ⁻⁸	0.0062306	CC
128	TTTTTC	< 10 ⁻⁸	0.94616484	0.05383516	TT
129	TTCTCTCCCTTTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
130	CCCCCCCC	0.9953644	< 10 ⁻⁸	0.0046356	CC
131	CCCCCCCCCCCC	0.9999544	< 10 ⁻⁸	4.56 × 10 ⁻⁶	CC
132	TTTTTTTTTTTT	< 10 ⁻⁸	0.99793813	0.00206187	TT
133	TTTTCCCTCC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
134	TTTCTCCCTCCTCTCTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
135	TTCCCTCTCT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
136	CCTTCCCTCTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
137	TTCTCTCCTTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
138	CCCCCCCCCCCC	0.99706937	< 10 ⁻⁸	0.00293063	CC
139	CCCCCCCCCCCC	0.99843399	< 10 ⁻⁸	0.00156601	CC
140	CCCCCCCCCCCC	0.99995099	< 10 ⁻⁸	4.901 × 10 ⁻⁵	CC
141	CCT	0.0005349	0.00069435	0.99877075	CT
142	CCT	0.0005349	0.00069435	0.99877075	CT
143	TTTTTTTT	< 10 ⁻⁸	0.04241906	0.95758094	CT
144	CCCCCCCCCCCC	0.99893778	< 10 ⁻⁸	0.00106222	CC
145	TTTTTTTTTTTTTTTT	< 10 ⁻⁸	0.999988	1.2 × 10 ⁻⁵	TT
146	CCCCCTCCTCCCTCTC	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
147	TTTTCTCCTTTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
148	CCCCCCCCCCCC	0.99993294	< 10 ⁻⁸	6.706 × 10 ⁻⁵	CC
149	TTCTCTTCTCCTTTCTT	< 10 ⁻⁸	< 10 ⁻⁸	1	CT
150	TTTTTTTTTTTTTCTTT	< 10 ⁻⁸	0.99984886	0.00015114	TT
151	TTTTTTT	< 10 ⁻⁸	0.98213034	0.01786966	TT

152	CCCCCC	0.93122796	$< 10^{-8}$	$< 10^{-8}$	0.068877204	CC
153	TTTTTTTTTTTTTTTT	$< 10^{-8}$	0.99992957	$< 10^{-8}$	7.043×10^{-5}	TT
154	CTCTCTCCT	$< 10^{-8}$	$< 10^{-8}$	1	1	CT
155	TTTTTTTTTTTT	$< 10^{-8}$	0.99774973	$< 10^{-8}$	0.00225027	TT
156	CTCTCCCCCTCTCCT	$< 10^{-8}$	$< 10^{-8}$	1	1	CT
157	TTTT	$< 10^{-8}$	0.87447095	$< 10^{-8}$	0.12552905	TT
158	CCCCCCCC	0.9932276	$< 10^{-8}$	$< 10^{-8}$	0.0067724	CC
159	TTTTTTTTTTTT	$< 10^{-8}$	0.9977553	$< 10^{-8}$	0.0022447	TT
160	TTTTTTTTTTTTTTTT	$< 10^{-8}$	0.99971465	$< 10^{-8}$	0.00028535	TT

A.2.1 Illustrating how the likelihoods are calculated

To illustrate how the likelihoods are calculated, consider the same SNP in locus 59754. One individual at this site is picked, and then using the sequence of reads for that individual, the probabilities that the individual has a particular genotype will be calculated.

Consider the 11th individual at this SNP. The sequence of reads for this individual is:

CCCTT.

The sequence of probabilities of errors for this individual ($\hat{p}_{11,j}$) is:

$$\begin{aligned} &3.981072 \times 10^{-5}, \\ &7.943282 \times 10^{-4}, \\ &5.011872 \times 10^{-5}, \\ &1.584893 \times 10^{-4}, \\ &1.000000 \times 10^{-4}. \end{aligned}$$

The major and minor alleles at this SNP are C and T , respectively, and the relative frequency of the minor allele is 0.465625. Therefore, there are three possible genotypes for this individual at this site, CC (two copies of the major allele), TT (two copies of the minor allele) and CT (one copy of the major allele and one copy of the minor allele). Using the *MLE* of the minor allele frequency, the prior probabilities of the above genotypes can be calculated as follows:

$$P(CC) = (1 - \gamma)^2 = (1 - 0.465625)^2 = 0.2855566.$$

$$P(TT) = \gamma^2 = (0.465625)^2 = 0.2168066.$$

$$P(CT) = 2\gamma(1 - \gamma) = 2 \times 0.465625 \times (1 - (0.465625)) = 0.4976367.$$

Using the above equations, Bayesian scores can be calculated for each genotype. As explained in section 4.2.1, firstly the likelihood of the data (the sequence of reads) for the individual given the genotype is calculated. The number of reads of the major and the minor

alleles together at this site is 5, i.e., $r_{11} = 5$.

If the genotype of this individual is CT , then the probability of a read being C (or T) is approximately 0.5. Hence,

$$L_{11}(\mathbf{x}_{11}|CT) \cong (0.5)^5 \prod_{1 \leq j \leq 5: x_{11,j} \notin \{C,T\}} 0.5 = 0.5^5 = 0.03125.$$

If the genotype of this individual is CC , so there are no alleles but the major allele C , any reads of anything other than the major allele are errors. Hence,

$$L_{11}(\mathbf{x}_{11}|CC) = \prod_{1 \leq j \leq 5: x_{11,j} = C} 1 - \hat{p}_{11,j} \prod_{1 \leq j \leq 7: x_{3,j} \neq C} \hat{p}_{11,j} =$$

$$(1 - 3.981072 \times 10^{-5}) \times (1 - 7.943282 \times 10^{-4}) \times (1 - 5.011872 \times 10^{-5})$$

$$\times (1.584893 \times 10^{-4}) \times (1.000000 \times 10^{-4}) = 1.583492 \times 10^{-8}.$$

Similarly, if there are no alleles but the prospective minor allele T , any reads of anything other than the minor allele is an error, hence:

$$L_{11}(\mathbf{x}_{11}|TT) = \prod_{1 \leq j \leq 5: x_{11,j} = T} 1 - \hat{p}_{11,j} \prod_{1 \leq j \leq 5: x_{11,j} \neq C} \hat{p}_{11,j}$$

$$= (3.981072 \times 10^{-5}) \times (7.943282 \times 10^{-4}) \times (5.011872 \times 10^{-5})$$

$$\times (1 - 1.584893 \times 10^{-4}) \times (1 - 1.000000 \times 10^{-4}) = 1.584484 \times 10^{-12}.$$

Now, the likelihood of the sequence of reads for this individual, $L_{11}(\mathbf{x}_{11})$, can be calculated using:

$$L_{11}(\mathbf{x}_{11}) = L_{11}(\mathbf{x}_{11}|CC)P(CC) + L_{11}(\mathbf{x}_{11}|CT)P(CT) +$$

$$L_{11}(\mathbf{x}_{11}|TT)P(TT) =$$

$$\begin{aligned}
& (1.583492 \times 10^{-8} \times 0.2855566) + (0.03125 \times 0.4976367) + \\
& (1.584484 \times 10^{-12} \times 0.2168066) = \\
& 0.01555115.
\end{aligned}$$

Using the likelihood function of the data, the Bayesian score for each of the genotypes at a SNP can be calculated. The genotype of an individual is inferred to be the one that is most likely given the data for that individual.

$$\begin{aligned}
P_{11}(TT|\mathbf{x}_{11}) &= \frac{L_{11}(\mathbf{x}_{11}|TT)P(TT)}{L_{11}(\mathbf{x}_{11})} = \\
& \frac{1.584484 \times 10^{-12} \times 0.2168066}{0.01555115} = 2.20901 \times 10^{-11},
\end{aligned}$$

$$\begin{aligned}
P_{11}(CT|\mathbf{x}_{11}) &= \frac{L(\mathbf{x}_{11}|CT)P(CT)}{L(\mathbf{x}_{11})} = \\
& \frac{0.03125 \times 0.4976367}{0.01555115} = 0.9999997,
\end{aligned}$$

$$\begin{aligned}
P(CC|\mathbf{x}_{11}) &= 1 - P(TT|\mathbf{x}_{11}) - P(CT|\mathbf{x}_{11}) = \\
& 1 - 2.20901 \times 10^{-11} - 0.9999997 = 2.907673 \times 10^{-7}.
\end{aligned}$$

Among the three probabilities, $P_{11}(CT|\mathbf{x}_{11})$ is greater than the other two, which means given the sequence of reads, it is most likely that the genotype of this individual is CT , i.e., the individual has

got one copy of the major allele (C), and one copy of the minor allele (T).

Appendix B

R Programmes

B.1 programme 1: SNP Detection from the Real Data

B.1.1 A brief note about the programme

The following programme has been written and used to detect the SNPs in the real data obtained from a sequencing laboratory in Dublin. Programme 2 should be run after this programme to genotype the data.

It should be noted that in order to speed up the detection of SNPs, only the likelihood scores for minor allele frequencies ≤ 0.05 were calculated (w goes from 1 to 16, i.e., 1 to 16 copies of the minor allele in the sample). The assumption is that if the minor allele frequency is > 0.05 , then a SNP will be detected with a probability very close to 1 (this is in agreement with the results of the simulations). So, once the programme was run and the SNPs were detected, the programme was run again for the sites where the minor allele frequency was initially inferred to be 0.05, knowing that the minor allele frequency in these sites is probably > 0.05 .

```
p_value <- rep(NA, 59999)
genotype <- rep(NA, 59999)
major <- rep(NA, 59999)
prosmminor <- rep(NA, 59999)
Ldata <- mat.or.vec(59999, 160)
Pmmdata <- mat.or.vec(59999, 160)
PMmmdata <- mat.or.vec(59999, 160)
PMMdata <- mat.or.vec(59999, 160)
nr <- 59999
nc <- 322
papr <- 1/300
pest <- 0
pe <- 0
Bayes_Score <- 0
M <- 0
data2 <- mat.or.vec(1, nc)
s <- h
j <- 4
L0 <- mat.or.vec((nc-2)/2, 1)
LAA <- mat.or.vec((nc-2)/2, 1)
LAC <- mat.or.vec((nc-2)/2, 1)
LAG <- mat.or.vec((nc-2)/2, 1)
LAT <- mat.or.vec((nc-2)/2, 1)
```



```

}
if ((tdata1[k]=="A" | (tdata1[k]==r)) {
  prodra <- prodra*(0.5)
}
else {
  prodra <- prodra*(10^(-q[k]/10))
}
}
LAA[c] <- (((1-p)^2)*L0[c])+(2*p*(1-p)*prodra)+((p^2)*prodaa)
}
if (r=="C") {
  LAC[c] <- L0[c]
}
else {
  for (k in 1:n) {
    if ((tdata1[k]=="w" | (r=="w"))) break
    if (tdata1[k]=="C") {
      prodcc <- prodcc*(1-(10^(-q[k]/10)))
    }
    else {
      prodcc <- prodcc*(10^(-q[k]/10))
    }
    if ((tdata1[k]=="C" | (tdata1[k]==r)) {
      prodrC <- prodrC*(0.5)
    }
    else {
      prodrC <- prodrC*(10^(-q[k]/10))
    }
  }
  LAC[c] <- (((1-p)^2)*L0[c])+(2*p*(1-p)*prodrC)+((p^2)*prodcc)
}
if (r=="G") {
  LAG[c] <- L0[c]
}
else {
  for (k in 1:n) {
    if ((tdata1[k]=="w" | (r=="w"))) break
    if (tdata1[k]=="G") {
      prodgg <- prodgg*(1-(10^(-q[k]/10)))
    }
    else {
      prodgg <- prodgg*(10^(-q[k]/10))
    }
    if ((tdata1[k]=="G" | (tdata1[k]==r)) {
      prodrG <- prodrG*(0.5)
    }
    else {
      prodrG <- prodrG*(10^(-q[k]/10))
    }
  }
}
}

```

```

LAG[c] <- (((1-p)^2)*L0[c])+(2*p*(1-p)*prodr) + ((p^2)*prodgg)
}
if (r=="T") {
LAT[c] <- L0[c]
}
else {
for (k in 1:n) {
if ((tdata1[k]=="w" | (r=="w"))) break
if (tdata1[k]=="T") {
prodtt <- prodtt*(1-(10^(-q[k]/10)))
}
else {
prodtt <- prodtt*(10^(-q[k]/10))
}
if ((tdata1[k]=="T" | (tdata1[k]==r)) {
prodr <- prodr*(0.5)
}
else {
prodr <- prodr*(10^(-q[k]/10))
}
}
LAT[c] <- (((1-p)^2)*L0[c])+(2*p*(1-p)*prodr) + ((p^2)*prodtt)
}
j <- j+2
}
LO[which(L0 < 0)] <- 1 ; LO[which(is.na(L0))] <- 1
LAA[which(LAA < 0)] <- 1 ; LAA[which(is.na(LAA))] <- 1
LAC[which(LAC < 0)] <- 1 ; LAC[which(is.na(LAC))] <- 1
LAG[which(LAG < 0)] <- 1 ; LAG[which(is.na(LAG))] <- 1
LAT[which(LAT < 0)] <- 1 ; LAT[which(is.na(LAT))] <- 1
LO <- log(L0)
LAA <- log(LAA)
LAC <- log(LAC)
LAG <- log(LAG)
LAT <- log(LAT)
site0 <- 0
siteAA <- 0
siteAC <- 0
siteAG <- 0
siteAT <- 0
site0 <- sum(L0)
siteAA <- sum(LAA)
siteAC <- sum(LAC)
siteAG <- sum(LAG)
siteAT <- sum(LAT)
M <- max(siteAA, siteAC, siteAG, siteAT)
if (M>Mp & M!=site0) {
Mp <- M
Stat[s] <- max(0,2*(Mp-site0))
pe <- p

```

```
}  
j <- 4  
}  
p_value[s] <- 1-pchisq(Stat[s], df=1)  
pest[s] <- pe/320  
Bayes_Score[s] <- (exp(Mp-site0)*papr)/((1-papr)+(exp(Mp-site0)*papr))  
s <- s+1  
}
```

B.2 programme 2: Genotyping of the SNPs

```

s <- s-1
j <- 2
nA <- 0
nC <- 0
nG <- 0
nT <- 0
nAC <- 0
nAG <- 0
nAT <- 0
nCA <- 0
nCG <- 0
nCT <- 0
nGA <- 0
nGC <- 0
nGT <- 0
nTA <- 0
nTC <- 0
nTG <- 0
major[s] <- r
error <- 1
LMM <- rep(1, 160)
Lmm <- rep(1, 160)
LMm <- rep(1, 160)
ord <- c(siteAA, siteAC, siteAG, siteAT)
ord<-order(ord)
dec=0
prosminor[s] <- ord[4]
if (prosminor[s]==1 & r!="A")
{prosminor[s] <- "A"
dec=1}
if (prosminor[s]==2 & r!="C")
{prosminor[s] <- "C"
dec=1}
if (prosminor[s]==3 & r!="G") {
prosminor[s] <- "G"
dec=1
}
if (prosminor[s]==4 & r!="T") {
prosminor[s] <- "T"
dec=1}
# if (dec==0) {for (jjj in 1:4)
# {
# if (ord[jjj]==3) {prosminor[s]=jjj}}
if (dec==0) prosminor[s] <- ord[3]
if (prosminor[s]==1) {prosminor[s] <- "A"}
if (prosminor[s]==2) {prosminor[s] <- "C"}
if (prosminor[s]==3) {prosminor[s] <- "G"}

```

```

if (prosmminor[s]==4) {prosmminor[s] <- "T"}
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="A") {nA <- nA+1}
if (tdata1[i]=="C") {nC <- nC+1}
if (tdata1[i]=="G") {nG <- nG+1}
if (tdata1[i]=="T") {nT <- nT+1}
}
j <- j+2
}
PMM <- (1-pe)^2
PMm <- 2*pe*(1-pe)
Pmm <- pe^2
j <- 2
if (major[s]=="A" & prosmminor[s]=="C") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="A") {
LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="C") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
}
if (tdata1[i]!="A" & tdata1[i]!="C") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*(10^(-q[i]/10))
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
j <- j+2
}
}

if (major[s]=="A" & prosmminor[s]=="G") {

```

```

for (c in 1:((nc-2)/2)) {
  n <- nchar(data1[s, j])
  q <- mat.or.vec(1, n)
  q <- as.numeric(charToRaw(data1[s, j+1]))
  q <- q-33
  tdata1 <- mat.or.vec(1, n)
  tdata1 <- data1[s, j]
  tdata1 <- strsplit(tdata1, split="")[[1]]
  for (i in 1:n) {
    if (tdata1[i]=="A") {
      LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
      LMm[c] <- LMm[c]*0.5
      Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
    }
    if (tdata1[i]=="G") {
      LMM[c] <- LMM[c]*(10^(-q[i]/10))
      LMm[c] <- LMm[c]*0.5
      Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
    }
    if (tdata1[i]!="A" & tdata1[i]!="G") {
      LMM[c] <- LMM[c]*(10^(-q[i]/10))
      LMm[c] <- LMm[c]*(10^(-q[i]/10))
      Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
    }
  }
  j <- j+2
}

if (major[s]=="A" & prosminor[s]=="T") {
  for (c in 1:((nc-2)/2)) {
    n <- nchar(data1[s, j])
    q <- mat.or.vec(1, n)
    q <- as.numeric(charToRaw(data1[s, j+1]))
    q <- q-33
    tdata1 <- mat.or.vec(1, n)
    tdata1 <- data1[s, j]
    tdata1 <- strsplit(tdata1, split="")[[1]]
    for (i in 1:n) {
      if (tdata1[i]=="A") {
        LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
        LMm[c] <- LMm[c]*0.5
        Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
      }
      if (tdata1[i]=="T") {
        LMM[c] <- LMM[c]*(10^(-q[i]/10))
        LMm[c] <- LMm[c]*0.5
        Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
      }
    }
    if (tdata1[i]!="A" & tdata1[i]!="T") {

```

```

                                LMM[c] <- LMM[c]*(10^(-q[i]/10))
                                LMm[c] <- LMm[c]*(10^(-q[i]/10))
                                Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
j <- j+2
}
}

if (major[s]=="C" & prosminor[s]=="A") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
      q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="C") {
  LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="A") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="C" & tdata1[i]!="A") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*(10^(-q[i]/10))
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j<-j+2
}
}

if (major[s]=="C" & prosminor[s]=="G") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
      q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {

```

```

if (tdata1[i]=="C") {
  LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="G") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="C" & tdata1[i]!="G") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*(10^(-q[i]/10))
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

if (major[s]=="C" & prosminor[s]=="T") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
  q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="C") {
  LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="T") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="C" & tdata1[i]!="T") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*(10^(-q[i]/10))
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

```

```

if (major[s]=="G" & prosminor[s]=="A") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="G") {
LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="A") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="G" & tdata1[i]!="A") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*(10^(-q[i]/10))
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

if (major[s]=="G" & prosminor[s]=="C") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q<-as.numeric(charToRaw(data1[s, j+1]))
q<-q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="G") {
LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="C") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
}
}
}

```

```

if (tdata1[i]!="G" & tdata1[i]!="C") {
    LMM[c] <- LMM[c]*(10^(-q[i]/10))
    LMm[c] <- LMm[c]*(10^(-q[i]/10))
    Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
j <- j+2
}
}

if (major[s]=="G" & prosminor[s]=="T") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
    q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="G") {
    LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
    LMm[c] <- LMm[c]*0.5
    Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="T") {
    LMM[c] <- LMM[c]*(10^(-q[i]/10))
    LMm[c] <- LMm[c]*0.5
    Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
}
if (tdata1[i]!="G" & tdata1[i]!="T") {
    LMM[c] <- LMM[c]*(10^(-q[i]/10))
    LMm[c] <- LMm[c]*(10^(-q[i]/10))
    Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

if (major[s]=="T" & prosminor[s]=="A") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
    q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {

```

```

if (tdata1[i]=="T") {
  LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="A") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="T" & tdata1[i]!="A") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*(10^(-q[i]/10))
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

if (major[s]=="T" & prosmenor[s]=="C") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="T") {
  LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="C") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*0.5
  Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="T" & tdata1[i]!="C") {
  LMM[c] <- LMM[c]*(10^(-q[i]/10))
  LMm[c] <- LMm[c]*(10^(-q[i]/10))
  Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

```

```

if (major[s]=="T" & prosmminor[s]=="G") {
for (c in 1:((nc-2)/2)) {
n <- nchar(data1[s, j])
q <- mat.or.vec(1, n)
q <- as.numeric(charToRaw(data1[s, j+1]))
q <- q-33
tdata1 <- mat.or.vec(1, n)
tdata1 <- data1[s, j]
tdata1 <- strsplit(tdata1, split="")[[1]]
for (i in 1:n) {
if (tdata1[i]=="T") {
LMM[c] <- LMM[c]*(1-(10^(-q[i]/10)))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
if (tdata1[i]=="G") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*0.5
Lmm[c] <- Lmm[c]*(1-(10^(-q[i]/10)))
}
if (tdata1[i]!="T" & tdata1[i]!="G") {
LMM[c] <- LMM[c]*(10^(-q[i]/10))
LMm[c] <- LMm[c]*(10^(-q[i]/10))
Lmm[c] <- Lmm[c]*(10^(-q[i]/10))
}
}
}
j <- j+2
}
}

for (c in 1:((nc-2)/2)){
Ldata[s, c] <- (LMM[c]*PMM)+(LMm[c]*PMm)+(Lmm[c]*Pmm)
Pmmdata[s, c] <- (Lmm[c]*Pmm)/Ldata[s, c]
PMmdata[s, c] <- (LMm[c]*PMm)/Ldata[s, c]
PMMdata[s, c] <- 1-Pmmdata[s, c]-PMmdata[s, c]
}
}

for (i in snps) {
genotype[i] <- paste(major[i], prosmminor[i])
}
}

```

B.3 programme 3: SNP Detection (Tagged Data - Poisson Distribution)

```

nlanes <- 160 # no. of individuals
nsim <- 500 # no. of sites
p <- 0.0 # minor allele frequency
nsd <- rep(0, 50) # no. of SNPs detected
nmsd <- rep(0, 50) # no. of non-SNPs inferred to be SNPs
nsd2 <- rep(0, 50) # no. of SNPs detected using threshold rule
nmsd2 <- rep(0, 50) # no. of non-SNPs inferred to be SNPs using threshold rule
cdfe <- 1:21 # empirical distribution for the probability of error
points <- nlanes+1
p0 <- 1:points
h <- 0.5/nlanes
lognum1 <- 1:points
# cdfe is the empirical distribution for the error rate
cdfe[1] <- 0.004
cdfe[2] <- 0.004
cdfe[3] <- 0.005
cdfe[4] <- 0.013
cdfe[5] <- 0.027
cdfe[6] <- 0.046
cdfe[7] <- 0.06
cdfe[8] <- 0.102
cdfe[9] <- 0.155
cdfe[10] <- 0.223
cdfe[11] <- 0.305
cdfe[12] <- 0.388
cdfe[13] <- 0.491
cdfe[14] <- 0.621
cdfe[15] <- 0.768
cdfe[16] <- 0.879
cdfe[17] <- 0.966
cdfe[18] <- 0.998
cdfe[19] <- 1
cdfe[20] <- 1
cdfe[21] <- 1
for (i in 1:points){
p0[i] <- h*(i-1)
}
lambda <- 5 # this can be changed to adjust the programme
alpha <- 0.05 # this can be changed to adjust the programme
err <- 0.000891047 # this can be changed to adjust the programme
# m-pool size, mmax - max. pool size, k[j]-no. of minor alleles in lane j,
# read[j] - no. of reads of minor allele in lane
# r[j] - total no. of reads in lane
# pp[a] - prob. of a read being for the minor allele given that there are a minor alleles,
# test1[m] - no. of sig. results with pool size m

```

```

# lograt1 - realisation of test statistic
# lr1[b,j] - likelihood measurement of results from lane j for p=p0[i]
m <- 2
pval <- 1:nsim
pval2 <- 1:nsim
maxread <- 1:nsim
lograt1 <- 1:nsim
k <- 1:nlanes
read <- 1:nlanes
r <- 1:nlanes
pp <- 1:(m+1)
lr1 <- array(0,dim=c(points,nlanes))
for (o in 1:50) {
  for (i in 1:nsim){
    if (i <= 5) {p <- 0.01}
    totr <- 0
    totread <- 0
      maxread[i] <- 0
    lograt1[i] <- 0
      logden <- 0
    for (b in 1:points){
      lognum1[b] <- 0
    }
    for (j in 1:nlanes){
      probMM <- 1
      probMm <- 1
      probmm <- 1
      read[j] <- 0
    r[j] <- rpois(1,lambda)
    k[j] <- rbinom(1,m,p)
    for (kk in 1:r[j]){
      x <- runif(1, min=0, max=1)
    v <- 1
    while(x > pdf[v]){
      v <- v+1
    }
    qualscore <- v+19
    perr <- 10^(-qualscore/10)
    y <- runif(1, min=0, max=1)
    if(y < perr+(0.5-perr)*k[j]){
      read[j] <- read[j]+1
      logden <- logden+log(perr)
      probMM <- probMM*perr
      probMm <- probMm*0.5
      probmm <- probmm*(1-perr)
    }
    if (y > perr+(0.5-perr)*k[j]){
      logden <- logden+log(1-perr)
      probMM <- probMM*(1-perr)
      probMm <- probMm*0.5
    }
  }
}

```

```

        probmm <- probmm*perr
      }
    }
    for (b in 1:points){
      prob <- (1-p0[b])**2*probMM+2*p0[b]*(1-p0[b])*probMm+p0[b]**2*probmm
      lognum1[b] <- lognum1[b]+log(prob)
    }
  totr <- totr+r[j]
  if (read[j] > maxread[i]) {maxread[i] <- read[j]}
  totread <- totread+read[j]
}
  logmax <- lognum1[1]
for (b in 2:points){
  if (lognum1[b] > logmax){
    logmax <- lognum1[b]
  }
}
lograt1[i] <- max(0,2*(logmax-logden))
  pval[i] <- 1-pchisq(lograt1[i],1)
  pval2[i] <- 1-ppois(maxread[i]-1,lambda*err)**nlanes
p <- 0
}
#dev.new()
#par(mfrow=c(1,2))
#hist(pval)
#hist(lograt1)
alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {
  for (i in 1:cn) {
    if (pval[i] < alpha1) {
      p_value_new[counter] <- pval[i]
      position1[counter] <- position[i]
      counter <- counter+1
    }
  }
}
if (counter==1){
  st <-1
  nsd[o]<-0
  nnsd[o]<-0
}
else if (counter-1==cn) {

```

```

st <- 1
nsd[o] <- length(pval)
for (z in 1:nsd[o]){
  if (position[z] > 5) nnsd[o] <- nnsd[o]+1
}
}
else{
  cn <- counter-1
  counter <- 1
  alpha1 <- 0.05*cn/nsim
}
pval <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
  alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {
  for (i in 1:cn) {
    if (pval2[i] < alpha1) {
      p_value_new[counter] <- pval2[i]
      position1[counter] <- position[i]
      counter <- counter+1
    }
  }
}
if (counter==1){
  st <-1
  nsd2[o]<-0
  nnsd2[o]<-0
}
else if (counter-1==cn) {
  st <- 1
  nsd2[o] <- length(pval2)
  for (z in 1:nsd2[o]){
    if (position[z] > 5) nnsd2[o] <- nnsd2[o]+1
  }
}
else{
  cn <- counter-1
  counter <- 1
  alpha1 <- 0.05*cn/nsim

```

```
}  
pval2 <- p_value_new  
p_value_new <- 0  
position <- position1  
position1 <- 0  
}  
}
```

B.4 programme 4: SNP Detection (Tagged Data - Negative Binomial Distribution)

```

nlanes <- 160 # no. of individuals
nsim <- 500 # no. of sites
p <- 0.0 # minor allele frequency
nsd <- rep(0, 50) # no. of SNPs detected
nnsd <- rep(0, 50) # no. of non-SNPs inferred to be SNPs
nsd2 <- rep(0, 50) # no. of SNPs detected using threshold rule
nnsd2 <- rep(0, 50) # no. of non-SNPs inferred to be SNPs using threshold rule
pdf <- 1:21 # empirical distribution for the probability of error
points <- nlanes+1
p0 <- 1:points
h <- 0.5/nlanes
lognum1 <- 1:points
pdf[1] <- 0.004
pdf[2] <- 0.004
pdf[3] <- 0.005
pdf[4] <- 0.013
pdf[5] <- 0.027
pdf[6] <- 0.046
pdf[7] <- 0.06
pdf[8] <- 0.102
pdf[9] <- 0.155
pdf[10] <- 0.223
pdf[11] <- 0.305
pdf[12] <- 0.388
pdf[13] <- 0.491
pdf[14] <- 0.621
pdf[15] <- 0.768
pdf[16] <- 0.879
pdf[17] <- 0.966
pdf[18] <- 0.998
pdf[19] <- 1
pdf[20] <- 1
pdf[21] <- 1
for (i in 1:points){
p0[i] <- h*(i-1)
}
lambda <- 20 # this can be changed to adjust the programme
alpha <- 0.05 # this can be changed to adjust the programme
err <- 0.000891047 # this can be changed to adjust the programme
# m-pool size, mmax - max. pool size, k[j]-no. of minor alleles in lane j,
# read[j] - no. of reads of minor allele in lane
# r[j] - total no. of reads in lane
# pp[a] - prob. of a read being for the minor allele given that there are a minor alleles,
# test1[m] - no. of sig. results with pool size m
# lograt1 - realisation of test statistic

```

```

# lr1[b,j] - likelihood measurement of results from lane j for p=p0[i]
m <- 2
pval <- 1:nsim
pval2 <- 1:nsim
maxread <- 1:nsim
lograt1 <- 1:nsim
k <- 1:nlanes
read <- 1:nlanes
r <- 1:nlanes
pp <- 1:(m+1)
lr1 <- array(0,dim=c(points,nlanes))
for (o in 1:50) {
  for (i in 1:nsim){
    if (i <= 5) {p <- 0.01}
    totr <- 0
    totread <- 0
      maxread[i] <- 0
    lograt1[i] <- 0
      logden <- 0
    for (b in 1:points){
      lognum1[b] <- 0
    }
    for (j in 1:nlanes){
      probMM <- 1
      probMm <- 1
      probmm <- 1
      read[j] <- 0
    r[j] <- rbinom(1, 4, 0.2)+4
    k[j] <- rbinom(1,m,p)
    for (kk in 1:r[j]){
      x <- runif(1, min=0, max=1)
      v <- 1
      while(x > pdf[v]){
        v <- v+1
      }
      qualscore <- v+19
      perr <- 10^(-qualscore/10)
      y <- runif(1, min=0, max=1)
      if(y < perr+(0.5-perr)*k[j]){
        read[j] <- read[j]+1
          logden <- logden+log(perr)
          probMM <- probMM*perr
          probMm <- probMm*0.5
          probmm <- probmm*(1-perr)
        }
      if (y > perr+(0.5-perr)*k[j]){
        logden <- logden+log(1-perr)
        probMM <- probMM*(1-perr)
        probMm <- probMm*0.5
        probmm <- probmm*perr
      }
    }
  }
}

```

```

    }
  }
  for (b in 1:points){
    prob <- (1-p0[b])**2*probMM+2*p0[b]*(1-p0[b])*probMm+p0[b]**2*probmm
    lognum1[b] <- lognum1[b]+log(prob)
  }
totr <- totr+r[j]
  if (read[j] > maxread[i]) {maxread[i] <- read[j]}
totread <- totread+read[j]
  }
  logmax <- lognum1[1]
for (b in 2:points){
  if (lognum1[b] > logmax){
    logmax <- lognum1[b]
  }
}
lograt1[i] <- max(0,2*(logmax-logden))
  pval[i] <- 1-pchisq(lograt1[i],1)
  pval2[i] <- 1-ppois(maxread[i]-1,lambda*err)**nlanes
p <- 0
}
#dev.new()
#par(mfrow=c(1,2))
#hist(pval)
#hist(lograt1)
alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {
  for (i in 1:cn) {
    if (pval[i] < alpha1) {
      p_value_new[counter] <- pval[i]
      position1[counter] <- position[i]
      counter <- counter+1
    }
  }
}
if (counter==1){
  st <-1
  nsd[o]<-0
  nmsd[o]<-0
}
else if (counter-1==cn) {
  st <- 1

```

```

nsd[o] <- length(pval)
for (z in 1:nsd[o]){
if (position[z] > 5) nnsd[o] <- nnsd[o]+1
}
}
else{
cn <- counter-1
counter <- 1
alpha1 <- 0.05*cn/nsim
}
pval <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
      alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
position[i] <- i
}
while (st==0) {
for (i in 1:cn) {
if (pval2[i] < alpha1) {
p_value_new[counter] <- pval2[i]
position1[counter] <- position[i]
counter <- counter+1
}
}
}
if (counter==1){
st <-1
nsd2[o]<-0
nnsd2[o]<-0
}
else if (counter-1==cn) {
st <- 1
nsd2[o] <- length(pval2)
for (z in 1:nsd2[o]){
if (position[z] > 5) nnsd2[o] <- nnsd2[o]+1
}
}
}
else{
cn <- counter-1
counter <- 1
alpha1 <- 0.05*cn/nsim
}

```

```
pval2 <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
}
```

B.5 programme 5: SNP Detection (Untagged Data - Poisson Distribution)

```

nlanes <- 8 # no. of lanes
nsim <- 500 # no. of sites
p <- 0.0 # minor allele frequency
nsd <- rep(0, 50) # no. of SNPs detected
nnsd <- rep(0, 50) # no. of non-SNPs inferred to be SNPs
nsd2 <- rep(0, 50) # no. of SNPs detected using threshold rule
nnsd2 <- rep(0, 50) # no. of non-SNPs inferred to be SNPs using threshold rule
pdf <- 1:21 # empirical distribution for the probability of error
points <- (m*nlanes/2)+1
p0 <- 1:points
h <- 0.5/(points-1)
lognum1 <- 1:points
pdf[1] <- 0.004
pdf[2] <- 0.004
pdf[3] <- 0.005
pdf[4] <- 0.013
pdf[5] <- 0.027
pdf[6] <- 0.046
pdf[7] <- 0.06
pdf[8] <- 0.102
pdf[9] <- 0.155
pdf[10] <- 0.223
pdf[11] <- 0.305
pdf[12] <- 0.388
pdf[13] <- 0.491
pdf[14] <- 0.621
pdf[15] <- 0.768
pdf[16] <- 0.879
pdf[17] <- 0.966
pdf[18] <- 0.998
pdf[19] <- 1
pdf[20] <- 1
pdf[21] <- 1
for (i in 1:points){
p0[i] <- h*(i-1)
}
rri <- 20 # this can be changed to adjust the programme
ps <- 5 # this can be changed to adjust the programme
lambda <- rri*ps
alpha <- 0.05
err <- 0.01 # this can be changed to adjust the programme
# m-pool size, mmax - max. pool size, k[j]-no. of minor alleles in lane j,
# read[j] - no. of reads of minor allele in lane
# r[j] - total no. of reads in lane
# pp[a] - prob. of a read being for the minor allele given that there are a minor alleles,

```

```

# test1[m] - no. of sig. results with pool size m
# lograt1 - realisation of test statistic
# lr1[b,j] - likelihood measurement of results from lane j for p=p0[i]
m <- 2*ps
pval <- 1:nsim
pval2 <- 1:nsim
maxread <- 1:nsim
lograt1 <- 1:nsim
k <- 1:nlanes
read <- 1:nlanes
r <- 1:nlanes
pp <- 1:(m+1)
lr1 <- array(0,dim=c(points,nlanes))
prob <- 0
pro <- 0
probm <- 1
for (o in 1:50) {
  for (i in 1:nsim){
    if (i <= 5) {p <- 0.01}
    totr <- 0
    tothead <- 0
      maxread[i] <- 0
    lograt1[i] <- 0
      logden <- 0
    for (b in 1:points){
      lognum1[b] <- 0
    }
    for (j in 1:nlanes){
      for (kkk in 1:points){
        prob[kkk] <- 1
      }
      read[j] <- 0
    r[j] <- rpois(1, lambda)
    k[j] <- rbinom(1,m,p)
      for (kk in 1:(m+1)){
        probm[kk] <- 1
      }
    for (kk in 1:r[j]){
      x <- runif(1, min=0, max=1)
    v <- 1
    while(x > pdf[v]){
      v <- v+1
    }
    qualscore <- v+19
    perr <- 10^(-qualscore/10)
    for (kkk in 1:(m+1)){
      pro[kkk] <- ((kkk-1)/m)+perr*(1-(2*(kkk-1)/m))
    }
    y <- runif(1, min=0, max=1)
    if(y < pro[k[j]+1]){

```

```

read[j] <- read[j]+1
                                logden <- logden+log(perr)
for (kkk in 1:(m+1)){
  probm[kkk] <- probm[kkk]*pro[kkk]
                                }
                                }
                                if (y > pro[k[j]+1]){
                                  logden <- logden+log(1-perr)
for (kkk in 1:(m+1)){
  probm[kkk] <- probm[kkk]*(1-pro[kkk])
                                }
}
                                }
                                for (b in 2:points){
                                  prob[b]<-0
for (kkk in 1:(m+1)){
  prob[b] <- prob[b]+dbinom(kkk-1, m, p0[b])*probm[kkk]
                                }
                                lognum1[b] <- lognum1[b]+log(prob[b])
                                }
totr <- totr+r[j]
                                if (read[j] > maxread[i]) {maxread[i] <- read[j]}
totread <- totread+read[j]
                                }
                                logmax <- logden
for (b in 2:points){
  if (lognum1[b] > logmax){
    logmax <- lognum1[b]
  }
}
lograt1[i] <- max(0,2*(logmax-logden))
  pval[i] <- 1-pchisq(lograt1[i],1)
  pval2[i] <- 1-ppois(maxread[i]-1,lambda*err)**nlanes
p <- 0
}
#dev.new()
#par(mfrow=c(1,2))
#hist(pval)
#hist(lograt1)
alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {

```

```

for (i in 1:cn) {
  if (pval[i] < alpha1) {
    p_value_new[counter] <- pval[i]
    position1[counter] <- position[i]
    counter <- counter+1
  }
}
if (counter==1){
  st <-1
  nsd[o]<-0
  mnsd[o]<-0
}
else if (counter-1==cn) {
  st <- 1
  nsd[o] <- length(pval)
  for (z in 1:nsd[o]){
    if (position[z] > 5) mnsd[o] <- mnsd[o]+1
  }
}
else{
  cn <- counter-1
  counter <- 1
  alpha1 <- 0.05*cn/nsim
}
pval <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
  alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {
  for (i in 1:cn) {
    if (pval2[i] < alpha1) {
      p_value_new[counter] <- pval2[i]
      position1[counter] <- position[i]
      counter <- counter+1
    }
  }
}
if (counter==1){
  st <-1
  nsd2[o]<-0

```

```
nnsd2[o]<-0
}
else if (counter-1==cn) {
st <- 1
nnsd2[o] <- length(pval2)
for (z in 1:nnsd2[o]){
if (position[z] > 5) nnsd2[o] <- nnsd2[o]+1
}
}
else{
cn <- counter-1
counter <- 1
alpha1 <- 0.05*cn/nsim
}
pval2 <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
}
```

B.6 programme 6: SNP Detection (Untagged Data - Negative Binomial Distribution)

```

nlanes <- 8 # no. of lanes
nsim <- 500 # no. of sites
p <- 0.0 # minor allele frequency
nsd <- rep(0, 50) # no. of SNPs detected
nnsd <- rep(0, 50) # no. of non-SNPs inferred to be SNPs
nsd2 <- rep(0, 50) # no. of SNPs detected using threshold rule
nnsd2 <- rep(0, 50) # no. of non-SNPs inferred to be SNPs using threshold rule
pdf <- 1:21 # empirical distribution for the probability of error
points <- (m*nlanes/2)+1
p0 <- 1:points
h <- 0.5/(points-1)
lognum1 <- 1:points
pdf[1] <- 0.004
pdf[2] <- 0.004
pdf[3] <- 0.005
pdf[4] <- 0.013
pdf[5] <- 0.027
pdf[6] <- 0.046
pdf[7] <- 0.06
pdf[8] <- 0.102
pdf[9] <- 0.155
pdf[10] <- 0.223
pdf[11] <- 0.305
pdf[12] <- 0.388
pdf[13] <- 0.491
pdf[14] <- 0.621
pdf[15] <- 0.768
pdf[16] <- 0.879
pdf[17] <- 0.966
pdf[18] <- 0.998
pdf[19] <- 1
pdf[20] <- 1
pdf[21] <- 1
for (i in 1:points){
p0[i] <- h*(i-1)
}
rri <- 2.5 # this can be changed to adjust the programme
ps <- 40 # this can be changed to adjust the programme
lambda <- rri*ps
alpha <- 0.05 # this can be changed to adjust the programme
err <- 0.000891047 # this can be changed to adjust the programme
# m-pool size, mmax - max. pool size, k[j]-no. of minor alleles in lane j,
# read[j] - no. of reads of minor allele in lane
# r[j] - total no. of reads in lane
# pp[a] - prob. of a read being for the minor allele given that there are a minor alleles,

```

```

# test1[m] - no. of sig. results with pool size m
# lograt1 - realisation of test statistic
# lr1[b,j] - likelihood measurement of results from lane j for p=p0[i]
m <- 2*ps
pval <- 1:nsim
pval2 <- 1:nsim
maxread <- 1:nsim
lograt1 <- 1:nsim
k <- 1:nlanes
read <- 1:nlanes
r <- 1:nlanes
pp <- 1:(m+1)
lr1 <- array(0,dim=c(points,nlanes))
prob <- 0
pro <- 0
probm <- 1
for (o in 1:50) {
  for (i in 1:nsim){
    if (i <= 5) {p <- 0.01}
    totr <- 0
    tothead <- 0
      maxread[i] <- 0
    lograt1[i] <- 0
      logden <- 0
    for (b in 1:points){
      lognum1[b] <- 0
    }
    for (j in 1:nlanes){
      for (kkk in 1:points){
        prob[kkk] <- 1
      }
      read[j] <- 0
    r[j] <- rbinom(1, 20, 0.2)+20
    k[j] <- rbinom(1, m, p)
      for (kk in 1:(m+1)){
        probm[kk]<-1
      }
    for (kk in 1:r[j]){
      x <- runif(1, min=0, max=1)
    v <- 1
    while(x > pdf[v]){
      v <- v+1
    }
    qualscore <- v+19
    perr <- 10^(-qualscore/10)
    for (kkk in 1:(m+1)){
      pro[kkk] <- ((kkk-1)/m)+perr*(1-(2*(kkk-1)/m))
    }
    y <- runif(1, min=0, max=1)
    if(y < pro[k[j]+1]){

```

```

read[j] <- read[j]+1
                                logden <- logden+log(perr)
for (kkk in 1:(m+1)){
  probm[kkk] <- probm[kkk]*pro[kkk]
                                }
                                }
                                if (y > pro[k[j]+1]){
                                logden <- logden+log(1-perr)
for (kkk in 1:(m+1)){
  probm[kkk] <- probm[kkk]*(1-pro[kkk])
                                }
}
                                }
                                for (b in 2:points){
                                prob[b]<-0
for (kkk in 1:(m+1)){
  prob[b] <- prob[b]+dbinom(kkk-1, m, p0[b])*probm[kkk]
                                }
                                lognum1[b] <- lognum1[b]+log(prob[b])
                                }
totr <- totr+r[j]
                                if (read[j] > maxread[i]) {maxread[i] <- read[j]}
totread <- totread+read[j]
                                }
                                logmax <- logden
for (b in 2:points){
  if (lognum1[b] > logmax){
  logmax <- lognum1[b]
  }
}
lograt1[i] <- max(0,2*(logmax-logden))
  pval[i] <- 1-pchisq(lograt1[i],1)
  pval2[i] <- 1-ppois(maxread[i]-1,lambda*err)**nlanes
p <- 0
}
#dev.new()
#par(mfrow=c(1,2))
#hist(pval)
#hist(lograt1)
alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {

```

```

for (i in 1:cn) {
  if (pval[i] < alpha1) {
    p_value_new[counter] <- pval[i]
    position1[counter] <- position[i]
    counter <- counter+1
  }
}
if (counter==1){
  st <-1
  nsd[o]<-0
  mnsd[o]<-0
}
else if (counter-1==cn) {
  st <- 1
  nsd[o] <- length(pval)
  for (z in 1:nsd[o]){
    if (position[z] > 5) mnsd[o] <- mnsd[o]+1
  }
}
else{
  cn <- counter-1
  counter <- 1
  alpha1 <- 0.05*cn/nsim
}
pval <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
  alpha1 <- 0.05
position <- 0
position1 <- 0
p_value_new <- 0
cn <- nsim
st <- 0
counter <- 1
for (i in 1:nsim) {
  position[i] <- i
}
while (st==0) {
  for (i in 1:cn) {
    if (pval2[i] < alpha1) {
      p_value_new[counter] <- pval2[i]
      position1[counter] <- position[i]
      counter <- counter+1
    }
  }
}
if (counter==1){
  st <-1
  nsd2[o]<-0

```

```
nnsd2[o]<-0
}
else if (counter-1==cn) {
st <- 1
nnsd2[o] <- length(pval2)
for (z in 1:nnsd2[o]){
if (position[z] > 5) nnsd2[o] <- nnsd2[o]+1
}
}
else{
cn <- counter-1
counter <- 1
alpha1 <- 0.05*cn/nsim
}
pval2 <- p_value_new
p_value_new <- 0
position <- position1
position1 <- 0
}
}
```

Bibliography

- [1] Balding, D.J., Bishop, M., Cannings, C. 2001. *Handbook of Statistical Genetics*.
John Wiley & Sons, LTD.
- [2] Mange, E. J., Mange A. P. 1990. *Basic Human Genetics*.
- [3] Lario, A., Gonzalez, A. and Dorado, G. 1997. *Automated Laser-Induced Fluorescence DNA Sequencing: Equalizing Signal-to-Noise Ratios Significantly Enhances Overall Performance*.
Analytical Biochemistry 247: 30-33.
- [4] Rosenblum, B., Lee, L., Spurgeon, S., Khan, S., Menchen, S., Heiner, C. and Chen, S. 1997. *TNew Dye-Labeled Terminators for Improved DNA Sequencing Patterns*.
Nucleic Acids Research 25(22): 4500-4504.
- [5] Sanders, J. Z., Petterson, A. A., Hughes, P. J., Connell, C. R., Raff, M., Menchen, S., Hood, L. E. and Teplow, D. B. 1991. *Imaging as a Tool for Improving Length and Accuracy of Sequence Analysis in Automated Fluorescence-Based DNA Sequencing*.
Electrophoresis 12: 3-11.
- [6] Berno, A. J. 1996. *A Graph Theoretic Approach to the Analysis of DNA Sequencing Data*.
Genome Research 6: 80-91.
- [7] Ewing, B., Hillier, L., Wendl, M. C. and Green, P. 1998. *Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment*.
Genome Research 8: 175-185.
- [8] Ewing, B. and Green, P. 1998. *Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities*. *Genome Research* 8: 186-194.

- [9] Sanger, F., Nicklen, S., Coulson, A. R. 1977. *DNA Sequencing with Chain-Terminating Inhibitors*. Proc. Natl. Acad. Sci. USA, 74; 12: 5463-5467.
- [10] Sham, P., Bader, J. S., Craig, I., O'Donovan, M., Owen, M. 2012. *DNA Pooling: a Tool for Large-Scale Association Studies*. Nature Reviews Genetics 3: 862-871.
- [11] Ramsey, D., Futschik, A. 2012. *DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms*. Statistical Applications in Genetics and Molecular Biology, 11; 5: Article 1.
- [12] Futschik, A., Schlötterer, C. 2010. *Massively Parallel Sequencing of Pooled DNA Samples - the Next Generation of Molecular Markers*. Genetics 186: 207-218.
- [13] Benjamini, Y. and Hochberg, Y. (2000). *The Adaptive Control of the False Discovery Rate in Multiple Hypotheses Testing*. J. Behav. Educ. Statist, 25: 6083.
- [14] Cox, D. R. (1965). *A Remark on Multiple Comparison Methods*. Technometrics, 2: 149-156.
- [15] Tukey, J. W. (1977). *Some Thoughts on Clinical Trials, Especially Problems of Multiplicity*. Science, 198: 697-684.
- [16] Benjamini, Y., Hochberg, Y. 1995. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society, Series B, 57: 289-300.
- [17] Storey, J. D. and Tibshirani, R. 2003. *Statistical Significance for Genomewide Studies*. Proc Natl Acad Sci USA, 5; 100(16): 9440-9445.
- [18] Miller, R. G. 1981. *Simultaneous Statistical Inference*. 2nd Edition, pp. 6-8.
- [19] Benjamini, Y., Yekutieli, D. 2001. *The Control of the False Discovery Rate in Multiple Testing under Dependency*. The Annals of Statistics, 29; 4: 1165-1188.
- [20] Benjamini, Y., Hochberg, Y. and Kling, Y. (1997). *False Discovery Rate Control in Multiple Hypotheses Testing Using Dependent Test Statistics*. Research Paper 97-1, Dept. Statistics and O.R., Tel Aviv Univ.

- [21] Barreiro, LB., Laval, G., Quach, H., Patin, E., Quintana-Murci, L. 2008. *Natural Selection Has Driven Population Differentiation in Modern Humans*. *Nature Genetics* 40: 340-345.
- [22] Kelly, S., Reed, J., Kramer, S., Ellis, L., Webb, H., Sunter, J., Salje, J., Marinsek, N., Gull, K., Wickstead, B., and Carrington, M. 2007. *Functional Genomics in Trypanosoma Brucei: a Collection of Vectors for the Expression of Tagged Proteins from Endogenous and Ectopic Gene Loci*. *Mol Biochem Parasitol* 154: 103-9.
- [23] ACHAZ, G. 2008. *Testing for Neutrality in Samples with Sequencing Errors*. *Genetics* 179: 1409-1424.
- [24] DAVISON, A. 2008. *Statistical Models*. Cambridge: Cambridge University Press.
- [25] Engle, M. and Burks, C. 1994. *GenFrag 2.2: New Features for More Robust Fragment Assembly Benchmarks*. *Computer Applications in the Bioscience* 10: 567-568.
- [26] Lipshutz, R. J., Taverner, F., Henessy, K., Hartzell, G. and Davis, R. 1994. *DNA Sequence Confidence Estimation*. *Genomics* 19: 417-424.
- [27] Engle, M. and Burks, C. 1993. *Artificially Generated Data Sets for Testing DNA Fragment Assembly Algorithms*. *Genomics* 16: 286-288.
- [28] Richterich, P. 1998. *Estimation of Errors in "Raw" DNA Sequences: A Validation Study (Letter)*. *Genome Research* 8: 251-259.
- [29] Tajima F. 1989. *Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism*. *Genetics* 123: 585-595.
- [30] Nielsen R., Williamson S., Kim Y., Hubisz M. J., Clark A. G. and Bustamante C. 2005. *Genomic Scans for Selective Sweeps Using SNP Data*. *Genome Res* 15: 1566-1575.
- [31] Durrett R. and Schweinsberg J. 2004. *Approximating Selective Sweeps*. *Theor. Popul. Biol.* 66: 129-138.
- [32] Lario, A., Gonzalez, A. and Dorado, G. 1997. *Automated Laser-Induced Fluorescence DNA Sequencing: Equalizing Signal-to-*

- Noise Ratios Significantly Enhances Overall Performance.* Analytical Biochemistry. 247: 30-33.
- [33] Berno, A. J. 1996. *A Graph Theoretic Approach to the Analysis of DNA Sequencing Data.* Genome Research. 6: 80-91.
- [34] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., Hirschhorn, J. N. 2008. *Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.* Nat Rev Genet. 9(5):356-69.
- [35] Ioannidis, J. P. 2007. *Non-Replication and Inconsistency in the Genome-Wide Association Setting.* Hum Heredity. 64: 203213.
- [36] Kwok, P. Y. and Chen, X. 2003. *Detection of Single Nucleotide Polymorphisms.* Curr. Issues Mol. Biol. 5: 43-60.
- [37] Xu, J. Y., Xu, G. B. and Chen, S. L. 2003. *A New Method for SNP Discovery.* BioTechniques. 46: 201-208.
- [38] Bansal, V. 2010. *A Statistical Method for the Detection of Variants from Next-Generation Resequencing of DNA Pools.* Bioinformatics. 26: 318-324.
- [39] Muralidharan, O., Natsoulis, G., Bell, J., Newburger, D., Xu, H., Kela, I., Ji, H. and Zhang, N. 2011. *A Cross-Sample Statistical Model for SNP Detection in Short-Read Sequencing Data.* Nucleic Acids Research. 18.
- [40] Efron, B. 2004. *Large-Scale Simultaneous Hypothesis Testing: the Choice of a Null Hypothesis.* J. Am. Stat. Assoc. 99: 96-104.