

South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2017

Threshold Models for Genome-wide Association Mapping of Familial Breast Cancer Incidence in Humans

Nasir Elmesmari
South Dakota State University

Follow this and additional works at: <https://openprairie.sdstate.edu/etd>



Part of the [Statistics and Probability Commons](#)

Recommended Citation

Elmesmari, Nasir, "Threshold Models for Genome-wide Association Mapping of Familial Breast Cancer Incidence in Humans" (2017). *Electronic Theses and Dissertations*. 1724.
<https://openprairie.sdstate.edu/etd/1724>

This Dissertation - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

THRESHOLD MODELS FOR GENOME-WIDE ASSOCIATION MAPPING OF
FAMILIAL BREAST CANCER INCIDENCE IN HUMANS

BY


NASIR ELMESMARI

A dissertation submitted in partial fulfillment of the requirements for the
Doctor of Philosophy
Major in Computational Science and Statistics
South Dakota State University
2017

THRESHOLD MODELS FOR GENOME-WIDE ASSOCIATION MAPPING OF
FAMILIAL BREAST CANCER INCIDENCE IN HUMANS

NASIR ELMESMARI


This dissertation is approved as a creditable and independent investigation by a candidate for Doctor of Philosophy in Computational Science and Statistics and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

 Genechis Djira, Ph.D.
Dissertation Advisor

Date

Kurt Cogswell, Ph.D.
Head, Mathematics & Statistics

Date

 Dean Graduate School

Date

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Gemechis D. Djira for accepting me as his advisee in January, 2017. Without his instruction, help, and support, I cannot imagine where my Ph.D. study would have ended up. He has helped me to successfully bring my doctoral study to completion. I would like to thank him for his patience, dedication, and guidance. In particular, he helped me to develop the methodology and analysis chapters. I also worked with Dr. Yunpeng Pan for three years and I appreciate his guidance and support.

I also gratefully acknowledge Dr. Kurt Cogswell, Head of the Mathematics and Statistics Department, and Dr. Donald Vestal, graduate program coordinator in the Department of Mathematics and Statistics, for providing me with a graduate teaching assistantship. They also provided me with all the necessary resources to conduct my research at South Dakota State University.

I am also grateful to my Ph.D. advisory committee members Dr. Semhar Michael, and Dr. Gary Hatfield, Department of Mathematics and Statistics, and my graduate school representative, Dr. Jane Christopher-Hennings, from the Department of Veterinary and Biomedical Sciences Department. They provided me with constructive comments, which greatly improved the quality of my dissertation. I would also like to thank Dr. Yeong C. Kim and Dr. San Ming Wang from the University of Nebraska Medical Center for providing me with the data used in this research. My thanks also goes to Dr. Xijin Ge for helping me get the data in this research. I would also thank Dr. Emhimad Abdalla from the University of Wisconsin for his constructive comments and his help with the Fortran programs.

My special thanks goes to the Ministry of Higher Education, Libya, for sponsoring my Ph.D. study abroad through the Libyan North American Scholarship Program (LNASP). I am also indebted to Dr. Mohamed. M. Mekaeil for his superb mentoring.

Last but not least, my sincere thanks to my family, especially to my mother, father, wife, brothers, and sons, for their love, sacrifices, and encouragement that helped me to complete my research. Finally, I must say that in the journey of my life, I am indebted to so many family members, friends, and well-wishers who have provided invaluable advice during uncertain and difficult times, and have helped me keep my dream alive. I wish I could thank every person but, nevertheless, they are always in my heart.

TABLE OF CONTENTS

LIST OF FIGURES	viii
LIST OF TABLES	ix
ABBREVIATIONS	x
ABSTRACT.....	xi
CHAPTER 1	1
GENERAL INTRODUCTION.....	1
1.1 Introduction.....	1
1.2 Data.....	4
1.3 The Objects of This Study	4
CHAPTER 2	6
CANCER AND GENOME-WIDE ASSOCIATION.....	6
2.1 Introduction.....	6
2.2 Genes Related to Breast Cancer.....	7
2.3 Genome-Wide Association Analysis for Breast Cancer.....	9
CHAPTER 3	12
DATA PREPARATION.....	12
3.1 Introduction.....	12
3.2 The Structure and Function of DNA.....	12
3.3 What is a Gene.....	14
3.4 Single Nucleotide Polymorphisms.....	15
3.5 Data.....	16
3.6 Paired-End Reads.....	17
3.7 Pipelines.....	19

3.7.1 Mapping with Bowtie 2	20
3.7.2 Getting Rid of Duplicates using Picard	22
3.7.3 Recalibration and Interrelation	22
3.7.4 SNPs Calling via SAMtools and BCFtools	23
3.7.5 Converting SNPs to Integers	24
3.8 Quality Control for Genomic Data	25
CHAPTER 4	27
METHODOLOGY	27
4.1 Introduction.....	27
4.2 Literature Review on GWAS.....	27
4.3 Linear Mixed Effects Models	29
4.4 Maximum Likelihood Estimation of Parameters.....	31
4.4.1 Case I: Ω and R are known.....	31
4.4.2 Case II: Ω and R are unknown.....	34
4.5 The Numerator Relationship Matrix.....	38
4.6 The Heritability Coefficient.....	41
4.7 Genomic Relationships Matrix	41
4.8 Threshold model	45
4.9 Gibbs Sampling Method	50
4.9.1 Prior and Posterior Distributions	51
4.10 Predicting SNP effects	57
CHAPTER 5	58
MUTITRAITS MODELS.....	58
5.1 Introduction.....	58
5.2 Multiple Trait Model	58

5.3 Equal Design Matrices	59
5.4 Unequal Design Matrices.....	62
CHAPTER 6	65
ANALYSIS.....	65
6.1 Introduction.....	65
6.2 The MCMCglmm R Package	65
6.3 BLUPF90 in Fortran 90/95	72
6.3.1 The THRGIBBS1F90.....	74
6.4 Plots of Estimators	74
6.5 History and Autocorrelation Function Plots	75
6.6 Manhattan Plot.....	77
CHAPTER 7	79
DISCUSSION AND CONCLUSION	79
Future works:	82
APPENDIX A.....	92
APPENDIX B	96
APPENDIX C	107
APPENDIX D.....	115
APPENDIX E	116

LIST OF FIGURES

Figure 1.1 Flowchart of the data and model matrices.....	5
Figure 3.1 The two strands of DNA.....	13
Figure 3.2 A part of DNA molecule coding.....	13
Figure 3.3 Single nucleotide polymorphism in six persons.....	15
Figure 3.4 Paired-end sequencing.....	17
Figure 3.5 The pedigree information of family 1.	18
Figure 3.6 The pedigree information of family 2.	18
Figure 3.7 The pedigree information of family 3.	19
Figure 3.8 Flowchart of the data preparing steps.....	20
Figure 4.1 Pedigree of seven individuals.....	39
Figure 6.1 Trace of the fixed effects using inverse-Gamma (0.001, 0.001).	68
Figure 6.2 Trace of the variance components inverse-Gamma (0.001, 0.001).....	69
Figure 6.3 Trace of the fixed effects using chi-square.....	70
Figure 6.4 Trace of the variance component using chi-square.	71
Figure 6.5 Histogram with density of the variances and heritability estimates.	75
Figure 6.6 Sample series plots of the variances and heritability estimates.....	76
Figure 6.7 Autocorrelation functions of the variances and heritability estimates.	76
Figure 7.1 Manhattan Plot of SNPs effects.....	77

LIST OF TABLES

Table 3.1 Variant Call Format.	24
Table 3.2 SNPs coded data.	25
Table 4.1 Pedigree for a family of seven individuals.	39
Table 5.1 Pre-weaning gain and post-weaning gain (kg) for five beef calves.....	61
Table 5.2 Solutions of multiple traits model.....	62
Table 6.1 The autocorrelation of samples from <i>inverse – Gamma</i>	70
Table 6.2 The autocorrelation of samples from Chi-square.	71
Table 6.3 Convergence test of samples generated from inverse-Gamma.....	72
Table 6.4 Convergence test of sample generated from Chi-square.	72
Table 6.5 Variance components and heritability estimates.	74
Table 6.6 Ordered SNP effects with locations.....	78

ABBREVIATIONS

BAM	A compressed binary version file of the Sequence Alignment/Map
BRCAx	familial breast cancer without known mutation in BRCA1 and BRCA 2 genes
ER	Estrogen Receptor
HBOC	Hereditary Breast and Ovarian Cancer
HWE	Hardy-Weinberg Equilibrium
GWAS	Genome Wide Association
GL	Likelihood of each possible call
GP	Probability of each possible genotype
GT	Genotype
MAF	Minor Allele Frequency
PCR	Polymerase Chain Reaction
LME	Linear Mixed Effects Model
SAM	A Sequence Alignment/Map format
SIMD	Single Instruction Multiple Data

ABSTRACT

THRESHOLD MODELS FOR GENOME-WIDE ASSOCIATION MAPPING OF
FAMILIAL BREAST CANCER INCIDENCE IN HUMANS

NASIR ELMESMARI

2017

Breast cancer is the second most fatal cancer in the world and one of the most highly harmful cancers from which people suffer. Breast cancer studies have been able to uncover some knowledge about genetic susceptibility for familial breast cancer in humans. Hence, determining genetic factors may potentially help track the disease, as well as discover the cancer in early stages, or perhaps before it starts. In addition, this may allow early determination of possible treatment strategies which will make it easier to prevent the disease. In this context, it is important to determine whether the heritability of breast cancer incidence is greater than zero, which can be investigated if there is a potential genetic component playing a role in the incidence of the disease. Traits with zero heritability are said to be completely subject to environmental factors, so genetics has no effect at all. Heritability is important because it indicates the extent of genetic variations which could provide a reason for the infection. In the case that heritability is found to be greater than zero, it is useful to estimate the single nucleotide polymorphism (SNP) effects, which may potentially determine the genes or the genomic regions that are associated with the incidence of breast cancer.

This study used data for three families with BRCAx as exome sequences provided by the University of Nebraska Medical Center and the Institutional Review Boards of

Creighton University. Specifically, the data consisted of pedigree information for 167 individuals from three families, including information on whether each person had breast cancer or not (binary trait, positive or negative). Genomic data was available for 22 individuals among the 167. Theoretically, heritability as well as SNP effects can be estimated using a variety of approaches, but given the data available for this study, the best strategy was to combine both the pedigree-based data and the genomic data in one matrix. This matrix offers an advantage over other approaches that use only one of these datasets. The data was analyzed using a threshold model and Gibbs sampling algorithm to estimate the heritability of breast cancer incidence, as well as to predict SNP effects. The binary response variable for breast cancer incidence was modeled such that gender (2 levels) and family (3 levels) were the fixed effects. The effect of the subjects was the only random effect in the model.

The heritability estimate was approximately 28%, indicating that there is a considerable genetic component underlying the incidence of breast cancer. In addition, the Genome-Wide Association Study (GWAS) analysis revealed that breast cancer is a complex trait, possibly controlled by many genes. However, some areas on the genome (specifically, chromosomes 1, 2, 4, 8, 14 and 16) may include candidate genes associated with breast cancer incidence. These genes might be responsible for this type of cancer and play important roles in susceptibility for the disease. The 20 SNPs with highest effects explained more than 3.5 % of the genetic variance, which is a good indicator that their genes are associated with breast cancer. The results of this study open the door for more research on breast cancer incidence. Despite the limitations related to the small sample used, the results of this study could be considered a first step for future work and

investigation. Further studies using larger data sets may reveal more information on this complex trait.

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

According to the World Health Organization (WHO), the number of patients who die from cancer has raised up from 8.2 million in 2012 to 8.8 million in 2015. This high number of deaths has been attributed to late diagnosis and lack of proper treatment. The delay of diagnosis occurs even in countries that have excellent health systems [1]. In the United States, the 2016 annual report noted 1,685,210 new cases of cancer diagnosed, as well as 595,690 cancer deaths [2].

Developing a cure for various cancers has been a worldwide initiative in the past two decades because it is considered to be one of the leading causes of death [3]. For instance, breast cancer is ranked the second deadliest cancer in the world. In order to appropriately determine cures or treatments for any disease, it is imperative that the cause is uncovered. Understanding a disease at the point of origin will also enable earlier diagnosis. Additionally, various types of cancer are more easily eradicated in the early stages. Some studies have indicated that heritability is a major contributing factor for the disease. Measuring heritability is, therefore, an important step in revealing the extent of the variation in response attributed to biological effect. Many breast cancer studies have been able to uncover the significance of genetic influences [4].

The history of heritability goes back to at least the 19th century. However, the ideas surrounding heritability estimates were developed by Wright (1920) [5]. He used the concept of heritability in his study on the coat color of guinea pigs [6]. This term is

significant because scientists use it to assess the interaction of genes and environment to determine the survivor species under the law of natural selection. Moreover, it is vital to estimating the potential for diseases in plants, animals and humans [7]. In a study conducted by Taylor (1975), he found a strong relationship between cancer and a predisposition due to exposure to ion radiation. Many patients showed extensive cell damage which led to breast cancer. In addition, the age of the patients did not have any role in causing cancer [8].

Roberts et al (1999) used a mixed model based on lognormal distributions to estimate the heritability of breast cancer [9]. Two studies of twins were used to evaluate this hypothesis, showing that the density of tissues in mammography at a given age had high heritability [10]. Moreover, other researchers have stated that the variance and covariance component models are useful in evaluating the heritability of breast density measures, serum sex-hormone levels, and volumetric mammographic density [11, 12].

The most common method used to estimate heritability is the linear mixed effect model (LME). Speed et al (2012) used an innovative approach GWAS to estimate heritability, discovering that kinship coefficients can be computed from genome-wide SNP genotypes, rather than from a known pedigree [13]. Cheng et al (2014) proposed a new inferential model that was able to evaluate the heritability coefficient [14].

Heckerman et al (2016) used a linear mixed effect model for the genomic random effect by using the kinship matrix (identity-by-descent estimates) from accurately phased genome-wide data [15]. Fong et al (2010) presented a GWAS Analyzer; this helps to limit a huge amount of phenotypic and genotypic data to predict the candidate gene that causes the severity of a disease by using SNPs [16].

In animal science, Misztal et al (2009) proposed a methodology to incorporate genomic information in addition to pedigree and phenotypic information, in one matrix [17]. A year later, Aguilar et al (2010) used this matrix to predict breeding value [18]. El-Dien et al (2016) presented a study in plant science as a first attempt to replace pedigree information (numerator matrix) with the genomic relationship matrix (G-matrix). They were then able to calculate more realistic variance components and heritability estimates for forest trees [19]. The use of animal and plant information may help us better understand genetic structure in humans as well.

Based on previous GWAS studies, approximately 100 common breast cancer susceptibility alleles have been identified [20]. Genetic studies on breast cancer have been conducted by a group from the University of Nebraska Medical Center. The authors Wen et al (2014) observed specific and novel variances in each family related to familial breast cancer. Moreover, they strongly recommended that adding and analyzing phenotypic data might dramatically enhance genetic prediction and the accuracy of disease diagnosis [21].

By following the recommendation stated in Wen et al (2014), we have used the genomic and phenotypic combination data to predict the SNP effect. To assure that this procedure is a novel study, we searched and compared all the studies that were cited in Aguilar et al (2010) by using Google Scholar and the Web of Science. As far as we know, this work is the first study focusing on human breast cancer that has used this innovative methodology.

1.2 Data

The data in this study was gathered from three families with BRCAx as exome sequences, as examined in a study by Wen et al (2014). This data has also been described in another study from the University of Nebraska Medical Center by Lynch et al (2013) [22]. The use of this data in both studies was approved by the University of Nebraska Medical Center and the Institutional Review Boards of Creighton University. All members of these families had signed a consent form to participate in cancer research, with the understanding that personal information would be kept confidential, as required by institutional privacy policies. A detailed discussion of the data can be found in the data preparation section of Chapter 3.

1.3 The Objects of This Study

The goal of this study was to estimate the genetic parameters of this disease by combining both the phenotypic and genomic information available to us. In addition, this study aims to find genomic regions and specific genes with major effects associated with familial breast cancer incidences. The general objectives this study were:

1. Prepare appropriate data for GWAS. This includes mapping to the reference genome, removing duplicate reads, calling variants for each individual, and merging them with the reference genome.
2. To estimate variance components and heritability for familial breast cancer incidence.
3. To find genomic regions and specific genes with major effects associated with familial breast cancer incidences using both genomic and phenotypic data.

To achieve these goals, the dissertation is organized as follows:

In Chapter 2, we present GWAS. Chapter 3 deals with the data preparation and also defines standard terminologies in genomic studies, such as the structure of DNA, gene, SNP, data and families. Additionally, the data preparation steps include mapping, removing duplicates, SNPs calling, converting SNPs to integers and quality control for genomic data. Chapter 4 describes the methods used, including linear mixed effects models, maximum likelihood estimation, numerator relationship matrix, the heritability coefficient, genomic relationship matrix, threshold model and the Gibbs sampling method. In Chapter 5, multi-trait models will be presented, while Chapter 6 explains the analysis using the software FORTRAN90/95 for variance components heritability and for prediction of SNP effects and the R software for graphics. The last Chapter is the discussion and conclusion. Figure 1.1 shows the version datasets and matrices used in the data analysis.

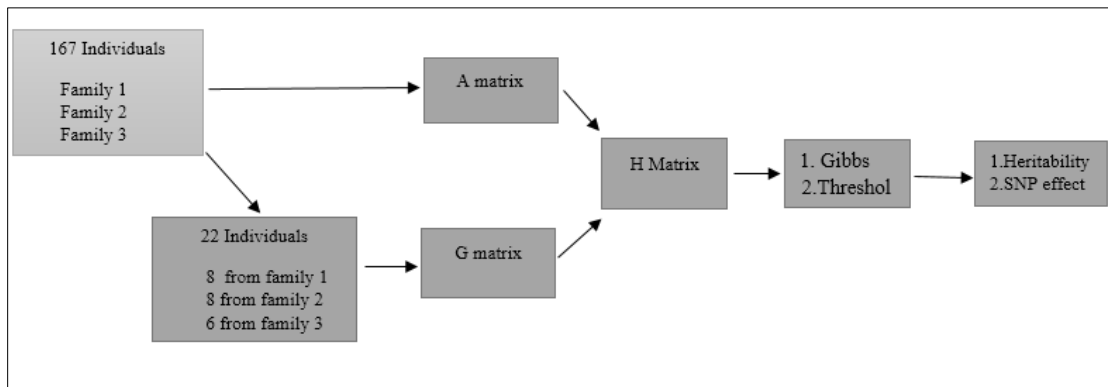


Figure 1.1 Flowchart of the data and model matrices.

CHAPTER 2

CANCER AND GENOME-WIDE ASSOCIATION

2.1 Introduction

Cancer is the general name used to refer to a variety of similar diseases, all of which involve some cells in the body that are continuously dividing and taking over the tissues surrounding the affected area. Cancer can start just about anywhere in the human body. In a healthy body, human cells divide to produce replacement cells when needed. These new cells replace old damaged ones when they stop working. The development of cancer in the body, however, breaks down this otherwise methodical progression. As abnormality takes over, the cells that should die and be replaced actually endure, and replacement cells continue to form even when they are not required. The extra cells multiply endlessly and become tumors.

Most common types cancer produce solid tumors or masses of tissue. However, blood cancers, such as leukemia, do not typically form tumors. When a tumor is cancerous, it is referred to as “malignant”, which means it can attack nearby tissues. As these tumors grow, some cancer cells can even break away from the original mass and migrate to other areas in the body through the bloodstream or the lymph system. They then start new tumors in locations other than the original site. Benign tumors differ from malignant tumors. Though they can also be quite large, they do not take over or attack nearby tissues. Benign tumors can be successfully removed and typically there is a big chance that will not come back again. However, malignant cancers can be removed but can continue to be produced in the body after removal. Benign brain tumors can be life threatening but most other benign tumors are not [23].

Remarkable advancements in cancer study into the past 50 years have provided insights into the development of cancer cells. Cancer is now defined as a disease in which there are changes or mutations in the cell genome. These changes, or Deoxyribonucleic acid (DNA) mutations, produce a protein that upsets the protocol between the cell division stage and the dormancy stage, allowing continuous cell division and the formation of cancer. At the point when cancer cells relocate to different parts of the body where they form new tumors and crowd out normal cells, it is called "metastasis" [25].

2.2 Genes Related to Breast Cancer

Breast cancer is a main cancer in women. It affects nearly a quarter of a million women each year in the United States alone. About 10–15% of breast cancers have genetic ties, affecting multiple family members over generations. Recognizing that a genetic predisposition increases susceptibility is a big step in unraveling the cause of breast cancer for early detection, diagnosis, prognosis, and treatment [26].

Broca was the person to recognize a family with a high prevalence of breast cancer in 1866. His wife was inflicted with early onset breast cancer. When Broca trailed her family tree, he found four generations with a history of breast [24]. The "Broca" report was of many that tied breast cancer to genetic predisposition, passing from one generation to the next. He also discovered 16 cases of cancer across five generations in one family. Fifteen were female, nine of which had breast cancers [27].

In 1972, Lynch et al. linked breast cancer with a predisposition to colon cancer. His findings showed that the members of some families with breast cancer also had a comparatively high predilection for cancer of the colon; some showed higher predispositions for gastric, ovarian, and endometrial carcinoma, and some demonstrated

higher odds for brain tumors, sarcoma, and leukemia. Hall et al (1990) started mapping the genes accountable for hereditary breast cancer, which allowed for the identification of early lesions that are a signature of the development of breast cancer. He theorized that Chromosome 17q21 appeared to be the position of the gene indicating a predisposition to breast cancer in families with a history of early-onset diseases. Shortly thereafter, in 1991, Lenoir et al. demonstrated an association to this same gene position with HBOC syndrome. The gene has since become known as "BRCA1". In 1994, Miki and Swensen noticed a strong potential for the BRCA1 gene, as it affects the predisposition to breast and ovarian cancer. It has also been known for its positional cloning method, which is a method to identify genes [29, 32].

A study by Wooster et al (1994) provided evidence of another locus to breast cancer susceptibility, BRCA2, to a 6-centimorgan interval on chromosome 13q12-13. Initial findings suggest that BRCA2 assumes a higher risk of breast cancer but, unlike BRCA1, does not impose any elevated risk of ovarian cancer. Mutations in BRCA 1 and BRCA 2 are found in most of the families having six or more cases of breast cancer, which is aligned with dominant inheritance [34]. However, Shih et al (2002) and Easton et al (1999) said that, overall, the identified susceptibility genes are estimated to be accountable for less than 25% of familial breast cancer, demonstrating a strong possibility that other susceptibility genes are yet to be discovered [36]. Antoniou et al (2003) said genetic mutations in BRCA1 and BRCA2 genes can increase the risk of breast cancer (60-85%) and ovarian cancer (15-40%) over a lifetime. Researchers have attempted to discover an assumed BRCAx gene using linkage analysis to support the genetic context for high-risk families. However, the results seem to show that many

genes are probably contributors to a predisposition to breast cancer [22, 37]. Others hypothesized that the genetic predispositions in many BRCAx family breast cancer cases may be specific variations [21]. Correspondingly, several studies showed that the risk of breast cancer could be increased with mutations in the following genes: P53, PTEN, CHEK2, ATM, PALB2, FGFR2 and TNRC9; other genes are still being vetted [21].

Current evidence suggest that three types of genetic predisposition exist for familial breast cancer: (a) high-risk genes with rare mutations, but high penetrance causing high risk of breast cancer, e.g., BRCA1 and BRCA2; (b) intermediate-risk genes with rare mutations causing intermediate risk of breast cancer, e.g., CHEK2, ATM, BRIP1, and PALB2; and (c) common modest risk genetic variants, such as the SNPs in or close to FGFR2, TNRC9, MAP3K1, and LSP1[22].

2.3 Genome-Wide Association Analysis for Breast Cancer

Identifying and understanding genetic risk factors for complex diseases is the main goal of human genetics. GWAS is one of several beneficial technologies used to study designs and examine the results of analytical tools in order to identify genetic risk factors. GWAS examines the whole genome in different individuals to see if there are any genetic regions related to a specific trait. Determining a genetic factor will help in tracking the disease and in discovering cancer in early stages. In addition, this can lead the way to additional possible treatments. This information can be used in fine mapping genes [38].

The first scientific report using genome-wide screening was in 2005 [39]. The researchers examined patients with age-related macular degeneration and found two SNPs with significantly different allele frequencies compared to healthy control subjects.

Many similar studies that have been reported successfully using the genome-wide association analysis to determine genetic variations that might contribute to a risk of type 2 diabetes, Parkinson's disease, heart disorders, obesity, Crohn's disease and prostate cancer, as well as genetic variations that influence responses to antidepressant medications [39, 40].

In recent years, several genetic SNPs related to breast cancer risk have also been identified via GWAS [22]. Haiman et al (2011) combined GWAS data from women of African ancestry (1,004 ER-negative cases from 2,745 controls) and European ancestry (1,718 ER-negative cases out of 3,670 controls), with replication testing conducted in an additional 2,292 ER-negative cases and 16,901 controls of European ancestry. A common risk variant for ER-negative breast cancer at the TERT-CLPTM1L locus on chromosome 5p15 (rs10069690) was discovered. The variant was also implicitly linked to triple-negative (ER-negative, progesterone receptor (PR)-negative and human epidermal growth factor-2 (HER2)-negative) breast cancer, particularly in younger women (<50 years of age) [41].

Another study in GWAS cited single-nucleotide polymorphisms at 1p11.2 and 14q24.1 as loci for breast cancer susceptibility. The early GWAS leaned in the direction of strong effects for both loci for ER-positive tumors. Using data from the Breast Cancer Association Consortium (BCAC), Figueroa et al (2011) sought to determine whether risks differ by ER, progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), grade, node status, tumor size, and ductal or lobular morphology. The data was derived from 46,036 invasive breast cancer cases and 46,930 controls from 39 studies. Analyses by tumor characteristics focused on subjects identifying as white women of

European ancestry and were based on 25,458 cases, of which 87% had ER data. The SNP at 1p11.2 revealed pointedly stronger associations with ER-positive tumors [42].

In the Triple Negative Breast Cancer Consortium (TNBCC) Stevens et al (2011) explored 22 common breast cancer susceptibility SNPs in 2,980 Caucasian women with triple negative breast cancer and 4,978 healthy controls. Six SNPs significantly related with risk of triple negative breast cancer were also identified, including: rs2046210 (ESR1), rs12662670 (ESR1), rs3803662 (TOX3), rs999737 (RAD51L1), rs8170 (19p13.11) and rs8100241 (19p13.11) [43].

CHAPTER 3

DATA PREPARATION

3.1 Introduction

In GWAS studies, researchers use complicated datasets, with the data needed for the final statistical modeling usually prepared in multiple steps. In this chapter, we tried to describe the various data preparation steps we used in our study. These steps include mapping, remove duplicates, SNPs calling, converting SNPs to integers and quality control for genomic data. We also give a short overview to the standard terminology used in genomic studies.

3.2 The Structure and Function of DNA

In the 1940s, biologists struggled to understand how DNA could be the key to our genetic make-up, due to the assumed simplicity of its configuration. There was awareness that DNA was comprised of four similar types of subunits laced together on a long polymer, and that each of the four subunits resembled one another chemically. In the early 1950s, DNA was studied by way of a procedure known as x-ray diffraction analysis, that distinguished the three-dimensional atomic structure of molecules. The preliminary findings of the x-ray diffraction exposed that the configuration of DNA was formed by two polymer strands spiraled into a helix (Figure 3.1 [147]). The detection of the two-stranded structure of DNA was revolutionary, becoming one of the most prominent pieces of evidence that led to the Watson-Crick Model for DNA structure. It was only once this model was proposed in 1953 that DNA's capacity for replication and material encoding become evident [44].

DNA is the hereditary material in humans and almost all other living organisms; in fact, the same DNA is present in almost every cell in a human body. DNA is most generally located in the cell nucleus. This type of DNA is called nuclear DNA. Small amounts of DNA are also found in mitochondria; this type of DNA called mtDNA.

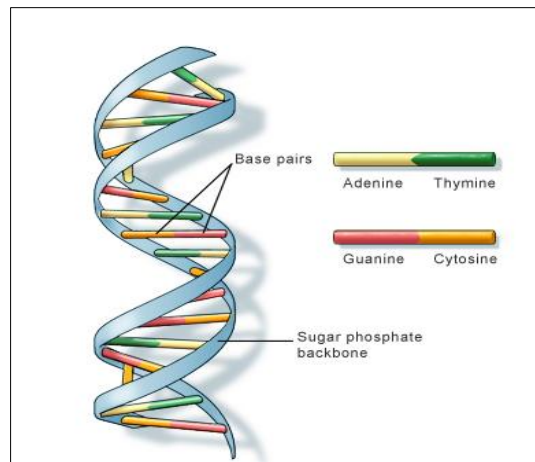


Figure 3.1 The two strands of DNA.

The information in DNA is stored as a code made up of four chemical bases: The four nucleobases that comprise the chemistry of DNA are Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). More than 99 percent of the roughly 3 billion bases of which human DNA is comprised are identical in all people. The bases line up in different arrangements in order to regulate the information needed to create and sustain an organism, much like individual letters of the alphabet can combine in different orders to form a variety of words and phrases [45].

Base pairs occur when two DNA bases attach to each other, such as A joins with T and C joins with G. Additionally, every base pair adds a sugar and a phosphate molecule. So the nucleotide is a combination of sugar and phosphate. Nucleotides join together to create the famous double helix building of DNA. The double helix is formed

in a ladder-like spiral, with the base pairs connecting horizontally, much like ladder rungs, while the sugar and phosphate molecules create long vertical strands. Possibly the most important characteristic of DNA is its ability to produce exact duplicates of itself.

The critical process of cell division relies on the exact replication of the DNA to create a new cell so each new cell is a perfect copy of the original. Each double-stranded DNA molecule has the ability to reproduce its base pairs in sequence [44].

3.3 What is a Gene

DNA molecules (comprised of base pairs) make up the standard material and practical element of heredity, known as a gene. Genes, in turn, contain the information necessary to create molecules called proteins, which are essential for maintaining the body's muscles, tissues, and organs (Figure 3.2 [148]). Human genes, for example, show a great variety in size, ranging from just a few hundred base pairs to more than 2 million. Research from the Human Genome Project indicates humans may have up to 25,000 genes. Over 99 percent of the genes are identical in every human. Each parent passes on a complete set of its genes; as a result, each person has two copies of every gene. In fewer than 1 percent of genes, the gene pairs have different patterns in their DNA bases; those gene pairs are known as alleles. Alleles are responsible for each person's individual corporeal differences [44].

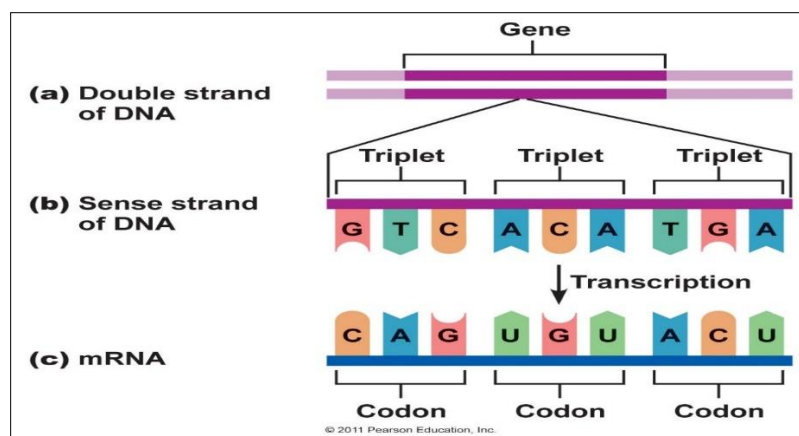


Figure 3.2 A part of a DNA molecule coding.

3.4 Single Nucleotide Polymorphisms

As discussed above, DNA patterns are made up of chains of nucleotide bases (A, C, G and T). A difference at a specific location in a DNA pattern of an individual is known as a single nucleotide polymorphism, or SNP (Figure 3.3 [149]). This deviation can be classified as an SNP only if it is present in less than 1 percent of a given population. A gene can be considered to have two alleles if an SNP occurs in that gene. An SNP inside a gene can alter the pattern of the amino acids. SNPs do not only occur within genes, but can also be found in DNA molecules that do not carry instructions for producing proteins, which named noncoding region.

	SNP 1	SNP 2	SNP 3
Person 1:	acggtta g ctacaattattttaa a cgggaggagggattttattaacc g atgtg		
Person 2:	acggtta t ctacaattattttaa a cgggaggagggattttattaacca a atgtg		
Person 3:	acggtta a ctacaattattttaa t gggaggagggattttattaacc g atgtg		
Person 4:	acggtta a ctacaattattttaa t gggaggagggattttattaacca a atgtg		
Person 5:	acggtta t ctacaattattttaa t gggaggagggattttattaacca a atgtg		
Person 6:	acggtta t ctacaattattttaa t gggaggagggattttattaacca a atgtg		

Figure 3.3 Single nucleotide polymorphism in six persons.

Although a particular SNP may not create a disorder, some SNPs are related to the occurrence of certain diseases. These associations enable scientists to search for SNPs in order to assess an individual's genetic predisposition to developing a disease. Additionally, if certain SNPs are understood to be correlated with a trait, then scientists may inspect stretches of DNA near these SNPs in an effort to detect the gene or genes responsible for that trait. Researchers are hopeful that awareness of an individual's SNP genotype will afford a basis for evaluating susceptibility to diseases and the ideal choice of therapies [46]. A key challenge in understanding these potentials is comprehending how and when the variants may cause a disease.

3.5 Data

This study used exome sequencing data. The data was previously analyzed in the study by Wen et al (2014) to test genetic predispositions; in their study, only genomic data was used. They matched all variants that were identified in the databases, while all known variants were removed. The study includes data from three families with BRCAx familial breast cancer (refer to Figure 3.5-3.7 and Appendix A); the families included seventeen members with cancer, and five members without. In the first family, seven members had cancer and one member was healthy, while in the second family, five members had cancer and three did not. The third family included five members who had cancer and one who remained unaffected. Data in this work were collected as blood samples from family that have members affected by cancer, as well as those who were not affected. The thorough genetic testing of individuals in each family displayed no mutation in BRCA1 or BRCA2. Family members over two generations were chosen for exome sequencing based on the pedigree's hereditary pattern of breast cancer, as well as

the accessibility of DNA samples. Exome sequences were collected with a HiSeq™ 2000 sequencer (Illumina, San Diego, CA) with a paired-end (2×100).

3.6 Paired-End Reads

Paired-end reads permit users to sequence both ends of a fragment, which generates superior sequence data that can be easily aligned (Figure 3.4 [150]). In addition to gene fusions and unusual transcripts, paired-end sequencing enables detection of genomic changes and elements of a repetitive sequence.

Subsequently, paired-end reads are more able to line up to a reference, and the value of the total data set is much improved. All Illumina next-generation sequencing (NGS) structures are capable of paired-end sequencing.

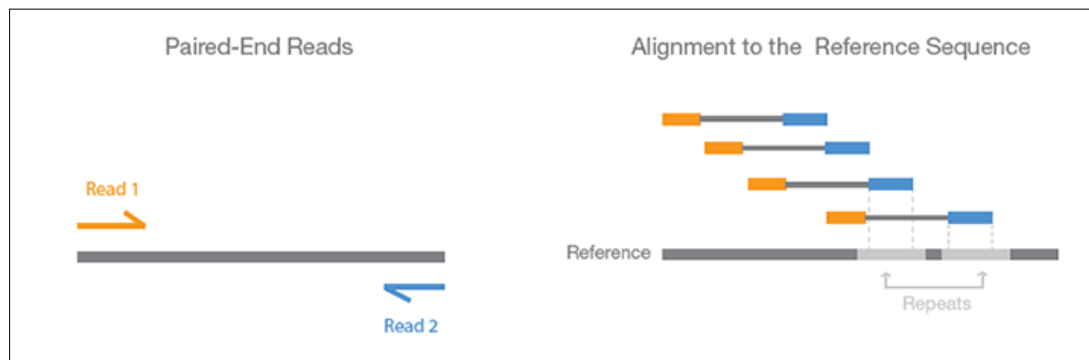


Figure 3.4 Paired-end sequencing.

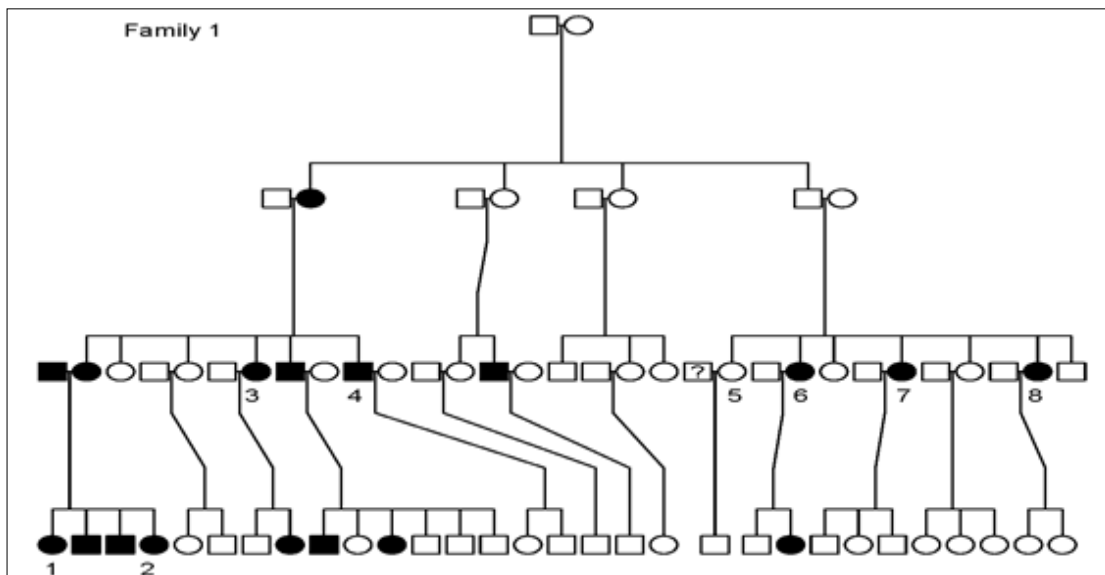


Figure 3.5 The pedigree information of family 1.

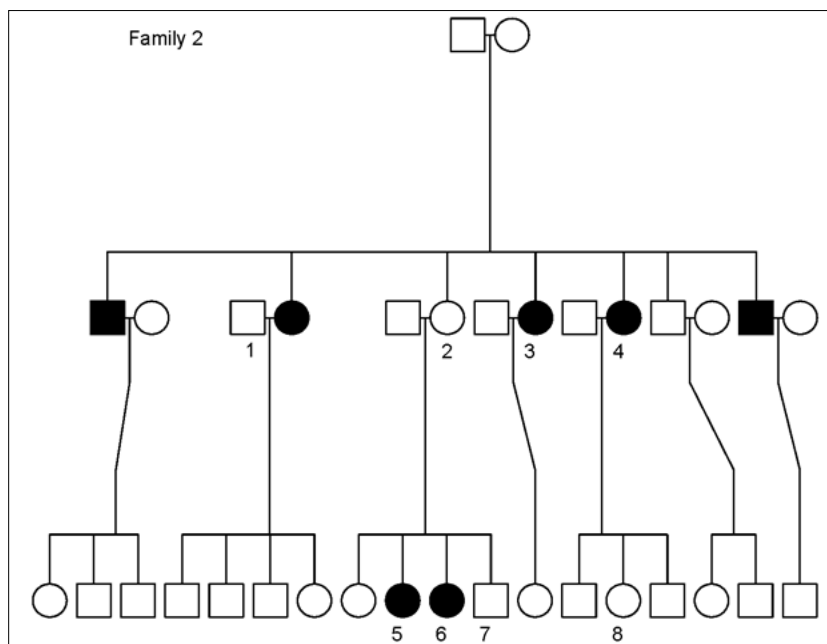


Figure 3.6 The pedigree information of family 2.

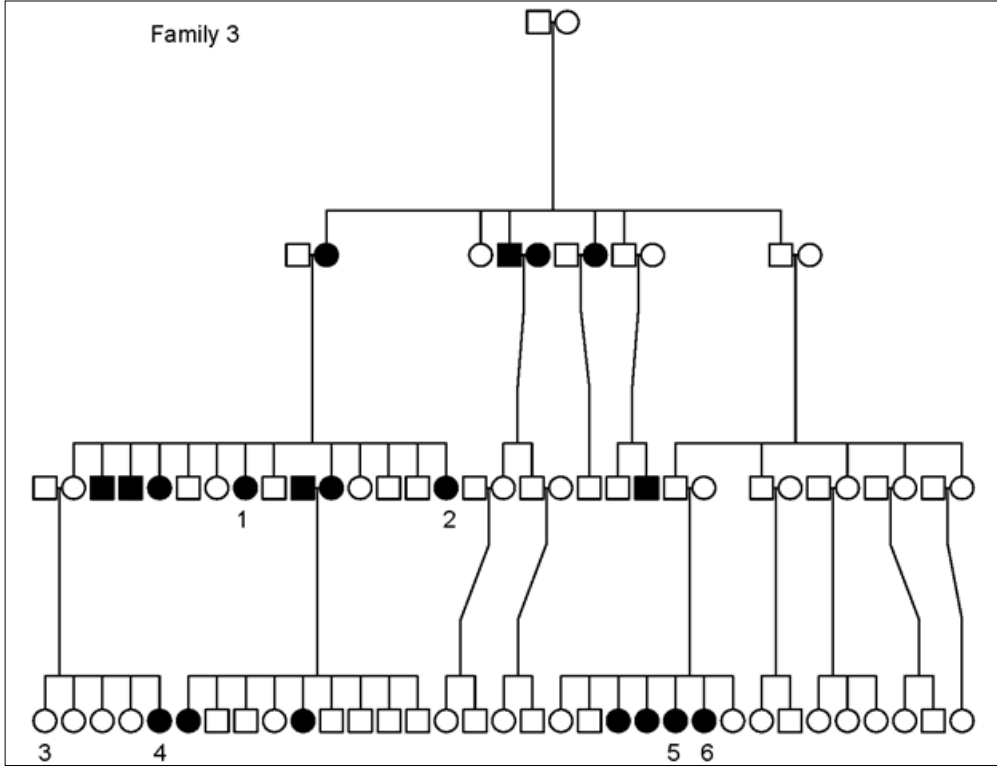


Figure 3.7 The pedigree information of family 3.

3.7 Pipelines

Pipelines are series of computational or data manipulation steps that result with specific data called “variants”, which are analyzable in the next-generation sequencing studies, all these steps are shown in (Figure 3.8). Before we go to detail let us define some terms; *Reads*: are sequences obtained of DNA, where each nucleotide sequences called read. Usually each read has 100 base pair; *Bowtie 2*: is a tool to align sequencing reads to reference sequences; *Picard*: is a tool to operate specific format such as SAM, bam and VCF file; *SAMtools*: is a tool that use alignments in BAM format; *BCFtools*: this tool to call variant as VCF and BCF format.

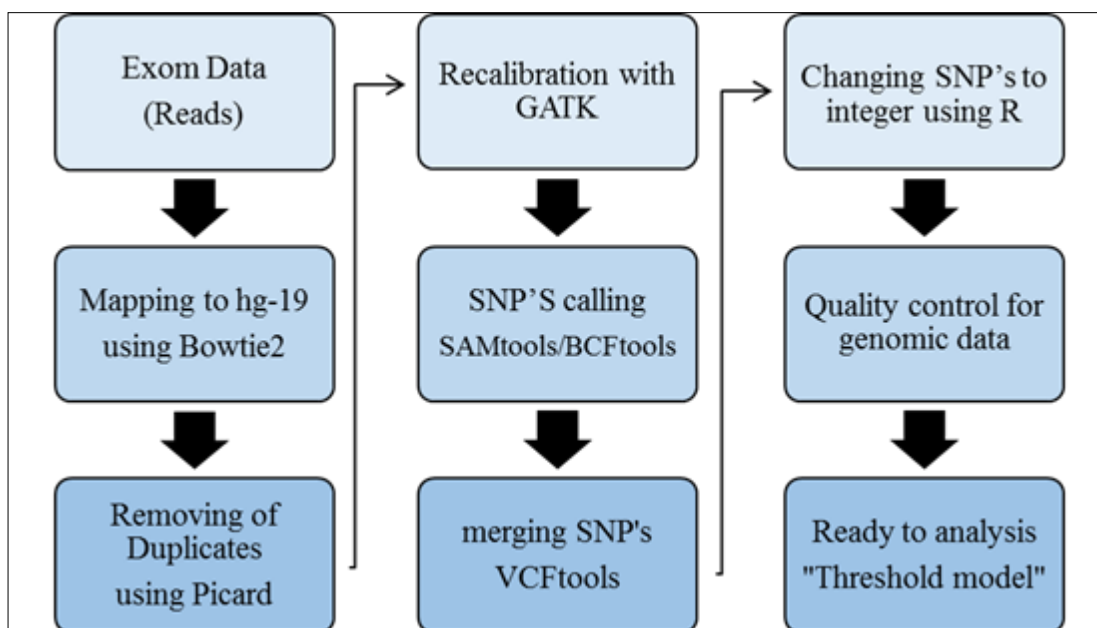


Figure 3.8 Flowchart of data preparing steps.

3.7.1 Mapping with Bowtie 2

The primary measure in several relative genomic pipelines is aligning sequencing reads to a reference genome, such as variant calling, isoform quantitation, and differential gene expression. In many instances, the alignment step is the longest, since for each read, the aligner has to resolve the difficult computational problem of ascertaining the read's most likely point of origination regarding a reference genome. Most aligners use a genome index to quickly narrow down the list of potential alignment locales. The full-text minute index is a fast and memory-efficient index that is being utilized by aligners. Index-assisted aligners function by looking for all the potential ways to mutate the read string into a string. This also happens to the reference point, which is of course, an alignment rule, and thereby limits the number of changes. Even though this search space is vast, several areas can be bypassed (pruned) without loss of precision. Pruning

approaches such as double indexing and bidirectional Burrows-Wheeler Transform (BWT) allow a rapid and thorough alignment of short reads [47].

For each read, Bowtie 2 is conducted in four steps (Supplementary Note and Supplementary). Step 1 of Bowtie 2 involves extracting ‘seed’ substrings from the read and its reverse complement. In step 2, the removed substrings are aligned to the reference without gaps, using the full-text minute index. For step 3, the seed alignments are ranked, and their locations in the reference genome are determined from the index. Finally, in step 4, the seeds are stretched into full alignments by executing SIMD-accelerated dynamic programming.

Langmead and Salzberg (2012) likened Bowtie 2 to four other full-text minute index-based read aligners. Included in their comparison was Burrows-Wheeler Aligner (BWA), BWA’s Smith-Waterman Alignment (BWA-SW) and Short Oligonucleotide Alignment Program 2 (SOAP2), and Bowtie. They attained 100-by-100 nucleotide (nt) paired-end HiSeq (2000) reads from a human resequencing study and took out an arbitrary subsection of 2 million pairs. They found that the Bowtie 2 default mode was quicker than the other BWA modes and was faster by more than 2.5 times over the BWA default mode. All of the Bowtie 2 modes aligned a larger number of reads than either BWA or SOAP2. In summary, they found in all cases, Bowtie 2 and BWA found more accurate alignments than SOAP2 and Bowtie. Bowtie 2 also presented more accurate and fewer inaccurate alignments for the unpaired reads than BWA presented over an area of charting quality limits. For paired-end reads, the disparity was reduced.

We have mapped data in this study while considering the human genome (hg 19) as a reference sequence using Bowtie 2 with default parameters. The rate of overall

alignment for each subject was around 92% see Appendix B for statistics from Bowtie 2. The output was formatted as a SAM file, then we converted it to the BAM format using SAMtools utility [47].

3.7.2 Getting Rid of Duplicates using Picard

It has been observed that some reads pile up with the same beginning and ending coordinates. These may be a product of PCR duplicates. The occurrence of these replicates injected by PCR expansion is a key problem in paired short reads from next generation sequencing. These replicates should be removed from BAM files as they may have a critical and adverse effect on research applications. Even more crucial is the fact that the precision of paired reads alignment could be compromised by genomic variations that are broadly dispersed among individuals, such as copy number variations, extensive structural variations, minor insertion or deletion (indels) variations, and SNPs. Duplicates for this study were removed using Picard Mark Duplicates, which is the favored method for this purpose [48].

3.7.3 Recalibration and Interrelation

The Genome Analysis ToolKit (GATK) is a software capable of incorporating the evidence for variants from several samples with joint genotyping. It enables the use of validated SNPs and indels to augment the accuracy of variant calling [49]. However, many research communities lack the large, validated collections of SNPs and indels needed to test using GATK's Best Practices procedures due to the investment required to obtain and curate such collections [50]. To work around the necessity for large-scale variant validation studies, McCormick et al (2015) created the Recalibration and Interrelation of genomic sequence data with the GATK workflow. This development

integrated data from multiple genomic sources and identified reliable sets of variants. A variety of factors exist that could cause erroneous results: inadequate or incorrect reference assemblies, mistaken realignment of reads to the reference genome (mainly in lower complexity regions and around indels), inexact base quality scores, and suboptimal variant filtration parameters [49, 51]. Additionally, the recalibration tool tries to improve variation quality with the machine cycle and sequence context, and by doing so, provides not only more precise quality scores but also more broadly dispersed ones [52].

We recalibrated the base quality scores of the sequencing by synthesizing reads into an aligned BAM file. Once completed, the quality scores in the QUAL field of each read in the output BAM were more precise, in that the reported quality score was closer to its actual probability of mismatching the reference genome.

3.7.4 SNPs Calling via SAMtools and BCFtools

One of the standard tasks for NGS studies is the detection of SNPs and indels in an individual sample. SAMtools enables users to call both types of variants concurrently using the *mpileup* output. In this study, variants have been called as BCF format using a specific parameter (`--skip-indels`). At each position, the SAMtools searches for variants that qualify under user-defined minimum conditions for sequence reporting. The conditions included the quantity of supporting reads. After using BCFtools to call SNP's variants with the parameters `-c -v` (call genotypes and output variant sites only), files are output in the Variant Call Format (VCF) format [53]. See Appendix C for output from SAMtools to call SNP.

3.7.5 Converting SNPs to Integers

VCF is a text file format used for the storage of marker and genotype data. The following example (Table [54]) explains how VCF encodes data for SNPs.

Table 3.1 Variant Call Format.

```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMP001 SAMP002
20 1291018 rs11449 G A . PASS . GT 0/0 0/1
20 2300608 rs84825 C T . PASS . GT:GP 0/1:. 0/1:0.03,0.97,0
20 2301308 rs84823 T G . PASS . GT:PL ./.:. 1/1:10,5,0
```

Each column included after first nine has genomic information called genotype. For sample data, GT alleles are numeric; the reference REF is 0, the first ALT is 1, and so on. As an example, the genotypes for SAMP001 (The tenth column in Table) are a homozygous reference first, followed by a heterozygous and third, a missing record.

Missing observations with levels of SNP data in some genetic analysis can offer good estimations of genetic parameters; however, little is known about the influence of a missing level that is higher than 90%. Traditionally, if the missing level is 15% or below, it is known to eliminate genetic markers with incomplete observations [55, 56]. Missing genotypes data will not include the calculations, so but we coded them 5 following VanRaden (2008) method [57].

Our data included 22 subjects (columns) with approximately 16 million SNPs (rows), so the rows with more than two missing SNPs were removed. This resulted in more than 10% of the missing levels being eliminated to obtain estimations of genetic parameters reliability. Refer to the following table (Table 3.2) which shows some of the SNP data used. This sample represents family 1, which has eight individuals x1- x8. The

numbers 0 and 2 represent recessive and dominant homozygous genes, with 1 being heterozygous and 5 denoting missing alleles. We will use this data set in matrix \mathbf{M} which is a genomic relationships matrix (\mathbf{G}). Refer to Section 4.7 for additional details.

Table 3.2 SNPs coded data.

CHROM	POS	x1	x2	x3	x4	x5	x6	x7	x8
1	808922	2	2	2	0	0	0	2	2
1	880238	0	0	0	0	0	0	0	0
1	883625	0	2	0	0	0	5	0	5
1	887560	2	0	1	1	1	1	0	1
1	887801	0	0	0	0	0	0	0	0
1	888639	2	0	0	0	2	0	0	2
1	888659	0	0	2	0	0	2	0	0
1	889158	0	0	0	0	0	0	0	0
1	889159	2	0	2	0	2	0	0	0
.
.
.
.
.
22	43520263	0	2	0	0	2	0	0	1
22	43528793	2	2	2	0	2	0	0	1

3.8 Quality Control for Genomic Data

Quality control (QC) of genotypes is vital to avoid false results in genome-wide association studies. Maintaining long-term data integrity is also of the utmost importance, especially in situations with ongoing genotyping. Some SNP data was problematic to score because of an irregular genotype clustering pattern. Procedures were performed to improve SNP data (25394 SNP's), including the following steps:

- 1) 22 markers were removed from the mitochondrial chromosome. 390 were removed from chromosome X because some noteworthy biological insights could be garnered by the exclusion of the X chromosome in the GWAS analysis.

- 2) SNPs with a low Minor Allele Frequency (MAF) were removed. As a rule, an MAF threshold of 1-2% is used, but studies with smaller sample sizes might require a higher threshold [58]. 25% was the threshold used to determine the MAF for calling SNPs.
- 3) Because the precision of genotyping is extremely reliant on the quality of the genotypes, repeated genotyping of the same samples could be misread as originating from different individuals when call rates drop under 90% [59]. Therefore, all SNPs with a call rate $< 90\%$ were ignored.
- 4) Most GWAS studies choose to eliminate markers that display significant deviation from Hardy-Weinberg Equilibrium (HWE) because it can indicate a genotyping or genotype calling error. On the other hand, variances from HWE might also show selection; a sample can display deviations from HWE at loci related to a disease; it would clearly be counter-productive to exclude these loci from a deeper inquiry. In practice, many SNPs with an HWE p-value less than 0.001 will be extracted. However, robustly genotyped SNPs, even if under this threshold, will remain under study. Upon checking, the departure of heterozygous from Hardy-Weinberg Equilibrium using default value 0.15 resulted in 17857 SNPs that were ready to use [58, 60].

CHAPTER 4

METHODOLOGY

4.1 Introduction

This chapter focuses on the methodologies that are used in the estimation of heritability coefficient and prediction of SNP's effects. Because the response variable in this study is binary (affected or normal), threshold model is employed. The idea of this technique is based on linear mixed effects models (LME) which are suitable methods in case of continuous response variables. In order to estimate the variance components, we utilize Gibbs Sampling.

4.2 Literature Review on GWAS

For modeling population relationships in Genome-Wide Association studies, LME models have been recommended [61]. LME models expand on the work initially described in the literature dealing with animal breeding. Later it was developed in the human genetics literature, in which, an interesting genetic effect (e.g., the number of copies of a specific allele at a specific test SNP) as a random effect, with an added fixed effect is incorporated to model the genetic relationship between individuals [62]. In the first few years of GWAS, linear mixed effects models were not used much due to computational issues [63]. In more recent years, a wide array of LME methods/software packages have surfaced [62].

In statistical analysis, needless to mention, it is critical to know the nature of our dependent variable and model it using appropriate methods. Tong et al (1976, 1977) and Berger and Freeman (1978) translated the categorical response into a quantitative

response by assigning m ordered numerical values or "scores" to the m categories. They then proceeded as though this discrete quantitative response followed a linear mixed effects models. Unfortunately, the assumptions implicit in many common linear mixed effects models, including those of homogeneity of variances, may be much less reasonable when the model is applied to a discrete response than when it is applied to a continuous response [67].

Linear mixed effects models have commonly been used in the study of continuous traits (e.g., see Anderson et al (2010)). They are grounded in the supposition of normality and are easily implemented using the software that is available to the public. In the case of longitudinal data we refer to Molenberghs and Verbeke (2000) and Fitzmaurice et al (2011). However, results from a linear mixed effects models may be unreliable under certain circumstances, such as, if the assumed Gaussian distribution of the response variable is not met, for instance, due to the occurrence of outliers or skewness. Alternatives include data transformation or a more adaptable modeling tactic [71, 72]. Ordered categorical traits are usually studied applying the threshold liability model, which was first used by Wright (1934) in the analysis of the quantity of digits in guinea pigs (normally have four front feet). It was also used by Bliss (1935) in toxicology trials. In the threshold model, it is hypothesized that there is an underlying or latent variable (liability) that has a continuous distribution [73]. A response in a specified category is noted if the value of liability lies between the thresholds, determining the suitable category [74-76].

When variability originates from two sources, linear mixed effects models and threshold model with two variance components are commonly used. In genetic studies,

variation in observations can be credited to biological and environmental influences. The heritability coefficient is an essential quantity that gauges the percentage of overall variability owing to biological influences [14]. Recently, Bayesian methods have been developed for variance components estimation [77-81]. As Broemeling (1985) had observed, each of these reviews discovered analytically intractable joint posterior distributions of variance components. Additional marginalization regarding dispersion parameters appears to be problematic or impossible by means of analytic. New developments in computing have fostered the use of numerical methods in Bayesian inference. For instance, following studies by Hammersley and Handscomb (1964), Kloek and Van Dijk (1978), and Rubinstein and Kroese (1981) have been used these numerical methods in econometric and Bauwens (1984,1988) in binary responses models [90].

Markov Chain Monte Carlo (MCMC) methods have granted computation of multidimensional integrals so analytic approximations can thus be avoided. Usage of the Gibbs sampler to analyze ordered categorical traits has been described by Zeger and Karim (1991) and Albert and Chib (1993). It was used by McCulloch et al (1994) and Sorensen (1995) to estimate variance components for binary data [75].

4.3 Linear Mixed Effects Models

Linear mixed effects models are beneficial in a variety of physical, biological, and social scientific applications with variability coming from multiple sources [94, 95].

Linear mixed effects models are extensions of standard linear models (e.g., linear regression and ANOVA). Linear mixed effects models contain fixed and random effects hence called linear mixed effects.

A linear mixed effects model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon} \quad (4.1)$$

where \mathbf{Y} is a vector of observations with dimension $n \times 1$, $\boldsymbol{\beta}$ is a vector of fixed effects with dimension $p \times 1$, \mathbf{d} is a vector of random effects with dimension $q \times 1$. \mathbf{X} and \mathbf{Z} are design matrices for the fixed and random effects with dimensions $n \times p$ and $n \times q$, respectively. The vector of random effects is distributed as normal with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Omega}$, $\mathbf{d} \sim N(\mathbf{0}, \boldsymbol{\Omega})$. And $\boldsymbol{\varepsilon}$ is a vector of normal random errors with means $\mathbf{0}$ and variance-covariance matrix \mathbf{R} , $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$. The vectors of random effects \mathbf{d} and error term $\boldsymbol{\varepsilon}$ are assumed to be independent. Therefore,

$$\text{Var} \begin{bmatrix} \mathbf{d} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

The expected value of \mathbf{Y} is

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} \quad ; \quad \text{since } E(\mathbf{d}) = E(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

The variance-covariance matrix of \mathbf{Y} is

$$\begin{aligned} \mathbf{V} &= \text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon}) \\ &= \text{Var}(\mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon}) = \mathbf{Z}\boldsymbol{\Omega}\mathbf{Z}' + \mathbf{R} \quad ; \text{ since } \mathbf{d} \text{ and } \boldsymbol{\varepsilon} \text{ are independent.} \end{aligned}$$

The distribution of \mathbf{Y} can be defined in two ways:

- i. The marginal distribution of \mathbf{Y} not knowing the random effects \mathbf{d} is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix \mathbf{V} . In other words $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. Hence, the pdf of \mathbf{Y} is:

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right].$$

One difficulty in linear mixed effects models is that $\mathbf{V} = \mathbf{Z}\boldsymbol{\Omega}\mathbf{Z}' + \mathbf{R}$ is large and often non-diagonal. Therefore, \mathbf{V}^{-1} is difficult or impossible to compute by common methods [96]. In general, the inverse of \mathbf{V} is given by:

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Omega}^{-1})\mathbf{Z}'\mathbf{R}^{-1}$$

This is true since for three nonsingular matrices \mathbf{A} , \mathbf{B} and \mathbf{C}

$$(\mathbf{A} + \mathbf{CBC}')^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})\mathbf{C}'\mathbf{A}^{-1}$$

and

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$

- ii. The conditional distribution of the dependent variable \mathbf{Y} given the random effects \mathbf{d} is normal with mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d}$ and variance-covariance matrix \mathbf{R} . In other words, $\mathbf{Y}|\mathbf{d} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d}, \mathbf{R})$. Therefore, the pdf is given by

$$g(\mathbf{y}|\mathbf{d}) = \frac{1}{(2\pi)^{n/2} |\mathbf{R}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right].$$

4.4 Maximum Likelihood Estimation of Parameters

4.4.1 Case I: $\boldsymbol{\Omega}$ and \mathbf{R} are known

If matrices $\boldsymbol{\Omega}$ and \mathbf{R} are known, then $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimation (BLUE) of $\boldsymbol{\beta}$ and $\hat{\mathbf{d}}$ is the best linear unbiased prediction (BLUP) of \mathbf{d} [97-100]. The estimates of $\boldsymbol{\beta}$ and \mathbf{d} can be found by maximizing the likelihood in $\boldsymbol{\beta}$ and \mathbf{d} in the joint density function of \mathbf{Y} and \mathbf{d} as follows.

$$f(\mathbf{y}, \mathbf{d}) = g(\mathbf{y}|\mathbf{d})h(\mathbf{d}),$$

where $h(\mathbf{d})$ is the pdf of \mathbf{d} . The likelihood function is

$$L(\boldsymbol{\beta}, \mathbf{d}) = \text{constant} \times \exp \left[\frac{-1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d}) - \frac{1}{2} \mathbf{d}' \boldsymbol{\Omega}^{-1} \mathbf{d} \right].$$

The log-likelihood function is

$$l(\boldsymbol{\beta}, \mathbf{d}) = \log(\text{constant}) - \frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d}) + \mathbf{d}' \boldsymbol{\Omega}^{-1} \mathbf{d}].$$

For $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \mathbf{d})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (-2\mathbf{X})] \\ &= [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (\mathbf{X})] \end{aligned}$$

By equating this partial derivative to zero and taking transpose, we get:

$$\mathbf{X}' \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta} + \mathbf{X}' \mathbf{R}^{-1} \mathbf{Z} \mathbf{d} = \mathbf{X}' \mathbf{R}^{-1} \mathbf{Y} \quad (4.2)$$

For \mathbf{d} :

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \mathbf{d})}{\partial \mathbf{d}} &= -\frac{1}{2\sigma^2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (-2\mathbf{Z}) - 2\mathbf{d}' \boldsymbol{\Omega}^{-1}] \\ &\quad - \frac{1}{2\sigma^2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' \mathbf{R}^{-1} (-2\mathbf{Z}) - 2\mathbf{d}' \boldsymbol{\Omega}^{-1}] \end{aligned}$$

Equating this partial derivative to zero and taking transpose we get:

$$\mathbf{Z}' \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta} + (\mathbf{Z}' \mathbf{R}^{-1} \mathbf{Z} + \boldsymbol{\Omega}^{-1}) \mathbf{d} = \mathbf{Z}' \mathbf{R}^{-1} \mathbf{y} \quad (4.3)$$

From Equations (4.2) and (4.3), the mixed model equations in matrix notation will be:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + \Omega^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad (4.4)$$

For further technical details, we refer to [101-103].

From Equation (4.3) we have

$$\begin{aligned} (Z'R^{-1}Z + \Omega^{-1})\hat{d} &= Z'R^{-1}(Y - X\hat{\beta}) \\ \Leftrightarrow \hat{d} &= (Z'R^{-1}Z + \Omega^{-1})^{-1} Z'R^{-1}(Y - X\hat{\beta}) \end{aligned} \quad (4.5)$$

and from Equation (4.2) we have

$$X'R^{-1}X\hat{\beta} + X'R^{-1}Z\hat{d} = X'R^{-1}y.$$

Substituting the solution for \hat{d} into this equation gives:

$$X'R^{-1}X\hat{\beta} + X'R^{-1}Z [(Z'R^{-1}Z + \Omega^{-1})^{-1} Z'R^{-1}(y - X\hat{\beta})] = X'R^{-1}y.$$

Let $U = (Z'R^{-1}Z + \Omega^{-1})^{-1}$, then we have

$$\begin{aligned} X'R^{-1}X\hat{\beta} + X'R^{-1}Z U Z'R^{-1}(Y - X\hat{\beta}) &= X'R^{-1}Y \\ \Leftrightarrow X'R^{-1}X\hat{\beta} - X'R^{-1}Z U Z'R^{-1}X\hat{\beta} &= X'R^{-1}Y - X'R^{-1}Z U Z'R^{-1}Y \\ \Leftrightarrow X'(R^{-1} - R^{-1}Z U Z'R^{-1})X\hat{\beta} &= X'(R^{-1} - R^{-1}Z U Z'R^{-1})Y \\ \Leftrightarrow X'V^{-1}X\hat{\beta} &= X'V^{-1}Y, \end{aligned}$$

where $V^{-1} = R^{-1} - R^{-1}Z U Z'R^{-1}$

$$\Leftrightarrow \hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y. \quad (4.6)$$

Hence $\widehat{\boldsymbol{\beta}}$ is called the generalized least-square estimate of $\boldsymbol{\beta}$. By plugging this estimator into Equation (4.5), we obtain estimate for the vector of random effects \mathbf{d} .

In case of general linear models with homoscedastic variance and independent residuals, note that $\mathbf{V} = \sigma_\varepsilon^2 \mathbf{I}_n$. Therefore, the least square estimator of $\boldsymbol{\beta}$ will be

$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If the matrix $\mathbf{X}'\mathbf{X}$ is singular, we use a generalized inverse (e.g., see Penrose (1955)) [103, 105-107].

If the matrix in Equation (4.4) is singular, then the solution for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{d}}$ will be based on generalized inverse [100]

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Omega}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where \mathbf{M}^{-1} denote the generalized inverse of a matrix \mathbf{M} .

4.4.2 Case II: $\boldsymbol{\Omega}$ and \mathbf{R} are unknown

The matrices \mathbf{X} and \mathbf{Z} in model (4.1) are known, but the elements of the matrices $\boldsymbol{\Omega}$ and \mathbf{R} maybe functions of an unobserved parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$. In ordinary mixed effects and random ANOVA models, there is some number c of random factors, with the i th factor having q_i levels. These levels are uncorrelated with each other. Associated with the i th random factor is a parameter σ_i^2 which represent the common variance of its levels. Also, the residual effects have common variance σ_{c+1}^2 . The variances $\sigma_1^2, \dots, \sigma_{c+1}^2$ are called variance components. Let $m = c + 1$ denote the number of variance components then,

$$\theta_i = \sigma_i^2, (i = 1, 2, \dots, m), \quad \mathbf{R} = \theta_m \mathbf{I}, \quad \boldsymbol{\Omega} = \text{diag}[\theta_1 \mathbf{I}, \dots, \theta_{m-1} \mathbf{I}],$$

$$\mathbf{V} = \theta_m \mathbf{I} + \sum_{i=1}^{m-1} \theta_i \mathbf{Z}_i \mathbf{Z}_i'$$

where \mathbf{Z}_i is a $n \times q_i$ matrix defined by the partitioning $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_{m-1})$.

The ANOVA models are sometimes parameterized in terms of $\gamma_{c+1} = \sigma_{c+1}^2$ and $\gamma_i = \frac{\sigma_i^2}{\sigma_{c+1}^2}$

($i = 1, \dots, c$) rather than in terms $\sigma_1^2, \dots, \sigma_{c+1}^2$. If we had taken $\gamma_i = \theta_i$, ($i = 1, 2, \dots, m$),

instead of taking $\boldsymbol{\theta}$, we would have had

$$\boldsymbol{\Omega} = \theta_m \text{diag}[\theta_1 \mathbf{I}, \dots, \theta_{m-1} \mathbf{I}],$$

and

$$\mathbf{V} = \theta_m \left(\mathbf{I} + \sum_{i=1}^{m-1} \theta_i \mathbf{Z}_i \mathbf{Z}_i' \right).$$

We will consider both maximum likelihood [28] and restricted maximum likelihood (REML) estimators for the variance components. We maximize the full log-likelihood function l_F in σ_{c+1}^2 , and $\boldsymbol{\theta}$:

$$l_F(\boldsymbol{\beta}, \sigma_{c+1}^2, \boldsymbol{\theta} | \mathbf{y}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

A criticism of the ML estimators for the variance component is that they are biased downward because they do not consider the loss of degrees of freedom from the estimation of $\boldsymbol{\beta}$. The ML estimator for the single “variance component” θ_1 has expectation $(n - p^*)/n$, that it is biased downward by an amount $\theta_1 p^*/n$, which can be significant if the degree of freedom $n - p^*$ is sufficiently small. Here p^* is the rank of the design matrix \mathbf{X} .

The REML method corrects for this by defining estimators of the variance components as the maximizers of the log-likelihood based on linearly independent error contrasts, where n is the total number of observations from all individuals. This log-likelihood l_R that is derived by Harville (1974) is given as,

$$\begin{aligned} l_R(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \sigma_{c+1}^2, \boldsymbol{\theta} | \mathbf{y}) &= -\frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + l_F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \sigma_{c+1}^2, \boldsymbol{\theta} | \mathbf{y}) \\ &= -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

For more details on maximum likelihood and related problems we refer to Harville et al (1977) [108].

There are computational methods that can be used to obtain ML, and REML. For example Newton-Raphson (NR) and Fisher Scoring algorithms [60]. The NR algorithm is an iterative procedure that computes new parameter values σ_{c+1}^2 and $\boldsymbol{\theta}$ from their current values [109]. After estimating the variance-covariance matrices $\boldsymbol{\Omega}$ and \mathbf{R} , the mixed model equations will become [110]:

$$\begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\boldsymbol{\Omega}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Y} \end{bmatrix} \quad (4.7)$$

To model biological data, there is often interest in genetic relationships that arise from the biology of a situation. A matrix that counts for whatever genetic relationships that exists among individuals is called the relationship matrix or numerator relationship matrix; we denote this matrix by (\mathbf{A}) as described in Henderson (1976) [111]. Given the matrix \mathbf{A} , then, for the variance of the vector of random effects \mathbf{d} becomes

$var(\mathbf{d}) = \sigma_d^2 \mathbf{A}$. The use of \mathbf{A} also extends to the case of multiple traits. For example, for two traits, with data vectors \mathbf{y}_1 , and \mathbf{y}_2 , we have

$$var(\mathbf{d}) = \begin{bmatrix} \sigma_{d_1}^2 \mathbf{A} & \sigma_{d_1 d_2} \mathbf{A} \\ \sigma_{d_2 d_1} \mathbf{A} & \sigma_{d_2}^2 \mathbf{A} \end{bmatrix} = \begin{bmatrix} \sigma_{d_1}^2 & \sigma_{d_1 d_2} \\ \sigma_{d_2 d_1} & \sigma_{d_2}^2 \end{bmatrix} \otimes \mathbf{A},$$

where \mathbf{d}_1 and \mathbf{d}_2 are the corresponding vectors of random effects [112].

In genetics the matrices $\mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}$ and $\mathbf{\Omega} = \sigma_d^2 \mathbf{A}$, where the variances σ_ε^2 and σ_d^2 denote residuals and additive variances. Mammals are genetically related to each other and undergoing asexual to produce an offspring that is expected to be correlated, unless σ_d^2 is 0 [113]. Thus, if d_i is the genetic effect for subject i then $v(d_i) = a_{ii} \sigma_d^2$ and $cov(d_i, d_j) = a_{ij} \sigma_d^2$, where the values a_{ii} and a_{ij} can be calculated in different ways depending on whether the parents of subject i are known or not.

Note that \mathbf{A} is a positive-definite matrix (unless identical twins or clones are in the pedigree, in which case it would be positive semi-definite) [114]. Next section will show how to calculate the elements of \mathbf{A} .

If the matrices \mathbf{R} and $\mathbf{\Omega}$ are nonsingular, since \mathbf{R}^{-1} is an identity matrix, we can rewrite the system (4.4) as the following:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \delta\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where $\delta = \frac{\sigma_\varepsilon^2}{\sigma_d^2}$. Therefore:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \delta\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}.$$

4.5 The Numerator Relationship Matrix

The numerator relationship (\mathbf{A}) is a component of BLUP. When we multiply this matrix by the additive genetic variance, it will give the variance-covariance matrix among random subject effects. Each element (a_{ij}) is a probability that a random gene from a given subject i is identical by descent (IBD) to a gene in the same locus from a subject j . In some studies termed the coefficients of kinship [115]. The dimensions of this matrix let say we have q subjects, then the dimensions will be $(q \times q)$, also it is symmetric. The diagonal element (a_{ii}) represents twice the probability that two gametes taken at random from animal i will carry identical alleles by descent. The off-diagonal element (a_{ij}) equals the probability that an allele selected randomly from subject i and an allele selected randomly from subject j at the same locus are identical alleles by descent [116]. The matrix \mathbf{A} can be computed using a recursive method which was described by Henderson (1976). Initially, subjects in the pedigree are coded 1 to n and ordered such that parents precede their progeny. The following are rules to calculate the elements of this matrix.

If both parents (f and m) of subject i are known:

$$a_{ji} = a_{ij} = 0.5(a_{jf} + a_{jm}); \quad j = 1 \text{ to } (j - 1)$$

$$a_{ii} = 1 + 0.5(a_{fm})$$

If only one parent f is known and assumed unrelated to the mate:

$$a_{ji} = a_{ij} = 0.5(a_{fm}); \quad j = 1 \text{ to } (j - 1)$$

$$a_{ii} = 1$$

If both parents are unknown and are assumed unrelated:

$$a_{ji} = a_{ij} = 0; \quad j = 1 \text{ to } (j - 1)$$

$$a_{ii} = 1$$

Example:

The pedigree for a family included seven individuals

Table 4.1 Pedigree for a family of seven individuals.

Individual	Father	Mother
3	1	2
4	Unknown	Unknown
5	3	4
6	3	4
7	3	4

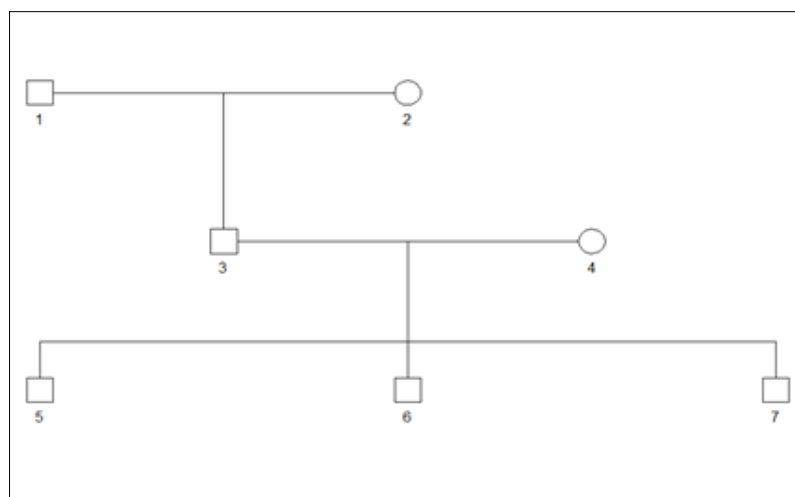


Figure 4.1. Pedigree of seven individuals.

The calculations to find the numerator relationship matrix for the pedigree table are:

$a_{11} = 1$ Both parents of individual 1 are unknown

$a_{12} = a_{21} = 0$ both parents of individual 1 or 2 are unknown

\vdots

$a_{77} = 1 + 0.5(a_{fm}) = 1 + 0.5(0) = 1$ both parents of individual are known

The \mathbf{A} matrix is:

$$\begin{bmatrix} 1 & 0 & 0.5 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 1 & 0.5 & 0 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.5 & 1 & 0 & 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0.5 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.25 & 0.25 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$$

In this matrix, the diagonal element for subject i (a_{ii}) is equal to $1 + F_i$, where F_i is called the inbreeding coefficient of subject i [107, 116]. We are going to use the numerator relationship matrix \mathbf{A} in LME and threshold model in next section. The matrix \mathbf{A} can also be calculated using *pedigreemm* package in R (see Appendix D for R code):

In our case, we have \mathbf{Y} is a vector which represents breast cancer incidence (affected = 1 and normal = 0) with dimension 167×1 , and factors namely gender and family which will treated as fixed effects, and \mathbf{X} is design matrix with dimensions 167×5 , where 5 is the number of levels for fixed effects (male and female and 3 families), $\boldsymbol{\beta}$ is 5×1 vector of parameters of fixed effects. \mathbf{Z} is 167×167 design matrix associated with the vector of genetic effects \mathbf{d} .

4.6 The Heritability Coefficient

In biological applications, the quantities \mathbf{d} and $\boldsymbol{\beta}$ in (4.1) denote the genetic and environmental effects, respectively. Given that “a central question in biology is whether observed variation in a particular trait is due to environmental or biological factors” [57].

The heritability coefficient is $h^2 = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_\varepsilon^2}$ which represents the proportion of phenotypic variance attributed to variation in genotypic values, is a fundamentally important quantity. Indeed, linear mixed effects models and inference on the heritability coefficient has been applied recently in genome-wide association studies [117, 118].

4.7 Genomic Relationships Matrix

In linear mixed effects models that include genomic relationships matrix (\mathbf{G}) will be more accurate than those that use expected relationships from pedigrees matrix. The numerator relationship matrix \mathbf{A} uses the only pedigree information to obtain probabilities that gene pairs are identical by descent [116]. But Genomic relationship matrix \mathbf{G} uses the genotypic information to estimate the segment of DNA that two individuals share [119].

To obtain the matrix, \mathbf{G} , let \mathbf{M} be the matrix that stipulates which marker alleles were inherited by everyone, with dimensions of $n_g \times s$, where n_g is number of individuals with genomic information and s is number of SNP's. \mathbf{M} has elements from 0, 1, 2 and 5, representative of homozygote, heterozygote, other homozygote, and missing SNP marker, correspondingly. Let \mathbf{P} be $n_g \times s$ matrix contain frequencies p_i of the second allele at each locus. Therefore, column i of \mathbf{P} is $2(p_i - 0.5)$ where $i = 1, 2, \dots, S$

[57, 119]. A \mathbf{G} is genomic relationship matrix can be found with three different methods.

First method is:

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \sum p_i(1 - p_i)}$$

Dividing by $2 \sum p_i(1 - p_i)$ scales matrix \mathbf{G} to be analogous to the numerator relationship matrix \mathbf{A} . In frequency estimates, missing genotypes (coded as 5 in the data) will not be included in our calculations. The elements of matrix $\mathbf{M} - \mathbf{P}$ set to be zero, which is the mean of frequency of missing genotypes, other elements of $\mathbf{M} - \mathbf{P}$ will standardized for each individual's proportion of missing genotypes [57].

The genomic inbreeding coefficient for individual j is simply $G_{jj} - 1$, and genomic relationships between individuals j and k , which are analogous to the relationship coefficients of Wright (1922), these coefficients are obtained by dividing elements G_{jk} by square roots of diagonals G_{jj} and G_{kk} .

The second method for obtaining \mathbf{G} weights markers by reciprocals of their expected variance instead of summing expectations across loci and then dividing:

$$\mathbf{G} = (\mathbf{M} - \mathbf{P})\mathbf{D}(\mathbf{M} - \mathbf{P})', \text{ where } \mathbf{D} \text{ is diagonal with } \mathbf{D}_{jj} = \frac{1}{m[2p_i(1-p_i)]}. \text{ This formula was}$$

proposed by Amin et al (2007) and Leutenegger et al (2003) [120, 121].

The third method to obtain \mathbf{G} does not require all ele frequencies and instead adjusts for mean homozygosity by regressing \mathbf{MM}' on \mathbf{A} and \mathbf{G} using the following model:

$$\mathbf{MM}' = \gamma_0 \mathbf{1}\mathbf{1}' + \gamma_1 \mathbf{A} + \mathbf{E},$$

Where γ_0 and γ_1 are the intercept and slope, respectively.

Here is an example showed how to calculate **G** matrix, suppose we have three individuals and two SNP's and suppose we have encoded AA=2, Aa=1, and aa=0.

So, matrix **M** with dimensions 3×2 .

$$\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \\ 1 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} aa & aa \\ aa & AA \\ aA & AA \end{bmatrix}$$

Also, matrix **P**_{3×2}

$$\text{Frequency A for locus 1 (SNP1)} = \hat{p}_1 = \frac{1}{6} = 0.16$$

$$\text{Frequency A for locus 2 (SNP2)} = \hat{p}_2 = \frac{4}{6} = 0.667$$

$$\text{First column in } \mathbf{P} \text{ matrix is } 2(\hat{p}_1 - 0.5) = 2(0.16 - 0.5) = -0.68$$

$$\text{Second column in } \mathbf{P} \text{ matrix is } 2(\hat{p}_2 - 0.5) = 2(0.667 - 0.5) = 0.32$$

$$\mathbf{P} = \begin{bmatrix} -0.68 & 0.32 \\ -0.68 & 0.32 \\ -0.68 & 0.32 \end{bmatrix}$$

$$\text{For the denominator } 2 \sum p_j(1 - p_j) = 2 * [0.16 * (1 - 0.16) + 0.667 * (1 - 0.667)]$$

Misztal et al (2009) proposed that it is possible to modify a numerator based relationship matrix **A** to a matrix **H** that takes in both pedigree-based relationships and genomic information **A**_Δ [17]:

$$\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta.$$

Let indices 1 and 2 denote ungenotyped and genotyped animals. Then

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$

$$\text{And} \quad \mathbf{A}_{\Delta} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$

The inverse of \mathbf{H} can be calculated as following:

Let the inverse of matrix \mathbf{A} be

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix},$$

To derive an inverse function of the combined relationship matrix of [122], using the properties of the inverse of a partitioned matrix, identities from $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ are:

$$\mathbf{A}^{11}\mathbf{A}_{11} + \mathbf{A}^{12}\mathbf{A}_{21} = \mathbf{I} \quad \text{A1}$$

$$\mathbf{A}^{21}\mathbf{A}_{12} + \mathbf{A}^{22}\mathbf{A}_{22} = \mathbf{I} \quad \text{A2}$$

$$\mathbf{A}^{11}\mathbf{A}_{12} + \mathbf{A}^{12}\mathbf{A}_{21} = \mathbf{0} \quad \text{A3}$$

$$\mathbf{A}^{21}\mathbf{A}_{11} + \mathbf{A}^{22}\mathbf{A}_{21} = \mathbf{0}, \text{ and} \quad \text{A4}$$

$$(\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} = \mathbf{A}^{11} \quad \text{A5}$$

Using [A1] through [A4], then multiplying the whole-population matrix [18],

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22}) \\ (\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

4.8 Threshold model

This model is totally flexible. The distribution of the elements of error can be taken to be some non-normal (e.g. logistic) distribution. Threshold model can be simply extended to multiple responses; some are ordered categorical and others that are quantitative. The previous study on the threshold model seems to have been confined primarily to the case where the underlying linear model is a fixed-effects model. McCullagh (1980) gave a comprehensive discussion of this case. Curnow and Smith (1975) reviewed genetic applications of threshold models. Thompson and Baker (1981) presented a device, termed a composite link function that allows threshold models to be embedded in the framework of the generalized linear models due to Nelder and Wedderburn (1972). The threshold model in which it is assumed that the observed category is determined by the value of an underlying unobservable continuous response. [67].

Suppose we have n individuals, and λ_i denote an underlying continuous-response variable associated with the i th of these individuals. Take $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)'$.

Let $\boldsymbol{\lambda}$ follow a linear mixed effects model;

$$\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon} \quad (4.8)$$

matrices \mathbf{X} and \mathbf{Z} and vectors $\boldsymbol{\beta}$, \mathbf{d} and $\boldsymbol{\varepsilon}$ were described in model (4.1); the only difference is the vector $\boldsymbol{\lambda}$ which contains $n \times 1$ unobserved continuous variables. Give the ordered categories numbers from 1 to m . It is assumed that λ_i is unobserved, but we observe the

category into which the i th individual falls or the category number O_i , $i = 1, 2, \dots, n$.

These categories determined into following relationship:

$$\tau_{k_{i-1}} < \lambda_i \leq \tau_{k_i} \Leftrightarrow O_i = k, \quad (4.9)$$

where $k \in \{1, 2, \dots, m\}$, $\tau_0 = -\infty$, $\tau_m = +\infty$ and $\tau_1, \dots, \tau_{m-1}$ are unknown thresholds which partition the real numbers into m categories. Therefore, when the realized value of λ_i belongs to the k interval, the observed values $O_i = k$.

Following this assumption, the probability-mass function of O_1, \dots, O_n is

$$\begin{aligned} p(O_1, \dots, O_n) &= pr\{O_i = o_i (i = 1, \dots, n)\} \\ &= pr\{\tau_{o_{i-1}} < \lambda_i \leq \tau_{o_i} (i = 1, \dots, n)\} \end{aligned}$$

Harville (1984) indicate to this model for the vector $\mathbf{o} = (O_1, \dots, O_n)'$ of categorical responses as the “threshold model”.

In implementing the threshold model, there is a request to make inferences about various functions of $\tau_1, \dots, \tau_{m-1}$, $\boldsymbol{\beta}$, and sample of the random vector \mathbf{d} say \mathbf{d}_c . In special, we need to make inferences about the quantity as following:

$$p_{k,k'} = pr\{\tau_k < x'\boldsymbol{\beta} + z'\mathbf{d} + \varepsilon \leq \tau_{k'} \mid \mathbf{d} = \mathbf{d}_c\} \quad (4.10)$$

where $k < k'$, x and z are specified vectors, and ε is distributed $N(0, \sigma^2)$ and independent on \mathbf{d} . The best choice of x and z , which represents the condition probability when $(\mathbf{d} = \mathbf{d}_c)$ that an $(n + 1)$ individual (which follows the same model as the n observable individuals) will belong to one of the categories $k + 1, \dots, k'$.

Note that we can re-express the inequality (4.10) as the following:

$$\tau_{k_{i-1}} < \lambda_i \leq \tau_{k_i}$$

$$(\tau_{k_{-1}} - \tau_1)/\sigma < W_i \leq (\tau_k - \tau_1)/\sigma \quad (4.11)$$

Where $W_i = (\lambda_i - \tau_1)/\sigma$, $i = 1, 2, \dots, n$, $k = 1, 2, \dots, m$.

In matrices notations if we have $\mathbf{w} = (W_1, W_2, \dots, W_n)'$ and using vector of ones $\mathbf{1}$, we can write;

$$\mathbf{w} = (\mathbf{1}, \mathbf{X}) \begin{bmatrix} -\sigma^{-1}\tau_1 \\ -\sigma^{-1}\boldsymbol{\beta} \end{bmatrix} + \mathbf{Z}(\sigma^{-1}\mathbf{d}) + \sigma^{-1}\boldsymbol{\varepsilon} \quad (4.12)$$

Therefore, the threshold model can be reformulated in terms of a second threshold model whose underlying continuous response variables correspond to $(\lambda_i - \tau_1)/\sigma$, $i = 1, 2, \dots, n$, whose boundaries correspond to $(\tau_k - \tau_1)/\sigma$, $k = 1, 2, \dots, m - 1$, and whose vectors of 'fixed', 'random' and 'residual' effects correspond to $\sigma^{-1}(-\tau_1, \boldsymbol{\beta}')$ to $\sigma^{-1}\mathbf{d}$, and to $\sigma^{-1}\boldsymbol{\varepsilon}$, respectively. In the latter, the threshold model, the residual variance equals 1, the first boundary point equals 0.

From now, we assume that $\sigma = 1$ and $\tau_1 = 0$, also the first column of \mathbf{X} equals $\mathbf{1}$. While these assumptions are met, we refer to the threshold model as the 'standardized threshold model' [67]. If we suppose that $\boldsymbol{\lambda}$ is observed, we could use Henderson's Best linear unbiased prediction (BLUP) procedure to estimate $x'\boldsymbol{\beta} + z'\mathbf{d}$. The BLUP would be $x'\tilde{\boldsymbol{\beta}} + z'\tilde{\mathbf{d}}$ where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{d}}$ denote any solution to the system of linear equations,

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\lambda} \\ \mathbf{Z}'\boldsymbol{\lambda} \end{bmatrix}$$

There is one way to obtain the mixed model equations and hence at the BLUP, that is finding values of $\boldsymbol{\beta}$ and \mathbf{d} that maximize

$$\begin{aligned}
\phi_{N+q} \left(\begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\beta} \end{bmatrix}; \begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{Z}\boldsymbol{\Omega} \\ \boldsymbol{\Omega}\mathbf{Z}' & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \right) &= \phi_N(\boldsymbol{\lambda}; \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d}_c, I) \phi_q(\mathbf{d}_c; \mathbf{0}, \boldsymbol{\Omega}) \\
&= \phi_q(\mathbf{d}_c; \mathbf{0}, \boldsymbol{\Omega}) \prod_{i=1}^n \phi_N(\lambda_i - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c), \tag{4.13}
\end{aligned}$$

where x_i' and z_i' represent the i th rows of \mathbf{X} and \mathbf{Z} , respectively. Note that (4.13) is the joint probability density function of $\boldsymbol{\lambda}$ and \mathbf{d} .

Now let us apply to the standardized threshold model ($\boldsymbol{\lambda}$ is unobserved) an approach analogous to the maximization (4.12).

$$\begin{aligned}
\psi(o_1, o_2, \dots, o_n; \boldsymbol{\tau}, \boldsymbol{\beta}, \mathbf{d}_c) &= \prod_{i=1}^n \int_{\tau_{o_{i-1}}}^{\tau_{o_i}} \phi(\lambda_i - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c) d\lambda_i \\
&= \prod_{i=1}^n \{ \Phi(\tau_{o_i} - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c) - \Phi(\tau_{o_{i-1}} - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c) \}. \tag{4.14}
\end{aligned}$$

This procedure consists of estimating $\mathbf{w}'\boldsymbol{\tau} - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c$ by $\mathbf{w}'\hat{\boldsymbol{\tau}} - x_i'\hat{\boldsymbol{\beta}} - z_i'\hat{\mathbf{d}}_c$, and $\hat{\boldsymbol{\tau}} = (\hat{\tau}_2, \hat{\tau}_3, \dots, \hat{\tau}_{m-1})$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{d}}_c$ denote any values of $\boldsymbol{\tau}$, $\boldsymbol{\beta}$ and \mathbf{d}_c that maximize

$$\psi(o_1, o_2, \dots, o_n; \boldsymbol{\tau}, \boldsymbol{\beta}, \mathbf{d}_c) \phi_q(\mathbf{d}_c; \mathbf{0}, \boldsymbol{\Omega}) \tag{4.15}$$

We consider the problem of computing $\hat{\boldsymbol{\tau}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{d}}_c$ that maximize the function (4.15).

Given that $\boldsymbol{\omega}' = [\boldsymbol{\tau}', \boldsymbol{\beta}', \mathbf{d}_c']$, Gianola and Foulley (1983) proceeded to find the estimator $\hat{\boldsymbol{\omega}}$ that maximizes the log of the density $L(\boldsymbol{\omega})$. They therefore provided the following non-linear iterative system of equations based on the first and second derivatives, assuming a normal distribution, to obtain solutions for $\Delta\boldsymbol{\tau}$, $\Delta\boldsymbol{\beta}$ and $\Delta\mathbf{d}_c$

$$\begin{bmatrix} \mathbf{Q} & \mathbf{L}'\mathbf{X} & \mathbf{L}'\mathbf{Z} \\ \mathbf{X}'\mathbf{L} & \mathbf{X}'\mathbf{T}\mathbf{X} & \mathbf{X}'\mathbf{T}\mathbf{Z} \\ \mathbf{Z}'\mathbf{L} & \mathbf{Z}'\mathbf{T}\mathbf{X} & \mathbf{Z}'\mathbf{T}\mathbf{Z} + \mathbf{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\tau} \\ \Delta\boldsymbol{\beta} \\ \Delta\mathbf{d}_c \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X}'\mathbf{V} \\ \mathbf{Z}'\mathbf{V} - \mathbf{\Omega}^{-1}\Delta\mathbf{d}_c \end{bmatrix} \quad (4.16)$$

The calculation of some of these matrices involves p_{ik} , which defines the probability of a response being observed in category k assuming a normal distribution of the i th row.

$$p_{ik} = \Phi(\tau_k - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c) - \Phi(\tau_{k-1} - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c)$$

This distribution is a function of the distance between $x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c$ and the threshold, Likewise, the height of the normal curve at τ_k under the conditions of the i th row becomes:

$$\phi_{ik} = \phi(\tau_k - x_i'\boldsymbol{\beta} - z_i'\mathbf{d}_c)$$

To compute the various matrices and vectors in (10.3):

$$V_i = \sum_{k=1}^m n_{ik} \left(\frac{\phi_{ik-1} - \phi_{ik}}{p_{ik}} \right)$$

Where n_{ik} is the number of counts in category k of response in row i , and for the elements of the matrix \mathbf{T} , which is a weighting factor, are computed as:

$$T_{ii} = \sum_{k=1}^m n_{i.} \left(\frac{\phi_{ik-1} - \phi_{ik}}{p_{ik}} \right)$$

And $n_{i.} = \sum_{k=1}^m n_{ik}$

The matrix \mathbf{Q} is an $(m-1)$ by $(m-1)$ banded matrix and the diagonal elements are calculated as:

$$q_{kk} = \sum_{i=1}^m n_i \left(\frac{p_{ik} - p_{i(k+1)}}{p_{ik}p_{i(k+1)}} \right) \phi_{ik}^2, k = 1, \dots, (m-1)$$

and for the off-diagonal elements are:

$$q_{(k+1)k} = - \sum_{i=1}^m n_i \frac{\phi_{i(k+1)}\phi_{ik}}{p_{i(k+1)}}, \text{ for } k = 1, \dots, (m-2)$$

and $q_{(k+1)k} = q_{k(k+1)}$.

The matrix \mathbf{L} is of order n by $(m-1)$ and its elements are calculated as:

$$l_{ik} = -n_i \phi_{ik} \left(\frac{\phi_{ik} - \phi_{i(k-1)}}{p_{ik}} - \frac{\phi_{i(k+1)} - \phi_{ik}}{p_{i(k+1)}} \right)$$

The elements of the vector \mathbf{P} are:

$$p_k = \left\{ \sum_{i=1}^n \left[\frac{n_{ik}}{p_{ik}} - \frac{n_{i(k+1)}}{p_{i(k+1)}} \right] \phi_{ik} \right\}; k = 1, \dots, m-1.$$

The matrices \mathbf{L} and \mathbf{P} are then substituted in the system of equation (4.16) [107].

4.9 Gibbs Sampling Method

This is a numerical integration method based on all possible conditional posterior distributions. It was first implemented by Geman and Geman (1984). In this method, the marginal posterior distributions generate random drawings by taking iterative samples [107]. Wang et al (1993) used Gibbs sampling to estimate variance components in univariate linear mixed effects model [124].

4.9.1 Prior and Posterior Distributions

Suppose we have the following univariate linear mixed effects model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon}$$

Where \mathbf{Y} is a vector of observations with dimension $n \times 1$; $\boldsymbol{\beta}$ is a vector of fixed effects with dimension $p \times 1$; \mathbf{d} is a vector of random effects with dimension $q \times 1$; $\boldsymbol{\varepsilon}$ is a vector of errors; \mathbf{X} and \mathbf{Z} are design matrices for the fixed and random effects with dimensions $n \times p$ and $n \times q$ respectively. The conditional distribution that generates the vector \mathbf{Y} is:

$$\mathbf{Y}|\boldsymbol{\beta}, \mathbf{d}, \sigma_e^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d}, \mathbf{R}\sigma_e^2)$$

The matrix \mathbf{R} here assumed to be identity with $n \times n$ dimension and σ_e^2 is residual variance.

Bayesian analysis required to determine prior distributions to the unidentified parameters. Nonetheless, the prior distributions of $\boldsymbol{\beta}$, \mathbf{d} , σ_e^2 and σ_d^2 are needed to fulfill the Bayesian condition of the model [124]. Commonly a flat, improper prior distribution was used for vector $\boldsymbol{\beta}$, as follows:

$$P(\boldsymbol{\beta}) \sim \text{constant} \quad (4.17)$$

Also, a multivariate normal distribution (MVN) was used for \mathbf{d} :

$$\mathbf{d} \sim \text{MVN}(\mathbf{0}, \mathbf{A}\sigma_d^2) \quad (4.18)$$

Where \mathbf{A} is a numerator relationship matrix. In addition, for variance components σ_e^2 and σ_d^2 the priors are the scaled inverted *Chi*-squared distribution form:

$$p(\sigma_e^2 | v_e, s_e^2) \propto (\sigma_e^2)^{\frac{-v_e}{2}-1} \exp\left(\frac{-1}{2} v_e s_e^2 / \sigma_e^2\right) \quad (4.19)$$

And

$$p(\sigma_d^2 | v_d, s_d^2) \propto (\sigma_d^2)^{\frac{-v_d}{2}-1} \exp\left(\frac{-1}{2} v_d s_d^2 / \sigma_d^2\right) \quad (4.20)$$

The parameters are degrees of belief v_e, v_d , and s_e^2, s_d^2 for (4.19) and (4.20) respectively. The assumption for the degrees of belief parameters v_e and v_d equal to zero to find the improper priors:

$$p(\sigma_e^2) \propto (\sigma_e^2)^{-1}; \quad p(\sigma_d^2) \propto (\sigma_d^2)^{-1} \quad (4.21)$$

The proper uniform prior is:

$$p(\sigma^2) \propto \begin{cases} k & 0 \leq \sigma^2 \leq \sigma_{max}^2 \\ 0 & otherwise \end{cases}$$

The common technique for calculating the conditional densities required for Gibbs sampling is to use the conditional independence in models (4.17-4.21) to write the joint posterior density as:

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{d}, \sigma_e^2, \sigma_d^2 | \mathbf{y}) &\propto \frac{1}{(\sigma_e^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right) \\ &\times \frac{1}{(\sigma_d^2)^{\frac{q}{2}+1}} \exp\left(-\frac{1}{2\sigma_d^2} \mathbf{d}' \mathbf{A}^{-1} \mathbf{d}\right) \end{aligned} \quad (4.22)$$

In Gibbs sampling method, the full conditional posterior for each parameter is needed; the scaled inverted *Chi*-square is the full conditional posterior for residual variance σ_e^2 which given by:

$$p(\sigma_e^2 | \boldsymbol{\beta}, \mathbf{d}, \sigma_d^2, \mathbf{y}) \propto \frac{p(\sigma_e^2, \boldsymbol{\beta}, \mathbf{d}, \sigma_d^2 | \mathbf{y})}{\int p(\sigma_e^2, \boldsymbol{\beta}, \mathbf{d}, \sigma_d^2 | \mathbf{y}) d\sigma_e^2}$$

$$\begin{aligned} & \frac{\frac{1}{(\sigma_e^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right)}{\frac{1}{(\sigma_d^2)^{\frac{q}{2}+1}} \exp\left(-\frac{1}{2\sigma_d^2} \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}\right)} \frac{1}{(\sigma_d^2)^{\frac{q}{2}+1}} \exp\left(-\frac{1}{2\sigma_d^2} \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}\right) \\ & \propto \frac{1}{(\sigma_d^2)^{\frac{q}{2}+1}} \exp\left(-\frac{1}{2\sigma_d^2} \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}\right) \int \frac{1}{(\sigma_e^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right) d\sigma_e^2 \\ & \propto \frac{\frac{1}{(\sigma_e^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right)}{\int \frac{1}{(\sigma_e^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right) d\sigma_e^2} \end{aligned}$$

Because the denominator is constant with respect to σ_e^2 then:

$$p(\sigma_e^2 | \boldsymbol{\beta}, \mathbf{d}, \sigma_d^2, \mathbf{y}) \propto (\sigma_e^2)^{\frac{-n}{2}-1} \exp\left(-\frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})\right) \quad (4.23)$$

Where the parameters are $v_e = n$ and $s_e^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})/n$. Each of the Gibbs conditionals can be computed in this manner [125]. So, the full conditional posterior for additive variance σ_d^2 is the scaled inverted *Chi*-squared:

$$p(\sigma_d^2 | \boldsymbol{\beta}, \mathbf{d}, \sigma_e^2, \mathbf{y}) \propto (\sigma_d^2)^{\frac{-q}{2}-1} \exp\left(-\frac{1}{2\sigma_d^2} \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}\right) \quad (4.24)$$

With the parameters $v_e = q$ and $s_e^2 = \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}/q$.

The full condition posterior of $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} | \mathbf{d}, \sigma_e^2, \sigma_d^2, \mathbf{y} \sim N[\tilde{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2] \quad (4.25)$$

Where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{Z}\mathbf{d})$

For the random vector \mathbf{d} :

$$\mathbf{d} | \boldsymbol{\beta}, \sigma_e^2, \sigma_d^2, \mathbf{y} \sim N[\tilde{\mathbf{d}}, \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_d^2}\right)^{-1} \sigma_e^2] \quad (4.26)$$

Where $\tilde{\mathbf{d}} = \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1} \frac{\sigma_e^2}{\sigma_d^2} \right)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{d})$

After defining these four posterior distributions, implementing Gibbs sampling steps can be shown below:

- i. setting arbitrary initial values of $\boldsymbol{\beta}$, \mathbf{d} , σ_e^2 , σ_d^2 ;
- ii. generate σ_e^2 from (4.23), and update it;
- iii. generate σ_d^2 from (4.24), and update it;
- iv. generate \mathbf{d} from (4.26), and update it;
- v. generate $\boldsymbol{\beta}$ from (4.25), and update it;
- vi. repeat steps (ii to v) k times, and update the values each time, where k denoted the length of Gibbs chain (iteration). Let $(\sigma_e^2)^{(k)}$, $(\sigma_d^2)^{(k)}$, $(\mathbf{d})^{(k)}$ and $(\boldsymbol{\beta})^{(k)}$ are the sample points with k th iteration. Then:
- vii. repeat steps (i-vi) m times, to get m Gibbs samples. Then we have:

$$(\sigma_e^2)_1^{(k)}, (\sigma_e^2)_2^{(k)}, \dots, (\sigma_e^2)_m^{(k)} \sim p(\sigma_e^2 | \mathbf{y})$$

$$(\sigma_d^2)_1^{(k)}, (\sigma_d^2)_2^{(k)}, \dots, (\sigma_d^2)_m^{(k)} \sim p(\sigma_d^2 | \mathbf{y})$$

$$(\mathbf{d})_1^{(k)}, (\mathbf{d})_2^{(k)}, \dots, (\mathbf{d})_m^{(k)} \sim p(\mathbf{d} | \mathbf{y})$$

$$(\boldsymbol{\beta})_1^{(k)}, (\boldsymbol{\beta})_2^{(k)}, \dots, (\boldsymbol{\beta})_m^{(k)} \sim p(\boldsymbol{\beta} | \mathbf{y})$$

Since the interest is to estimate the variance components σ_e^2 and σ_d^2 , we will not monitor vectors \mathbf{d} and $\boldsymbol{\beta}$ [124].

After getting samples, and as noted by Casella and George (1992) and Gelfand and Smith (1990), the estimator of the marginal density of σ_e^2 is:

$$\hat{p}(\sigma_e^2|\mathbf{y}) = \frac{1}{m} \sum_{j=1}^m p(\sigma_e^2 | (\boldsymbol{\beta})_j^{(k)}, (\mathbf{d})_j^{(k)}, (\sigma_d^2)_j^{(k)}, \mathbf{y})$$

Similarly, the estimator of the marginal density of σ_d^2 is:

$$\hat{p}(\sigma_d^2|\mathbf{y}) = \frac{1}{m} \sum_{j=1}^m p(\sigma_d^2 | (\boldsymbol{\beta})_j^{(k)}, (\mathbf{d})_j^{(k)}, \mathbf{y})$$

For threshold model; the response variable O_i for each individual takes on two possible values 1 or 0; The variable is the expression of latent continuous variable λ_i ; this liability of individual i . The variable O_i takes 1 if λ_i exceeds an unknown fixed threshold τ and 0 otherwise. The variable λ_i is distributed normal with mean θ and variance 1.

Hence:

$$\lambda_i | \theta \sim N(w_i' \theta, 1)$$

Where $\theta' = (\boldsymbol{\beta}', \mathbf{d}')$ and $\boldsymbol{\beta}, \mathbf{d}$ as defined in model (4.8), also w_i' is a row incidence vector related to individual i . The conditional posterior of thresholds and liabilities are uniforms and truncated normals, respectively. The remaining parameters $(\boldsymbol{\beta}, \mathbf{d}, \sigma_e^2, \sigma_d^2)$ are the same in linear mixed effect model [87].

For given θ , the conditional distributions of λ_i are independent. Therefore, the joint density is given by

$$p(\boldsymbol{\lambda}|\theta) = \prod_{i=1}^n \phi_{\lambda_i}(w_i'\theta, 1) = \phi_{\boldsymbol{\lambda}}(\mathbf{W}\boldsymbol{\theta}, \mathbf{1}) \quad (4.27)$$

Where $\phi_{\lambda}(\cdot)$ is a normal density, also in (4.27) put $\mathbf{W}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon}$. However, given the model, we have:

$$p(O_i = 1|\theta, \tau) = p(\lambda_i > \tau|\boldsymbol{\theta}, \tau) = \int_{\tau - w_i'\theta}^{\infty} \phi(x) dx = \Phi(-(\tau - w_i'\theta)) \quad (4.28)$$

Where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. We can set the constant τ to 0. So the model (4.28) can be write as:

$$p(O_i = 1|\theta) = \Phi(w_i'\theta) \quad (4.29)$$

Distribution of λ_i conditional on θ and on $O_i = o_i$ follows a truncated normal distribution. So for $o_i = 1$:

$$p(\lambda_i|\theta, O_i = 1) = \frac{\phi_{\lambda_i}(w_i'\theta, 1)}{\Phi(w_i'\theta)} \mathbf{1}(\lambda_i > 0) \quad (4.30)$$

Where $\mathbf{1}(X \in R)$ is the indicator function that takes 1 if X is contained in the set R , and 0 otherwise [87]. For $o_i = 0$, the density is

$$p(\lambda_i|\theta, O_i = 0) = \frac{\phi_{\lambda_i}(w_i'\theta, 1)}{\Phi(-w_i'\theta)} \mathbf{1}(\lambda_i \leq 0) \quad (4.31)$$

4.10 Predicting SNP effects

Let the random effects vector \mathbf{d} be decomposed into those for genotyped (\mathbf{d}_g) and ungenotyped (\mathbf{d}_n) individuals. The random effects of genotyped individuals are a function of SNP effects \mathbf{g} :

$$\mathbf{d}_g = \mathbf{Z}_g \mathbf{g}$$

SNP effects can be projected using genomic liability $\hat{\mathbf{d}}$, \mathbf{D} which is a diagonal matrix of weights for the variances of SNPs. \mathbf{Z}_g is a matrix linking the genotype of each locus per the following:

$$\hat{\mathbf{g}} = \mathbf{D} \mathbf{Z}_g' [\mathbf{Z}_g \mathbf{D} \mathbf{Z}_g']^{-1} \hat{\mathbf{d}}_g$$

Dimensions: $\hat{\mathbf{g}}$ is $S \times 1$, \mathbf{D} is $S \times S$ and $\hat{\mathbf{d}}_g$ is $n_g \times 1$. This is the best predictor for SNP effects [102].

CHAPTER 5

MULTITRAITS MODELS

5.1 Introduction

The excellent method to evaluate the subjects on different traits is a multiple trait analysis, because it considers the relationship between these traits. Analysis of multiple traits includes the simultaneous evaluation of subjects for more than one trait and makes use of the genetic and phenotypic correlations between the traits. Henderson and Quaas (1976) applied the first BLUP in multiple traits. We will present the multivariate BLUP (MBLUP) in this chapter and give examples in its application [107].

5.2 Multiple Trait Model

The method is based on an extension of Henderson's method for practical of relatives records in single trait model. Furthermore, the main advantage of MBLUP is that the increase of accuracy. This accuracy depends on the absolute difference between the residual and genetic correlations between the traits. The larger the enhancement in efficiency is due the larger of differences in these correlations.

Suppose we have n related subjects each with records on t traits, for each of the traits we have a stack of the univariate models as following:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_t \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & X_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & X_t \end{bmatrix} \begin{bmatrix} B_1 \\ \vdots \\ B_q \end{bmatrix} + \begin{bmatrix} Z_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & Z_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Z_t \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_t \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_t \end{bmatrix}$$

Where \mathbf{Y}_i is a vector of the response observations for trait i , \mathbf{B}_i and \mathbf{d}_i are vectors of fixed and random effects for the trait i , \mathbf{e}_i is vector of residuals effects for trait i , and matrices \mathbf{X}_i and \mathbf{Z}_i are design matrices related to fixed and random effects, respectively [107].

5.3 Equal Design Matrices

If all traits are affected by the same fixed effect and records of all subjects are taken for all traits, then the same design matrices \mathbf{X} and \mathbf{Z} are use for all traits.

Forexample, we have two traits and for each trait we can write the model as follows:

For the first trait:

$$\mathbf{Y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_1\mathbf{d}_1 + \boldsymbol{\varepsilon}_1$$

The second trait is:

$$\mathbf{Y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_2\mathbf{d}_2 + \boldsymbol{\varepsilon}_2$$

If we order the subjects with respect traits, we could write the multivariate model in the following form:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \quad (5.1)$$

It is assumed that the variance of random effects \mathbf{d}_i and residuals \mathbf{e}_i given by:

$$\text{Var} \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} g_{11}\mathbf{A} & g_{12}\mathbf{A} & 0 & 0 \\ g_{21}\mathbf{A} & g_{22}\mathbf{A} & 0 & 0 \\ 0 & 0 & r_{11}\mathbf{I} & r_{12}\mathbf{I} \\ 0 & 0 & r_{21}\mathbf{I} & r_{22}\mathbf{I} \end{bmatrix}$$

Where $\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$ is additive variance-covariance matrix for random effects. The

elements of matrix \mathbf{G} are; g_{11} is additive variance for direct effect of trait 1; $g_{12} = g_{21}$ is

additive covariance between traits 1 and 2; g_{22} is additive variance for direct effect of trait 2; \mathbf{A} is the numerator relationship matrix (see section 4.5); $\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$ is variance-covariance matrix of residuals.

The heritability for trait i can be calculated using the following formula [128]:

$$h_i^2 = \frac{g_{ii}}{g_{ii} + r_{ii}}$$

To solve the multivariate model (5.1) we can use mixed model equations system as the following:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1} \otimes \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (5.2)$$

Where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix}, \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}, \hat{\mathbf{d}} = \begin{bmatrix} \hat{\mathbf{d}}_1 \\ \hat{\mathbf{d}}_2 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$$

We can rewrite MME for multivariate model with two traits as follows:

$$\begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{11}\mathbf{X}_1 & \mathbf{X}'_1\mathbf{R}^{12}\mathbf{X}_2 & \mathbf{X}'_1\mathbf{R}^{11}\mathbf{Z}_1 & \mathbf{X}'_1\mathbf{R}^{12}\mathbf{Z}_2 \\ \mathbf{X}'_2\mathbf{R}^{12}\mathbf{X}_1 & \mathbf{X}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{X}'_2\mathbf{R}^{21}\mathbf{Z}_1 & \mathbf{X}'_2\mathbf{R}^{22}\mathbf{Z}_2 \\ \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{X}_1 & \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{X}_2 & \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{Z}_1 + \mathbf{A}^{-1}g^{11} & \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{Z}_2 + \mathbf{A}^{-1}g^{12} \\ \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{X}_1 & \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{Z}_1 + \mathbf{A}^{-1}g^{21} & \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{Z}_2 + \mathbf{A}^{-1}g^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\mathbf{d}}_1 \\ \hat{\mathbf{d}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{11}\mathbf{y}_1 + \mathbf{X}'_1\mathbf{R}^{12}\mathbf{y}_2 \\ \mathbf{X}'_2\mathbf{R}^{21}\mathbf{y}_1 + \mathbf{X}'_2\mathbf{R}^{22}\mathbf{y}_2 \\ \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{y}_1 + \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{y}_2 \\ \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{y}_1 + \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{y}_2 \end{bmatrix}$$

Where g^{ij} are the elements of the inverse of the additive variance-covariance matrix, \mathbf{G}^{-1} . Note that if the two traits are uncorrelated then \mathbf{R}^{12} and g^{12} will be zero. In this case the matrices in the equation above reduce to the single trait model [107].

$$\begin{bmatrix} \mathbf{X}'_1 r^{11} \mathbf{X}_1 & 0 & \mathbf{X}'_1 r^{11} \mathbf{Z}_1 & 0 \\ 0 & \mathbf{X}'_2 r^{22} \mathbf{X}_2 & 0 & \mathbf{X}'_2 r^{22} \mathbf{Z}_2 \\ \mathbf{Z}'_1 r^{11} \mathbf{X}_1 & 0 & \mathbf{Z}'_1 r^{11} \mathbf{Z}_1 + \mathbf{A}^{-1} g^{11} & 0 \\ 0 & \mathbf{Z}'_2 r^{22} \mathbf{X}_2 & 0 & \mathbf{Z}'_2 r^{22} \mathbf{Z}_2 + \mathbf{A}^{-1} g^{22} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{d}_1 \\ \hat{d}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 r^{11} \mathbf{y}_1 \\ \mathbf{X}'_2 r^{22} \mathbf{y}_2 \\ \mathbf{Z}'_1 r^{11} \mathbf{y}_1 \\ \mathbf{Z}'_2 r^{22} \mathbf{y}_2 \end{bmatrix}$$

Example:

The data below are the pre-weaning gain (WWG) and postweaning gain (PWG) for five beef calves, we interested to estimate the fixed effect (sex) and random effect (breeding values) for all animals:

Table 5.1. Pre-weaning gain and post-weaning gain (kg) for five beef calves.

Calf	Sex	Sire	Dam	WWG	PWG
4	Male	1	-	4.5	6.8
5	Female	3	2	2.9	5.0
6	Female	1	2	3.9	6.8
7	Male	4	5	3.5	6.0
8	Male	3	6	5.0	7.5

Using MBLUP analysis. Assume that the additive genetic covariance \mathbf{G} Matrix is:

$$\mathbf{G} = \begin{matrix} WWG & [20 & 18] \\ PWG & [18 & 40] \end{matrix}$$

And the residual covariance matrix is:

$$\mathbf{R} = \begin{matrix} WWG & [40 & 11] \\ PWG & [r_{21} & 30] \end{matrix}$$

Since the matrices are too large, the MME have not presented. The solutions of the system given below:

Table 5.2 Solutions of multiple traits model.

Effects	WWG	PWG
Sex		
1	4.361	6.800
2	3.397	5.880
Animal		
1	0.151	0.280
2	-0.015	-0.008
3	-0.078	-0.170
4	-0.010	-0.013
5	-0.270	-0.478
6	0.276	0.517
7	-0.316	-0.479
8	0.244	0.392

Where for fixed effect, 1 = male and 2 = female.

5.4 Unequal Design Matrices

In multivariate analysis, the model called unequal (unbalanced) if traits are affected through different fixed or random effects. For example, the multiple traits model of yields in different locations. However, in this case we can apply Henderson et al (1976) for evaluation [128].

Example:

We have fat yield in each parity as different traits shown in the table below:

Cow	Sire	Dam	HYS1	HYS2	FAT1	FAT2
4	1	2	1	1	201	280
5	3	2	1	2	150	200
6	1	5	2	1	160	190
7	3	4	1	1	180	250
8	1	7	2	2	285	300

For parity 1 and 2 the herd-year-season are HYS1 and HYS2, respectively; FAT1 and FAT2 fat yield in parity 1 and 2. To estimate the breeding values (FAT1 and FA2), assume the genetic and environmental parameters are:

$$\mathbf{G} = \begin{bmatrix} 35 & 28 \\ 28 & 30 \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} 65 & 27 \\ r_{21} & 70 \end{bmatrix}$$

For the inverses are:

$$\mathbf{G}^{-1} = \begin{bmatrix} 0.113 & -0.105 \\ -0.105 & 0.132 \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} 0.018 & -0.007 \\ -0.007 & 0.017 \end{bmatrix}$$

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{d}_1 \\ \widehat{d}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 \mathbf{R}^{11} \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{R}^{12} \mathbf{X}_2 & \mathbf{X}'_1 \mathbf{R}^{11} \mathbf{Z}_1 & \mathbf{X}'_1 \mathbf{R}^{12} \mathbf{Z}_2 \\ \mathbf{X}'_2 \mathbf{R}^{12} \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{R}^{22} \mathbf{X}_2 & \mathbf{X}'_2 \mathbf{R}^{21} \mathbf{Z}_1 & \mathbf{X}'_2 \mathbf{R}^{22} \mathbf{Z}_2 \\ \mathbf{Z}'_1 \mathbf{R}^{11} \mathbf{X}_1 & \mathbf{Z}'_1 \mathbf{R}^{12} \mathbf{X}_2 & \mathbf{Z}'_1 \mathbf{R}^{11} \mathbf{Z}_1 + \mathbf{A}^{-1} g^{11} & \mathbf{Z}'_1 \mathbf{R}^{12} \mathbf{Z}_2 + \mathbf{A}^{-1} g^{12} \\ \mathbf{Z}'_2 \mathbf{R}^{21} \mathbf{X}_1 & \mathbf{Z}'_2 \mathbf{R}^{22} \mathbf{X}_2 & \mathbf{Z}'_2 \mathbf{R}^{21} \mathbf{Z}_1 + \mathbf{A}^{-1} g^{21} & \mathbf{Z}'_2 \mathbf{R}^{22} \mathbf{Z}_2 + \mathbf{A}^{-1} g^{22} \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} \mathbf{X}'_1 \mathbf{R}^{11} \mathbf{y}_1 + \mathbf{X}'_1 \mathbf{R}^{12} \mathbf{y}_2 \\ \mathbf{X}'_2 \mathbf{R}^{21} \mathbf{y}_1 + \mathbf{X}'_2 \mathbf{R}^{22} \mathbf{y}_2 \\ \mathbf{Z}'_1 \mathbf{R}^{11} \mathbf{y}_1 + \mathbf{Z}'_1 \mathbf{R}^{12} \mathbf{y}_2 \\ \mathbf{Z}'_2 \mathbf{R}^{21} \mathbf{y}_1 + \mathbf{Z}'_2 \mathbf{R}^{22} \mathbf{y}_2 \end{bmatrix}$$

The matrix X_1 relates HYS1 effect and X_2 relates to HYS2 effect. The transposes of these unequal matrices are:

$$X_1' = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad X_2' = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

The solutions of MME are:

Effects	FAT1	FAT2
HYS		
1	175.7	243.2
2	219.6	240.6
Animal		
1	8.969	8.840
2	-2.999	-2.777
3	-5.970	-6.063
4	11.754	11.658
5	-16.253	-15.824
6	-17.314	-15.719
7	8.690	8.138
8	22.702	20.931

CHAPTER 6

ANALYSIS

6.1 Introduction

The variance components and heritability estimates for familial breast cancer incidence were calculated with a threshold model using the recorded binary observations linked to the probit function (transformation). This chapter will present these programs and the steps used for analysis. Four programs in *Fortran 90/95* and *R software* were used to analyze the data. All graphs in this chapter were created using *R software*, in order to show the distributions and the autocorrelation functions of variance components and heritability estimates. We estimated heritability using two methods; first, we used the MCMCglmm Package in R for phenotypic data only and second, we included genomic data (SNP) using the THRGIBBS1F90 program.

6.2 The MCMCglmm R Package

For the binary phenotype variable, the model is defined on an underlying latent variable:

$$\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\varepsilon} \quad (6.1)$$

Where \mathbf{X} and \mathbf{Z} are design matrices relating to fixed and random effects, respectively.

These matrices have associated parameter vectors $\boldsymbol{\beta}$ and \mathbf{d} , while $\boldsymbol{\varepsilon}$ is a residual vector.

The distribution of vectors \mathbf{d} and $\boldsymbol{\varepsilon}$ are assumed to be multivariate normal distribution as:

$\mathbf{d} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$. The matrices \mathbf{G} and \mathbf{R} are (co)variances of the random effects and residuals, respectively. The structure form of \mathbf{G} in MCMCglmm is:

$$\mathbf{G} = (\mathbf{V}_1 \otimes \mathbf{A}_1) \oplus (\mathbf{V}_2 \otimes \mathbf{A}_2) \oplus \dots \quad (6.2)$$

Typically, the (co)variance matrices (\mathbf{V}) are low-dimensional and the structured matrices are (\mathbf{A}) high dimensional. The terms are separated by a direct sum (\oplus) as component terms and each component term is formed through the Kronecker product (\otimes). So if we have two component terms, we can write matrix \mathbf{G} as:

$$\mathbf{G} = \begin{bmatrix} (\mathbf{V}_1 \otimes \mathbf{A}_1) & \mathbf{0} \\ \mathbf{0} & (\mathbf{V}_2 \otimes \mathbf{A}_2) \end{bmatrix}$$

By the same manner, we can get \mathbf{R} :

$$\mathbf{R} = (\mathbf{V}_{e1} \otimes \mathbf{I}_1) \oplus (\mathbf{V}_{e2} \otimes \mathbf{I}_2) \oplus \dots \quad (6.3)$$

Where \mathbf{V}_{e1} , \mathbf{V}_{e2} are residual variances and \mathbf{I}_1 , \mathbf{I}_2 are identity matrices [129].

To fit the binary data y_i for one trait, we use a probit link and a Bernoulli distribution:

$$y_i = B(\text{probit}^{-1}(\lambda_i))$$

Since we want to calculate heritability for the model (6.1), it is necessary to take into account a supplementary source of variance coming from the probit link.

$$\hat{h}^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2 + 1}$$

This is justified by the fact that, to be strictly equivalent to a threshold model (where y is 1 if $\lambda > 0$), we need to include the “variance” of the link transformation into the total variance, which is 1 for a probit link [130].

In MCMCglmm, the prior distribution of the (co)variances is an inverse-Wishart and a normal prior for the fixed effect. For a single variance component, an inverse-Gamma distribution with parameters nu and v (*inverse – Gamma* $(\frac{nu}{2}, \frac{nu \times v}{2})$) is a

common choice. The possible set of parameters would be $nu=0.002$ and $v=1$, which is actually *inverse – Gamma*(0.001,0.001)[131] because $\alpha = \frac{nu}{2}$, $\beta = \frac{nu v}{2}$. In a binary variable, the residual variance (V_R) will be fixed to 1 and an estimate additive variance (V_G). Also, when we want to estimate the heritability of binary data, it is advised to use *chi-square* prior [132].

Our estimation of heritability is 0.03947 using *inverse – Gamma*(0.001,0.001) which is the common prior. Using *chi-square* with 1 degree of freedom as prior, heritability is 0.3369; refer to the R code in Appendix E. The diagnostic of the results from both priors is described as follows.

Before we consider our estimation as a final estimation, we need to check the convergence and autocorrelation of samples. In MCMCglmm output, there are two main components (model\$Sol and model\$VCV) where Sol is the posterior distribution solution, including fixed effects, and VCV is a posterior distribution of (co)variance matrices. First, let us look at the trace of samples from *inverse – Gamma*(0.001,0.001) prior (Figure 6.1 and Figure 6.2).

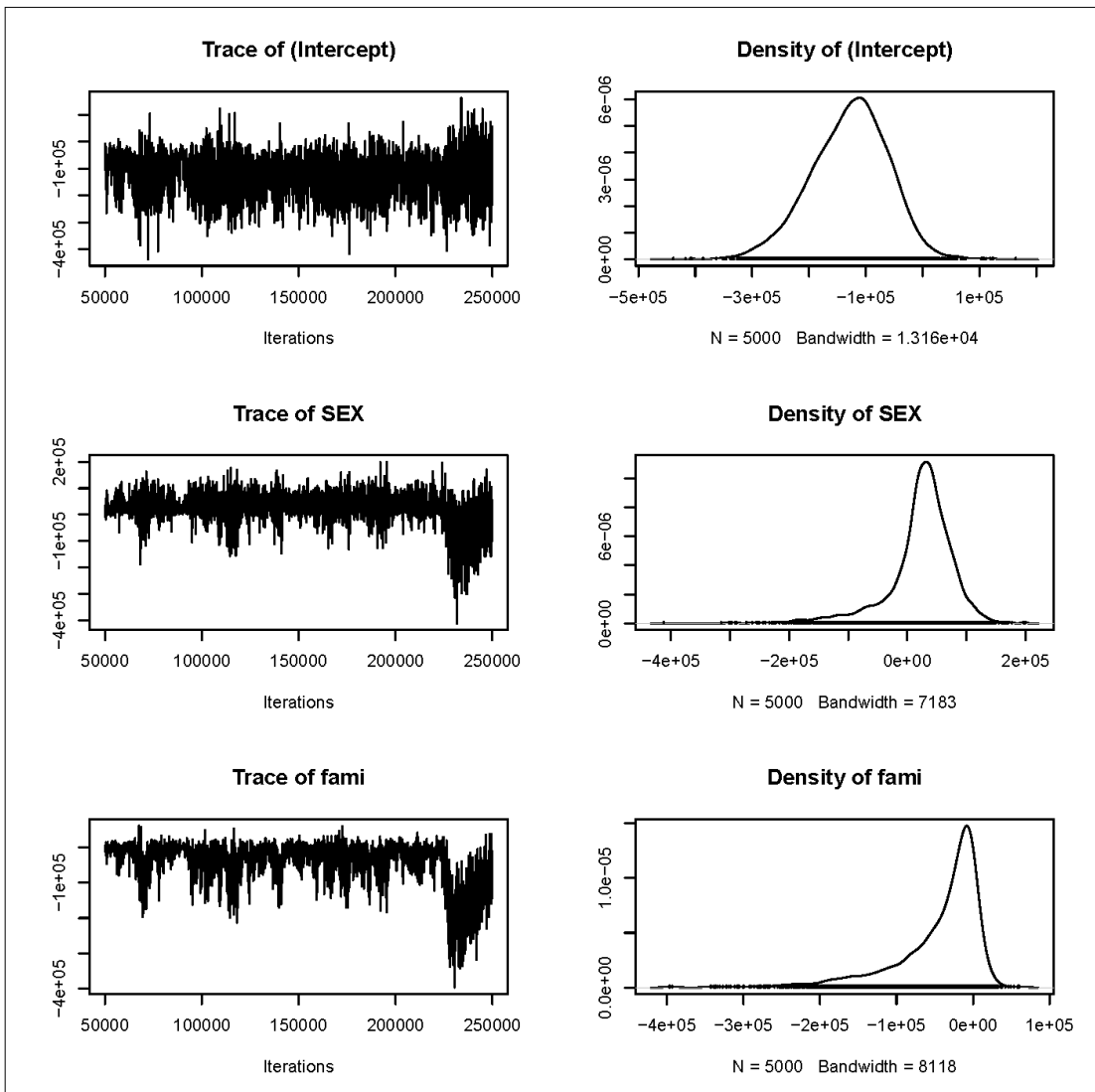


Figure 6.1 Trace of the fixed effects using inverse-Gamma (0.001, 0.001).

As we can see for each graph of the samples (Figures 6.1 and 6.2), the values are widely spread. While the values in the graphs of samples (Figures 6.3 and 6.4) are spread (fluctuated) in a small range, these samples were generated using *chi-square* with 1 degree of freedom as prior.

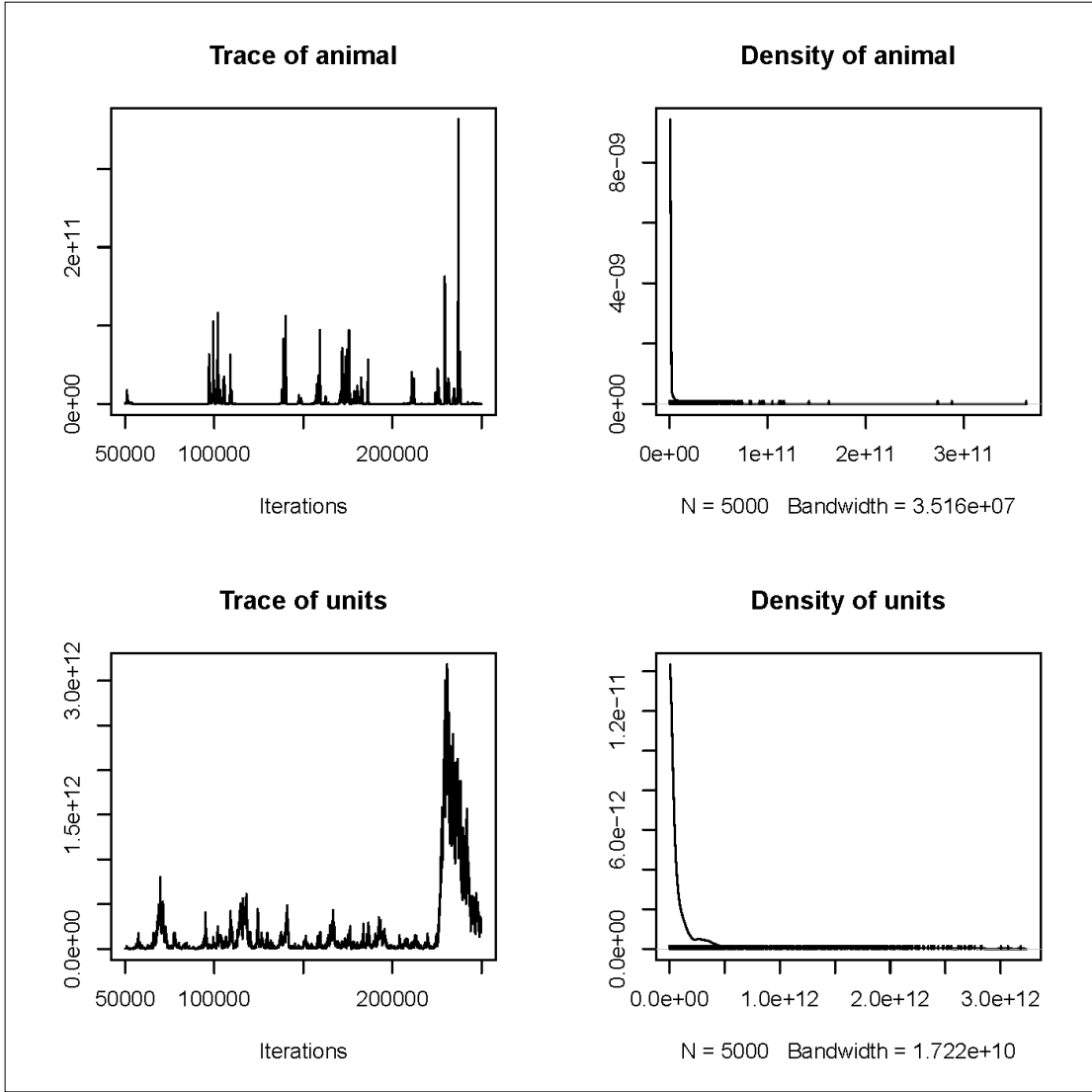


Figure 6.2 Trace of the variance components inverse-Gamma (0.001, 0.001).

Also, little autocorrelation is found in the samples that were generated from *inverse – Gamma*(0.001,0.001), especially in the family, animal and units (Table 1.1). However, the autocorrelation is reasonable for the sample of additive variance generated from *Chi-square* (6.2). Note that we fixed the residual variance to 1 in this prior (Figure 6.3 and Figure 6.4).

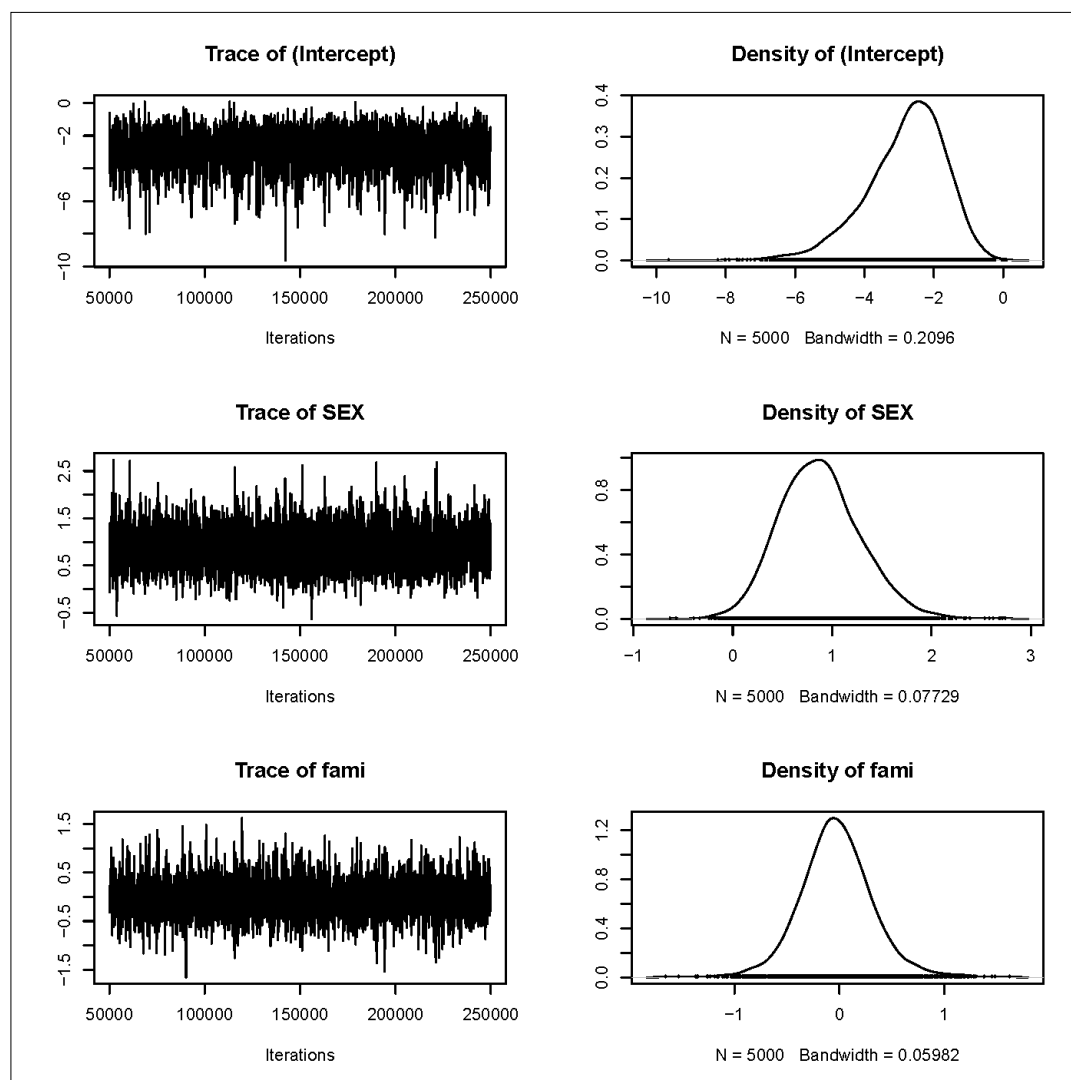


Figure 6.3 Trace of the fixed effects using chi-square.

Table 1.1 The autocorrelation of samples from *inverse – Gamma*.

	Intercept	sex	Family	Animal
Lag 0	1.000000	1.000000	1.000000	1.000000
Lag 40	0.107780	0.227881	0.581473	0.782403
Lag 200	0.077046	0.168276	0.520311	0.531336
Lag 400	0.045379	0.141118	0.433896	0.335605
Lag 2000	0.032968	0.038969	0.125210	0.024426

Table 6.2 The autocorrelation of samples from Chi-square.

	Intercept	sex	Family	Animal
Lag 0	1.000000	1.000000	1.000000	1.000000
Lag 40	0.171016	0.072454	0.007925	0.438784
Lag 200	0.006327	0.004607	-0.010548	0.043452
Lag 400	0.011712	0.015504	0.001981	-0.004161
Lag 2000	-0.008358	0.004110	-0.013991	0.010373

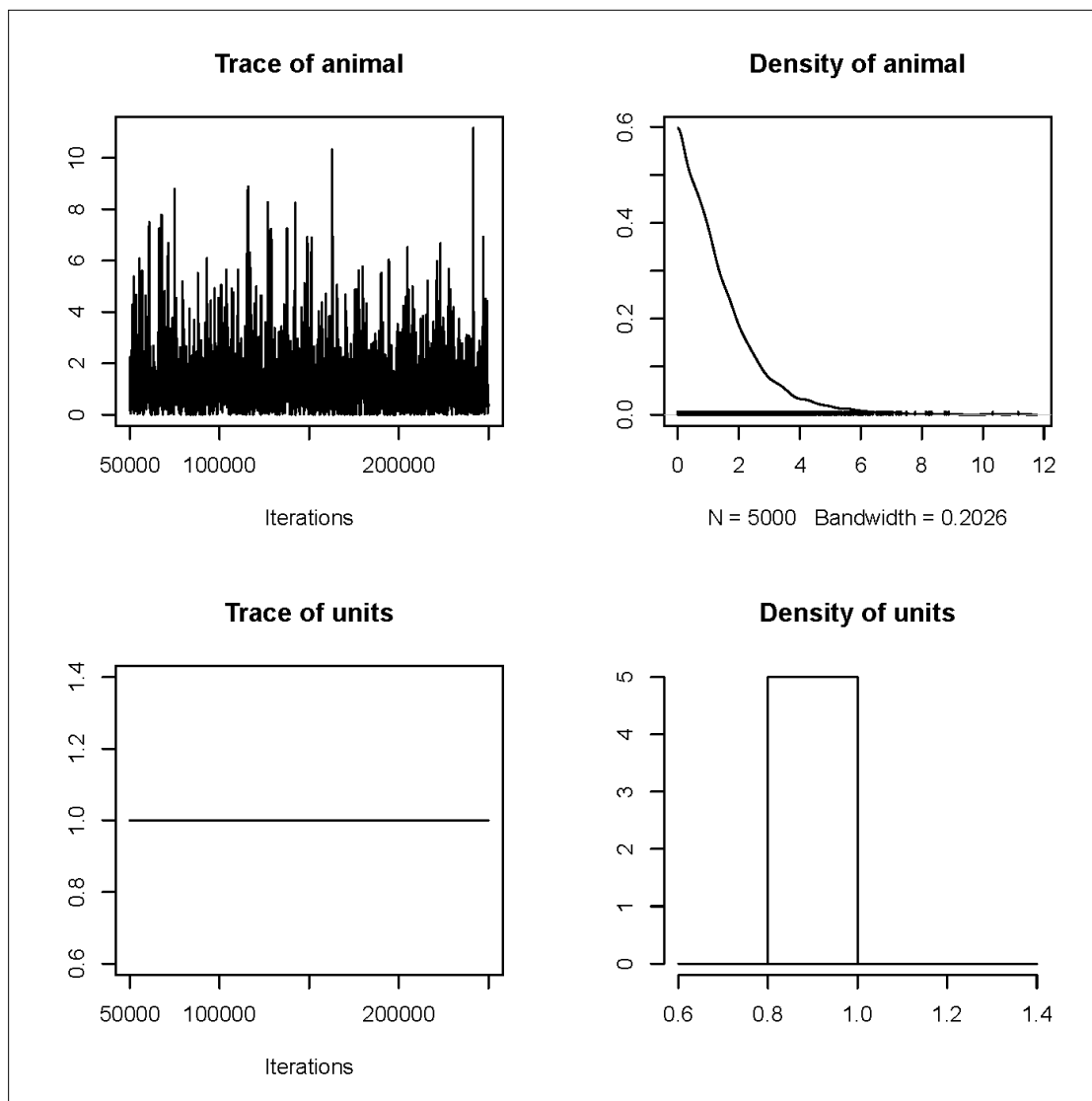


Figure 6.4. Trace of the variance component using chi-square.

In MCMCglmm, the convergence is often very fast. Therefore, we used the Heidelberg stationarity test to see if the convergence is reasonable; if the p-value exceeds 0.05, that means the sample is from a stationary distribution.

Table 6.3 Convergence test of samples generated from inverse-Gamma.

	Stationarity test	Start iteration	p-value
animal	passed	1	0.513
Units	passed	1	0.399

Table 6.4 Convergence test of sample generated from Chi-square.

	Stationarity test	Start iteration	p-value
animal	passed	1	0.435

6.3 BLUPF90 in Fortran 90/95

The analysis in this research was also completed using BLUPF90 in Fortran 90/95. These programs were originally developed as exercises for a class taught by Ignacy Misztal at the University of Georgia [133]. Gradually, they have been upgraded and improved by many contributors. This a family of programs for mixed-models computations. The programs can execute data conditioning, estimate variances using several methods, compute BLUP for very large data sets, calculate approximate accuracy, and use SNP data for improved accuracy of breeding values for GWAS studies [134].

The programs have been designed with three goals:

- i. Flexibility to support a large set of models found in animal breeding applications.
- ii. Software simplicity to abate errors and enable modifications.

- iii. Productivity at the algorithmic level.

The four programs used in this study are RENUMF90, THRGIBBSxF90, POSTGIBBSF90 and POSTGSF90 [133].

- a. RENUMF90

This is a program for renumbering in the BLUPF90 family of programs. It supports multiple traits, dissimilar effects per trait, alphanumeric and numeric fields. The program provides data statistics, performs comprehensive pedigree checking, and supports unknown parent groups.

- b. THRGIBBS1F90

This program performs a Gibbs sampler for threshold-linear mixed effect models involving multiple categorical and linear variables. The original program THRGIBBSF90 was composed by DeukHwan Lee in 2001, and rewritten by Shogo Tsuruta in 2004. This program should be used after the renumbering program RENUMF90.

- c. POSTGIBBSF90

This program was designed to calculate posterior means, standard deviation and diagnosis of convergence. The program reads gibbs_samples and fort.99 files from Gibbs sampling (THRGIBBS1F90) programs. The output files from this program are postgibbs_samples, postmean, postsd and postout.

- d. POSTGSF90

This program is used to obtain the prediction of SNP effects. The output files from this program are snp_sol, chrsnp, chrsnpvar and snp_pred (the second file contains the values of SNP effects used in Manhattan plots).

6.3.1 The THRGIBBS1F90

The joint posterior distribution was estimated using the MCMC method based on Gibbs sampling with the THRGIBBS1F90 software [134]. The number of iterations is 250000, with 50000 samples discarded as burn-in for convergence, with the remaining samples being thinned using intervals of 20 samples. Ten thousand (10000) samples were stored and used to calculate the posterior features of interest, such as posterior means and standard deviations. The derived variance components and heritability estimates for familial breast cancer incidence are shown in Table 6.5.

Table 6.5 Variance components and heritability estimates.

	Additive Effect		Residuals variance		Heritability	
Mean	0.0590174		0.1471117		0.2812776	
SD	0.02835517		0.02496394		0.1166282	
Confidence Interval	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
	0.0147	0.1225	0.1019	0.19910	0.0765	0.5187

The heritability coefficient is an estimate of the amount of variation in a phenotypic trait in a population due to genetic variation among individuals in that population. Here it is 28% variation in the response variable, as explained by genetic variations (SNPs).

6.4 Plots of Estimators

If we look at the histogram graphs (see Figure 6.5), each graph clearly shows more data around the average, where the highest "bump" is located, indicating this as "symmetrical." This is a good sign that the averages (for each estimator) are approximately in the center of the data.

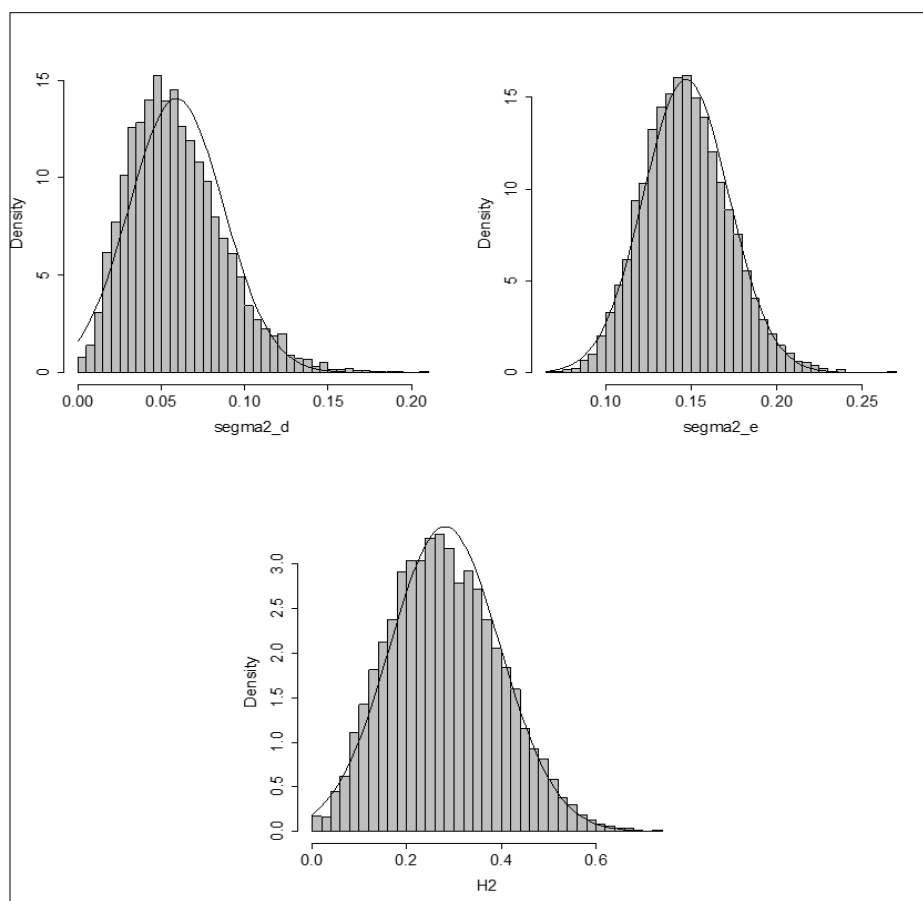


Figure 6.5 Histogram with density of the variances and heritability estimates.

6.5 History and Autocorrelation Function Plots

The series of estimated values show a consistent trend, as noted in Figure 6.6. The series of additive variance estimates fluctuate around the same value, which is approximately 0.059. Likewise, the same behavior is seen for the other two estimated values, where the residuals variance fluctuates around 0.147 and heritability is close to 0.28. Furthermore, the autocorrelation function (ACF) measures the correlation between pairs of random variables. The optimal plot of ACF decays reasonably rapidly to zero, either from above or below. We can see that all ACFs (Figure 6.7) have decayed to zero before lag 10, which indicates that the series of samples are independent.

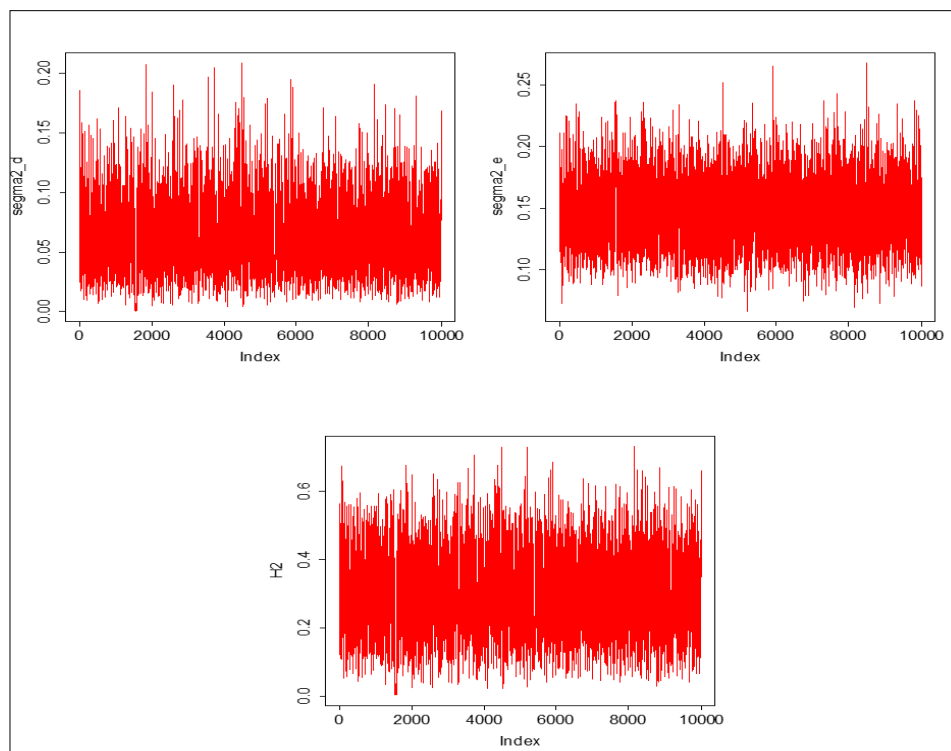


Figure 6.6 Sample series plots of the variances and heritability estimates.

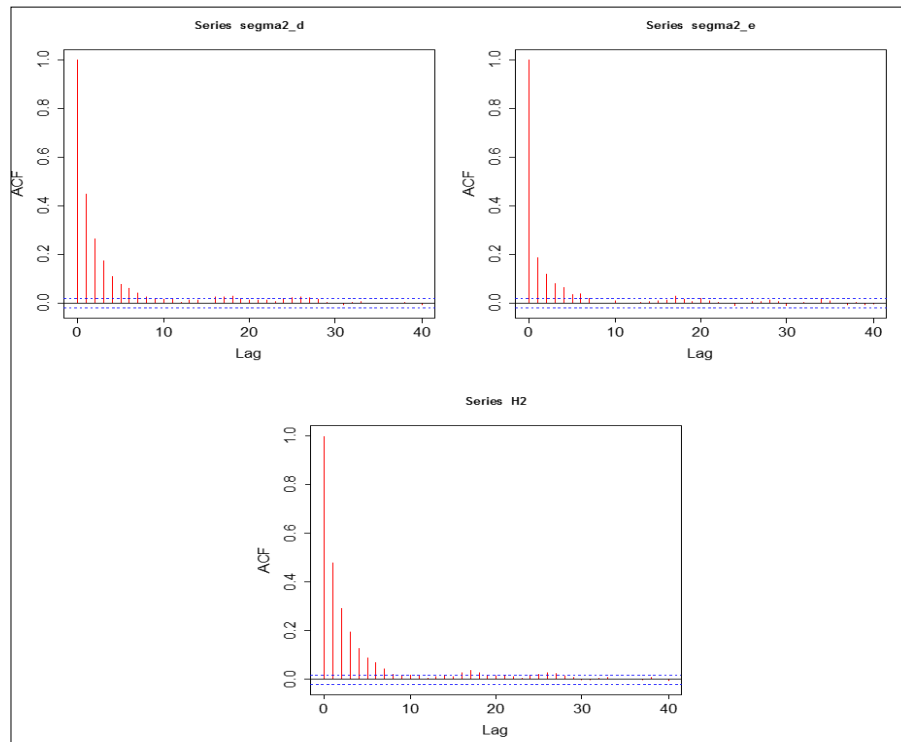


Figure 6.7 Autocorrelation functions of the variances and heritability estimates.

It is essential to check the convergence to determine whether we are as close to the actual value as possible. The burn-in value, at 50000, was sufficient to reach convergence. Also, to avoid an autocorrelation issue, we used a thinning interval of 20 and noted that the autocorrelation function went to zero after a few lags, which was good.

6.6 Manhattan Plot

GWAS analysis typically reports a visualization of genome-wide association of SNPs using a Manhattan plot. The SNP effects are displayed as their absolute values (y-axis) against SNP location (x-axis: chromosomes and SNP positions). The higher dots represent high SNP effects at a given position associated with familial breast cancer [96]. In this context, gene mapping based on a Manhattan plot (Figure 6.8) showed that there are some areas on the genome (chromosomes 1, 2, 4, 8, 14 and 16) that may include candidate genes associated with breast cancer incidence and susceptibility.

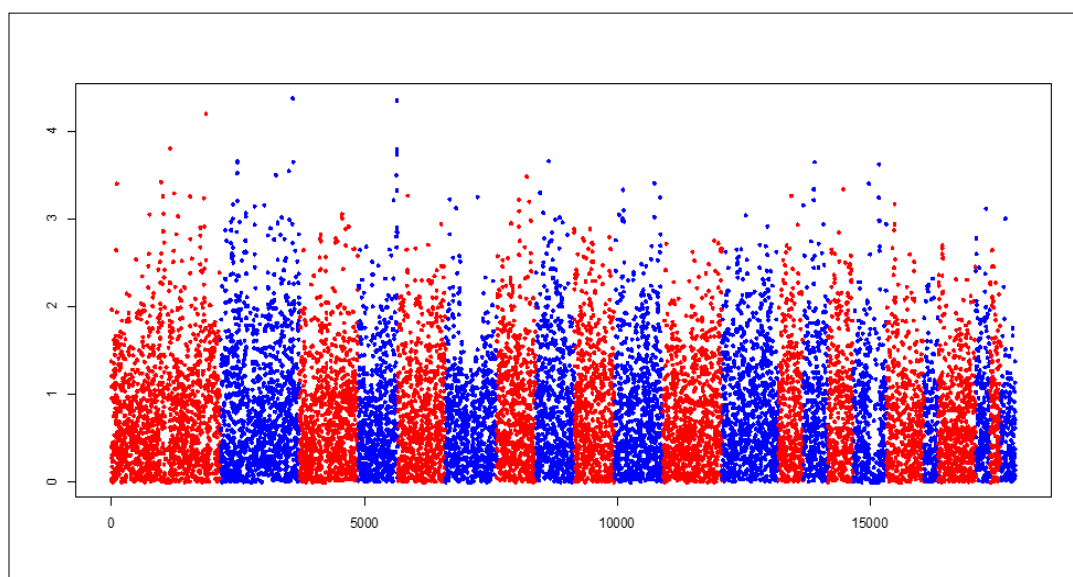


Figure 6.8 Manhattan Plot of SNPs effects.

In the Manhattan plot (Figure 6.8), we do not use thresholds, or p-values, because two problems may arise when we use a single SNP regression. First, thousands of tests will be run and SNP effects will be over-estimated. Second, it is difficult to define a genomic region with the true mutation [135, 136]. To avoid these two problems, we fitted SNP effects simultaneously.

The 20 markers with the highest effects are shown in Table . These SNPs explained more than 3.5% of the genetic variance, which indicates that this group of SNPs, and hence their genes, are highly associated with the incidence of this disease.

Table 6.6 Ordered SNP effects with locations.

No	SNP Effects	CHROM	POS	GENE
1	4.371465	2	223793480	ACSL3
2	4.349453	4	187534542	FAT1
3	4.198865	1	225611661	LBR
4	3.800145	1	150999863	PRUNE
5	3.787771	4	187534375	FAT1
6	3.736460	4	187535169	FAT1
7	3.661992	8	27147716	TRIM35
8	3.661992	8	27147939	TRIM35
9	3.653323	2	44547909	SLC3A1
10	3.649030	2	224569668	PREPL
11	3.647262	14	50094913	NEMF/ DNAAF2
12	3.641489	2	44547574	SLC3A1
13	3.641489	2	44550290	PREPL
14	3.641489	2	44570804	PREPL
15	3.641489	2	44617324	CAMKMT
16	3.617063	16	71668800	MARVELD3
17	3.544095	2	217123958	MARCH4
18	3.51585	2	44550233	PREPL
19	3.498177	2	178936373	PDE11A
20	3.496521	4	187077393	FAM149A

CHAPTER 7

DISCUSSION AND CONCLUSION

Among various types of cancer, breast cancer has been reported as the second most deadly cancer in the world. In addition, patients who have been diagnosed as positive for this disease face the issue that no treatment is currently available to prevent or reduce the occurrence of this cancer [4]. For this reason, more consideration is being given to genomic studies that are based on advanced statistical models and methodologies to track the cause of this disease, with promising progress being achieved [26]. Finding a genetic component, as well as a genomic region, that could be associated with breast cancer will facilitate the diagnosis of those most likely to be infected with breast cancer in the future. This does not just raise caution, but will also allow for early treatment, which may lead to preventing the disease from progressing to advanced stages. In this study, we estimated the heritability of breast cancer incidence based on the presence of breast cancer in 167 people, of which 22 had genomic information. Since the response variable was binary, a threshold model was used [21]. In order to maximize the usage of this small data set, we decided to combine both the pedigree data and the genomic data in one analysis [17]. This approach has been widely used with accurate results [18]. The challenge was to scale the genomic data to the pedigree data, so the two sources of information can be reasonably combined.

The heritability estimate for breast cancer incidence obtained in this study indicates that genetic tracking can be used to investigate the probability of the incidence of the disease. Furthermore, a genetic component becomes a simple indicator that a genomic region on the DNA is associated with the incidence of this disease. For any trait with a non-zero heritability, the genetic effects must be located somewhere on the chromosomes [137].

The heritability estimates showed that genetics plays a key role in the incidence of this type of cancer. In other words, there is a genetic component that can be further investigated to determine the genes or genomic regions that might be associated with breast cancer. It is then a matter of locating that genetic effect and genotyping the appropriate DNA.

Since the heritability estimate was higher than zero, we decided to apply a GWAS, which will predict the effect of each marker (SNP) we have in the genomic data and provide the extent of the genetic variation that can be explained by markers. The GWAS revealed that several genes, with small effects for each, might be responsible for breast cancer incidence. Specifically, SNPs on chromosomes 1, 2, 4, 8, 14 and 16 could be related to the incidence of the disease. Some of these results confirm findings from other similar studies [138]. Seven genes were mentioned in previous studies as related to breast cancer. The ACSL3 (acyl-CoA synthetase long chain family member 3) gene may be one that enhances the amount of cytotoxicity if it gets suppressed, while SREBPs are responsible for aberrant proliferation of breast cancer cells [139]. FAT1 (FAT atypical cadherin 1) is one of the FAT family of genes which are vital to suppress cancer cells based on their ability of homozygous deletion; they can also help to determine oncogenic status [140]. The gene PRUNE (prune exopolyphosphatase 1) is related to metastasis in breast cancer; when h-PRUNE gets suppressed by dipyrindamole, the connection with nm23-H1 increases, which leads to increasing the cellular movement in the metastasis process [141]. TRIM35 (tripartite motif containing 35) cooperates in the process of the formation and increase in the size of cell cancers [142]. SLC3A1 (solute carrier family 3 member 1) is crucial because it promotes the cysteine uptake and antioxidant N-acetylcysteine. Also it is considered therapeutic in the treatment of breast cancer [143]. A strong relationship has been noted

between SNPs and breast cancer risk, with one of these SNPs called "rs8410," located in the PREPL (prolyl endopeptidase-like) gene [144]. Gene MARVELD3 (MARVEL domain containing 3) is very helpful for paracellular ion connections [145].

We also found new seven genes that could be related to breast cancer. These genes are: LBR (lamin B receptor); NEMF (nuclear export mediator factor); DNAAF2 (dynein axonemal assembly factor 2); CAMKMT (calmodulin-lysine N-methyltransferase); MARCH4 (membrane associated ring-CH-type finger 4); PDE11A (phosphodiesterase 11A) and FAM149A (family with sequence similarity 149 member A).

The results of this study demonstrate that breast cancer is a complex disease, probably controlled by many genes with small effects. Finding genes that show major effects is uncommon for disease traits. Such traits are usually influenced by many biological components and function factors controlled by a large number of genes. Furthermore, disease traits are affected by environmental factors.

The main limitation of this work was the sample size. Collecting a larger sample, as well as pedigree information in humans, is not an easy task. Further studies depend upon the availability of larger datasets, which may then reveal more information on this complex trait. It may also help to reduce the number of genes potentially associated with breast cancer. Also, including the SNPs that have higher effects through fine mapping and pathway analysis may uncover more knowledge about breast cancer incidence and help in determining various ways to reduce and prevent the disease in the future [146].

Future works:

- 1) Estimating heritability coefficient using Aguilar et al (2010) approach in R software.
- 2) Increase the size of data by adding family or more available to get more knowledge about breast cancer.
- 3) Making check these SNPs that have higher effects in pathway analysis to see if they biological related to familial breast cancer.

LITERATURE CITED

1. World Health Organization. *World Cancer Day 2017*. 2017; Available from: <http://www.who.int/cancer/world-cancer-day/2017/en/>.
2. American Cancer Society. *Cancer Facts & Figures 2016*. 2017; Available from: <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2016.html>.
3. Gill, J., R. Sullivan, and D. Taylor, *Overcoming cancer in the 21st century*. 2015: UCL School of Pharmacy.
4. Ferlay, J., H.R. Shin, F. Bray, D. Forman, C. Mathers, and D.M. Parkin, *Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008*. International journal of cancer, 2010. **127**(12): p. 2893-2917.
5. Wright, S., *The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs*. Proceedings of the National Academy of Sciences, 1920. **6**(6): p. 320-332.
6. Stoltenberg, S.F., *Coming to terms with heritability*. Genetica, 1997. **99**(2): p. 89-96.
7. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era—concepts and misconceptions*. Nature Reviews Genetics, 2008. **9**(4): p. 255-266.
8. Taylor, A., D. Harnden, C. Arlett, S. Harcourt, A. Lehmann, S.u. Stevens, and B. Bridges, *Ataxia telangiectasia: a human mutation with abnormal radiation sensitivity*. Nature, 1975. **258**(5534): p. 427-429.
9. Roberts, S.A., A.R. Spreadborough, B. Bulman, J.B. Barber, D. Evans, thG, thR, and D. Scott, *Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer?* The American Journal of Human Genetics, 1999. **65**(3): p. 784-794.
10. Boyd, N.F., G.S. Dite, J. Stone, A. Gunasekara, D.R. English, M.R. McCredie, G.G. Giles, D. Trichler, A. Chiarelli, and M.J. Yaffe, *Heritability of mammographic density, a risk factor for breast cancer*. New England Journal of Medicine, 2002. **347**(12): p. 886-894.
11. Brand, J.S., K. Humphreys, D.J. Thompson, J. Li, M. Eriksson, P. Hall, and K. Czene, *Volumetric mammographic density: heritability and association with breast cancer susceptibility loci*. Journal of the National Cancer Institute, 2014. **106**(12): p. dju334.
12. Varghese, J.S., P.L. Smith, E. Folkard, J. Brown, J. Leyland, T. Audley, R.M. Warren, M. Dowsett, D.F. Easton, and D.J. Thompson, *The heritability of mammographic breast density and circulating sex-hormone levels; two independent breast cancer risk factors*. Cancer Epidemiology and Prevention Biomarkers, 2012: p. cebp. 0789.2012.
13. Speed, D., G. Hemani, M.R. Johnson, and D.J. Balding, *Improved heritability estimation from genome-wide SNPs*. The American Journal of Human Genetics, 2012. **91**(6): p. 1011-1021.
14. Cheng, Q., X. Gao, and R. Martin, *Exact prior-free probabilistic inference on the heritability coefficient in a linear mixed model*. Electronic Journal of Statistics, 2014. **8**(2): p. 3062-3076.
15. Heckerman, D., D. Gurdasani, C. Kadie, C. Pomilla, T. Carstensen, H. Martin, K. Ekoru, R.N. Nsubuga, G. Ssenyomo, and A. Kamali, *Linear mixed model for heritability estimation that explicitly addresses environmental variation*. Proceedings of the National Academy of Sciences, 2016. **113**(27): p. 7377-7382.

16. Fong, C., D.C. Ko, M. Wasnick, M. Radey, S.I. Miller, and M. Brittnacher, *GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis*. Bioinformatics, 2010. **26**(4): p. 560-564.
17. Misztal, I., A. Legarra, and I. Aguilar, *Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information*. Journal of Dairy Science, 2009. **92**(9): p. 4648-4655.
18. Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor, *Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score*. Journal of Dairy Science, 2010. **93**(2): p. 743-752.
19. El-Dien, O.G., B. Ratcliffe, J. Klápště, I. Porth, C. Chen, and Y.A. El-Kassaby, *Implementation of the Realized Genomic Relationship Matrix to Open-Pollinated White Spruce Family Testing for Disentangling Additive from Nonadditive Genetic Effects*. G3: Genes | Genomes | Genetics, 2016. **6**(3): p. 743-753.
20. Wen, W., X.-o. Shu, X. Guo, Q. Cai, J. Long, M.K. Bolla, K. Michailidou, J. Dennis, Q. Wang, and Y.-T. Gao, *Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry*. Breast Cancer Research, 2016. **18**(1): p. 124.
21. Wen, H., Y.C. Kim, C. Snyder, F. Xiao, E.A. Fleissner, D. Becirovic, J. Luo, B. Downs, S. Sherman, and K.H. Cowan, *Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for BRCAx familial breast cancer*. BMC cancer, 2014. **14**(1): p. 470.
22. Lynch, H., H. Wen, Y.C. Kim, C. Snyder, Y. Kinarsky, P.X. Chen, F. Xiao, D. Goldgar, K.H. Cowan, and S.M. Wang, *Can Unknown Predisposition in Familial Breast Cancer be Family-Specific?* The breast journal, 2013. **19**(5): p. 520-528.
23. Institute, N.C., *What Is Cancer?* National Cancer Institute, 2015: p. <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
24. Broca, P., *Traite des tumeurs [Treatise on tumors], vols. I and II*. Paris: Asselin, 1866.
25. Hejmadi, M., *Introduction to cancer biology*. 2009: Bookboon.
26. Lynch, H., C. Synder, and S.M. Wang, *Considerations for comprehensive assessment of genetic predisposition in familial breast cancer*. The breast journal, 2015. **21**(1): p. 67-75.
27. van der Groep, P., E. van der Wall, and P.J. van Diest, *Pathology of hereditary breast cancer*. Cellular oncology, 2011. **34**(2): p. 71-88.
28. Lynch, H.T., A.J. Krush, H.M. Lemon, A.R. Kaplan, P.T. Condit, and R.H. Bottomley, *Tumor variation in families with breast cancer*. Jama, 1972. **222**(13): p. 1631-1635.
29. Hall, J.M., M.K. Lee, B. Newman, J.E. Morrow, and L.A. Anderson, *Linkage of early-onset familial breast cancer to chromosome 17q21*. Science, 1990. **250**(4988): p. 1684.
30. Lenoir, G., H. Lynch, P. Watson, T. Conway, J. Lynch, S. Narod, and J. Feunteun, *Familial breast-ovarian cancer locus on chromosome 17q12-q23*. The Lancet, 1991. **338**(8759): p. 82-83.
31. Miki, Y. and J. Swensen, *Breast and ovarian cancer susceptibility gene brca1*. Science, 1994. **266**: p. 7.
32. Gene, S., *Breast and ovarian cancer susceptibility gene brca1*. Science, 1994. **266**: p. 7.
33. Wooster, R., S.L. Neuhausen, J. Mangion, Y. Quirk, D. Ford, N. Collins, K. Nguyen, S. Seal, T. Tran, and D. Averill, *Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13*. Science, 1994. **265**(5181): p. 2088-2091.
34. Ford, D., D. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. Bishop, B. Weber, G. Lenoir, and J. Chang-Claude, *Genetic heterogeneity and penetrance analysis of the*

- BRCA1 and BRCA2 genes in breast cancer families.* The American Journal of Human Genetics, 1998. **62**(3): p. 676-689.
35. Easton, D.F., *How many more breast cancer predisposition genes are there?* Breast Cancer Research, 1999. **1**(1): p. 14.
 36. Smith, P., L. McGuffog, D.F. Easton, G.J. Mann, G.M. Pupo, B. Newman, G. Chenevix-Trench, C. Szabo, M. Southey, and H. Renard, *A genome wide linkage search for breast cancer susceptibility genes.* Genes, Chromosomes and Cancer, 2006. **45**(7): p. 646-655.
 37. Melchor, L., E. Honrado, J. Huang, S. Álvarez, T.L. Naylor, M.J. García, A. Osorio, D. Blesa, M.R. Stratton, and B.L. Weber, *Estrogen receptor status could modulate the genomic pattern in familial and sporadic breast cancer.* Clinical Cancer Research, 2007. **13**(24): p. 7305-7313.
 38. Bush, W.S. and J.H. Moore, *Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
 39. Klein, R.J., C. Zeiss, E.Y. Chew, J.-Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, and S.T. Mayne, *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-389.
 40. Manolio, T.A. and F.S. Collins, *The HapMap and genome-wide association studies in diagnosis and therapy.* Annual review of medicine, 2009. **60**: p. 443-456.
 41. Haiman, C.A., G.K. Chen, C.M. Vachon, F. Canzian, A. Dunning, R.C. Millikan, X. Wang, F. Ademuyiwa, S. Ahmed, and C.B. Ambrosone, *A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer.* Nature genetics, 2011. **43**(12): p. 1210-1214.
 42. Figueroa, J.D., M. Garcia-Closas, M. Humphreys, R. Platte, J.L. Hopper, M.C. Southey, C. Apicella, F. Hammet, M.K. Schmidt, and A. Broeks, *Associations of common variants at 1p11. 2 and 14q24. 1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: findings from the Breast Cancer Association Consortium.* Human molecular genetics, 2011. **20**(23): p. 4693-4706.
 43. Stevens, K.N., C.M. Vachon, A.M. Lee, S. Slager, T. Lesnick, C. Olswold, P.A. Fasching, P. Miron, D. Eccles, and J.E. Carpenter, *Common breast cancer susceptibility loci are associated with triple negative breast cancer.* Cancer research, 2011: p. canres. 1266.2011.
 44. Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell. new york: Garland science; 2002.* Classic textbook now in its 5th Edition, 2002.
 45. Watson, J.D. and F.H. Crick. *The structure of DNA.* in *Cold Spring Harbor symposia on quantitative biology.* 1953. Cold Spring Harbor Laboratory Press.
 46. Masood, E., *As consortium plans free SNP map of human genome.* 1999, Nature Publishing Group.
 47. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
 48. Xu, H., X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen, *FastUniq: a fast de novo duplicates removal tool for paired short reads.* PloS one, 2012. **7**(12): p. e52249.
 49. DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, and M. Hanna, *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. **43**(5): p. 491-498.
 50. Auwera, G.A., M.O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, and J. Thibault, *From FastQ data to high-confidence variant*

- calls: the genome analysis toolkit best practices pipeline*. Current protocols in bioinformatics, 2013: p. 11.10. 1-11.10. 33.
51. Li, H., *Towards better understanding of artifacts in variant calling from high-coverage samples*. Bioinformatics, 2014: p. btu356.
 52. McCormick, R.F., S.K. Truong, and J.E. Mullet, *RIG: recalibration and interrelation of genomic sequence data with the GATK*. G3: Genes | Genomes | Genetics, 2015. **5**(4): p. 655-665.
 53. Koboldt, D.C., D.E. Larson, and R.K. Wilson, *Using VarScan 2 for germline variant calling and somatic mutation detection*. Current protocols in bioinformatics, 2013: p. 15.4. 1-15.4. 17.
 54. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, *The sequence alignment/map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-2079.
 55. Excoffier, L. and H.E. Lischer, *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows*. Molecular ecology resources, 2010. **10**(3): p. 564-567.
 56. Fu, Y.-B., *Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment*. G3: Genes | Genomes | Genetics, 2014. **4**(5): p. 891-900.
 57. VanRaden, P., *Efficient methods to compute genomic predictions*. Journal of dairy science, 2008. **91**(11): p. 4414-4423.
 58. Anderson, C.A., F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, and K.T. Zondervan, *Data quality control in genetic case-control association studies*. Nature protocols, 2010. **5**(9): p. 1564-1573.
 59. Cooper, T., G. Wiggins, and P. VanRaden, *Relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle*. Journal of dairy science, 2013. **96**(5): p. 3336-3339.
 60. Wiggins, G., T. Sonstegard, P. VanRaden, L. Matukumalli, R. Schnabel, J. Taylor, F. Schenkel, and C. Van Tassell, *Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada*. Journal of dairy science, 2009. **92**(7): p. 3431-3436.
 61. Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman, *FaST linear mixed models for genome-wide association studies*. Nature methods, 2011. **8**(10): p. 833-835.
 62. Eu-Ahsunthornwattana, J., E.N. Miller, M. Fakiola, S.M. Jeronimo, J.M. Blackwell, H.J. Cordell, and W.T.C.C.C. 2, *Comparison of methods to account for relatedness in genome-wide association studies with family-based data*. PLoS Genet, 2014. **10**(7): p. e1004445.
 63. Zhang, Z., E. Ersoz, C.-Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, and J.M. Ordovas, *Mixed linear model approach adapted for genome-wide association studies*. Nature genetics, 2010. **42**(4): p. 355-360.
 64. Berger, P. and A. Freeman, *Prediction of Sire Merit for Calving Difficulty1*. Journal of Dairy Science, 1978. **61**(8): p. 1146-1150.
 65. Tong, A., J. Wilton, and L. Schaeffer, *Evaluation of ease of calving for Charolais sires*. Canadian Journal of Animal Science, 1976. **56**(1): p. 17-26.
 66. Tong, A., J. Wilton, and L. Schaeffer, *Application of a scoring procedure and transformations to dairy type classification and beef ease of calving categorical data*. Canadian Journal of Animal Science, 1977. **57**(1): p. 1-5.

67. Harville, D.A. and R.W. Mee, *A mixed-model procedure for analyzing ordered categorical data*. Biometrics, 1984: p. 393-408.
68. Anderson, C.J., J. Verkuilen, and T. Johnson, *Applied generalized linear mixed models: Continuous and discrete data*. 2010, New York: Springer.
69. Fitzmaurice, G.M., N.M. Laird, and J.H. Ware, *Applied longitudinal analysis*. Vol. 998. 2012: John Wiley & Sons.
70. Molenberghs, G. and G. Verbeke, *Linear mixed models for longitudinal data*. 2000: Springer.
71. Rosa, G., C.R. Padovani, and D. Gianola, *Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation*. Biometrical Journal, 2003. **45**(5): p. 573-590.
72. Strandén, I. and D. Gianola, *Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach*. Genetics Selection Evolution, 1999. **31**(1): p. 25.
73. Abdalla, E., G. Rosa, K. Weigel, and T. Byrem, *Genetic analysis of leukosis incidence in United States Holstein and Jersey populations*. Journal of dairy science, 2013. **96**(9): p. 6022-6029.
74. Bliss, C.I., *The calculation of the dosage-mortality curve*. Annals of Applied Biology, 1935. **22**(1): p. 134-167.
75. Moreno, C., D. Sorensen, L. García-Cortés, L. Varona, and J. Altarriba, *On biased inferences about variance components in the binary threshold model*. Genetics Selection Evolution, 1997. **29**(2): p. 145-160.
76. Wright, S., *An analysis of variability in number of digits in an inbred strain of guinea pigs*. Genetics, 1934. **19**(6): p. 506.
77. Carriquiry, A.L., *Bayesian prediction and its application to the genetic evaluation of livestock*. 1989.
78. Foulley, J.L., S. Im, D. Gianola, and I. Höschele, *Empirical Bayes estimation of parameters for n polygenic binary traits*. Génétique sélection évolution, 1987. **19**(2): p. 197-224.
79. Gianola, D. and R.L. Fernando, *Bayesian methods in animal breeding theory*. Journal of Animal Science, 1986. **63**(1): p. 217-244.
80. Gianola, D., S. Im, R. Fernando, and J. Foulley, *Mixed model methodology and the Box-Cox theory of transformations: a Bayesian approach*, in *Advances in statistical methods for genetic improvement of livestock*. 1990, Springer. p. 15-40.
81. Macedo, F. and D. Gianola. *Bayesian analysis of univariate mixed models with informative priors*. in *European Association for Animal Production, 38th Annual Meeting, Lisbon, Portugal*. 1987.
82. Broemeling, L.D., *Bayesian analysis of linear models*. Vol. 60. 1985: M. Dekker.
83. Hammersley, J. and D. Handscomb, *Monte Carlo Methods, Methuen's Monographs on Applied Probability*. 1964, Wiley, New York.
84. Rubinstein, R.Y. and D. Kroese, *Simulation and the Monte Carlo Method*. John Wiley&Sons. Inc. Publication, 1981.
85. Kloek, T. and H.K. Van Dijk, *Bayesian estimates of equation system parameters: an application of integration by Monte Carlo*. Econometrica: Journal of the Econometric Society, 1978: p. 1-19.
86. Moreno, C., *Estudio de los componentes genéticos del modelo umbral*. 1993, Tesis doctoral. Universidad de Zaragoza, Spain.
87. Sorensen, D., S. Andersen, D. Gianola, and I. Korsgaard, *Bayesian inference in threshold models using Gibbs sampling*. Genetics Selection Evolution, 1995. **27**(3): p. 229-249.

88. Bauwens, L., *Bayesian Full Information Analysis of the Simultaneous Equation Models using Monte-Carlo Integration*. Springer-Verlag, Berlin, 1984.
89. Zellner, A., L. Bauwens, and H.K. Van Dijk, *Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods*. Journal of Econometrics, 1988. **38**(1-2): p. 39-72.
90. Zellner, A. and P.E. Rossi, *Bayesian analysis of dichotomous quantal response models*. Journal of Econometrics, 1984. **25**(3): p. 365-393.
91. Zeger, S.L. and M.R. Karim, *Generalized linear models with random effects; a Gibbs sampling approach*. Journal of the American statistical association, 1991. **86**(413): p. 79-86.
92. McCulloch, C.E., *Maximum likelihood variance components estimation for binary data*. Journal of the American Statistical Association, 1994. **89**(425): p. 330-335.
93. Albert, J.H. and S. Chib, *Bayesian analysis of binary and polychotomous response data*. Journal of the American statistical Association, 1993. **88**(422): p. 669-679.
94. Khuri, A.I. and H. Sahai, *Variance components analysis: a selective literature survey*. International Statistical Review/Revue Internationale de Statistique, 1985: p. 279-300.
95. Searle, S., G. Casella, and C. McCulloch, *Variance components: Wiley series in probability and mathematical statistics*. 1992, John Wiley & Sons, New York.
96. Henderson, C.R., *Applications of linear models in animal breeding*. Applications of linear models in animal breeding., 1984.
97. Harville, D.A., *BLUP (best linear unbiased prediction) and beyond*. 1990: Springer.
98. McLean, R.A., W.L. Sanders, and W.W. Stroup, *A unified approach to mixed linear models*. The American Statistician, 1991. **45**(1): p. 54-64.
99. Robinson, G.K., *That BLUP is a good thing: the estimation of random effects*. Statistical science, 1991: p. 15-32.
100. Searle, S.R., *Linear Models*. New York: John Wiley & Sons. 1971, Inc.
101. Henderson, C.R., *Estimation of genetic parameters*. Annals of Mathematical Statistics, 1950: p. 21, 309.
102. Henderson, C.R., *Sire evaluation and genetic trends*. Journal of Animal Science, 1973. **1973**(Symposium): p. 10-41.
103. Henderson, C.R., O. Kempthorne, S.R. Searle, and C. Von Krosigk, *The estimation of environmental and genetic trends from records subject to culling*. Biometrics, 1959. **15**(2): p. 192-218.
104. Penrose, R. *A generalized inverse for matrices*. in *Mathematical proceedings of the Cambridge philosophical society*. 1955. Cambridge Univ Press.
105. Henderson, C.R., *Best linear unbiased estimation and prediction under a selection model*. Biometrics, 1975: p. 423-447.
106. Henderson, C.R., *Selection index and expected genetic advance*. Statistical genetics and plant breeding, 1963. **982**: p. 141-163.
107. Mrode, R.A., *Linear models for the prediction of animal breeding values*. 2014: Cabi.
108. Harville, D.A., *Maximum likelihood approaches to variance component estimation and to related problems*. Journal of the American Statistical Association, 1977. **72**(358): p. 320-338.
109. Jennrich, R.I. and M.D. Schluchter, *Unbalanced repeated-measures models with structured covariance matrices*. Biometrics, 1986: p. 805-820.
110. Littell, R.C., W.W. Stroup, G.A. Milliken, R.D. Wolfinger, and O. Schabenberger, *SAS for mixed models*. 2006: SAS institute.

111. Henderson, C.R., *A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values*. Biometrics, 1976: p. 69-83.
112. Searle, S.R., G. Casella, and C.E. McCulloch, *Variance components*. Vol. 391. 2009: John Wiley & Sons.
113. Vazquez, A., D. Bates, G. Rosa, D. Gianola, and K. Weigel, *an R package for fitting generalized linear mixed models in animal breeding*. Journal of animal science, 2010. **88**(2): p. 497-504.
114. Vazquez, A., D. Bates, G. Rosa, D. Gianola, and K. Weigel, *Technical note: an R package for fitting generalized linear mixed models in animal breeding*. Journal of animal science, 2010. **88**(2): p. 497-504.
115. Falconer, D. and T. Mackay, *Introduction to quantitative genetics., 4th edn (Longman Group Ltd: Essex, UK)*. 1996.
116. Wright, S., *Coefficients of inbreeding and relationship*. The American Naturalist, 1922. **56**(645): p. 330-338.
117. Golan, D. and S. Rosset, *Accurate estimation of heritability in genome wide studies using random effects models*. Bioinformatics, 2011. **27**(13): p. i317-i323.
118. Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, and G.W. Montgomery, *Common SNPs explain a large proportion of the heritability for human height*. Nature genetics, 2010. **42**(7): p. 565-569.
119. VanRaden, P., *Genomic measures of relationship and inbreeding*. Interbull bull, 2007. **37**: p. 33-36.
120. Amin, N., C.M. Van Duijn, and Y.S. Aulchenko, *A genomic background based method for association analysis in related individuals*. PloS one, 2007. **2**(12): p. e1274.
121. Leutenegger, A.-L., B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E.A. Thompson, *Estimation of the inbreeding coefficient through use of genomic data*. The American Journal of Human Genetics, 2003. **73**(3): p. 516-523.
122. Legarra, A., I. Aguilar, and I. Misztal, *A relationship matrix including full pedigree and genomic information*. Journal of dairy science, 2009. **92**(9): p. 4656-4663.
123. Geman, S. and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. IEEE Transactions on pattern analysis and machine intelligence, 1984(6): p. 721-741.
124. Wang, C., J. Rutledge, and D. Gianola, *Marginal inferences about variance components in a mixed linear model using Gibbs sampling*. Genetics Selection Evolution, 1993. **25**(1): p. 41.
125. Hobert, J.P. and G. Casella, *The effect of improper priors on Gibbs sampling in hierarchical linear mixed models*. Journal of the American Statistical Association, 1996. **91**(436): p. 1461-1473.
126. Casella, G. and E. George, *An introduction to Gibbs sampling*. The American Statistician, 1992. **46**(46): p. 167-174.
127. Gelfand, A.E. and A.F. Smith, *Sampling-based approaches to calculating marginal densities*. Journal of the American statistical association, 1990. **85**(410): p. 398-409.
128. Henderson, C. and R. Quaas, *Multiple trait evaluation using relatives' records*. Journal of Animal Science, 1976. **43**(6): p. 1188-1197.
129. Hadfield, J.D., *MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package*. Journal of Statistical Software, 2010. **33**(2): p. 1-22.
130. de Villemereuil, P., *Estimation of a biological trait heritability using the animal model*. How to use the MCMCglmm R package, 2012.

131. Gelman, A., *Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)*. Bayesian analysis, 2006. **1**(3): p. 515-534.
132. Hadfield, J., M.J. Hadfield, and C. SystemRequirements, *Package 'MCMCglmm'*. 2016.
133. Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica, *Manual for BLUPF90 family of programs*. Athens: University of Georgia, 2014.
134. Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. Lee. *BLUPF90 and related programs (BGF90)*. in *Proceedings of the 7th world congress on genetics applied to livestock production*. 2002. Montpellier, Communication No. 28–27.
135. Hayes, B., *Overview of statistical methods for genome-wide association studies (GWAS)*. Genome-wide association studies and genomic prediction, 2013: p. 149-169.
136. Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. Muir, *Genome-wide association mapping including phenotypes from relatives without genotypes*. Genetics Research, 2012. **94**(02): p. 73-83.
137. VanRaden, P., C. Van Tassell, G. Wiggans, T. Sonstegard, R. Schnabel, J. Taylor, and F. Schenkel, *Invited review: Reliability of genomic predictions for North American Holstein bulls*. Journal of dairy science, 2009. **92**(1): p. 16-24.
138. Gonzalez-Neira, A., J.M. Rosa-Rosa, A. Osorio, E. Gonzalez, M. Southey, O. Sinilnikova, H. Lynch, R.A. Oldenburg, C.J. Van Asperen, and N. Hoogerbrugge, *Genomewide high-density SNP linkage analysis of non-BRCA1/2 breast cancer families identifies various candidate regions and has greater power than microsatellite studies*. BMC genomics, 2007. **8**(1): p. 299.
139. Padanad, M.S., G. Konstantinidou, N. Venkateswaran, M. Melegari, S. Rindhe, M. Mitsche, C. Yang, K. Batten, K.E. Huffman, and J. Liu, *Fatty acid oxidation mediated by acyl-CoA synthetase long chain 3 is required for mutant KRAS lung tumorigenesis*. Cell reports, 2016. **16**(6): p. 1614-1628.
140. Katoh, M., *Function and cancer genomics of FAT family genes (review)*. International journal of oncology, 2012. **41**(6): p. 1913-1918.
141. D'Angelo, A., L. Garzia, A. André, P. Carotenuto, V. Aglio, O. Guardiola, G. Arrigoni, A. Cossu, G. Palmieri, and L. Aravind, *Prune cAMP phosphodiesterase binds nm23-H1 and promotes cancer metastasis*. Cancer cell, 2004. **5**(2): p. 137-149.
142. Xue, W., T. Kitzing, S. Roessler, J. Zuber, A. Krasnitz, N. Schultz, K. Revill, S. Weissmueller, A.R. Rappaport, and J. Simon, *A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions*. Proceedings of the National Academy of Sciences, 2012. **109**(21): p. 8212-8217.
143. Jiang, Y., Y. Cao, Y. Wang, W. Li, X. Liu, Y. Lv, X. Li, and J. Mi, *Cysteine transporter SLC3A1 promotes breast cancer tumorigenesis*. Theranostics, 2017. **7**(4): p. 1036.
144. He, N., H. Zheng, P. Li, Y. Zhao, W. Zhang, F. Song, and K. Chen, *miR-485-5p binding site SNP rs8752 in HPGD gene is associated with breast cancer risk*. PloS one, 2014. **9**(7): p. e102093.
145. Steed, E., N.T. Rodrigues, M.S. Balda, and K. Matter, *Identification of MarvelD3 as a tight junction-associated transmembrane protein of the occludin family*. BMC cell biology, 2009. **10**(1): p. 95.
146. Abdalla, E., F. Peñagaricano, T. Byrem, K. Weigel, and G. Rosa, *Genome-wide association mapping and pathway analysis of leukosis incidence in a US Holstein cattle population*. Animal genetics, 2016. **47**(4): p. 395-407.
147. Medicine, N.L.o. *Cells and DNA*. 2017; Available from: <https://ghr.nlm.nih.gov/primer/basics/dna>.

148. studyblue. *Chapter 1 Code of Genes/Genomes*. 2017; Available from: <https://www.studyblue.com/notes/note/n/chapter-1-code-of-genes-genomes/deck/9365552>.
149. biologos. *Adam, Eve, and Human Population Genetics: Signature in the SNPs*. 2015; Available from: <http://biologos.org/blogs/dennis-venema-letters-to-the-duchess/adam-eve-and-human-population-genetics-part-4-signature-in-the-snps>.
150. illumina. *Advantages of paired-end and single-read sequencing*. 2017; Available from: <https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html>.

APPENDIX A

ID	SEX	DAD	MAM	Y	family
1	M	NA	NA	0	1
2	F	NA	NA	0	1
3	M	NA	NA	0	1
4	F	1	2	1	1
5	F	1	2	0	1
6	F	1	2	0	1
7	M	1	2	0	1
8	F	NA	NA	0	1
9	M	NA	NA	1	1
10	F	3	4	1	1
11	F	3	4	0	1
12	F	3	4	0	1
13	F	3	4	1	1
14	M	NA	NA	0	1
15	M	3	4	1	1
16	F	NA	NA	0	1
17	M	3	4	1	1
18	F	NA	5	0	1
19	M	NA	5	1	1
20	M	NA	6	0	1
21	F	NA	6	0	1
22	F	NA	6	0	1
23	F	7	8	0	1
24	F	7	8	1	1
25	M	7	8	0	1
26	M	7	8	0	1
27	F	7	8	1	1
28	M	7	8	0	1
29	F	7	8	0	1
30	F	7	8	1	1
31	M	7	8	0	1
32	M	7	8	0	1
33	F	9	10	1	1
34	M	9	10	1	1
35	M	9	10	1	1
36	F	9	10	1	1
37	F	0	12	0	1
38	M	0	12	0	1
39	M	14	13	0	1

40	F	14	13	0	1
41	M	15	16	1	1
42	F	15	16	0	1
43	F	15	16	1	1
44	M	15	16	0	1
45	M	15	16	0	1
46	M	15	16	0	1
47	F	17	NA	0	1
48	M	17	NA	0	1
49	M	NA	18	0	1
50	M	19	NA	0	1
51	F	NA	21	0	1
52	M	NA	23	0	1
53	M	25	24	0	1
54	F	25	24	1	1
55	M	28	27	0	1
56	F	28	27	0	1
57	M	28	27	0	1
58	F	NA	29	0	1
59	F	NA	29	0	1
60	F	NA	29	0	1
61	F	31	30	0	1
62	F	31	30	0	1
63	M	NA	NA	0	2
64	F	NA	NA	0	2
65	M	63	64	1	2
66	F	NA	NA	0	2
67	F	63	64	1	2
68	M	NA	NA	0	2
69	F	63	64	0	2
70	M	NA	NA	0	2
71	F	63	64	1	2
72	M	NA	NA	0	2
73	F	63	64	1	2
74	M	NA	NA	0	2
75	M	63	64	0	2
76	F	NA	NA	0	2
77	M	63	64	1	2
78	F	NA	NA	0	2
79	F	65	66	0	2
80	M	65	66	0	2
81	M	65	66	0	2
82	M	68	67	0	2

83	M	68	67	0	2
84	M	68	67	0	2
85	F	68	67	0	2
86	F	70	69	0	2
87	F	70	69	1	2
88	F	70	69	1	2
89	M	70	69	0	2
90	F	72	71	0	2
91	M	74	73	0	2
92	F	74	73	0	2
93	M	74	73	0	2
94	F	75	76	0	2
95	M	75	76	0	2
96	M	77	78	0	2
97	M	NA	NA	0	3
98	F	NA	NA	0	3
99	M	NA	NA	0	3
100	F	97	98	1	3
101	F	97	98	0	3
102	M	97	98	1	3
103	F	NA	NA	1	3
104	F	97	98	1	3
105	M	NA	NA	0	3
106	M	97	98	0	3
107	F	NA	NA	0	3
108	M	97	98	0	3
109	F	99	100	0	3
110	M	NA	NA	0	3
111	M	99	100	1	3
112	M	99	100	1	3
113	F	99	100	1	3
114	M	99	100	0	3
115	F	99	100	0	3
116	F	99	100	1	3
117	M	99	100	0	3
118	F	99	100	1	3
119	M	NA	NA	1	3
120	F	99	100	0	3
121	M	99	100	0	3
122	M	99	100	0	3
123	F	99	100	1	3
124	F	102	103	0	3
125	M	102	103	0	3

126	M	105	104	0	3
127	M	106	107	0	3
128	M	106	107	1	3
129	M	108	NA	0	3
130	F	NA	NA	0	3
131	M	108	NA	0	3
132	F	108	NA	0	3
133	F	108	NA	0	3
134	F	108	NA	0	3
135	F	110	109	0	3
136	F	110	109	0	3
137	F	110	109	0	3
138	F	110	109	0	3
139	F	110	109	1	3
140	F	119	118	1	3
141	M	119	118	0	3
142	M	119	118	0	3
143	F	119	118	0	3
144	F	119	118	1	3
145	M	119	118	0	3
146	M	119	118	0	3
147	M	119	118	0	3
148	M	119	118	0	3
149	F	NA	124	0	3
150	M	NA	124	0	3
151	F	125	NA	0	3
152	M	125	NA	0	3
153	F	129	130	0	3
154	M	129	130	0	3
155	F	129	130	1	3
156	F	129	130	1	3
157	F	129	130	1	3
158	F	129	130	1	3
159	F	129	130	0	3
160	F	131	NA	0	3
161	M	131	NA	0	3
162	F	NA	132	0	3
163	F	NA	132	0	3
164	F	NA	132	0	3
165	F	NA	133	0	3
166	M	NA	133	0	3
167	F	NA	134	0	3

APPENDIX B

bowtie2 mapping statistics

Family_1

Individual 1

17496012 reads; of these:

17496012 (100.00%) were paired; of these:

756753 (4.33%) aligned concordantly 0 times

12814966 (73.25%) aligned concordantly exactly 1 time

3924293 (22.43%) aligned concordantly >1 times

756753 pairs aligned concordantly 0 times; of these:

162244 (21.44%) aligned discordantly 1 time

594509 pairs aligned 0 times concordantly or discordantly; of these:

1189018 mates make up the pairs; of these:

744617 (62.62%) aligned 0 times

216524 (18.21%) aligned exactly 1 time

227877 (19.17%) aligned >1 times

97.87% overall alignment rate

Individual 2

46240754 reads; of these:

46240754 (100.00%) were paired; of these:

4231915 (9.15%) aligned concordantly 0 times

32513660 (70.31%) aligned concordantly exactly 1 time

9495179 (20.53%) aligned concordantly >1 times

4231915 pairs aligned concordantly 0 times; of these:

1644471 (38.86%) aligned discordantly 1 time

2587444 pairs aligned 0 times concordantly or discordantly; of these:

5174888 mates make up the pairs; of these:

2709650 (52.36%) aligned 0 times

946097 (18.28%) aligned exactly 1 time

1519141 (29.36%) aligned >1 times

97.07% overall alignment rate

#####

Individual 3

23418595 reads; of these:

23418595 (100.00%) were paired; of these:

863355 (3.69%) aligned concordantly 0 times

17067338 (72.88%) aligned concordantly exactly 1 time

5487902 (23.43%) aligned concordantly >1 times

863355 pairs aligned concordantly 0 times; of these:

153832 (17.82%) aligned discordantly 1 time

709523 pairs aligned 0 times concordantly or discordantly; of these:

1419046 mates make up the pairs; of these:

914103 (64.42%) aligned 0 times

256193 (18.05%) aligned exactly 1 time

248750 (17.53%) aligned >1 times

98.05% overall alignment rate

Individual 4

40313161 reads; of these:

40313161 (100.00%) were paired; of these:

1583496 (3.93%) aligned concordantly 0 times

28742837 (71.30%) aligned concordantly exactly 1 time

9986828 (24.77%) aligned concordantly >1 times

1583496 pairs aligned concordantly 0 times; of these:

395581 (24.98%) aligned discordantly 1 time

1187915 pairs aligned 0 times concordantly or discordantly; of these:

2375830 mates make up the pairs; of these:

1615490 (68.00%) aligned 0 times

372749 (15.69%) aligned exactly 1 time

387591 (16.31%) aligned >1 times

98.00% overall alignment rate

#####

Individual 5

50944516 reads; of these:

50944516 (100.00%) were paired; of these:

3319899 (6.52%) aligned concordantly 0 times

37364337 (73.34%) aligned concordantly exactly 1 time

10260280 (20.14%) aligned concordantly >1 times

3319899 pairs aligned concordantly 0 times; of these:

1446095 (43.56%) aligned discordantly 1 time

1873804 pairs aligned 0 times concordantly or discordantly; of these:

3747608 mates make up the pairs; of these:

1709233 (45.61%) aligned 0 times

784899 (20.94%) aligned exactly 1 time

1253476 (33.45%) aligned >1 times

98.32% overall alignment rate

Individual 6

40158059 reads; of these:

40158059 (100.00%) were paired; of these:

1320920 (3.29%) aligned concordantly 0 times

29257869 (72.86%) aligned concordantly exactly 1 time

9579270 (23.85%) aligned concordantly >1 times

1320920 pairs aligned concordantly 0 times; of these:

270128 (20.45%) aligned discordantly 1 time

1050792 pairs aligned 0 times concordantly or discordantly; of these:

2101584 mates make up the pairs; of these:

1404688 (66.84%) aligned 0 times

353131 (16.80%) aligned exactly 1 time

343765 (16.36%) aligned >1 times

98.25% overall alignment rate

#####

Individual 7

42973730 reads; of these:

42973730 (100.00%) were paired; of these:

3057263 (7.11%) aligned concordantly 0 times

31345819 (72.94%) aligned concordantly exactly 1 time

8570648 (19.94%) aligned concordantly >1 times

3057263 pairs aligned concordantly 0 times; of these:

1164237 (38.08%) aligned discordantly 1 time

1893026 pairs aligned 0 times concordantly or discordantly; of these:

3786052 mates make up the pairs; of these:

2094072 (55.31%) aligned 0 times

646607 (17.08%) aligned exactly 1 time

1045373 (27.61%) aligned >1 times

97.56% overall alignment rate

Individual 8

27830687 reads; of these:

27830687 (100.00%) were paired; of these:

2927435 (10.52%) aligned concordantly 0 times

19500773 (70.07%) aligned concordantly exactly 1 time

5402479 (19.41%) aligned concordantly >1 times

2927435 pairs aligned concordantly 0 times; of these:

1062618 (36.30%) aligned discordantly 1 time

1864817 pairs aligned 0 times concordantly or discordantly; of these:

3729634 mates make up the pairs; of these:

2084863 (55.90%) aligned 0 times

753643 (20.21%) aligned exactly 1 time

891128 (23.89%) aligned >1 times

96.25% overall alignment rate

#####

Family_2

Individual 1

33419098 reads; of these:

33419098 (100.00%) were paired; of these:

5027688 (15.04%) aligned concordantly 0 times

24119293 (72.17%) aligned concordantly exactly 1 time

4272117 (12.78%) aligned concordantly >1 times

5027688 pairs aligned concordantly 0 times; of these:

1649133 (32.80%) aligned discordantly 1 time

3378555 pairs aligned 0 times concordantly or discordantly; of these:

6757110 mates make up the pairs; of these:

4756798 (70.40%) aligned 0 times

1127561 (16.69%) aligned exactly 1 time

872751 (12.92%) aligned >1 times

92.88% overall alignment rate

Individual 2

27261117 reads; of these:

27261117 (100.00%) were paired; of these:

4333860 (15.90%) aligned concordantly 0 times

19462395 (71.39%) aligned concordantly exactly 1 time

3464862 (12.71%) aligned concordantly >1 times

4333860 pairs aligned concordantly 0 times; of these:

1351131 (31.18%) aligned discordantly 1 time

2982729 pairs aligned 0 times concordantly or discordantly; of these:

5965458 mates make up the pairs; of these:

4132945 (69.28%) aligned 0 times

1061209 (17.79%) aligned exactly 1 time

771304 (12.93%) aligned >1 times

92.42% overall alignment rate

#####

Individual 3

29561523 reads; of these:

29561523 (100.00%) were paired; of these:

5644272 (19.09%) aligned concordantly 0 times

20213488 (68.38%) aligned concordantly exactly 1 time

3703763 (12.53%) aligned concordantly >1 times

5644272 pairs aligned concordantly 0 times; of these:

1434971 (25.42%) aligned discordantly 1 time

4209301 pairs aligned 0 times concordantly or discordantly; of these:

8418602 mates make up the pairs; of these:

5036006 (59.82%) aligned 0 times

1971671 (23.42%) aligned exactly 1 time

1410925 (16.76%) aligned >1 times

91.48% overall alignment rate

Individual 4

25790769 reads; of these:

25790769 (100.00%) were paired; of these:

3630652 (14.08%) aligned concordantly 0 times

18918333 (73.35%) aligned concordantly exactly 1 time

3241784 (12.57%) aligned concordantly >1 times

3630652 pairs aligned concordantly 0 times; of these:

1165758 (32.11%) aligned discordantly 1 time

2464894 pairs aligned 0 times concordantly or discordantly; of these:

4929788 mates make up the pairs; of these:

3560055 (72.22%) aligned 0 times

778019 (15.78%) aligned exactly 1 time

591714 (12.00%) aligned >1 times

93.10% overall alignment rate

#####

Individual 5

17433912 reads; of these:

17433912 (100.00%) were paired; of these:

3067745 (17.60%) aligned concordantly 0 times

12120755 (69.52%) aligned concordantly exactly 1 time

2245412 (12.88%) aligned concordantly >1 times

3067745 pairs aligned concordantly 0 times; of these:

894972 (29.17%) aligned discordantly 1 time

2172773 pairs aligned 0 times concordantly or discordantly; of these:

4345546 mates make up the pairs; of these:

2941382 (67.69%) aligned 0 times

807040 (18.57%) aligned exactly 1 time

597124 (13.74%) aligned >1 times

91.56% overall alignment rate

Individual 6

37657589 reads; of these:

37657589 (100.00%) were paired; of these:

6745900 (17.91%) aligned concordantly 0 times

25936201 (68.87%) aligned concordantly exactly 1 time

4975488 (13.21%) aligned concordantly >1 times

6745900 pairs aligned concordantly 0 times; of these:

1578708 (23.40%) aligned discordantly 1 time

5167192 pairs aligned 0 times concordantly or discordantly; of these:

10334384 mates make up the pairs; of these:

6314251 (61.10%) aligned 0 times

2247352 (21.75%) aligned exactly 1 time

1772781 (17.15%) aligned >1 times

91.62% overall alignment rate

#####

Individual 7

35977512 reads; of these:

35977512 (100.00%) were paired; of these:

2642581 (7.35%) aligned concordantly 0 times

28827029 (80.13%) aligned concordantly exactly 1 time

4507902 (12.53%) aligned concordantly >1 times

2642581 pairs aligned concordantly 0 times; of these:

848358 (32.10%) aligned discordantly 1 time

1794223 pairs aligned 0 times concordantly or discordantly; of these:

3588446 mates make up the pairs; of these:

1960475 (54.63%) aligned 0 times

964176 (26.87%) aligned exactly 1 time

663795 (18.50%) aligned >1 times

97.28% overall alignment rate

Individual 8

33662978 reads; of these:

33662978 (100.00%) were paired; of these:

4647026 (13.80%) aligned concordantly 0 times

24648332 (73.22%) aligned concordantly exactly 1 time

4367620 (12.97%) aligned concordantly >1 times

4647026 pairs aligned concordantly 0 times; of these:

1513512 (32.57%) aligned discordantly 1 time

3133514 pairs aligned 0 times concordantly or discordantly; of these:

6267028 mates make up the pairs; of these:

4614943 (73.64%) aligned 0 times

897139 (14.32%) aligned exactly 1 time

754946 (12.05%) aligned >1 times

93.15% overall alignment rate

#####

Family_3

Individual 1

35014538 reads; of these:

35014538 (100.00%) were paired; of these:

1079170 (3.08%) aligned concordantly 0 times

25523400 (72.89%) aligned concordantly exactly 1 time

8411968 (24.02%) aligned concordantly >1 times

1079170 pairs aligned concordantly 0 times; of these:

206693 (19.15%) aligned discordantly 1 time

872477 pairs aligned 0 times concordantly or discordantly; of these:

1744954 mates make up the pairs; of these:

1155676 (66.23%) aligned 0 times

270321 (15.49%) aligned exactly 1 time

318957 (18.28%) aligned >1 times

98.35% overall alignment rate

Individual 2

31736845 reads; of these:

31736845 (100.00%) were paired; of these:

1150292 (3.62%) aligned concordantly 0 times

22703841 (71.54%) aligned concordantly exactly 1 time

7882712 (24.84%) aligned concordantly >1 times

1150292 pairs aligned concordantly 0 times; of these:

277066 (24.09%) aligned discordantly 1 time

873226 pairs aligned 0 times concordantly or discordantly; of these:

1746452 mates make up the pairs; of these:

1084416 (62.09%) aligned 0 times

260096 (14.89%) aligned exactly 1 time

401940 (23.01%) aligned >1 times

98.29% overall alignment rate

#####

Individual 3

31798628 reads; of these:

31798628 (100.00%) were paired; of these:

3092650 (9.73%) aligned concordantly 0 times

22042378 (69.32%) aligned concordantly exactly 1 time

6663600 (20.96%) aligned concordantly >1 times

3092650 pairs aligned concordantly 0 times; of these:

1364125 (44.11%) aligned discordantly 1 time

1728525 pairs aligned 0 times concordantly or discordantly; of these:

3457050 mates make up the pairs; of these:

1670330 (48.32%) aligned 0 times

563809 (16.31%) aligned exactly 1 time

1222911 (35.37%) aligned >1 times

97.37% overall alignment rate

Individual 4

29648460 reads; of these:

29648460 (100.00%) were paired; of these:

2843326 (9.59%) aligned concordantly 0 times

20878239 (70.42%) aligned concordantly exactly 1 time

5926895 (19.99%) aligned concordantly >1 times

2843326 pairs aligned concordantly 0 times; of these:

1428820 (50.25%) aligned discordantly 1 time

1414506 pairs aligned 0 times concordantly or discordantly; of these:

2829012 mates make up the pairs; of these:

1049976 (37.11%) aligned 0 times

562948 (19.90%) aligned exactly 1 time

1216088 (42.99%) aligned >1 times

98.23% overall alignment rate

#####

Individual 5

38418769 reads; of these:

38418769 (100.00%) were paired; of these:

4234333 (11.02%) aligned concordantly 0 times

26159383 (68.09%) aligned concordantly exactly 1 time

8025053 (20.89%) aligned concordantly >1 times

4234333 pairs aligned concordantly 0 times; of these:

1956741 (46.21%) aligned discordantly 1 time

2277592 pairs aligned 0 times concordantly or discordantly; of these:

4555184 mates make up the pairs; of these:

2015700 (44.25%) aligned 0 times

772282 (16.95%) aligned exactly 1 time

1767202 (38.80%) aligned >1 times

97.38% overall alignment rate

Individual 6

53411156 reads; of these:

53411156 (100.00%) were paired; of these:

3214461 (6.02%) aligned concordantly 0 times

38192836 (71.51%) aligned concordantly exactly 1 time

12003859 (22.47%) aligned concordantly >1 times

3214461 pairs aligned concordantly 0 times; of these:

1503438 (46.77%) aligned discordantly 1 time

1711023 pairs aligned 0 times concordantly or discordantly; of these:

3422046 mates make up the pairs; of these:

1327096 (38.78%) aligned 0 times

688756 (20.13%) aligned exactly 1 time

1406194 (41.09%) aligned >1 times

98.76% overall alignment rate

#####

APPENDIX C

SNPs calling using SAMtools with default parameters

Family 1

```
##### Individual 1 #####
```

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm21911.pileup

196987678 bases in pileup file

108123 variant positions (96630 SNP, 11493 indel)

8081 were failed by the strand-filter

89264 variant positions reported (89264 SNP, 0 indel)

```
##### Individual 2 #####
```

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm21918.pileup

182913245 bases in pileup file

85939 variant positions (78697 SNP, 7242 indel)

5284 were failed by the strand-filter

73806 variant positions reported (73806 SNP, 0 indel)

```
##### Individual 3 #####
```

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm219120.pileup

195113638 bases in pileup file

105093 variant positions (95744 SNP, 9349 indel)

7651 were failed by the strand-filter

88704 variant positions reported (88704 SNP, 0 indel)

```
#####
```

Individual 4

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm219129.pileup

221707410 bases in pileup file

132614 variant positions (120249 SNP, 12365 indel)

12008 were failed by the strand-filter

109185 variant positions reported (109185 SNP, 0 indel)

Individual 5

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm219187.pileup

216475306 bases in pileup file

138907 variant positions (125734 SNP, 13173 indel)

12974 were failed by the strand-filter

113759 variant positions reported (113759 SNP, 0 indel)

Individual 6

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm219190.pileup

200240382 bases in pileup file

109139 variant positions (97904 SNP, 11235 indel)

7953 were failed by the strand-filter

90676 variant positions reported (90676 SNP, 0 indel)

#####

Individual 7

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm2191100.pileup

193071544 bases in pileup file

96626 variant positions (87003 SNP, 9623 indel)

6409 were failed by the strand-filter

81179 variant positions reported (81179 SNP, 0 indel)

Individual 8

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm2191116.pileup

168683406 bases in pileup file

45903 variant positions (41899 SNP, 4004 indel)

1140 were failed by the strand-filter

40853 variant positions reported (40853 SNP, 0 indel)

#####

Family 2

Individual 1

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm32521.pileup

1293693353 bases in pileup file

81253 variant positions (76248 SNP, 5005 indel)

6492 were failed by the strand-filter

70150 variant positions reported (70150 SNP, 0 indel)

Individual 2

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325220.pileup

1103088982 bases in pileup file

71228 variant positions (67001 SNP, 4227 indel)

5243 were failed by the strand-filter

62059 variant positions reported (62059 SNP, 0 indel)

Individual 3

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325230.pileup

1184582585 bases in pileup file

73998 variant positions (69588 SNP, 4410 indel)

3567 were failed by the strand-filter

66229 variant positions reported (66229 SNP, 0 indel)

#####

Individual 4

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325235.pileup

1110105513 bases in pileup file

71298 variant positions (66886 SNP, 4412 indel)

3999 were failed by the strand-filter

63132 variant positions reported (63132 SNP, 0 indel)

Individual 5

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325238.pileup

778112946 bases in pileup file

49103 variant positions (46533 SNP, 2570 indel)

3099 were failed by the strand-filter

43581 variant positions reported (43581 SNP, 0 indel)

Individual 6

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325239.pileup

778112946 bases in pileup file

82869 variant positions (78063 SNP, 4806 indel)

3519 were failed by the strand-filter

74387 variant positions reported (74387 SNP, 0 indel)

#####

Individual 7

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325240.pileup

962885434 bases in pileup file

51541 variant positions (47783 SNP, 3758 indel)

3999 were failed by the strand-filter

44039 variant positions reported (44039 SNP, 0 indel)

Individual 8

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm325243.pileup

732515495 bases in pileup file

40387 variant positions (37595 SNP, 2792 indel)

2794 were failed by the strand-filter

34935 variant positions reported (34935 SNP, 0 indel)

#####

Family 3

Individual 1

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm33951.pileup

225934110 bases in pileup file

137065 variant positions (124210 SNP, 12855 indel)

12408 were failed by the strand-filter

112822 variant positions reported (112822 SNP, 0 indel)

Individual 2

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm339549.pileup

225934110 bases in pileup file

137065 variant positions (124210 SNP, 12855 indel)

12408 were failed by the strand-filter

112822 variant positions reported (112822 SNP, 0 indel)

Individual 3

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm339560.pileup

181391721 bases in pileup file

69378 variant positions (62932 SNP, 6446 indel)

2775 were failed by the strand-filter

60410 variant positions reported (60410 SNP, 0 indel)

#####

Individual 4

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm3395167.pileup

181073229 bases in pileup file

70240 variant positions (63529 SNP, 6711 indel)

3135 were failed by the strand-filter

60668 variant positions reported (60668 SNP, 0 indel)

Individual 5

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm3395341.pileup

194972383 bases in pileup file

94105 variant positions (85056 SNP, 9049 indel)

5623 were failed by the strand-filter

79913 variant positions reported (79913 SNP, 0 indel)

Individual 6

Only SNPs will be reported

Warning: No p-value threshold provided, so p-values will not be calculated

Min coverage: 10

Min reads2: 2

Min var freq: 0.2

Min avg qual: 15

P-value thresh: 0.01

Reading input from sm3395343.pileup

198289411 bases in pileup file

120509 variant positions (107945 SNP, 12564 indel)

9297 were failed by the strand-filter

99507 variant positions reported (99507 SNP, 0 indel)

#####

APPENDIX D

```
#####  
#####  
####          #####  
####   Pedigree Matrix   ####  
####          #####  
#####  
  
install.packages("pedigreemm")  
library(pedigreemm)  
id<-1:7  
father<-c(NA,NA,1,NA,3,3,)  
mother<-c(NA,NA,2,NA,4,4,4)  
ped <- pedigree(label=id,sire=father,dam=mother)  
# Computes inbreeding coefficients  
F<-inbreeding(ped)  
A<- getA(ped)  
#####
```

APPENDIX E

MCMCglmm Package

```
#####
library(Matrix)
library(coda)
library(ape)
library(MCMCglmm)
databin<-read.csv("incedcor.csv",header=TRUE)
pedigreeb<-read.csv("pedigree.csv",header=TRUE)
prior <- list(R = list(V=1, nu=0.002), G = list(G1 = list(V=1, nu=0.002)))
modelbin <- MCMCglmm(phen ~ SEX + fami, random = ~animal, family = "ordinal",prior = prior,
pedigree = pedigreeb, data = databin, nitt = 250000,burnin = 50000, thin = 40)
heritbin <- modelbin$VVCV[, "animal"]/(modelbin$VVCV[, "animal"] + modelbin$VVCV[, "units"] + 1
mean(heritbin)

##### Diagnostic of the MCMC #####
plot(modelbin$Sol)
plot(modelbin$VVCV)
autocorr.diag(modelbin$Sol)
autocorr.diag(modelbin$VVCV)
effectiveSize(modelbin$Sol)
effectiveSize(modelbin$VVCV)
heidel.diag(modelbin$VVCV)
summary(modelbin)
effectiveSize(heritbin)
mean(heritbin)

#####
priora <- list(R = list(V = 1, fix = 1), G = list(G1 = list(V = 1, nu = 1000, alpha.mu = 0, alpha.V = 1)))
modelbin2 <- MCMCglmm(phen ~ SEX + fami, random = ~animal, family = "ordinal",prior =
priora, pedigree = pedigreeb, data = databin, nitt = 250000,burnin = 50000, thin = 40)

heritbin2 <- modelbin2$VVCV[, "animal"]/(modelbin2$VVCV[, "animal"] + modelbin2$VVCV[,
"units"] + 1)
#####
```

```
##### Diagnostic of the MCMC #####
```

```
plot(modelbin$Sol)
plot(modelbin$VCV)
autocorr.diag(modelbin$Sol)
autocorr.diag(modelbin$VCV)
effectiveSize(modelbin$Sol)
effectiveSize(modelbin$VCV)
heidel.diag(modelbin$VCV)
summary(modelbin)
effectiveSize(heritbin)
mean(heritbin)
```

```
#####
```