X Chromosome Association Testing in Genome-Wide Association Studies

Honours Thesis
November 6, 2009

Peter Hickey

Department of Mathematics and Statistics,
The University of Melbourne

Bioinformatics Division,
The Walter and Eliza Hall Institute of Medical Research

Supervisors:
Dr. Melanie Bahlo Professor Richard Huggins

Abstract

Genome wide association studies (GWAS) have revealed fascinating insights into the genetics of complex diseases. These studies provide many statistical challenges but one problem that has received surprisingly little attention is the testing of associations between phenotype and genotype on the X chromosome.

In this thesis we show that there are methods that perform significantly better than those in current wide-spread use for the analysis of X chromosome GWAS data. In particular we establish that the methods proposed by Clayton (2008) are amongst the most powerful for X chromosome analysis. We quantify these gains via a simulation study under a variety of genetic models and experimental designs, to compare eight existing analytical methods.

Using the knowledge gained from this simulation study we apply the most powerful method to the X chromosome data from a genome wide association study of multiple sclerosis (Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009). Our analysis identifies 11 genetic markers that warrant further study, an improvement upon the published analysis of this data.

Acknowledgements

I would like to thank my two wonderful supervisors Melanie and Richard. In particular to Melanie for suggesting such an interesting topic for my thesis and being a continual source of guidance and encouragement during my time at WEHI. Were it not for her support this work would not be of the level that I am proud to say it is.

I would also like to thank

> Ian Gordon, Aihua Xia, Richard Huggins, Hugh Miller and Kostya Borovkov for their challenging Honours courses, and all my lecturers during my time at the University of Melbourne;
>
> Everyone in the Bioinformatics division at WEHI for the innumerable useful chats over tea. Whether it be programming tips, statistical advice or answers to my questions on genetics, I know that someone here will be able to help. Their encouragement and friendship have made this year so enjoyable. Particular thanks to Gordon Smyth for his help with score tests;
>
> My fellow Honours students, both in Maths & Stats and at WEHI — the Honours year can be very challenging and daunting at times, but having such a great group of people to go through it with has made it a lot of fun;
>
> To all my other friends — I may not have seen as much of you all as I'd like this year, but the times when I can just hang out with you and forget about my study are some of the most important to me;
>
> The Maurice H. Belz Honours Scholarship and the Alan W Harris Honours Scholarship for the financial support that allowed me the freedom to focus on my study.
>
> The Australia and New Zealand Multiple Sclerosis Genetics Consortium for the data used in my case study

And finally, and most importantly, to my parents and to my brothers — thank you so much for your never-ending support and encouragement to pursue my ambitions.

Contents

List of Figures

List of Tables

Chapter 1

Human genetics

Modern biology is producing more and more data as advances in technology allow us to probe deeper than ever before into understanding the science of life. In particular, the advances in genetics using what are known as *high-throughput* technologies are producing gigabytes (1 GB = 10^9 bytes), and even terabytes (1 TB = 1000 GB), of data for a single experiment.

With these high-throughput experiments there is a real need for skilled analysts to help make sense of the masses of data and to distinguish significant results from the inevitable noise such experiments produce. Statisticians are ideally placed to assist in this regard, and there is a boom industry in the application of statistical methods to biological data; an area broadly defined as *bioinformatics*.

1.1 A short introduction to human genetics

Human genetics describes the study of heriditary variation as it occurs in human beings. This variation is inscribed in the *human genome* in the language of *DNA*. Over 99.9% of the genome is common between any two people (International HapMap Consortium, 2003) so it is not surprising that non-heriditary effects such as upbringing and environment also contribute to the large variation seen across the population. One of the main aims of the research in human genetics is to better understand the genetic mechanisms involved in human disease to improve treatments, find cures and ultimately prevent the disease.

In this section we will explain the necessary biological terms and concepts to be used in this project. This includes a brief introduction to human genetics, with a particular focus on the X chromosome, as well as an introduction to the world of the genome wide association study (GWAS).

The following material on human genetics makes close use of "An Introduction to Genetic Analysis" by Suzuki et al. (1989).

DNA

Deoxyribonucleic acid, or DNA, carries the information necessary for the development, maintenance, and reproduction of all organisms, from bacteria to humans (Lange, 2002). Chemically, DNA is composed of only four molecules called nucleotides. These nucleotides form in two long polymers, with backbones made of sugars and phosphate groups joined by ester bonds in what is now famously known as the double helix (shown in figure 1.1).

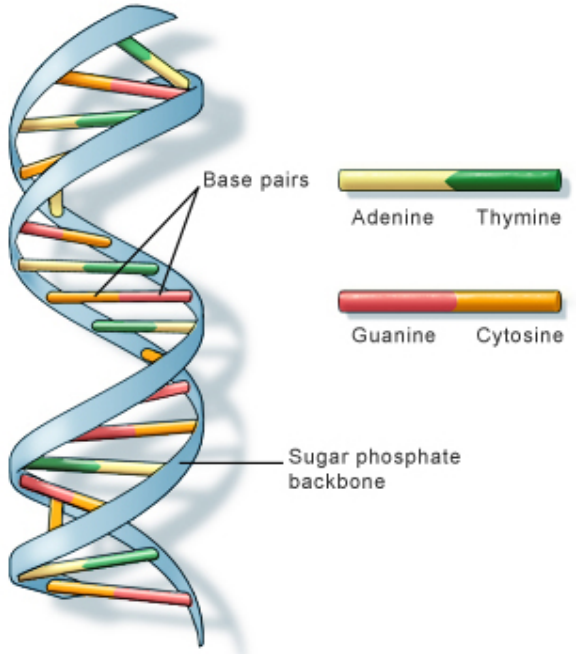

Figure 1.1: A schematic of the DNA double helix (Source: U.S. National Library of Medicine)

It is the sequence of these four bases along the backbone that encodes information. These 4 bases are adenine, guanine, cytosine and thymine, commonly abbreviated to A, G, C and T respectively. As as is shown in figure 1.1, the bases pair up in a complementary manner. Adenine always pairs with thymine while guanine always pairs with cytosine and each such pairing is known as a basepair (bp). The human genome is some 3.4 billion bp in length.

Chromosomes

The genome is broken down into smaller units known as the chromosomes which are found inside the nucleus in most cell types in the human body. In humans, there are 24 different chromosomes denoted $1, 2, \ldots, 22$, X and

Y. Chromosomes $1, 2, \ldots, 22$ are termed autosomes, and humans have two copies of each in what are known as homologous pairs. For each of the autosomes, a person will inherit one copy maternally and one copy paternally. Males and females both have 22 pairs of autosomes, thus where males and females differ are known as the sex chromosomes, the X and Y chromosomes. This characteristic chromosome complement is called a karyotype. We show in figure 1.2 an image of a normal human female karyotype highlighting the chromosomal pairings.

Females carry two X chromosomes and males carry one X and one Y chromosome. The inheritance pattern for the sex chromosomes is different from that of the autosomes. Males inherit their Y chromosome from their father and their X chromosome from their mother. Females receive one X chromosome from each of their mother and father. In both sexes, the maternally inherited X chromosome is a randomly selected copy of X_1 or X_2 the maternal pair of X chromosomes. The X chromosome will be discussed in further detail in section 1.4.

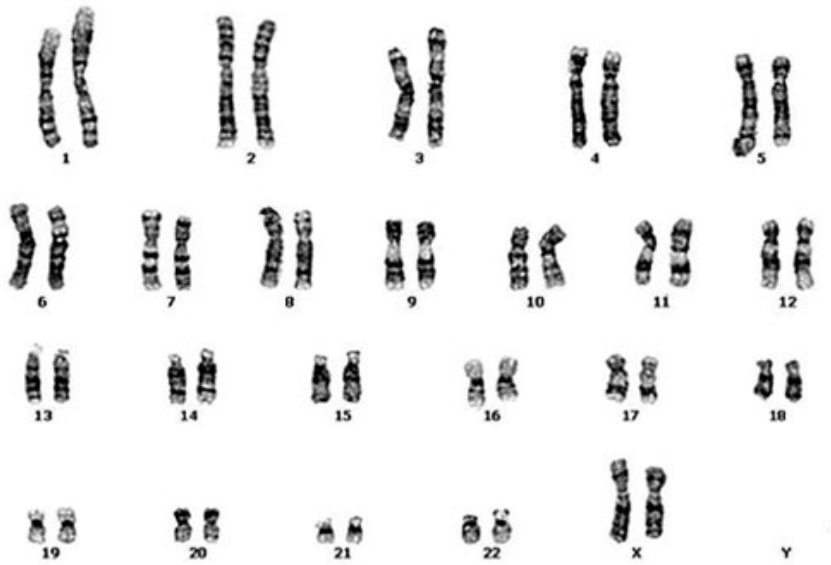

Figure 1.2: A karyotype image of a normal female human genome. Note the autosomal pairs $1, \ldots, 22$, the pair of X chromosomes and the absence of the Y chromosome. Source: `http://www.bio.miami.edu/~cmallery/150/mendel/karyotype.htm` accessed 9/10/2009

Loci, alleles, genotypes and haplotypes

A *locus* (pl. loci) is a fixed position on a chromosome. At some loci it is possible for there to be differences in the sequence of DNA between individuals.

Such loci are called polymorphic loci and the variants at that site are called *alleles*. As humans have two copies of each chromosome, they possess two alleles at each locus (slightly different for the sex chromosomes). If there are just two possible alleles at a locus then the allele that is more frequent (resp. less frequent) in the population is referred to as the major (resp. minor) allele.

The combination of the two alleles at a locus is referred to as the *genotype* at that locus. For example, suppose a person has at a locus the alleles A and B, then their genotype is denoted A/B (which is equivalent to B/A as genotypes are unordered). A genotype such as A/A or B/B is called homozygous and a genotype such as A/B is called heterozygous.

The sequence of alleles along the chromosome of a *gamete* (sperm and egg cells) constitutes a *haplotype*.

Genes

Genes are the functional units of the genome. They are sequences of DNA that provide the template for a protein. Humans have about 20,000-25,000 genes, a surprisingly small number, with over 98% of the genome being made up of non-coding DNA whose function is currently unknown. Genes are the functional units of heredity in a living organism (Suzuki et al., 1989).

Recombination

Recombination is a process whereby the chromosomes are "broken up" and the genetic material reshuffled. In humans recombination occurs during the production of the gametes and leads to the offspring having a different combination of genes to their parents. It can be thought of as a reshuffling of genetic material whereby a parent transmits to a child a chromosome that is a mosaic of their own homologous parental chromosomes.

During recombination, homologous chromosomes are broken up, intertwine and rejoin to form a new pair of homologous chromosomes that are a mosaic of the original homologous pair. The assortment of genetic material to this mosaic pair is random yet preserves the overall structure of the chromosome.

Lange (2002) gives a nice illustration of recombination. Consider a parent producing a gamete. One member of each chromosome pair is painted black and the other member is painted white. Instead of inheriting an all-black or all-white representative of a given pair, a gamete inherits a chromosome that alternates between black and white. The points of exchange are termed *crossovers*. Any given gamete will have just a few randomly positioned crossovers per chromosome.

As recombination occurs between homologous pairs of chromosomes (e.g. the autosomes) the process of recombination is more complicated for the sex

chromosomes (see section 1.4).

1.1.1 Genetic variation

Genetic variation can be classified into two types: inherited (aka germ-line) variation and somatic (aka de novo) variation. Germ-line variation is inherited from one's parents while somatic variation is due to a mutation in the genome during a person's lifetime.

Every time a cell divides the genome must be replicated for these new daughter cells. This process occurs millions of times per day and during the replication mutations can occur spontaneously. An example of a less "natural" mutation is a genetic mutation due to exposure to high doses of radiation. It is important to note that not all mutations are deleterious, indeed most are benign and equally some are advantageous. The simplest mutation is known as a point-mutation, where one base is changed to another.

More recently there has been considerable interest in copy number variations (CNVs) in which a person has more or less than the normal number of copies of a region of the genome. There is gathering evidence that CNVs have a role in a number of complex diseases (McCarroll and Altshuler, 2007). For example, Down's syndrome is caused by the presence of all or part of an extra 21st chromosome. Until recently, only relatively large-scale CNVs were detectable, such as whole chromosomal duplications or deletions, however with new sequencing technologies we are able to detect CNVs of a far smaller scale.

One genetic variant of particular interest to us is called the *single nucleotide polymorphism*, or SNP, and will be discussed in further detail.

Single nucleotide polymorphisms

Individuals may differ by a single nucleotide substitution at a SNP. For example, consider the two sequenced DNA fragments from different individuals shown in figure 1.3; they contain a difference at a single nucleotide (circled).

A SNP is generally bi-allelic, though up to 4 alleles are possible[1]. We will denote these two possible alleles by the generic symbols A and B. Thus at a SNP we consider 3 possible genotypes: A/A, A/B, B/B.

The International HapMap Consortium (2003) was established to "determine the common patterns of DNA sequence variation in the human genome (of which SNPs are a major contributor)". Of the approximately 0.1% of the genome that is different between any two humans, the HapMap project discovered up to 90% of this variation in the population is due to some 10 million common SNPs. The distribution of SNPs through the genome varies between different ethnic groups, an important point we will return to later.

[1] Technically a SNP must have a minor allele frequency $\geq 1\%$

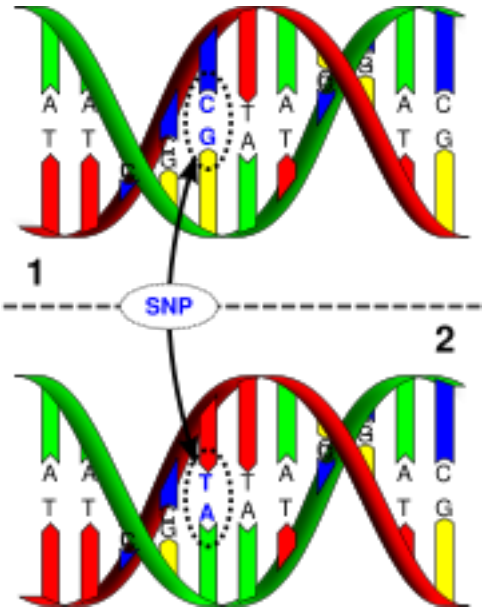

Figure 1.3: A schematic of a single nucleotide polymorphism. DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism). Source: `http://urgi.versailles.inra.fr/projects/GnpSNP/images/snp_dble_helice2.png` accessed 9/10/2009

1.2 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) is one of the fundamental concepts in populations genetics. Provided certain conditions are met, HWE states that both allele and genotype frequencies in a population remain constant from generation to generation. An equivalent probabilistic description for HWE is that the alleles for the next generation for any given individual are chosen randomly and independently of each other.

This model relies on the seven following explicit assumptions: (a) infinite population size, (b) discrete generations, (c) random mating, (d) no selection, (e) no migration, (f) no mutation, and (g) equal genotype frequencies in the two sexes (Lange, 2002). Clearly these conditions are not always satisfied in nature. However, as is common in statistics we often assume independence (i.e. HWE) when it is "near enough" to holding true. This will make our arguments more tractable and simplify the mathematical formulae. We can formally test for deviation from HWE at a locus, for example by using a χ^2 test or Fisher's exact test to compare the observed genotype frequencies to the expected genotype frequencies under HWE (Lange, 2002).

Following Lange (2002), we will now define HWE as a mathematical model for the autosomes (the case for the X chromosome will follow in

section 1.4.1).

Assume the seven conditions for HWE hold, and for simplicities sake we consider an autosomal loci with two possible alleles A and B. Suppose the initial proportions of the genotypes are u for A/A, v for A/B, and w for B/B. We then consider all possible crossings of these genotypes, i.e. the offspring genotypes for all pairings of parents. As an example, under HWE the result of crossing a A/A genotype with the A/B genotype would be an offspring frequency of $\frac{1}{2}$ with the genotype A/A and $\frac{1}{2}$ with the genotype A/B. These proportions of outcomes for the various possible crosses are known as segregation ratios. Under the stated assumptions, the next generation will be composed as shown in table 1.1

Mating Type	Nature of Offspring	Frequency
$A/A \times A/A$	A/A	u^2
$A/A \times A/B$	$\frac{1}{2}A/A + \frac{1}{2}A/B$	$2uv$
$A/A \times B/B$	A/B	$2uw$
$A/B \times A/B$	$\frac{1}{4}A/A + \frac{1}{2}A/B + \frac{1}{4}B/B$	v^2
$A/B \times B/B$	$\frac{1}{2}A/B + \frac{1}{2}B/B$	$2vw$
$B/B \times B/B$	B/B	w^2

Table 1.1: Mating outcomes for Hardy-Weinberg equilibrium

For the next generation we get from table 1.1 the frequencies for genotypes A/A, A/B, and B/B as

$$u^2 + uv + \frac{1}{4}v^2 = (u + \frac{1}{2}v)^2$$
$$uv + 2uw + \frac{1}{2}v^2 + vw = 2(u + \frac{1}{2}v)(\frac{1}{2}v + w)$$
$$\frac{1}{4}v^2 + vw + w^2 = (\frac{1}{2}v + w)^2$$

respectively.

If we define the frequencies of the two alleles A and B as $p_1 = u + \frac{v}{2}$ and $p_2 = \frac{v}{2} + w$, then A/A occurs with frequency p_1^2, A/B with frequency $2p_1p_2$, and B/B with frequency p_2^2 (with $p_1 + p_2 = 1$). After a second round of

random mating, the frequencies of the genotypes A/A, A/B, and B/B are

$$\begin{aligned}(p_1^2 + \frac{1}{2}2p_1p_2)^2 &= [p_1(p_1+p_2)]^2 \\ &= p_1^2 \\ 2(p_1^2 + \frac{1}{2}2p_1p_2)(\frac{1}{2}2p_1p_2 + p_2^2) &= 2p_1(p_1+p_2)p_2(p_1+p_2) \\ &= 2p_1p_2 \\ (\frac{1}{2}2p_1p_2 + p_2^2)^2 &= [p_2(p_1+p_2)]^2 \\ &= p_2^2\end{aligned}$$

Thus, after a single round of random mating, genotype frequencies stabilise at the Hardy-Weinberg proportions.

In essence, Hardy-Weinberg equilibrium corresponds to the random union of two gametes, one gamete being an egg and the other being a sperm. The importance of Hardy-Weinberg Equilibrium for genome wide association studies is that certain test statistics fail to work when HWE does not hold (we prove this in Theorem 3.1).

1.3 Linkage disequilibrium

Linkage disequilibrium (LD) is the correlation between alleles of unrelated individuals and occurs at the level of a population. The distribution of LD is non-uniform in the human population. Mathematically, consider two loci $i = 1, 2$ with corresponding alleles Y_i; the alleles are said to be *in linkage disequilibrium* if they possess the property

$$\Pr(Y_2 = y_2 | Y_1 = y_1) \neq \Pr(Y_2 = y_2), \quad \forall y_1, y_2.$$

That is, two alleles are in LD if they are not independent of one another. One measure of linkage disequilibrium between two loci is the standard sample correlation correlation coefficient, r, and its square, r^2.

Linkage disequilibrium plays an important role in GWA studies. We can exploit the correlation structure within the genome to reduce the number of SNPs we have to genotype as much information can be inferred using a smaller subset of SNPs. For this type of approach to be feasible however requires further knowledge of the LD structure for humans, and it was for this reason the International HapMap project was instigated in 2002.

HapMap

The International HapMap Project genotyped several million well-defined SNPs in 270 individuals from 4 populations. Of these 270 samples, 90 samples (30 trios of two parents and an adult child) were from a population of

Caucasians with European ancestry living in Utah, USA (aka CEU population); 90 samples (30 trios) were from the Yoruba people in Ibadan, Nigeria (aka YRI population); 45 samples were from unrelated Japanese people in Tokyo, Japan (aka JTY population); and 45 samples were from unrelated Han Chinese people living in Beijing, China (International HapMap Consortium, 2003).

Using the data generated by this project the researchers were able to construct what amounts to an emperical distribution of linkage disequilibrium for each of these 4 populations. The data from this massive project is available in the public domain from the project's website `http://www.hapmap.org`.

What makes the HapMap database so useful is that rather than having to genotype all 10 million SNPs, researchers can instead genotype so-called *tag SNPs*. Due to the correlation structure imposed by LD, tag SNPs capture most of the information on the pattern of genetic variation within that region resulting in huge savings in cost and time.

The data from the HapMap project can be explored using the UCSC Genome Browser `http://genome.ucsc.edu/`. One useful feature of the Genome Browser is the capability to represent the LD structure in a region of the genome for any of the HapMap populations in the form of a heat map. We present an example of such a plot and explain its interpretation in figure 1.4.

1.4 X chromosome

The X chromosome is the 8th longest chromosome in humans at more than 153 million bp, some 6 times longer than the Y chromosome. Recombination can occur along the length of the X chromosome in females, whereas in males recombination is restricted to the so-called pseudo-autosomal regions (PARs).

The PARs are two short regions at the tips of the X chromosome that recombine with similar pseudo-autosomal regions on the Y in the same manner as recombination in the autosomes (see figure 1.5 for a schematic of the PARs). For all intents and purposes we can treat loci in these PARs as we would any autosomal loci for our statistical analyses. Unless otherwise stated, when we describe a locus on the X chromosome we mean a locus outside the pseudo-autosomal regions.

Another biological phenomenon unique to the X chromosome is X inactivation (XCI) whereby one of the female X chromosomes is silenced early in development and remains inactive in somatic tissues thereafter. This can be thought of as a dosage compensation mechanism to equalize the expression of X-related characteristics in females and males. The choice of X chromosome to be silenced is a random one, but the inactive chromosome is

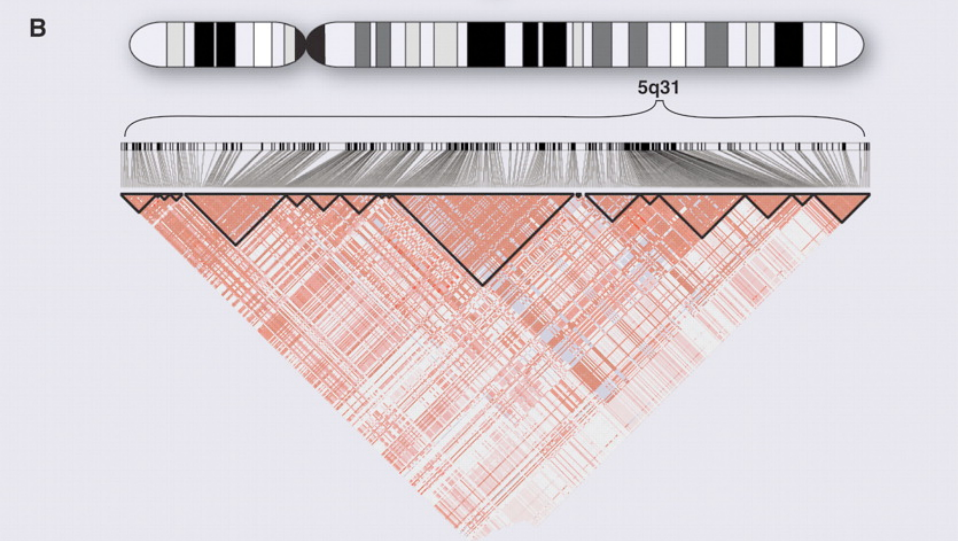

Figure 1.4: This diagram depicts actual data from the International HapMap Project, showing 420 genetic variants in a region of 500 kb on human chromosome 5q31. Positions of the variants and the pairwise correlations are shown below. Those regions which are brighter red correspond to regions containing higher levels of linkage disequilibrium. Blocks of strong correlation are indicated by the black outlines. Source: Altshuler et al. (2008)

reactivated and undergoes recombination with the second X chromosome at meiosis (Ross et al., 2005).

Importantly, we are unable to detect which of the X chromosomes has been inactivated from genotype data alone. The genotyping process is performed on a random subset of lymphocytes[2] (blood cells) and will therefore display a roughly $50:50$ split of the A and B alleles in female heterozygotes.

The details of X-inactivation are not yet fully understood and it is a far more complicated process than the presentation given here (see Chow and Heard, 2009, for a discussion of current understanding and research). We will assume, as is standard in all current methods for statistical analysis of X chromosome data, that the inactivation process is homogeneous. That is, we assume the inactive X chromosome is completely silenced in females, though in reality this is not true as the inactivation process is heterogeneous and some genes escape inactivation (see Carrel and Willard, 2005).

It is clear that due to these differences between the X chromosome and the autosomes that the statistical analysis of X chromosome data requires

[2]If the DNA sample is derived from a saliva sample then the genotyping is performed on a random subset of buccal epithelial cells and white blood cells. Source: `http://www.dnagenotek.com/DNA_Genotek_Support_FAQs_DNA.html`

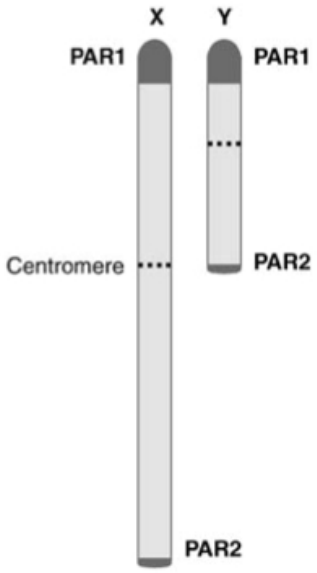

Figure 1.5: A schematic showing the locations of the pseudo-autosomal regions, PAR1 and PAR2, on the human X and Y chromosomes (not to scale). Source: Flaquer et al. (2008)

its own specialised methods.

1.4.1 Hardy-Weinberg equilibrium for the X chromosome

Hardy-Weinberg equilibrium for the X chromosome is more subtle than for the autosomes. For those loci in the PARs the derivation for HWE is identical to that of the autosomes, but for the remaining X loci we have a different situation.

Consider a biallelic locus on the X chromosome and either of the two alleles at that locus. At generation n let the frequency of the given allele in females be q_n and in males be r_n. Under our stated assumptions for HWE, one can show that q_n and r_n converge quickly to the value $p = \frac{2}{3}q_0 + \frac{1}{3}r_0$. Twice as much weight is attached to the initial female frequency since females have two X chromosomes while males have only one.

For a male we have the following recurrence relation

$$r_n = q_{n-1} \tag{1.1}$$

since a male always inherits his X chromosome from his mother, and his mother precedes him by one generation. Likewise, the frequency in females is the average frequency for the two sexes from the preceding generation:

$$q_n = \frac{1}{2}q_{n-1} + \frac{1}{2}r_{n-1}. \tag{1.2}$$

Equations 1.1 and 1.2 together imply

$$\begin{aligned}\frac{2}{3}q_n + \frac{1}{3}r_n &= \frac{2}{3}(\frac{1}{2}q_{n-1} + \frac{1}{2}r_{n-1}) + \frac{1}{3}q_{n-1} \\ &= \frac{2}{3}q_{n-1} + \frac{1}{3}r_{n-1}. \end{aligned} \tag{1.3}$$

It follows that the weighted average $\frac{2}{3}q_n + \frac{1}{3}r_n = p$ for all n.

From equations 1.2 and 1.3, we deduce that

$$\begin{aligned} q_n - p &= q_n - \frac{3}{2}p + \frac{1}{2}p \\ &= \frac{1}{2}q_{n-1} + \frac{1}{2}r_{n-1} - \frac{3}{2}(\frac{2}{3}q_{n-1} + \frac{1}{3}r_{n-1}) + \frac{1}{2}p \\ &= -\frac{1}{2}q_{n-1} + \frac{1}{2}p \\ &= -\frac{1}{2}(q_{n-1} - p) \end{aligned}$$

Continuing in this manner we get,

$$q_n - p = \Big(-\frac{1}{2}\Big)^n (q_0 - p). \tag{1.4}$$

Thus the difference between q_n and p diminishes by half at each generation, and q_n approach p in a zigzag manner. The male frequency r_n displays the same behaviour but lags behind by one generation due to (1.1). In contrast to the autosomal case, it takes more than one generation to achieve Hardy-Weinberg equilibrium for the X chromosome

In the extreme case that $q_0 = 0.75$ and $r_0 = 0.12$, figure 1.6 plots q_n and r_n for 10 generations and we see that equilibrium is still approached relatively fast.

Under HWE, the female genotypes A/A, A/B, and B/B have frequencies p_1^2, $2p_1p_2$, and p_2^2 respectively. The male *hemizygous* genotypes $A/-$ and $B/-$ have frequences p_1 and p_2.

1.5 Modern genetics

The field of genetics had advanced rapidly in the time since the publication of the first working draft of the human genome in by the International Human Genome Sequencing Consortium (2001). Beginning in 1990, it took this international team of hundreds of scientists 10 years to produce a map that covered approximately 94% of the human genome. The technological and scientific advances have been so great since that time that researchers are already talking of full genome sequencing as becoming a routine part of genetic research (Kahvejian et al., 2008).

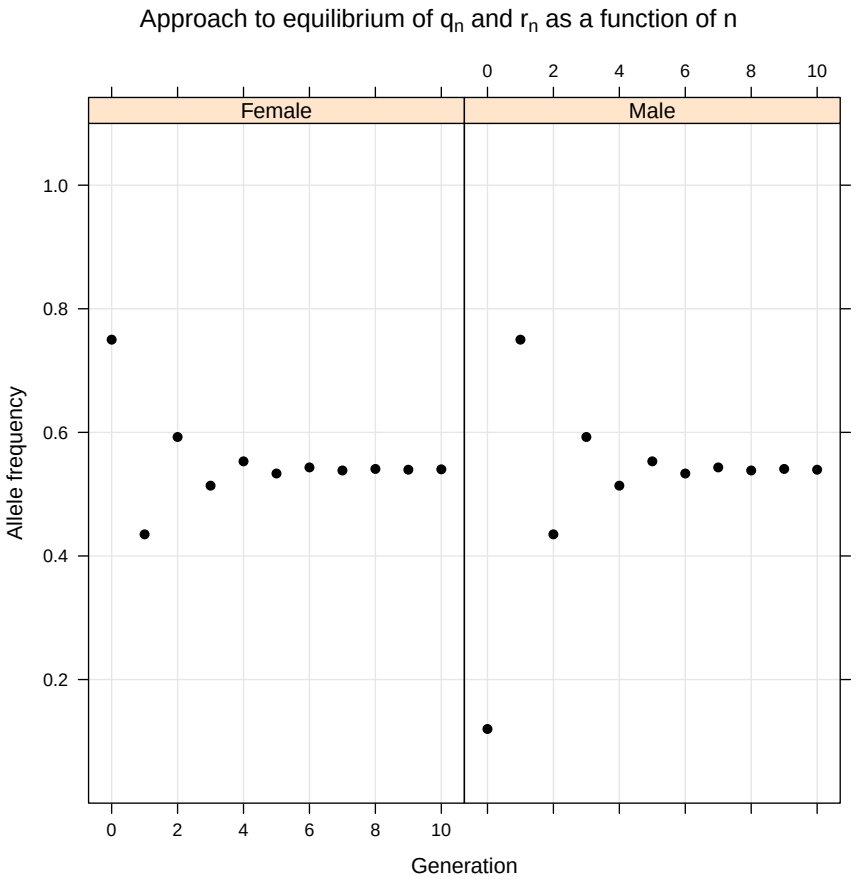

Figure 1.6: After just 6 generations both the allele frequency in females, q_n, and the allele frequency in males, r_n, are close to equilibrium. Note how the male allele frequency lags behind the female allele frequency by one generation

While the necessary reductions in cost and time are still a few years away for full sequencing to be feasible, there are existing technologies that are proving very powerful in helping us to understand complex genetic mechanisms.

One such technology is known as SNP genotyping which samples the genome at specific locations to provide a measurement of genetic variation between members of a species. We will explain the basic idea behind the genotyping technology and where it has been very useful in application.

1.5.1 SNP genotyping

SNPs have proven in recent years to be a powerful measure of genetic variation in humans due to their abundance in the genome, the knowledge we have of their distribution (from HapMap), and the relative ease with which a large number of individuals can be assayed, or genotyped, using the technology known as SNP chips.

Current SNP chips allow the assaying of over 1 million SNPs per individual, which provides a large data set for analyzing associations between genotype and phenotype. The two main producers of SNP chip technology are Illumina (`www.illumina.com`) and Affymetrix (`www.affymetrix.com`). The two platforms use different technologies and chemistries. We will focus on the Illumina platform, in particular the Illumina Hap370CNV chip used by the Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009).

Genotyping procedure

For each individual in the study a DNA sample is required. Generally this is obtained via a blood sample, though saliva samples are possible too[3] and produce DNA of a similar quality for genotyping (Bahlo et al., 2009). This DNA sample then undergoes various biochemical procedures to cut the DNA into strands that are then repeatedly copied, via process known as "amplification", so that there is a sufficient amount of DNA for the genotyping process. This processed sample is then "washed" across the SNP chip where the DNA sample hybridizes (binds) to the relavent probes.

The Illumina Hap370CNV chip has over 318,000 SNP probes to assay genotypes as well as some 52,000 probes designed to measure copy number variation. We know the precise location of these SNP markers in the genome, and their variation patterns in certain populations, due to the HapMap project. The following technical explanation of the Illumina BeadChip platform is from Ritchie et al. (2009).

"Illumina BeadChips are composed of a number of rectangular strips, each containing many randomly arranged, replicated beads. For Infinium

[3]Indeed some saliva samples are used in the ANZgene study

genotyping, beads are coupled with specific 50mer probes designed to be complementary to the sequence adjacent to the SNP site, and the two alleles (A, B) are discriminated using either a red or green dye (Steemers *et al.*, 2006). Data are acquired by scanning each strip at different wave lengths using Illumina's scanning device followed by automatic image analysis (Galinsky, 2003). A robust summary of the intensity in each chanel for each SNP assayed is reported in the proprietary idat files."

Using Illumina's proprietary GenCall algorithm, or an open source genotype calling algorithm such as CRLMM (Ritchie et al., 2009), the genotypes of the samples are then "called" from the idat files.

The clustering algorithms used to assign genotypes based on the idat files will not be discussed in detail here, suffice to say that each SNP assayed on the chip is assigned a genotype call. The SNPs are biallelic and thus the genotype calls are generically called as either A/A, A/B or B/B, with some measure of confidence also reported for this assignment procedure. Additionally a SNP may be called NC, or "no-call", if the calling algorithm cannot assign the marker to a genotype cluster with sufficient confidence. Poorly performing samples and markers are removed in the quality control procedures prior to analysis of the data.

1.6 Genome wide association studies

A genome wide association study (GWAS) can be thought of as a new form of the classical case/control study. Performing a GWAS has only become feasible in the past few years due to advances in SNP chip technology and the resulting savings in time and money. We will briefly outline the aim of a GWAS, the challenges it provides to statisticians, and the current consensus on "best practice" on a few key issues. For a nice review of the state of the art for GWA studies see the paper of McCarthy et al. (2008).

A GWAS is a retrospective study where samples (people) are selected based on the presence or absence of a particular phenotype of interest, for example multiple sclerosis. Each person is genotyped using SNP chip technology to examine the associations between genotype and phenotype.

What makes these studies such particularly fruitful ground for statisticians is the sheer size of the data generated by a GWAS. As McCarthy et al. note, "the initial wave of GWA studies has shown that, with rare exceptions, the effect sizes results from common SNP associations are modest, and that samples sizes in the thousands are essential". That is to say, each genetic variant is believed to have only a small effect on the overall phenotypic variation so we require large sample sizes to have any hope of detecting them. A typical GWAS may have 3000 samples with 10^6 observations per sample, and the data sets are only getting bigger.

After this initial "discovery" phase of a GWAS the results must be repli-

cated in an independent cohort to ensure their validity. As a result, a GWAS takes a large team of scientists, medical workers and statisticians some years to carry out — as well as considerable money. GWA studies are a particularly hot area in current genetics research, and have also generated plenty of controversy during their brief history.

Successes

The first major GWAS, performed by Sladek et al., was published in Nature in February 2007 and identified a novel risk locus for type 2 diabetes. A few months later in June 2007, perhaps the seminal paper on GWA studies was published in Nature by the Wellcome Trust Case Control Consortium (WTCCC).

The WTCCC studied 14,000 cases of seven common diseases — bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes — and used a shared pool of 3,000 controls. The original paper identified 24 independent association signals with $P < 5 \times 10^{-7}$, thus validating GWA studies as powerful research tools and establishing the experimental framework for performing them.

The popularity of GWA studies, and their power, is evidenced by the 398 published genome-wide associations through to March 2009 with $P \leq 5 \times 10^{-8}$ for phenotypes ranging from breast and prostate cancer, through to hair and iris colour, through to weight and nicotine dependence (Hindorff LA, Junkins HA, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: `www.genome.gov/gwastudies` accessed 30/08/2009).

Problems

Problems can arise in GWA studies when researchers overreach and try to claim too much from their results. It is important to remember that the **A** in GWAS stands for *association*, a point that boils down to the classic statistical rule "correlation does not imply causation". A GWAS can only identify associations between genotype and phenotype and much more "traditional" biological work is required to uncover the molecular mechanism that underpins the association, if indeed it exists.

In this sense, many think of GWA studies as hypothesis-free, or hypothesis-forming studies, rather than hypothesis driven research tools. A GWAS can motivate a new line of research from its results but on its own cannot determine genetic causality. Some statistical challenges particular to the design of a GWAS and the analysis of the data will be further discussed in chapter 3.

Results commonly reported for a GWAS are the odds ratios (see section 2.2.1) of having the disease against not having the disease given a particular genotype. These odds ratios for individual loci are typically quite

small, of the order $1 - 2$, so the question of their meaning or relevance for treatment of the disease naturally arises. Thus even once these statistical challenges are overcome we are still left with the difficult task of interpreting the results so as best to direct further research and to explain the results to the wider community.

The challenges of GWA studies are many and varied, but statisticians are frequently suited to tackling these challenges. One such challenge, and the focus of my thesis, is "the problem of testing for genotype-phenotype association with loci on the X chromosome, (which) has received surprisingly little attention" (Clayton, 2008).

Chapter 2

Overview of some statistical techniques relevant to GWAS data

We will review some statistical techniques necessary for this project including generalized linear models, score tests and Pearson's χ^2 test, and how these relate to one another.

2.1 Generalized linear models

Generalized linear models (GLMs) are a flexible extension to the classical linear model. This generalization allows us to model data from many distributional forms in a similar way to the classical linear model. GLMs bring together many statistical techniques into one cohesive framework and have proved a very powerful and popular tool since its initial formulation by Nelder and Wedderburn (1972). For an extensive study of GLMs we refer the reader to McCullagh and Nelder (1989) or for a more elementary introduction Dobson (2002); both of which we reference in the following.

Exponential family of distributions

For a GLM we assume that each component of the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ has a distribution in the exponential family. The exponential family is a class of probability distributions with many useful properties and includes most common distributions such as the Poisson, normal and binomial distributions. The distribution of a random variable Y belongs to the exponential family if it can be written in the form

$$f_Y(y; \theta, \phi) = \exp\left\{ \big(y\theta - b(\theta)\big)/\phi + c(y; \phi) \right\} \tag{2.1}$$

for some specific ϕ, $b(\cdot)$ and $c(\cdot)$ (McCullagh and Nelder, 1989). If ϕ is known the distribution is said to be in canonical form and θ is commonly referred to as the natural parameter. As an example, we show that the binomial distribution is a member of the exponential family of distributions.

Example 2.1. The binomial distribution is a member of the exponential family of distributions.

Proof. Let $Y \stackrel{d}{=} \text{Bin}(n, \pi)$, then

$$\begin{aligned} f_Y(y;\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y}, \\ &= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right\}, \end{aligned}$$

which is of the form of (2.1) with

$$\begin{aligned} &\theta = \log\left(\frac{\pi}{1-\pi}\right), \quad b(\theta) = -n\log(1-\pi) = n\log(1+e^\theta), \\ &\phi = 1, \quad c(y;\phi) = \log\binom{n}{y}. \end{aligned}$$

We see from this example that $\log\left(\frac{\pi}{1-\pi}\right)$ is the natural parameter for the binomial distribution. □

GLM formulation

For a GLM we have three key components:

1. Each element Y_i of the response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is independent and has a distribution of the same form from the exponential family (e.g. all binomial or all Poisson).

2. A set of parameters $\boldsymbol{\beta}$ and a vector of predictors $X_i = (x_{i1}, \ldots, x_{ip})$ for each of the Y_i which we combine together in the design matrix as

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}$$

3. A monotone link function g such that

$$g(\mu_i) = X_i^T\boldsymbol{\beta} = \eta_i$$

where

$$\mu_i = \mathbb{E}(Y_i)$$

We can then write the joint density of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ as

$$f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\Big\{ \sum_{j=1}^{n} \frac{y_j \theta_j - b(\theta_j)}{\phi_j} + c(y_j; \phi_j) \Big\} \tag{2.2}$$

where $\theta_j = \theta(\eta_j)$.

2.2 Logistic regression

One popular generalized linear model is logistic regression. Logistic regression models binary responses, such as binomial data, in terms of a set of continuous and categorical predictor variables.

The logistic regression model for data $Y_i \stackrel{d}{=} \text{Binomial}(n_i, \pi_i), \quad i = 1, \ldots, n$ is of the form

$$g(\pi_i) = \text{logit}(\pi_i) = \log\Big(\frac{\pi_i}{1-\pi_i}\Big) = \alpha + X_i \boldsymbol{\beta} \tag{2.3}$$

where α is the intercept term. At first glance g may not appear to be a function of μ_i but noting that $\mu_i = n_i \pi_i$ this can be re-written as $g(\mu_i) = \log\Big(\frac{\mu_i}{n_i - \mu_i}\Big)$.

We obtain the same estimates of $\boldsymbol{\beta}$ regardless of whether we group observations as frequencies by covariate patterns (i.e. $n_i =$ number of observations with the ith covariate pattern) or code each individual as 0 or 1 and write the individual's covariate pattern separately (i.e. $n_i = 1, \quad i = 1, \ldots, n$).

Logistic regression can also be used to analyse $2 \times k$ contingency tables since it models binary outcome variables. There is a wealth of literature on logistic regression so we just highlight a few relevant points here.

2.2.1 Odds ratio

The odds ratio is a relative measure of the odds of an event occurring in one group compared to it occurring in another group. It plays an important role in the interpretation of results from a logistic regression model. We will first introduce the concept using a simple example and then show how it is generalized in the logistic regression framework.

Example 2.2. Consider the simple setup where there is some event D of interest whose presence is studied in two groups denoted G_1 and G_0. For example, let D be the event of contracting a specific disease, and D^c (the complement of D) be the event of not contracting the disease, with G_1 the group of exposed individuals and G_0 the group of unexposed individuals.

We define the joint distribution of D and G by $p_{11} = \Pr(D, G_1), p_{10} = \Pr(D, G_0), p_{01} = \Pr(D^c, G_1)$ and $p_{00} = \Pr(D^c, G_0)$ where $p_{11} + p_{10} + p_{01} + p_{00} = 1$. The joint distribution of D and G can be summarised as in table 2.1.

		Disease status	
		D	D^c
Group	G_1	p_{11}	p_{01}
	G_0	p_{10}	p_{00}

Table 2.1: Joint distribution of disease status and group. For example, $\Pr(D, G_0) = p_{10}$

Alternatively, we can consider the conditional probabilities of contracting the disease, D, given the exposure level G_i $(i = 0, 1)$. These probabilities are defined as in table 2.2.

		Conditional probability of disease	
		$\Pr(D\|\cdot)$	$\Pr(D^c\|\cdot)$
Group	G_1	$p_{11}/(p_{11} + p_{01})$	$p_{01}/(p_{11} + p_{01})$
	G_0	$p_{10}/(p_{10} + p_{00})$	$p_{00}/(p_{10} + p_{00})$

Table 2.2: Conditional distribution of disease status given group. For example, $\Pr(D|G_0) = p_{10}/(p_{10} + p_{00})$

The odds ratio (OR) for this conditional table is given by

$$\begin{aligned} OR &= \frac{p_{11}/(p_{11} + p_{01})}{p_{01}/(p_{11} + p_{01})} \Big/ \frac{p_{10}/(p_{10} + p_{00})}{p_{00}/(p_{10} + p_{00})} \\ &= \frac{p_{11}}{p_{01}} \Big/ \frac{p_{10}}{p_{00}} \\ &= \frac{p_{11}p_{00}}{p_{01}p_{10}} \end{aligned} \tag{2.4}$$

i.e. the odds of contracting the disease when exposed compared to the odds of contracting the disease when not exposed.

The concept of odds ratios can be extended to more complex situations via logistic regression. Suppose we have a binary response variable Y and predictor variables $X, Z_1, \ldots, Z_p$, where X is binary and the $Z_1, \ldots, Z_p$ may or may not be binary. If we use logistic regression to model Y given $X, Z_1, \ldots, Z_p$, the estimated coefficient $\hat{\beta}_x$ for X is related to a conditional odds ratio by the following:

$$\exp(\hat{\beta}_x) = \frac{\Pr(Y = 1|X = 1, Z_1, \ldots, Z_p)/\Pr(Y = 0|X = 1, Z_1, \ldots, Z_p)}{\Pr(Y = 1|X = 0, Z_1, \ldots, Z_p)/\Pr(Y = 0|X = 0, Z_1, \ldots, Z_p)}.$$

The interpretation of $\exp(\hat{\beta}_x)$ is an estimate of the odds ratio at the population level between Y and X when the values of $Z_1, \ldots, Z_p$ are held fixed.

A very useful property of the odds ratio is that the estimate of the odds ratio is invariant under certain sampling schemes. This will be expanded upon in section 2.3.

2.2.2 Relative risk

Another important measure, particularly in epidemiology and medical statistics, is that of relative risk (RR). Returning to the conditional distribution in our simple example (see table 2.2) we define the relative risk as

$$RR = \frac{\Pr(D|G_1)}{\Pr(D|G_0)} = \frac{p_{11}/(p_{11}+p_{01})}{p_{10}/(p_{10}+p_{00})} \tag{2.5}$$

i.e. the probability of contracting the disease when exposed compared to the probability of contracting the disease when not exposed.

It is important to note the difference between equation (2.4) and equation (2.5). It is only for very small probabilities of having the disease, p_{11} and p_{10}, that $p_{11}/(p_{11}+p_{01}) \approx p_{11}/p_{01}$ and $p_{10}/(p_{10}+p_{00}) \approx p_{10}/p_{00}$ so that

$$RR = \frac{p_{11}/(p_{11}+p_{01})}{p_{10}/(p_{10}+p_{00})} \approx \frac{p_{11}/p_{01}}{p_{10}/p_{00}} = OR$$

Thus for very rare diseases the relative risk can be approximated by the odds ratio but it is important to remember that in general the odds ratio and the relative risk are quite different concepts.

2.3 Retrospective vs prospective sampling

Retrospective and prospective sampling are two fundamental sampling methods frequently used in statistics. Each sampling method has the aim of identifying associations between an outcome and a set of predictors, but there are substantial differences between the two sampling schemes. Both methods of study require *cohorts* of individuals who are ideally matched in as many ways as possible (such as sex, age, ethnicity, etc.) but differ by a certain key characteristic. This main characteristic on which they differ is normally the variable of interest, for example the incidence of lung cancer in a cohort of smokers versus a cohort of non-smokers.

The variables used for predicting outcomes are often referred to as exposure variables in the epidemiology literature — for our lung cancer example these may be the number of cigarettes smoked per week, weight, and family medical history for each person.

In a retrospective study the samples are assigned to cohorts based on their outcome variable, e.g. the presence or absence of lung cancer, and the exposure variables are then collected from *past records*. For a prospective study the samples are assigned to cohorts based on their exposure variables and followed over time to see how these factors affect their eventual outcome.

A retrospective study for the effects of smoking on lung cancer would be selecting a cohort of people with lung cancer and a similar cohort without lung cancer and determining their history of smoking. In contrast, the prospective study for this same experiment would be selecting a cohort of smokers and a similar cohort of non-smokers and then following both groups over a number of years to determine the rates of lung cancer in each group.

Each of these study designs has its advantages, however it is clear a retrospective study has the benefit of being less time consuming and cheaper to perform than a prospective study.

For a GWAS the retrospective case/control model is clearly the more efficient study design. But these efficiencies would not be worthwhile if the results of a retrospective study were unreliable when compared to those obtained from a prospective study. Fortunately there is a nice result, which we will prove in section 2.3.1, that shows the results are identical under certain conditions.

Furthermore, for a GWAS the interest is in *identifying* any genetic predictors; estimating the accuracy of the effect size is not so relevant at this stage. Indeed, it has been argued that a retrospective study design can result in inflated estimates of the odds ratio since the sampling may be performed to obtain a so-called *hypernormal* control cohort (see, for example, McCarthy et al., 2008; Cordell and Clayton, 2002). These hypernormal (resp. *hyperabnormal* for a case cohort) are not representative of the wider population but have been selected to increase the chance of finding the genetic predictors.

2.3.1 Invariance of odds ratios under retrospective and prospective sampling

We will show that the estimate of the odds ratio is identical under either prospective or retrospective sampling when logistic regression is used. More precisely, "one important property of the logistic function not shared by the other link functions is that differences on the logistic scale can be estimated regardless of whether the data are sampled *prospectively* or *retrospectively*" (McCullagh and Nelder, 1989). This idea will be illustrated using a simple example, followed by a proof for the general case.

Example 2.3. Suppose that a population is partitioned according to two binary variables, (D, D^c) referring to the presence or absence of disease, and (G, G^c) referring to the presence or absence of a specific genetic mutation. Suppose that the proportions of the population in the four categories thus formed are as shown in table 2.3.

In a prospective study, a group of subjects is selected with the genetic mutation together with a comparable group of subjects without the genetic mutation. The progress of each group is monitored, often over a prolonged

	G	G^c	Total
D	$\pi_{00} = 0.04$	$\pi_{01} = 0.01$	$\pi_{0.} = 0.05$
D^c	$\pi_{10} = 0.16$	$\pi_{11} = 0.79$	$\pi_{1.} = 0.95$
Total	$\pi_{.0} = 0.2$	$\pi_{.1} = 0.8$	1

Table 2.3: Hypothetical frequencies of disease and genetic status

period, with a view towards comparing the incidence of the disease in the two groups. In this design the column totals in table 2.3 can be thought of as being fixed by design while the row totals are random, reflecting the incidence of the disease in the overall population.

In a retrospective study the subjects are sampled based on their disease status and it is their genetic status that is now considered as random. In this way the row totals can now be thought of as fixed by design while the column totals are random, reflecting the frequency of the genetic mutation in the population.

Considering the prospective study first, the logits (logarithm of the odds) for the two genetic groups are

$$\begin{aligned} \log(\pi_{10}/\pi_{00}) &= \log(4) = 1.386 \\ \log(\pi_{11}/\pi_{01}) &= \log(79) = 4.369. \end{aligned}$$

The log odds ratio is thus

$$\begin{aligned} \log(OR) &= \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00}) \\ &= 2.983, \end{aligned}$$

from which we can find $\widehat{OR} = \exp(2.983) = 19.74697$. But this could equally be estimated by sampling retrospectively from the two disease groups D and D^c since

$$\begin{aligned} \log(OR) &= \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00}) \\ &= \log(\pi_{11}/\pi_{10}) - \log(\pi_{01}/\pi_{00}) \\ &= \log(4.9375) - \log(0.25) \\ &= 1.597 - (-1.386) \\ &= 2.983, \end{aligned}$$

also gives $\widehat{OR} = \exp(2.983) = 19.74697$. In fact, for this example a retrospective study design will be much more efficient than a prospective study. For a prospective study to be effective would require a very large number of healthy individuals to be involved and followed up for a long time in order for a sufficient number of subjects to fall victim to the disease.

However, for a retrospective study a large number of subjects with the disease can be identified, via hospital records for example, and their genotypes assayed using SNP chips. The advantages of a retrospective study design here are clear and they, along with the reasons given in section 2.3, are why GWA studies are generally performed using a retrospective sampling study design.

We will now extend the above argument to a vector of covariates $\mathbf{x}$ (such as multiple SNP genotypes). Provided the intercept is treated as a nuisance parameter we show that the estimates of the regression parameters are identical, regardless of whether the sample is obtained prospectively or retrospectively.

We may write the linear logistic model in the form

$$\Pr(D|\mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})/[1 + \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})] \tag{2.6}$$

for the probability of contracting the disease given that the subject has covariates $\mathbf{x}$.

Model (2.6) is appropriate for data that has been sampled prospectively; however suppose the data was instead collected retrospectively. It is essential that the sampling proportions depend on disease status only and not on the covariates, $\mathbf{x}$. We introduce a dummy variable Z to define whether an individual is sampled or not, and denote these sampling proportions by

$$\begin{aligned}\Pr(Z = 1|D, \mathbf{x}) &= \Pr(Z = 1|D) = q_0\\ \Pr(Z = 1|D^c, \mathbf{x}) &= \Pr(Z = 1|D^c) = q_1\end{aligned}$$

We can now apply Bayes' Theorem to compute the disease frequency among sampled individuals who have a specified covariate vector $\mathbf{x}$.

$$\begin{aligned}\Pr(D|Z = 1, \mathbf{x}) &= \frac{\Pr(Z = 1|D, \mathbf{x})\Pr(D|\mathbf{x})}{\Pr(Z = 1|D, \mathbf{x})\Pr(D|\mathbf{x}) + \Pr(Z = 1|D^c, \mathbf{x})\Pr(D^c|\mathbf{x})}\\ &= \frac{q_0\left[\frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right]}{q_0\left[\frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right] + q_1\left[1 - \frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right]}\\ &= \frac{q_0\left[\frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right]}{q_0\left[\frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right] + q_1\left[\frac{1}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{x})}\right]}\\ &= \frac{q_0\exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})}{q_0\exp(\alpha + \boldsymbol{\beta}^T\mathbf{x}) + q_1}\\ &= \frac{\exp(\alpha^* + \boldsymbol{\beta}^T\mathbf{x})}{1 + \exp(\alpha^* + \boldsymbol{\beta}^T\mathbf{x})}\end{aligned} \tag{2.7}$$

where $\alpha^* = \alpha + \log(q_0/q_1)$.

Comparing (2.7) with (2.6) we see that the two equations have identical $\boldsymbol{\beta}$ and only differ in their intercept terms, α^* and α respectively. In other words, the logistic model (2.6) continues to apply with the same coefficients $\boldsymbol{\beta}$ but a different intercept. It follows therefore, that the logistic models described here in the context of prospective studies can be applied to retrospective studies provided that the intercept is treated as a nuisance parameter.

2.4 Score tests

The score test, sometimes known as Rao's score test, is based on the asymptotic distribution of the score statistic $\partial\ell/\partial\boldsymbol{\theta}$. The following derivation of the score test is from Smyth (2003).

Let $\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})$ be a log-likelihood function depending on a response vector $\boldsymbol{y}$ and parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The score test is a test of the hypothesis $H_0 : \boldsymbol{\theta}_2 = 0$ against the alternative that $\boldsymbol{\theta}_2$ is unrestricted. We call the parameters $\boldsymbol{\theta}_1$ nuisance parameters as we are not interested in them but they must still be estimated for the score test statistic to be computed. The likelihood score vectors for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the partial derivatives

$$\dot{\ell}_1 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_1}$$

and

$$\dot{\ell}_2 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_2}$$

respectively. The observed information matrix for the parameters is $-\ddot{\ell}$ with

$$\ddot{\ell} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_1 \boldsymbol{\theta}_2^T} = \begin{pmatrix} \ddot{\ell}_{11} & \ddot{\ell}_{12} \\ \ddot{\ell}_{21} & \ddot{\ell}_{22} \end{pmatrix}.$$

The expected or Fisher information matrix is $\mathcal{I} = \mathbb{E}(-\ddot{\ell})$, which is partitioned conformally with $\ddot{\ell}$ as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}.$$

The score test statistic is based on the fact that the score vector $\dot{\ell}$ is normally distributed with mean zero and covariance matrix $\mathcal{I}$. If the nuisance parameter $\boldsymbol{\theta}_1$ is known, then the score test statistic of H_0 is

$$Z = \mathcal{I}_{22}^{-1/2} \dot{\ell}_2,$$

where $\mathcal{I}_{22}^{1/2}$ stands for any factor such that $\mathcal{I}_{22}^{1/2}(\mathcal{I}_{22}^{1/2})^T = \mathcal{I}_{22}$, or equivalently

$$S = Z^T Z = \dot{\ell}_2^T \mathcal{I}_{22}^{-1} \dot{\ell}_2$$

with ℓ_2 and $\mathcal{I}_{22}$ evaluated at $\boldsymbol{\theta}_2 = 0$. The score vector $\dot{\ell}$ is a sum of terms corresponding to individual observations and so is asymptotically normal under standard regularity conditions. It follows that Z is asymptotically a standard normal p_2-vector under the null hypothesis H_0 and that S is asymptotically chi-square distributed on p_2 degrees of freedom, where p_2 is the dimension of $\boldsymbol{\theta}_2$.

If the nuisance parameters are not known (as is often the case), the score test requires their substitution by their maximum likelihood estimators $\hat{\boldsymbol{\theta}}_1$ under the null hypothesis. Setting $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ is equivalent to setting $\dot{\ell}_1 = 0$, so we need the asymptotic distribution of $\dot{\ell}_2$ conditional on $\dot{\ell}_1 = 0$, which is normal with mean zero and covariance matrix

$$\mathcal{I}_{2.1} = \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}. \tag{2.8}$$

The score test becomes

$$S = \dot{\ell}_2^T \mathcal{I}_{2.1}^{-1} \dot{\ell}_2 \tag{2.9}$$

with $\dot{\ell}_2$ and $\mathcal{I}_{2.1}$ evaluated at $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2 = 0$.

If $\mathcal{I}_{12} = 0$ then $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be orthogonal. In that case, $\dot{\ell}_1$ and $\dot{\ell}_2$ are independent and $\mathcal{I}_{2.1} = \mathcal{I}_{22}$, meaning that the information matrix $\mathcal{I}_{22}$ does not need to be adjusted for estimation of $\boldsymbol{\theta}_1$.

As an example of where score tests may be useful in the GWAS context, consider the setting where we wish to regress (in a GLM) each sample's phenotype on their genotype at a particular SNP to asses the SNP's significance. In this case $\boldsymbol{\theta}_1$ (our nuisance parameter) is some overall mean of the sample and $\boldsymbol{\theta}_2$ (our parameter of interest) is the genotype at that SNP. The significance of the SNP can be assessed using a score test. This idea can be extended to multiple SNPs provided that p, the total number of SNPs in the full model, is less than n, the sample size.

The score test is an alternative to the likelihood ratio test or the Wald test, and as Smyth notes "the score test is often simpler than the likelihood ratio test because the statistic requires parameter estimators to be obtained only under the null hypothesis". This advantage comes to the fore in the analysis of GWAS data due to the hundreds of thousands of tests that need to be computed.

2.5 Pearson's χ^2 test

When people talk of a χ^2 test they are generally referring to Pearson's χ^2 test, perhaps the simplest and most common way to analyse data in the form of a contingency table. In what follows, when we refer to a χ^2 test we implicitly mean Pearson's test.

The χ^2 test can be used to test the hypothesis that paired observations expressed in a contingency table are independent of one another. It contrasts

the number of observed times a response occurs with the number of times you would expect if the two events were independent.

Let us consider a contingency table with two factors A and B, with c and r levels respectively. For each combination of factors, (i, j), we have an observed count O_{ij} which can be represented as cell (i, j) in a contingency table (see table 2.4).

	A_1	$\dots$	A_c	Total
B_1	O_{11}	$\dots$	O_{1c}	$\sum_{l=1}^{c} O_{1l} = O_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	O_{r1}	$\dots$	O_{rc}	$\sum_{l=1}^{c} O_{rl} = O_{r.}$
Total	$\sum_{k=1}^{r} O_{k1} = O_{.1}$	$\dots$	$\sum_{k=1}^{r} O_{kc} = O_{.c}$	$\sum_{k=1}^{r} O_{k.} = \sum_{l=1}^{c} O_{.l} = n$

Table 2.4: A general $r \times c$ contingency table

Under the null hypothesis of independence of rows and columns (i.e. no association between any levels of the two factors) we have an expected number of observations given by

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}.$$

The test statistic X^2 contrasts the observed counts with the expected counts for each cell in the form

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{2.10}$$

Under the null hypothesis, X^2 has an approximate χ^2 distribution on $rc - (r + c - 1) = (r - 1)(c - 1)$ degrees of freedom. The value of the X^2 statistic can then be compared against the corresponding $\chi^2_{(r-1)(c-1)}$ distribution to determine the significance of the result. For the 2×2 table there exists a well-known computational "shortcut" to calculate X^2, namely

$$X^2 = \frac{n\big(O_{11}O_{22} - O_{12}O_{21}\big)}{O_{.1}O_{.2}O_{1.}O_{2.}}. \tag{2.11}$$

It is worth noting that there exist so-called "exact tests" for contingency tables (see Agresti, 1992) but these will not be pursued here as for the most part it is Pearson's χ^2 that is used in GWA studies owing to the large sample sizes at play and the test's computational efficiency.

2.5.1 Relationship between Pearson's χ^2 statistic and the score test

Pearson's χ^2 statistic, in addition to its simple set up and calculation, has a deeper relationship with the score test for generalized linear models. Smyth

(2003) shows that for any generalized linear model, the Pearson goodness of fit statistic is the score test statistic for testing the current model against the saturated model.

A corollary of this result is that the χ^2 test for independence in a $r \times c$ contingency table is a score test statistic, based on the assumption that the counts are independent and Poisson distributed. This amounts to the logistic regression assumption that the Y_i are binomially distributed, conditional on the row totals.

Chapter 3

Statistical challenges particular to GWAS

GWA studies provide many statistical challenges due to the high dimensionality of the data. Many standard techniques fail, or must be modified, as we have many more predictor variables than observations. A typical GWAS has 300,000 - 1,000,000 observations (SNP genotypes) per sample, but only a few thousand samples (people), which puts us in the realm of $p \gg n$ dimensionality.

In addition to the challenges of the high dimensionality of the data, there is much work in the pre-processing of the data. This includes, for example, quality control procedures and the imputation of missing data, and will only be given a cursory overview here. A nice review of the range of statistical methods used in GWA studies can be found in Balding (2006).

3.1 Pre-processing of data

There is a large amount of time and effort spent on the so-called "cleaning" of the data that is produced by the genotyping procedure. Potential sources of bias and contamination include hidden population structure due to population stratification and "cryptic relatedness". Also, with such complicated technology involved in the genotyping process, errors inevitably occur and need to be dealt with prior to any analysis of the data due to the possibility of spurious associations.

There exist a variety of quality-control procedures including the removal of poorly performing SNPs, and the removal of poorly performing samples. We will highlight some of the quality control (QC) procedures implemented by the ANZgene study to give a sense for how these issues are handled in practice.

The first step of QC is the removal of entire samples that fail the genotyping process. In the ANZgene study (discussed in chapter 5), a criterion

for an entire sample to be removed was if the genotype call rate[1] was less than 98%. The call rate of a sample gives a strong indication of the quality of the genotyping procedure, though it will of course not detect samples that have been repeatedly miss-called as opposed to no-called.

The second stage of QC is the removal of individual SNPs across all the remaining samples. For the ANZgene study, SNPs were excluded if they had a minor allele frequency $< 1\%$ or were in significant Hardy-Weinberg disequilibrium ($P < 1 \times 10^{-7}$). If not excluded, poorly performing SNPs "might be disproportionately represented among the most extreme association signals" (McCarthy et al., 2008).

The missing genotype data can often be "recovered" using imputation methods, due to the phenomenon of linkage disequilibrium (see section 1.3). The imputation of missing genotypes was performed for the ANZgene study using the HapMap data as a reference, but only for the autosomes (see Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009, Methods).

The final stage of the quality-control procedure is the removal of samples that display cryptic relatedness, and to assess hidden population structure in the remaining case and control data. Cryptic relatedness is evidence — typically gained from analysis of GWA data — that, despite allowances for known family relationships, individuals in the study sample have residual, non-trivial degrees of relatedness, which can violate the independence assumptions of standard statistical techniques (McCarthy et al., 2008). It is surprisingly common to find closely related individuals enrolled in the same study despite the best efforts of researchers to obtain "independent" samples. Inadvertent duplication, swaps, or mislabeling of samples (such as a male sample being labeled female) are also frequently revealed in the quality-control stage of a GWAS. Samples can also be removed if they display hidden population structure which is assessed using principal components analysis on a subset of the SNPs that have been "pruned" for LD.

3.2 GWAS analysis software

Some of the key challenges provided by GWA studies are computational. These challenges range from data management issues of said large data sets, through to performing the statistical analysis in the most efficient way possible. Aside from the proprietary software of the SNP chip manufacturers (e.g. Illumina GenomeStudio and Affymetrix Genotyping Studio) there has been much development of third party software for the analysis of GWAS data. This software ranges from genotype calling algorithms such as CRLMM (Carvalho et al., 2007) through to software to perform statistical

[1]For each sample the call rate = (number of SNPs genotyped - number of no-calls)/(number of SNPs genotyped)

tests on the cleaned data.

Some of these tools are written as add-ons for existing statistical programming environments such as R and SAS, while others are stand-alone programs written specifically for the analysis of GWAS data. One of the most widely used third party programs is the stand-alone software PLINK (Purcell et al., 2007). In the words of the software's authors, "PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner". It does this job very well and is widely used; for example both the ANZgene consortium and International Multiple Sclerosis Genetics Consortium used PLINK in the analysis of their respective GWA studies. However, PLINK does have its shortcomings, particularly in relation to the analysis of X chromosome data as we shall see in section 3.6.1 and chapter 4

3.3 Multiple testing

It is clear that the significance level, α, for a GWAS requires adjustment for the hundreds of thousands of hypothesis tests being performed. The simplest approach is the Bonferroni correction, which gives the genome wide significance level, α, as

$$\alpha = \alpha^*/n,$$

where n is the number of tests performed and α^* is the target significance level. Using $\alpha^* = 0.05$ and $n = 10^6$ results in a genome wide significance level of $\alpha = 5 \times 10^{-8}$.

The application of a Bonferroni correction for GWAS results has been criticised by Cordell and Clayton (2002), who argue, "approaches such as the Bonferroni correction are not appropriate because it is not the number of tests in any one investigation that is important. Rather, it is that the vast majority of loci tested will not be associated, so that even a small false positive probability will mean that most positive results will turn out to be false." They thus advocate Bayesian, or empirical Bayes methods to allow calculation of the posterior probability that an association is genuine when a prior probability of association is known. As yet however, such Bayesian methods are not well developed and the frequentist paradigm remains the dominant one for assessing genome wide significance.

Dudbridge and Gusnanto (2008) also studied the question of significance thresholds for GWA studies. Using a complicated permutation and bootstrap method, the authors suggest a genome wide significance level of 7.2×10^{-8}, a result of the same order as the far simpler Bonferroni correction. In practice a genome wide significance level of $\alpha \approx 5 \times 10^{-8}$ is a popular choice (McCarthy et al., 2008).

Having discussed the choice of an appropriate significance level I will now say that adherence to it is often somewhat loose. This is not unreasonable

since a GWAS is frequently used to prioritize regions of the genome for further analysis rather than definitively determine risk-associated loci. A common approach for selecting SNPs for further study is to rank the SNPs by p-value and select the 500 SNPs[2] with the lowest p-values as SNPs to be used in the replication phase of the study. This selection of "highly significant" markers is supplemented by SNPs of some biological interest to the researchers, such as markers within genes suspected of having a role in the disease, or SNPs with a previously reported association to the disease or related disease.

3.4 Association tests

The majority of GWA studies use single point analyses where each SNP is analysed individually in turn. While this is not ideal, for reasons to be outlined below, it has thus far been the most powerful tool for the analysis of GWAS data (McCarthy et al., 2008). Firstly, at a local level, SNPs are frequently co-linear due to linkage disequilibrium. This will cause all the usual problems associated with co-linearity of predictor variables — such as highly variable parameter estimates if using the SNPs in a regression context — but this co-linearity can have its advantages. The genotyping of multiple SNPs from a local region in the genome can be desirable because not all SNPs are equally powerful, due to differences in allele frequencies for example. Furthermore, some SNPs will fail in the genotyping procedure and so having multiple genotyped SNPs in a region means that we can often use a nearby marker as a proxy for missing data.

On a more global scale, single point analyses are not ideal owing to much biological interest in interactions between loci. The high dimensionality of GWAS data makes the systematic testing of interactions between SNPs infeasible. Any testing of interactions in a GWAS is typically confined to testing for interactions between significant main effects.

3.4.1 Autosomal association testing

We will now introduce the most common association tests for autosomal data and derive some key results for these. This is in anticipation of our discussion of methods specific to the X chromosome in section 3.6.

We begin, of course, by introducing some notation. Consider a SNP with two possible alleles, A and B, resulting in the three possible genotypes A/A, A/B, and B/B (we ignore no-calls here). We have genotype data for R cases and S controls giving a total of $R + S = N$ genotypes. The respective genotype frequencies are given as in table 3.1 with the subscripts denoting

[2]Or another somewhat arbitrary number of SNPs

the number of B alleles. At each locus we have one observation per sample — the genotype.

	Genotype			
	A/A	A/B	B/B	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 3.1: Generic genotype table for autosomal data

As each person has two alleles at any autosomal locus, the data can also be summarised by counting the number of alleles present, rather than the frequencies of genotypes (see table 3.2). At each locus we now have *two* observations per sample — 1 observation for each of the alleles.

	Allele		
	A	B	Total
Cases	$2r_0 + r_1$	$2r_2 + r_1$	2R
Controls	$2s_0 + s_1$	$2s_2 + s_1$	2S
Total	$2n_0 + n_1$	$2n_2 + n_1$	2N

Table 3.2: Generic allele table for autosomal data

An obvious question is whether the two different approaches will lead us to the same results. A sensible answer would seem to be in the affirmative, provided that each of the alleles were independent, that is assuming Hardy-Weinberg equilibrium. Sasieni (1997) formalises this intuitive notion and we will derive this result in section 3.5.

3.4.2 Allelic test

The test statistic for the allele table is referred to as the allele based test (ABT) in the GWAS literature. The test statistic is given by

$$X_A^2 = \frac{2N\{2N(r_1 + 2r_2) - 2R(n_1 + 2n_2)\}^2}{(2R)2(N-R)\{2N(n_1 + 2n_2) - (n_1 + 2n_2)^2\}}. \tag{3.1}$$

We show in Remark A.1 of the appendix that X_A^2 is simply the Pearson's χ^2 test of the allele table, and thus X_A^2 has an approximate χ^2 distribution on 1 degree of freedom under the null.

It is important to note that the ABT assumes that the 2 chromosomes carried by each individual can be regarded as independently sampled from a population of chromosomes — the assumption of Hardy-Weinberg equilibrium.

3.4.3 Genotype tests

Genotype tests are designed to test a specific alternative hypothesis reflecting a biologically plausible genetic model. The three classical models are known as dominant, recessive and additive, and are most used in practice (though others have also been proposed). These three models can be constructed in terms of genotypic relative risks.

The genotypic relative risks (GRRs) are a 3-vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)$ of relative risks defined by

$$\begin{aligned}\lambda_0 &= \frac{\Pr(\text{Case}|AA)}{\Pr(\text{Case}|AA)} \equiv 1 \\ \lambda_1 &= \frac{\Pr(\text{Case}|AB)}{\Pr(\text{Case}|AA)} \\ \lambda_2 &= \frac{\Pr(\text{Case}|BB)}{\Pr(\text{Case}|AA)}.\end{aligned}$$

The GRRs corresponding to various genetic models are summarised in table 3.3. It is not hard to see that the vector of GRRs can be completely determined by specifying a model and $\lambda_2 = r$.

Model	$\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)$
Null	$(1, 1, 1)$
Dominant	$(1, r, r)$
Recessive	$(1, 1, r)$
Additive	$(1, \frac{r+1}{2}, r)$

Table 3.3: Genotypic relative risks, $\boldsymbol{\lambda}$, for the null model and the 3 classical genetic models. Here we assume B is the risk allele and so $r > 1$.

The interpretation of the classical autosomal models are as follows:

- For a dominant model the genetic risk is identical regardless of whether 1 or 2 copies of the risk allele are present
- For a recessive model the genetic risk exists only if 2 copies of the risk allele are present
- For the additive model the genetic risk for the heterozygous genotype lies halfway between the risks for the 2 homozygous genotypes

Unlike the ABT where we have two observations per sample the genotype tests have just the one observation per sample. The genotype tests are therefore not subject to the assumption of HWE as we are now conducting the analysis on the genotype level rather than the allele level.

The two classes of genotype tests are the conventional Pearsonian χ^2 on 2 degrees of freedom for the 2×3 genotype table and the Cochran-Armitage trend test on 1 degree of freedom for testing the hypothesis of a specific genetic model. As the true genetic model is usually not known procedures to test multiple genetic models at each SNP that include corrections for this multiple testing have been proposed (see, for example, Joo et al., 2009; Freidlin et al., 2002).

3.4.4 Cochran-Armitage trend test

The Cochran-Armitage (CA) trend test (Cochran, 1954; Armitage, 1955) can be used to test for a dose-response effect between the risk of having a disease and the genotype of a sample. It is the ubiquitous test in the GWAS literature (see McCarthy et al., 2008; Balding, 2006; Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009, for example). The test involves a parameter vector $\boldsymbol{x} = (x_{(0)}, x_{(1)}, x_{(2)})$, with the choice of $\boldsymbol{x}$ corresponding to a particular genetic model.

The general form of the Cochran-Armitage trend test for the 2×3 genotype table is

$$\widetilde{X}^2_{CA}(\boldsymbol{x}) = \frac{N\big(N\sum_{j=0}^{2} r_j x_{(j)} - R\sum_{i=0}^{2} n_j x_{(j)}\big)^2}{R(N-R)\Big\{N\sum_{j=0}^{2} n_j x_{(j)}^2 - \big(\sum_{j=0}^{2} n_j x_{(j)}\big)^2\Big\}} \tag{3.2}$$

where the $x_{(j)}$, $j = 0, 1, 2$ are weights or "scores" for each of the 3 genotypes A/A, A/B, and B/B respectively. Under the null hypothesis $\widetilde{X}^2_{CA}(\boldsymbol{x})$ has an approximate χ^2 distribution on 1 degree of freedom (we prove this in section 3.5).

In GWA studies, when B is the disease-associated allele (risk allele), the optimal choice of $\boldsymbol{x}$ is $(0, 0, 1)$, $(0, 1, 2)$, and $(0, 1, 1)$ for the recessive, additive and dominant genetic models respectively.

Zheng et al. (2009) show that the CA trend test is invariant to linear transformations of $\boldsymbol{x}$, i.e. $\widetilde{X}^2_{CA}(\boldsymbol{x}) \equiv \widetilde{X}^2_{CA}(0, (x_{(1)} - x_{(0)})/(x_{(2)} - x_{(0)}), 1)$, so that the test can be reduced to a one parameter form $X^2_{CA}(x)$ where $x \in [0, 1]$.

While we have noted that the Cochran-Armitage test can be parameterised to test the hypothesis of various genetic models, it is most commonly used to test the hypothesis of an additive model. McCarthy et al. (2008) note that in situations when few causal variants are likely to be genotyped (such as in a GWAS), the additive model is likely to perform well. The additive version of the test is reasonably robust to misspecification of the true genetic model, though Lettre et al. (2007) note it performs poorly when the true mode of inheritance is recessive with a low minor allele frequency.

In the additive model the alternative hypothesis is that of a monotonic trend in the case-control ratio ordered by the number of copies of a nomi-

nated allele (0, 1 or 2), reflecting risk in the underlying population (Clayton, 2008). The appropriate choice of weights for testing the additive model are $x_{(j)} = j,\ j = 0, 1, 2$ giving us the best-known, and most-widely applied, expression for the CA test in the GWAS literature,

$$\begin{aligned} X^2_G &:= \widetilde{X}^2_{CA}(0, 1, 2) \\ &\equiv X^2_{CA}(1/2) \\ &= \frac{N\{N(r_1 + 2r_2) - R(n_1 + 2n_2)\}^2}{R(N - R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}}. \end{aligned} \tag{3.3}$$

We show in section 3.5 that the Cochran-Armitage (CA) trend test is in fact a score test for a logistic regression model of phenotype given genotype.

3.4.5 Alternatives to the Cochran-Armitage trend test

There have been a variety of other approaches proposed for the analysis of autosomal data. An obvious downside to the CA trend test is the requirement to specify the correct genetic model to achieve (local) maximum power. In the complex genetic diseases that are of interest in GWA studies this genetic model is rarely known. This has prompted the development and use of more robust statistical tests.

Lettre et al. (2007) investigated the effects of model misspecification when using the Cochran-Armitage trend test for autosomal data. Based on their results they promote the conventional Pearsonian χ^2 test on 2 degrees of freedom for the 2×3 genotype table as a robust alternative to the CA trend test.

The advantage of the χ^2 test is that the alternative hypothesis is simply an association between disease and genotype rather than any specific genetic model. However owing to the additional degree of freedom in the χ^2 test, it is less powerful than the $X^2_{CA}(x)$ when the choice of x accurately captures the true genetic model.

Another robust alternative to the CA trend test is the MAX test procedure (Podgot et al., 1996; Freidlin et al., 2002). To calculate the MAX test for a single SNP the CA trend test is calculated under a number of genetic models (e.g. dominant, recessive, additive, etc.) and then the most significant test result is selected. The asymptotic null distribution of the MAX statistic is not known and so must be approximated via numerical methods. This is clearly a more computationally demanding procedure than the Pearsonian χ^2 test on 2 degrees of freedom and will not be pursued further here.

The statistical analysis of the Wellcome Trust Case Control Consortium (2007) (WTCC) GWAS applied Bayesian methods to account for the prior belief of the existence of some number of genes involved in the disease. This involves the reporting of Bayes factors, in addition to p-values, to assess the

significance of each association. Additional research by the those involved in the WTCCC study has been focused on multi-marker, imputation and haplotype based analyses of GWAS data.

This is by no means a complete list of analytical methods that have been applied to GWAS data (a variety of data-mining type approaches have been proposed for example) but it is important to bear in mind that there is much ongoing research into methods other than the popular single-marker Cochran-Armitage trend test.

3.5 Some key results for autosomal tests

The allele based test has received considerable criticism in the literature due to the requirement of HWE holding for the test to be valid. Zheng (2008) calls for the test to be "retired" from analysis of GWA studies and labels the ABT a "nuisance test". This criticism is based on the following two results which are due to Sasieni (1997).

Theorem 3.1 (Sasieni, 1997)**.** *For the allele based test X_A^2 to be valid requires HWE to hold in the population.*

Proof. As we have seen, Hardy-Weinberg equilibrium is equivalent to the 2 alleles a person receives being independently chosen. Suppose that a proportion p of the alleles in a population are of type A. The number of alleles of type A received by an individual will be binomially distributed (with $n = 2$ and probability p) if and only if sampling of the two alleles is independent.

Without the requirement of Hardy-Weinberg equilibrium, the χ^2 approximation for the allele based test X_A^2 is invalid. The reason for this invalidity is that it assumes that $s_1 + 2s_2$, the number of controls with the B allele, is binomially distributed from a sample size of $2S$. By the above argument, $s_1 + 2s_2$ will only be binomially distributed if the two alleles in a given individual are independent, i.e. if Hardy-Weinberg equilibrium holds. □

The ABT is locally most powerful if and only if the allele effect is additive *and* HWE holds in the population (Sasieni, 1997). The Cochran-Armitage trend test does not require HWE as it is a test at the genotype level rather than at the allele level. Sasieni further shows that X_A^2 is equivalent to X_G^2 under HWE.

Theorem 3.2 (Sasieni, 1997)**.** $X_A^2 \equiv X_G^2$ *if and only if HWE holds in the population.*

Proof. Looking at the definitions of X_A^2 (3.1) and X_G^2 (3.3) we see that the numerators, apart from a factor of 8, are identical. However the denomina-

tors (variances) differ. We define

$$\begin{aligned} Q &= \frac{2N\big\{N(r_1+2r_2)-R(n_1+2n_2)\big\}^2}{R(N-R)} \\ B &= 2N(n_1+2n_2)-(n_1+2n_2)^2 \\ C &= 2\big\{N(n_1+4n_2)-(n_1+2n_2)^2\big\} \end{aligned}$$

(note there is an typographical error in Sasieni's original definition of Q). Via some simple algebra we see that

$$\begin{aligned} \frac{Q}{B} &= \frac{2N\big\{N(r_1+2r_2)-R(n_1+2n_2)\big\}^2}{R(N-R)\big\{2N(n_1+2n_2)-(n_1+2n_2)^2\big\}} \times \frac{4}{4} \\ &= \frac{2N\big\{2N(r_1+2r_2)-2R(n_1+2n_2)\big\}^2}{(2R)2(N-R)\big\{2N(n_1+2n_2)-(n_1+2n_2)^2\big\}} \\ &= X_A^2 \end{aligned}$$

and

$$\begin{aligned} \frac{Q}{C} &= \frac{2N\big\{N(r_1+2r_2)-R(n_1+2n_2)\big\}^2}{R(N-R)2\big\{N(n_1+4n_2)-(n_1+2n_2)^2\big\}} \\ &= \frac{N\big\{N(r_1+2r_2)-R(n_1+2n_2)\big\}^2}{R(N-R)\big\{N(n_1+4n_2)-(n_1+2n_2)^2\big\}} \\ &= X_G^2 \end{aligned}$$

and thus $X_A^2/X_G^2 = C/B$. Noting that $N = n_0+n_1+n_2$, we then have

$$\begin{aligned} B &= (n_1+2n_2)\big\{2(n_0+n_1+n_2)-n_1-2n_2\big\} \\ &= (n_1+2n_2)(n_1+2n_0) \end{aligned}$$

and

$$\begin{aligned} C &= 2\big\{N(n_1+2n_2)-(n_1+2n_2)^2+2Nn_2\big\} \\ &= 2\big\{(n_1+2n_2)(n_0+n_1+n_2-n_1-2n_2)+(n_0+n_1+n_2)2n_2\big\} \\ &= 2\big\{n_1n_0+4n_0n_2+n_1n_2\big\} \\ &= B+4n_0n_2-n_1^2. \end{aligned}$$

So we have the result that

$$\frac{X_A^2}{X_G^2} = \frac{C}{B} = 1+\frac{4n_0n_2-n_1^2}{(n_1+2n_2)(2n_0+n_1)} \tag{3.4}$$

where (3.4) is equal to 1 if and only if $4n_0n_2-n_1^2=0$. The condition that $4n_0n_2-n_1^2=0$ can only occur if the observed genotypic proportions are strictly in equilibrium and so the result follows[3]. □

[3]See Guedj et al., 2008, for a formal probabilistic proof of this final statement

Not surprisingly, X^2_G is the locally most powerful test if and only if the allele effect is exactly additive (Zheng, 2008) and the above proof shows that X^2_A is locally most powerful if and only if the allele effect is additive and the population is in Hardy-Weinberg equilibrium (Sasieni, 1997). Similarly, $X^2_{CA}(x)$ is the locally most powerful test if the choice of x correctly specifies the true (and generally unknown) genetic model.

The Cochran-Armitage trend test is a score test statistic

We now show that the Cochran-Armitage trend test is simply a score test statistic from a logistic regression model. We begin the proof by calculating the score test statistic for a logistic regression of phenotype on genotype.

Let

$$Y_i = \begin{cases} 1 & \text{if individual } i \text{ is a case} \\ 0 & \text{if individual } i \text{ is a control} \end{cases}$$

and for each SNP we encode the genotypes for individual i as

$$a_i = \begin{cases} x_{(0)} & \text{if A/A} \\ x_{(1)} & \text{if A/B} \\ x_{(2)} & \text{if B/B} \end{cases}$$

where the $x_{(j)}$ are as in the $X^2_{CA}(x_{(0)}, x_{(1)}, x_{(2)})$ test.

Definition 3.3. *The logistic regression score test statistic for testing a linear trend in genotype/phenotype associations is* $S = \frac{\left(\sum_{i=1}^N a_i(Y_i - \bar{Y})\right)^2}{\bar{Y}(1-\bar{Y})\sum_{i=1}^N (a_i - \bar{a})^2}$

Proof. We treat the Y_i as independent Binomial(1, π_i) trials with $\pi_i = \frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}$ where a_i is the observed genotype of the ith sample and model $P(Y_i|a_i)$ using logistic regression. Thus

$$\begin{aligned}
L(\pi_1, \ldots, \pi_N) &= \prod_{i=1}^N \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \prod_{i=1}^N \Big(\frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}\Big)^{y_i} \Big(1 - \frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}\Big)^{1-y_i} \\
&= \prod_{i=1}^N \Big(\frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}\Big)^{y_i} \Big(\frac{1}{1+\exp(\alpha+\beta a_i)}\Big)^{1-y_i} \\
&= \prod_{i=1}^N \Big(\exp(\alpha+\beta a_i)\Big)^{y_i} \Big/ \Big(1+\exp(\alpha+\beta a_i)\Big) \\
&= L(\alpha, \beta)
\end{aligned}$$

giving

$$\begin{aligned}\ln(L(\alpha,\beta)) &= \ell(\alpha,\beta)\\ &= \sum_{i=1}^{N}\Big[y_i(\alpha+\beta a_i) - \ln\big(1+\exp(\alpha+\beta a_i)\big)\Big]\end{aligned}$$

Using the notation of section 2.4, we define $\alpha = \boldsymbol{\theta}_1$ as the nuisance parameter and $\beta = \boldsymbol{\theta}_2$ as our parameter of interest since a non-zero β corresponds to an association between genotype and phenotype.

Taking derivatives of the log-likelihood function $\ell(\alpha,\beta)$ with respect to α and β yields the score functions

$$\begin{aligned}\dot{\ell_1} &= \frac{\partial \ell}{\partial \alpha}\\ &= \sum_{i=1}^{N}[y_i - \frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}]\\ &= \sum_{i=1}^{N}(y_i - \pi_i) \qquad (3.5)\end{aligned}$$

and

$$\begin{aligned}\dot{\ell_2} &= \frac{\partial \ell}{\partial \beta}\\ &= \sum_{i=1}^{N}[y_i a_i - a_i\frac{\exp(\alpha+\beta a_i)}{1+\exp(\alpha+\beta a_i)}]\\ &= \sum_{i=1}^{N} a_i(y_i - \pi_i) \qquad (3.6)\end{aligned}$$

The observed information matrix for the parameters is $-\ddot{\ell}$ with

$$\begin{aligned}\ddot{\ell} &= \frac{\partial^2 \ell}{\partial\alpha\partial\beta}\\ &= \begin{pmatrix} \ddot{\ell}_{11} & \ddot{\ell}_{12} \\ \ddot{\ell}_{21} & \ddot{\ell}_{22} \end{pmatrix}\\ &= \begin{pmatrix} \sum_{i=1}^{N}\pi_i(\pi_i-1) & \sum_{i=1}^{N} a_i\pi_i(\pi_i-1) \\ \sum_{i=1}^{N} a_i\pi_i(\pi_i-1) & \sum_{i=1}^{N} a_i^2\pi_i(\pi_i-1) \end{pmatrix}\end{aligned}$$

and the Fisher information matrix is given by $\mathcal{I} = \mathbb{E}(-\ddot{\ell})$, which is partitioned conformally with $\ddot{\ell}$ as

$$\begin{aligned}\mathcal{I} &= \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}\\ &= \begin{pmatrix} \sum_{i=1}^{N}\pi_i(1-\pi_i) & \sum_{i=1}^{N} a_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{N} a_i\pi_i(1-\pi_i) & \sum_{i=1}^{N} a_i^2\pi_i(1-\pi_i) \end{pmatrix}.\end{aligned}$$

To calculate the score statistic S we need $\mathcal{I}_{2.1}$ the asymptotic variance of $\dot{\ell}_2$ conditional on $\dot{\ell}_1 = 0$. Using the formula given in section 2.4 we have

$$\begin{aligned}
\mathcal{I}_{2.1} &= \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12} \\
&= \Big[\sum_{i=1}^{N} a_i^2\pi_i(1-\pi_i)\Big] - \Big[\sum_{i=1}^{N} a_i\pi_i(1-\pi_i)\frac{1}{\sum_{i=1}^{N}\pi_i(1-\pi_i)}\sum_{i=1}^{N} a_i\pi_i(1-\pi_i)\Big] \\
&= \Big[\sum_{i=1}^{N} a_i^2\pi_i(1-\pi_i)\Big] - \Big[\frac{\big(\sum_i^N a_i\pi_i(1-\pi_i)\big)^2}{\sum_{i=1}^{N}\pi_i(1-\pi_i)}\Big]. \qquad (3.7)
\end{aligned}$$

The score test statistic for testing the hypothesis that $\beta \neq 0$ is given by

$$S = \dot{\ell}_2\mathcal{I}_{2.1}^{-1}\dot{\ell}_2 \qquad (3.8)$$

with $\dot{\ell}_2$ and $\mathcal{I}_{2.1}$ evaluated at $\alpha = \hat{\alpha}$ and $\beta = 0$.

Under these conditions the maximum likelihood estimator of π_i is $\hat{\pi}_i = \bar{Y}$, so the expression for $\dot{\ell}_2$ becomes

$$\dot{\ell}_2 = \sum_{i=1}^{N} a_i(Y_i - \bar{Y}) \qquad (3.9)$$

and similarly the expression for $\mathcal{I}_{2.1}$ becomes

$$\begin{aligned}
\mathcal{I}_{2.1} &= \bar{Y}(1-\bar{Y})\Big\{\Big(\sum_{i=1}^{N} a_i^2\Big) - \frac{(\sum_{i=1}^{N} a_i)^2}{N}\Big\} \\
&= \bar{Y}(1-\bar{Y})\Big\{\Big(\sum_{i=1}^{N} a_i^2\Big) - N\bar{a}^2\Big\} \\
&= \bar{Y}(1-\bar{Y})\Big\{\Big(\sum_{i=1}^{N} a_i^2\Big) - 2N\bar{a}^2 + N\bar{a}^2\Big\} \\
&= \bar{Y}(1-\bar{Y})\Big\{\Big(\sum_{i=1}^{N} a_i^2\Big) - 2\bar{a}\Big(\sum_{i=1}^{N} a_i\Big) + N\bar{a}^2\Big\} \\
&= \bar{Y}(1-\bar{Y})\sum_{i=1}^{N}\Big(a_i^2 - 2\bar{a}a_i + \bar{a}^2\Big) \\
&= \bar{Y}(1-\bar{Y})\sum_{i=1}^{N}(a_i - \bar{a})^2 \qquad (3.10)
\end{aligned}$$

Combining (3.9) and (3.10) we have

$$\begin{aligned}
S &= \dot{\ell}_2\mathcal{I}_{2.1}^{-1}\dot{\ell}_2 \\
&= \frac{\big(\sum_{i=1}^{N} a_i(Y_i - \bar{Y})\big)^2}{\bar{Y}(1-\bar{Y})\sum_{i=1}^{N}(a_i - \bar{a})^2}. \qquad (3.11)
\end{aligned}$$

the score test statistic for testing the hypothesis of a linear trend in the association between genotype and phenotype.

Since S is a score test statistic, it has an approximate χ^2 distribution on $p_2 = 1$ degree of freedom under the null.

□

To simplify the proof that the Cochran-Armitage trend test is equivalent to a score test statistic we first derive an equivalent definition of X^2_{CA}.

Definition 3.4. *The Cochran-Armitage trend test can be alternatively parametrised as* $X^2_{CA} = \frac{N\Big[\sum_{j=0}^{2} x_{(j)}(Sr_j - Rs_j)\Big]^2}{SR\Big[\sum_{j=0}^{2} x_{(j)}^2 n_j(N-n_j) - 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}n_k n_j\Big]}$.

Proof. Let

$$W = \sum_{j=0}^{2} x_{(j)}(Sr_j - Rs_j) \tag{3.12}$$

and

$$V = \frac{SR}{N}\Big[\sum_{j=0}^{2} x_{(j)}^2 n_j(N-n_j) - 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}n_k n_j\Big]. \tag{3.13}$$

Now

$$\begin{aligned} W &= S\sum_{j=0}^{2} x_{(j)}r_j - R\sum_{j=0}^{2} x_{(j)}s_j \\ &= (N-R)\sum_{j=0}^{2} x_{(j)}r_j - R\sum_{j=0}^{2} x_{(j)}s_j \\ &= N\sum_{j=0}^{2} x_{(j)}r_j - R\sum_{j=0}^{2} x_{(j)}(r_j + s_j) \\ &= N\sum_{j=0}^{2} x_{(j)}r_j - R\sum_{j=0}^{2} x_{(j)}n_j \end{aligned}$$

and

$$\begin{aligned} V &= \frac{(N-R)R}{N}\Big[N\sum_{j=0}^{2} x_{(j)}^2 n_j - \sum_{j=0}^{2} x_{(j)}^2 n_j^2 - 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}n_k n_j\Big] \\ &= \frac{(N-R)R}{N}\Big[N\sum_{j=0}^{2} x_{(j)}^2 n_j - \big(\sum_{j=0}^{2} x_{(j)}n_j\big)^2\Big]. \end{aligned}$$

Thus

$$\begin{aligned}\frac{W^2}{V} &= \frac{N\big[\sum_{j=0}^{2} x_{(j)}(Sr_j - Rs_j)\big]^2}{SR\big[\sum_{j=0}^{2} x_{(j)}^2 n_j(N-n_j) - 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}n_k n_j\big]} \\ &= \frac{N\big(N\sum_{j=0}^{2} r_j x_{(j)} - R\sum_{i=0}^{2} n_j x_{(j)}\big)^2}{R(N-R)\Big\{N\sum_{j=0}^{2} n_j x_{(j)}^2 - \big(\sum_{j=0}^{2} n_j x_{(j)}\big)^2\Big\}} \end{aligned} \tag{3.14}$$

which is the X^2_{CA} test as defined in (3.2) and the claim follows. □

Theorem 3.5. *The Cochran-Armitage trend test X^2_{CA} is equivalent to the score test statistic S.*

Proof. To prove this theorem we use the alternative form of $X^2_{CA} = W^2/V$ given in definition 3.4. One obvious difference between the score test statistic (3.11) and the Cochran-Armitage trend test statistic (3.14) are the units of summation. The score statistic S is calculated by summing over individuals $i = 1, \ldots, N$ while X^2_{CA} is calculated by summing over genotype groups $j = 0, 1, 2$. We need to show these two approaches yield equivalent results.

We begin by showing

$$N\dot{\ell}_2 = N\sum_{i=1}^{N} a_i(Y_i - \bar{Y}) = \sum_{j=0}^{2} x_{(j)}(Sr_j - Rs_j) = W. \tag{3.15}$$

Proof.

$$\begin{aligned} N\dot{\ell}_2 &= N\sum_{i=1}^{N} a_i(Y_i - \bar{Y}) \\ &= N\sum_{i=1}^{N} a_i(Y_i - R/N) \\ &= \sum_{i=1}^{N} a_i(NY_i - R) \end{aligned}$$

since $R = \sum_{i=1}^{N} Y_i$ is the number of cases.

Now we sum over genotype groups $x_{(j)}$ instead of individuals i. We can construct $(x_{(0)}, x_{(1)}, x_{(2)})$ such that there exists a $x_{(j)} \equiv a_i$ for all $i = 1, \ldots, N$ since each is just an encoding of a sample's genotype. That is, we can replace each a_i with an equivalent $x_{(j)}$, so for each individual i

$$a_i(NY_i - R) = \begin{cases} a_i(N-R) \equiv x_{(j)}(N-R) & \text{if individual } i \text{ is a case} \\ a_i(-R) \equiv x_{(j)}(-R) & \text{if individual } i \text{ is a control} \end{cases}$$

Since for each genotype group $x_{(j)}$ there will be r_j cases we will have r_j number of $x_{(j)}(N-R)$ terms in our sum. Similarly for each genotype group $x_{(j)}$ there will be s_j controls, and so s_j number of $x_{(j)}(-R)$ terms. Therefore

$$\begin{aligned} N\dot{\ell}_2 &= \sum_{i=1}^{N} a_i(NY_i - R) \\ &= \sum_{j=0}^{2} x_{(j)}\big[r_j(N-R) + s_j(-R)\big] \\ &= \sum_{j=0}^{2} x_{(j)}\big[Sr_j - Rs_j\big] \\ &= W \end{aligned}$$

as claimed. □

Since $W = N\dot{\ell}_2$ and $\mathrm{Var}(\dot{\ell}_2) = \mathcal{I}_{2.1}^{-1}$ we have $\mathrm{Var}(W) = N^2\mathcal{I}_{2.1}^{-1}$. To complete the proof of Theorem 3.5 we must show that $\mathrm{Var}(W) = V$.

We can consider the marginal totals S and R of the genotype table to be fixed and thus both $\boldsymbol{r} = (r_0, r_1, r_2)$ and $\boldsymbol{s} = (s_0, s_1, s_2)$ follow multinomial distributions with probabilities of success n_j/N for $j = 0, 1, 2$ under the null hypothesis. Now

$$\begin{aligned} \mathrm{Var}(W) &= \mathrm{Var}\Big(\sum_{j=0}^{2} x_{(j)}(Sr_j - Rs_j)\Big) \\ &= S^2\mathrm{Var}\big(\sum_{j=0}^{2} x_{(j)}r_j\big) + R^2\mathrm{Var}\big(\sum_{j=0}^{2} x_{(j)}s_j\big) \\ &= S^2\Big[\sum_{j=0}^{2} x_{(j)}^2\mathrm{Var}(r_j) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}\mathrm{Cov}(r_k, r_j)\Big] \\ &\quad + R^2\Big[\sum_{j=0}^{2} x_{(j)}^2\mathrm{Var}(s_j) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)}\mathrm{Cov}(s_k, s_j)\Big] \end{aligned}$$

(Under H_0)

$$
\begin{aligned}
&= S^2\Big[\sum_{j=0}^{2} x_{(j)}^2 R\Big(\frac{n_j}{N}\Big)\Big(\frac{N-n_j}{N}\Big) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)} R\Big(\frac{n_k}{N}\Big)\Big(\frac{n_j}{N}\Big)\Big] \\
&\quad + R^2\Big[\sum_{j=0}^{2} x_{(j)}^2 S\Big(\frac{n_j}{N}\Big)\Big(\frac{N-n_j}{N}\Big) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)} S\Big(\frac{n_k}{N}\Big)\Big(\frac{n_j}{N}\Big)\Big] \\
&= \frac{S}{N^2}\Big[\sum_{j=0}^{2} x_{(j)}^2 SRn_j(N-n_j) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)} SRn_kn_j\Big] \\
&\quad + \frac{R}{N^2}\Big[\sum_{j=0}^{2} x_{(j)}^2 RSn_j(N-n_j) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)} RSn_kn_j\Big] \\
&= \frac{SR}{N}\Big[\sum_{j=0}^{2} x_{(j)}^2 n_j(N-n_j) + 2\sum_{k=0}^{1}\sum_{j=k+1}^{2} x_{(k)}x_{(j)} n_kn_j\Big] \\
&= V
\end{aligned}
$$

and we have shown that $\mathrm{Var}(W) = V$.

Thus

$$
\begin{aligned}
X_{CA}^2 &= \frac{W^2}{V} \\
&= \frac{(N\dot{\ell}_2)^2}{N^2\mathcal{I}_{2.1}} \\
&= \frac{(\dot{\ell}_2)^2}{\mathcal{I}_{2.1}} \\
&= S
\end{aligned}
$$

and we have shown that X_{CA}^2 is a score test statistic as claimed. □

Corollary 3.6. *X_{CA}^2 has an approximate χ^2 distribution on 1 degree of freedom.*

Proof. This follows immediately from the preceding theorem. X_{CA}^2 is an equivalent way to write S and S has an approximate χ^2 distribution on 1 degree of freedom since it it a score test statistic. □

Remark 3.7. *Clayton's derivation of the Cochran-Armitage trend test is asymptotically equivalent to S.*

Proof. Clayton (2008) denotes the score test statistic for testing the effect

of genotype on phenotype by

$$\begin{aligned} S_C &:= \frac{(\sum_{i=1}^{N}(Y_i - \bar{Y})a_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \sum_{i=1}^{N}(a_i - \bar{a})^2} \\ &= \frac{U_A}{\widehat{\mathrm{Var}(U_A)}} \end{aligned}$$

where again the a_i encode genotypes according to some genetic model. Clayton encodes the genotype by $a_i = 0, 1, 2$ for $A/A, A/B, B/B$ respectively. That is, Clayton only consider the additive version of the Cochran-Armitage trend test, a somewhat subtle point when reading the paper. It is important to bear this in mind when applying Clayton's methods. We know of course that the a_i can be optimally chosen to test a specific genetic model (see section 3.4.4).

Comparing S_C and S we see that they have identical numerators but differ in their denominators (variances). Clayton sketches his derivation of the variance of the score test statistic according to the following theory.

Using the score statistic

$$U_A = \sum_{i=1}^{N}(Y_i - \bar{Y})a_i$$

Clayton considers the phenotypes $\{Y_i\}$ to be i.i.d. random variables, with the genotypes $\{a_i\}$ to be fixed and non-random. Since the variance of a sum of iid random variables is the sum of the individual variances, Clayton arrives at

$$\begin{aligned} \mathrm{Var}_C(U_A) &= \mathrm{Var}\Big(\sum_{i=1}^{N}(Y_i - \bar{Y})a_i\Big) \\ &= V_Y \sum_{i=1}^{N}(a_i - \bar{a})^2 \end{aligned}$$

where V_Y is the variance of Y and the C subscript is to denote Clayton's estimator. He then estimates V_Y by the unbiased estimator

$$\widehat{V_Y} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

to arrive at his score statistic

$$\begin{aligned} S_C &= \frac{(U_A)^2}{\widehat{\mathrm{Var}_C}(U_A)} \\ &= \frac{\Big(\sum_{i=1}^{N}(Y_i - \bar{Y})a_i\Big)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \sum_{i=1}^{N}(a_i - \bar{a})^2} \end{aligned}$$

Thus $U_A \equiv \dot{\ell}_2$ so S and S_C only differ in their respective denominators (variance estimators). It is easily shown that S_C is asymptotically equivalent to the score test statistic S (and thus X^2_{CA} by Theorem 3.5). We have

$$\begin{aligned}
\mathcal{I}_{2.1} &= \bar{Y}(1-\bar{Y})\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= (\bar{Y}-\bar{Y}^2)\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= (\bar{Y}-2\bar{Y}^2+\bar{Y}^2)\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= (\bar{Y}-\frac{2}{N}\bar{Y}N\bar{Y}+\bar{Y}^2)\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= \Big(\frac{1}{N}\sum_{i=1}^{N}Y_i-\frac{2}{N}\bar{Y}\sum_{i=1}^{N}Y_i+\frac{N}{N}\bar{Y}^2\Big)\sum_{i=1}^{N}(a_i-\bar{a})^2
\end{aligned}$$

and noting that since $Y_i = 0, 1$ then $\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} Y_i^2$

$$\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(Y_i^2-2\bar{Y}Y_i+\bar{Y}^2)\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2\sum_{i=1}^{N}(a_i-\bar{a})^2 \\
&= \frac{N-1}{N}\widehat{\mathrm{Var}_C(U_A)}
\end{aligned}$$

Thus we have

$$\begin{aligned}
S &= \frac{(\dot{\ell}_2)^2}{\mathrm{Var}(\dot{\ell}_2)} \\
&= \frac{(U_A)^2}{\frac{N-1}{N}\widehat{\mathrm{Var}_C(U_A)}} \\
&= \frac{N}{N-1}\times S_C
\end{aligned}$$

and the two score test statistics are asymptotically equivalent. □

Clayton further derives a score test statistic on 2 degrees of freedom that amounts to a Pearson χ^2 test of the 2×3 genotype table. We define Clayton's 2 degree of freedom score test here as it will be required in section 3.7.

Definition 3.8. *Clayton's score test for genotype/phenotype associations on 2 degrees of freedom*

Let Y_i and a_i be as in Definition 3.3 and further define

$$d_i = \begin{cases} 1 & \text{if individual } i \text{ is a heterozygote i.e. } A/B \\ 0 & \text{if individual } i \text{ is a homozygote i.e. } A/A \text{ or } B/B \end{cases}$$

We again have

$$U_A = \sum_{i=1}^{N}(Y_i - \bar{Y})a_i \tag{3.16}$$

and

$$\widehat{\text{Var}(U_A)} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \sum_{i=1}^{N}(a_i - \bar{a})^2. \tag{3.17}$$

Now we also have a score statistic for D given by

$$U_D = \sum_{i=1}^{N}(Y_i - \bar{Y})d_i \tag{3.18}$$

with corresponding variance

$$\text{Var}(U_D) = V_D \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{3.19}$$

where V_D is estimated by

$$\widehat{V_D} = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2. \tag{3.20}$$

We must also estimate the covariance of U_A and U_D which we do by

$$\text{Cov}(U_D, U_A) = V_{AD} \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{3.21}$$

where V_{AD} is estimated by

$$\widehat{V_{AD}} = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a})(d_i - \bar{d}). \tag{3.22}$$

The above can then be written into matrix form with

$$U = \begin{pmatrix} U_A \\ U_D \end{pmatrix}, \quad \widehat{V} = \begin{pmatrix} \widehat{V_A} & \widehat{V_{AD}} \\ \widehat{V_{AD}} & \widehat{V_D} \end{pmatrix} \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

giving the score test statistic $U^T\widehat{V}^{-1}U$ which under H_0 is approximately χ^2 distributed on 2 degrees of freedom. We know from Smyth (2003) that this score test is equivalent to the Pearsonian χ^2 test (see section 2.5.1).

3.6 X chromosome analysis

The X chromosome provides unique challenges for testing associations of genotype and phenotype. The main difference between the X chromosome and the autosomes is of course the number of copies each person has — females having two X chromosomes and males one X chromosome. Males are termed *hemizygous* for the X chromosome and their genotypes are written $A/-$ or $B/-$. Loci on the pseudo-autosomal region of the X chromosome can be treated in exactly the same way as autosomal loci so we restrict our study to those X loci not in the pseudo-autosomal regions.

Tables 3.4, 3.5 and 3.6 introduce various ways we can consider X chromosome data along with the necessary notation for this section.

	Genotype			
	A/A	A/B	B/B	Total
Cases	r_{f0}	r_{f1}	r_{f2}	R_f
Controls	s_{f0}	s_{f1}	s_{f2}	S_f
Total	n_{f0}	n_{f1}	n_{f2}	N_f

Table 3.4: Generic genotype table for female X chromosome data

	Genotype		
	$A/-$	$B/-$	Total
Cases	r_{m0}	r_{m2}	R_m
Controls	s_{m0}	s_{m2}	S_m
Total	n_{m0}	n_{m2}	N_m

Table 3.5: Generic genotype table for male X chromosome data

	Allele		
	A	B	Total
Cases	$2r_{f0} + r_{f1}$	$2r_{f2} + r_{f1}$	$2R_f$
Controls	$2s_{f0} + s_{f1}$	$2s_{f2} + s_{f1}$	$2S_f$
Total	$2n_{f0} + n_{f1}$	$2n_{f2} + n_{f1}$	$2N_f$

Table 3.6: Generic allele table for female X chromosome data

There are a number of ways to summarise this X chromosome data and multiple methods of analysis. We could analyse the male and female tables separately, that is, stratify the X chromosome analysis by sex. This would of course lead to a loss in power due to stratification. If we can sensibly

combine the data across sexes we hope to avoid this loss in power. We discuss several approaches to combining the two tables into a single table.

Firstly, we could take a simple allele counting approach to combine the results for the male and female samples, as in table 3.7. This is the simplest approach but, like all allele based tests, is only a valid approach when HWE holds. Furthermore, males will only have half the impact on the analysis as the females.

	Allele		
	A	B	Total
Cases	$2r_{f0} + r_{f1} + r_{m0}$	$2r_{f2} + r_{f1} + r_{m2}$	$2R_f + R_m$
Controls	$2s_{f0} + s_{f1} + s_{m0}$	$2s_{f2} + s_{f1} + s_{m2}$	$2S_f + S_m$
Total	$2n_{f0} + n_{f1} + n_{m0}$	$2n_{f2} + n_{f1} + n_{m2}$	$2N_f + N_m$

Table 3.7: Generic allele table for combined male and female X chromosome data

Instead of collapsing over alleles, we can collapse over genotypes in a number of different ways. We could consider each of A/A, A/B, B/B, $A/-$ and $B/-$ as separate genotypes, and then combine the data like that in table 3.8.

	Genotype					
	A/A	A/B	B/B	$A/-$	$B/-$	Total
Cases	r_{f0}	r_{f1}	r_{f2}	r_{m0}	r_{m2}	$R_f + R_m$
Controls	s_{f0}	s_{f1}	s_{f2}	s_{m0}	s_{m2}	$S_f + S_m$
Total	n_{f0}	n_{f1}	n_{f2}	n_{m0}	n_{m2}	$N_f + N_m$

Table 3.8: Generic genotype table for combined X chromosome data

Table 3.8 could then be analysed with a χ^2 test. However the penalty paid for such an approach is a loss in power due to the χ^2 statistic now being distributed on 4 degrees of freedom. Ideally, we would like to combine male and female genotype data in a way that is biologically sensible and meaningful without too great an increase in the degrees of freedom. Two such approaches are considered here.

Method 1

Due to the process of X-inactivation only one of the two X chromosomes is active in females (see section 1.4). It has been thus proposed, for example in Clayton (2008), that male hemizygotes be treated the same as female homozygotes for the X chromosome analysis. That is, consider a male $A/-$ as a female A/A and a male $B/-$ as a female B/B, since males and females

should be equivalent at such loci in the absence of interactions with other loci or environmental factors. How we then handle female A/B heterozygotes is not entirely clear, and will be discussed further in section 3.8.

Method 2

The second approach is to count the number of B alleles present in each sample. This is equivalent to assuming the effect for an $A/-$ male is the same as for an A/A female, and similarly a $B/-$ male is the same as a A/B female. However, it is not clear why an A/B female should be equivalent to a $B/-$ male, and not an $A/-$ male, since the female is equally likely to expressed the A allele as the B allele.

Neither of these two approaches is perfect as X-inactivation is far more complex than the idealised process described here. However, both approaches are reasonable models that are used in practice.

3.6.1 Current methods for the X chromosome

We review the methods used by the GWAS analysis package PLINK (Purcell et al., 2007) for X chromosome data. We also introduce some methods previously studied by Zheng et al. (2007) for X chromosome loci.

PLINK

PLINK offers several methods for the analysis of X chromosome data. Aside from the allele based test using the combined male and female sample we will discuss two other approaches available in PLINK.

The first method, and perhaps the most obvious, is to use logistic regression to model phenotype given genotype *and* sex. Following the notation in the PLINK manual (Purcell, 2009), male genotypes are coded as

$$\text{GENOTYPE} = \begin{cases} 0 & \text{if } A/- \\ 1 & \text{if } B/- \end{cases}$$

while female genotypes are

$$\text{GENOTYPE} = \begin{cases} 0 & \text{if } A/A \\ 1 & \text{if } A/B \\ 2 & \text{if } B/B \end{cases},$$

with sex encoded

$$\text{SEX} = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}.$$

We see that male genotypes are encoded so that a $B/-$ male is equivalent to a A/B female. This leads to the question, described in **Method 2**, of why a female A/B should be the equivalent to a $B/-$ male. Furthermore, it

is not clear from the PLINK manual how genotype/phenotype associations are assessed; for example, whether it is via a likelihood ratio test, a Wald test or a score test, though the results should be asymptotically equivalent.

The second method for X chromosome data in PLINK is to apply a standard association test, such as the CA trend test or a χ^2 test of the genotype table, but using *only the female samples*. The loss in power associated with this approach is clear (unless of course all samples are female). The data for a GWAS is expensive to obtain and so it appears wasteful to use only the female samples for X chromosome analysis. There are also many diseases where the sex distribution of cases is skewed and so for a male-biased disease this method would fare particularly poorly.

Zheng et al's proposed tests

Zheng et al. (2007) propose a series of tests that combine separate tests of males and females to make use of all the samples. They also assess the performance of these tests via a simulation study — to be discussed further in section 4.1. However I have been unable to find any examples of their proposed statistics being used on real data outside the original paper.

In the notation of Zheng et al., let Z^2_{fG} be the CA trend test for the female genotype table 3.4, Z^2_m be the ABT for male genotype table 3.5, and Z^2_{fA} be the ABT for the female allele table 3.6. Zheng et al., then introduce the test statistics described in table 3.9.

Test	Definition and Description	Null Distribution
Z^2_A	The ABT for table 3.7	χ^2 on 1 df
Z^2_C	$Z^2_C = Z^2_m + Z^2_{fG}$	χ^2 on 2 df
Z_{mfA}	$Z^2_{mfA} = \left(\sqrt{\frac{N_m}{N_m+2N_f}}Z_m + \sqrt{\frac{2N_f}{N_m+2N_f}}Z_{fA}\right)^2$	χ^2 on 1 df
Z_{mfG}	$Z_{mfG} = \left(\sqrt{\frac{N_m}{N_m+N_f}}Z_m + \sqrt{\frac{N_f}{N_m+2N_f}}Z_{fG}\right)^2$	χ^2 on 1 df

Table 3.9: Zheng et al.'s proposed tests

In other words, Z^2_A is simply the allele based test when combining male and female samples while the remaining 3 test statistics are various combinations of tests within the male and female samples. For further details on these 4 statistics, as well as a further two tests designed to be optimal when the risk allele differs between males and females, we direct the reader to the original paper (Zheng et al., 2007).

3.7 Clayton's corrected tests for X chromosome loci

Clayton (2008) proposes two new statistics for the analysis of X chromosome data in GWA studies. These are modifications of the autosomal 1 degree of freedom CA trend test and 2 degree of freedom χ^2 test described in Remark 3.7 and Definition 3.8 respectively.

Assuming that the allele frequency does not vary between males and females, a purported benefit to Clayton's approach is that it avoids the attendant loss in power due to stratification of testing by sex since it combines male and female samples. However, there is no analysis comparing Clayton's proposed method to existing methods using simulated (or otherwise) X chromosome data. We address this in chapter 4.

3.7.1 Derivation of Clayton's X chromosome test statistics

Clayton (2008) seeks to modify the autosomal tests to make them appropriate for X chromosome loci. We first consider the 2 degree of freedom genotype test (see definition 3.8) to make it appropriate for X chromosome loci. We assume that due to the process of X-inactivation that male hemizygotes are equivalent to female homozygotes. Accordingly, for X loci in males we code the genotypes a_i either as 0 or 2 corresponding to $A/-$ and $B/-$ respectively, and d_i should be coded 0.

As Clayton notes, this has several consequences which require modifications to the statistic defined in 3.8. Namely,

1. Under HWE, if the allele frequency does not vary between sexes, then in both sexes $\mathbb{E}(A) = 2P$, where $P \in (0, 0.5]$ is the frequency of allele B (the minor allele).

   Proof. In females

   $$A = \begin{cases} 0 & \text{if A/A, w.p. } (1-P)^2 \\ 1 & \text{if A/B, w.p. } 2P(1-P) \\ 2 & \text{if B/B, w.p. } P^2 \end{cases}$$

   and in males

   $$A = \begin{cases} 0 & \text{if A/A, w.p. } (1-P) \\ 2 & \text{if B/B, w.p. } P \end{cases}$$

   thus $\mathbb{E}(A) = 2P$ in both males and females. □

   Thus, the expectations of U_A will remain at 0 under H_0, even when the phenotype, Y, is related to sex.

2. The variance of A differs between males and females. For example, under Hardy-Weinberg equilibrium $\text{Var}(A) = 2P(1-P)$ in females and $4P(1-P)$ in males. This means that, in general, an alternative variance estimate of U_A must be used.

3. Only females contribute to the dominance score, U_D. Without loss of generality, we assume that samples are arranged so that samples $1, \ldots, F$ are female and samples $(F+1), \ldots, N$ are male. Then,

$$U_D = \sum_{i=1}^{F} (Y_i - \bar{Y}_F) d_i,$$

where $\bar{Y}_F$ is the mean of Y in females. Using this formulation U_D also has mean 0 under H_0.

With the above points in mind Clayton derives a modified covariance matrix of $U = (U_A, U_D)^T$. Firstly, for females the covariance matrix of U_A and U_D is estimated by

$$\widehat{V}_F = \frac{1}{F-1} \sum_{i=1}^{F} \begin{pmatrix} (a_i - \bar{a})^2 & (a_i - \bar{a})(d_i - \bar{d}_F) \\ (a_i - \bar{a})(d_i - \bar{d}_F) & (d_i - \bar{d}_F)^2 \end{pmatrix}, \tag{3.23}$$

where $\bar{d}_F$ is the mean of d_i in females. Since we are assuming the allele frequencies are equal between males and females we can use the whole sample to calculate $\bar{a}$.

Since males have only a single copy of the allele, the covariance matrix of U_A and U_D in males is estimated by

$$\widehat{V}_M = \begin{pmatrix} 4P(1-P) & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.24}$$

where $P = \Pr(B)$ is the minor allele frequency[4].

Again, P can be estimated in the entire sample since we assume the allele frequency is equal between sexes. We use $\widehat{P} = \bar{a}/2$ to estimate P as it is simpler to compute than the equally valid estimator of P based on allele counting.

Combining equations (3.23) and (3.24), our estimator of the covariance matrix of the 2-vector of scores, U, is given by

$$\widehat{V} = \widehat{V}_F \sum_{i=1}^{F} (Y_i - \bar{Y})^2 + \widehat{V}_M \sum_{i=F+1}^{N} (Y_i - \bar{Y})^2. \tag{3.25}$$

[4]NB: Clearly $\widehat{V}_M$ is invariant to whether P is defined as the major or minor allele frequency

As before the score test statistic $S^{(2)} = U^T \widehat{V}^{-1} U$ has an approximate χ^2 distribution on 2 degrees of freedom under the null hypothesis. As $S^{(2)}$ is based on the χ^2 test on 2 degrees of freedom for the genotype table, $S^{(2)}$ should be robust to misspecification of the genetic model as it simply tests for an association between genotype and phenotype rather than a specific association due to a genetic model (see section 3.4.5).

The 1 degree of freedom test (i.e. the CA trend test adjusted for X chromosome loci) is given by $S^{(1)} = U_A^2 / \widehat{V}_{11}$ and again has an approximate χ^2 distribution on 1 degree of freedom under the null hypothesis. As $S^{(1)}$ is based on the Cochran-Armitage trend test for when the trend is *additive*, it should perform best when the true genetic model is additive. The choice of a_i could perhaps be altered to test a recessive or dominant model, but, as we will see in section 3.8, the genetic models are rather more complicated for X chromosome loci and so we only consider the $a_i = 0, 1, 2$ case for Clayton's statistics. Both Clayton's 1 and 2 degree freedom tests are invariant to whether allele A or allele B is the minor allele.

To empirically check the null distribution of $U_A^2 / \widehat{V}_{11}$ and $U^T \widehat{V}^{-1} U$ I simulated 10,000 male and female genotype tables under the null hypothesis of no association between genotype and phenotype. Each table was simulated[5] using 2000 cases and 2000 controls, evenly split across sexes, with a disease prevalence of 1/1000 and a minor allele frequency of 0.1.

For each of these 10000 pairs of genotype tables I computed Clayton's 1 and 2 degree of freedom tests. In figures 3.1 and 3.2 I present quantile-quantile plots comparing the observed values of the test statistics to their respective theoretical distributions. These plots, generated using the `qq.chisq` function in the R package `snpMatrix` (Clayton and Leung, 2009), include a 95% confidence band to assist in their interpretation.

[5]For a complete description of the simulation methods see chapter 4

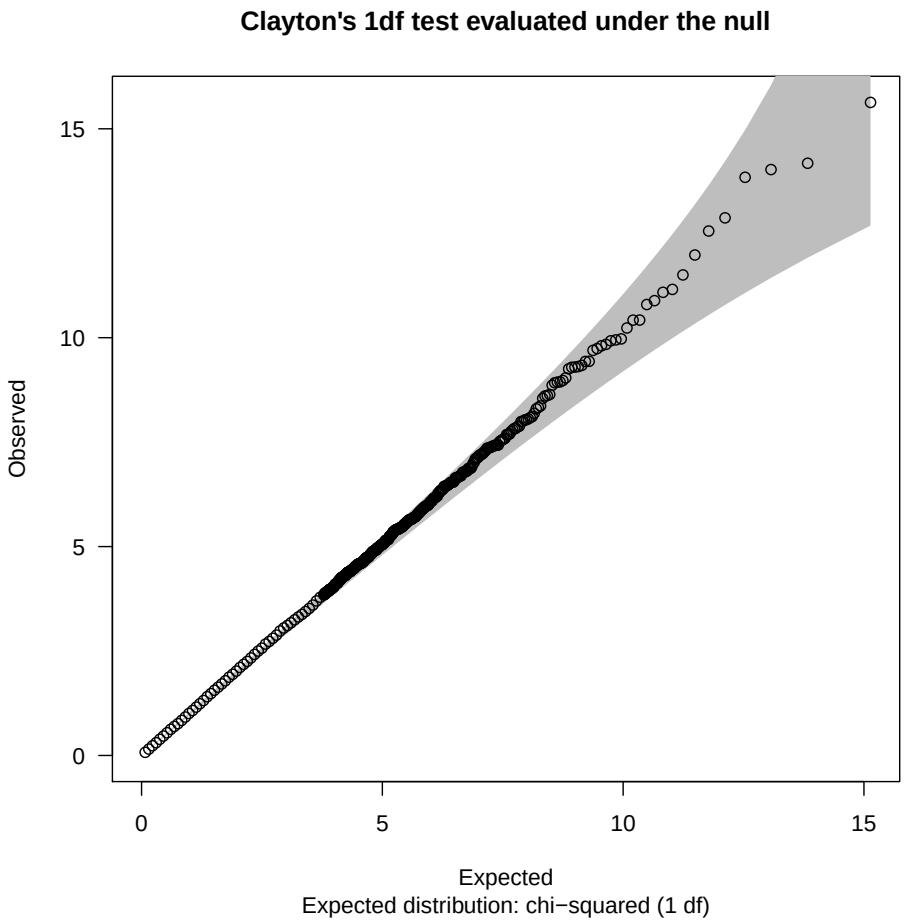

Figure 3.1: A q-q plot comparing the observed values of Clayton's 1 degree of freedom statistic under the null hypothesis against its expected χ^2 distribution. The grey band is a 95% confidence band and we see that the plot is consistent with the test statistic having an approximate null distribution of χ^2 on 1 degree of freedom.

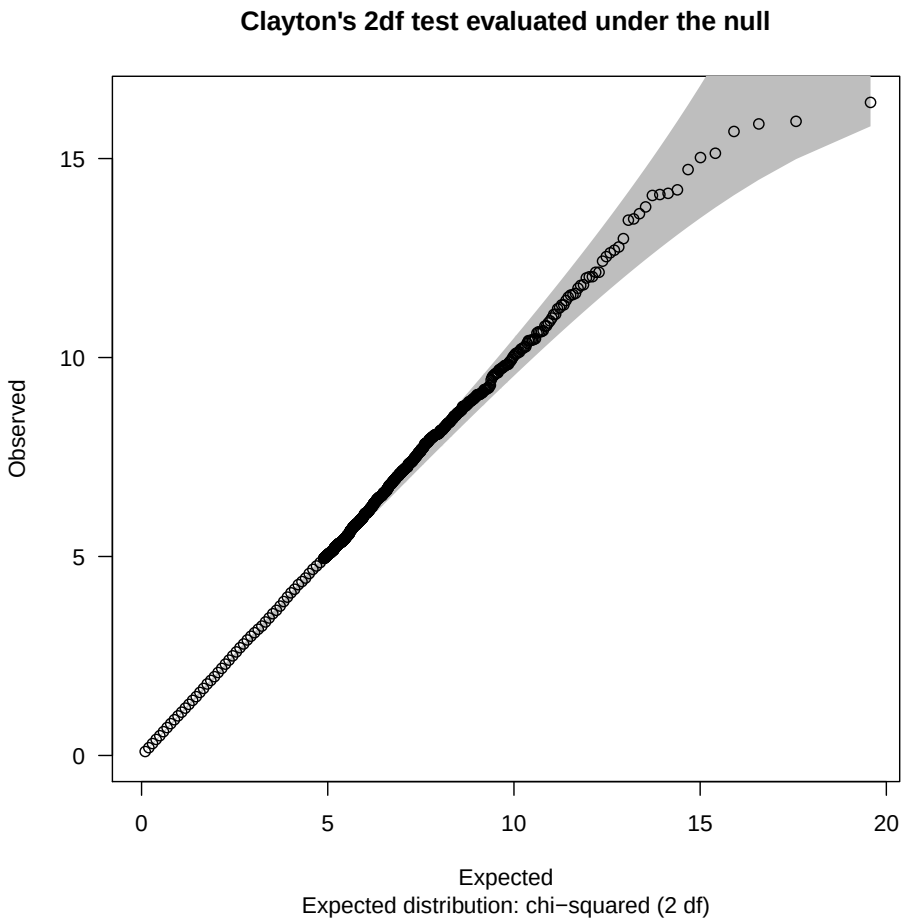

Figure 3.2: A q-q plot comparing the observed values of Clayton's 2 degree of freedom statistic under the null hypothesis against its expected χ^2 distribution. The grey band is a 95% confidence band and we see that the plot is consistent with the test statistic having an approximate null distribution of χ^2 on 2 degree of freedom.

3.8 Biologically plausible hypotheses for the X chromosome

One of the challenges for testing association on X chromosome loci is choosing biologically meaningful alternate hypotheses. The classical additive, dominant and recessive genetic models have rather different interpretations for X chromosome loci.

Under Clayton's scheme we code male hemizygotes as we would female homozygotes, and so the genetic risk to a male hemizygote should be identical to the genetic risk to the equivalent female homozygote. A difficulty lies in the treatment of the female A/B heterozygote. Do we treat A or B as the active allele in the heterozygotes?

Unfortunately we do not know which of the X chromosomes in females is inactivated from genotype data alone. X inactivation is a random process in each cell and the genotyping is performed on a random subset of cells. It will therefore display a roughly $50:50$ distribution of the A and B alleles in female heterozygotes (see section 1.4).

One approach to tackling this problem would be to randomly assign female heterozygotes to either of the female homozygous genotypes thereby mimicking the process of X-inactivation. This would introduce further complexity to the model and so we have instead taken the following approach.

Female X chromosome loci

The three classical genetic models can still be considered to apply for the homologous pair of female X chromosomes. For example, red-green colour blindness, Haemophilia A, and fragile X syndrome are all X-linked recessive diseases while hypophosphatemia is an example of an X-linked dominant disorder. The genotoypic relative risks for female X chromosome genotypes under the classical genetic models are given in 3.10.

These models are complicated by the inactivation process unique to the female X chromosome and so no longer have strictly the same interpretation as for the autosomal models. It is nevertheless common to assume that one of the classical genetic models holds for female X chromosome data in practice (see, for example, Zheng et al., 2007).

Male X chromosome loci

Since we code male hemizygotes as we would female homozygotes, the male hemizygotes have the same GRRs as their female homozygous counterparts. We see from table 3.11 that for X chromosome loci in males it does not matter which genetic model is proposed — for fixed r the GRRs are identical for all models.

Genetic model	λ_0	λ_1	λ_2
Dominant	1	r	r
Additive	1	$\frac{r+1}{2}$	r
Recessive	1	1	r
Genotype	A/A	A/B	B/B

Table 3.10: Genotypic relative risks for female X chromosome genotypes under the three classical genetic model

Genetic model	λ_0	λ_2
Dominant	1	r
Additive	1	r
Recessive	1	r
Genotype	$A/-$	$B/-$

Table 3.11: Genetic risk ratios for male X chromosome loci under the three classical genetic model

Given that the GRRs in males are constant regardless of the model we would expect the distribution of male cases to be identical regardless of the genetic model. Therefore, for fixed r, we should have identical power to detect genotype/phenotype associations in male samples regardless of the genetic model specified.

This result is consistent with what we would expect when we realise that for male samples we are just performing an allele based test (there being only 2 genotypes, $A/-$ and $B/-$) and therefore we have no power to discriminate between genetic models.

It is clear that to specify a biologically plausible model for X chromosome data is quite challenging. As with all models, the ones considered here make a number of simplifying assumptions that may not hold in reality. Our approach is to note that for male X chromosome data there is no need to specify a genetic model, and to assume that one of the classical genetic models holds for female data, with a somewhat altered interpretation.

Chapter 4

Simulation study

There is currently no study comparing the methods proposed in Clayton (2008) with those methods already in use for the analysis of X chromosome GWAS data. To address this issue, and to give a more thorough study of the existing methods, I performed a simulation study.

I wrote the code for the simulation in the statistical programming language R (R version 2.9.0, R Development Core Team, 2009) with plots produced using the R package `lattice` (Sarkar, 2009).

There is one previously published simulation study comparing association tests for X chromosome loci (Zheng et al., 2007) however Clayton's methods are not included. We begin with a brief critique of this paper to highlight the differences between it and my simulation study.

4.1 Zheng et al's simulation study

Zheng et al. (2007) examine the power of 6 different tests to detect associations between genotype and phenotype on the X chromosome in a GWAS using simulated data. One of the strengths of the study is that it simulates data not only under HWE but also when HWE does not hold. This allows the authors to make conclusions on the impact of departure from HWE for these statistics as well as which test is locally most powerful for various genetic models.

Zheng et al. consider a simulation with only 200 cases and 200 controls. As was discussed in section 1.6, a GWAS typically requires sample sizes of at least 1000 cases and 1000 controls in order to achieve sufficient power to detect the small genetic effect sizes typical of complex diseases such as multiple sclerosis. It is therefore unclear how these test statistics would perform in the context of a realistic study.

Another oddity of the study is the significance level applied in the simulation. The significance level $\alpha = 2.772 \times 10^{-5}$ is 3 orders of magnitude larger than the consensus Type I error threshold of $\alpha = 10^{-8}$ (see section 3.3). This

is due to the authors performing a Bonferroni correction that only adjusts for testing of the markers on the X chromosome ($\alpha = 0.05/1804 = 2.772 \times 10^{-5}$). It is rare, particularly in a *genome-wide* association study, that only the X chromosome SNPs would be analysed for genetic associations. The significance level, α, is therefore too liberal and would likely result in many false positives if applied to real data.

The study also assumes an equal number of males and females in both the cases and controls. While this may be a reasonable assumption for simulation purposes, it will not always be true in practise (see the ANZgene multiple sclerosis study discussed in chapter 5 for example). The disease prevalence K is set at 0.1, or 1 case per 10 people in the wider population. This disease prevalence is too high to be realistic for many "common" complex disease (see table 4.1), let alone for less common diseases such as multiple sclerosis with $K \approx \frac{1}{1000}$ (Oksenberg et al., 2008).

Disease	Prevalence in Australia
Type 1 diabetes	0.4%
Type 2diabetes	3.52%
Asthma	9.9%
Heart, stroke and vascular disease	5.2%

Table 4.1: A table of some common complex diseases and their estimated prevalence rates in the Australian population (Source: National Health Survey: Summary of Results, 2007-2008 (Reissue), Australian Bureau of Statistics).

The simulation is performed assuming a minor allele frequency of 0.1 or 0.3. This does not allow for accurate interpolation or extrapolation of the results to other minor allele frequencies. We know that the minor allele frequency varies across the X chromosome (see figure 4.1), and so it is of interest how these methods perform across the full range of MAFs.

All of the above criticisms are easily addressed by simple alterations to the simulation parameters. However, there is a more fundamental problem with the way the data is simulated in Zheng et al. that I believe significantly biases the results and requires modification. This problem lies in how the distribution of male genotypes is simulated.

Zheng et al. define the distribution of male genotypes with respect to the distribution of female genotypes by

$$\Pr(A/-|\text{Case}) = \Pr(A/A|\text{Case}) + \Pr(A/B|\text{Case})/2$$
$$\Pr(B/-|\text{Case}) = \Pr(B/B|\text{Case}) + \Pr(A/B|\text{Case})/2$$

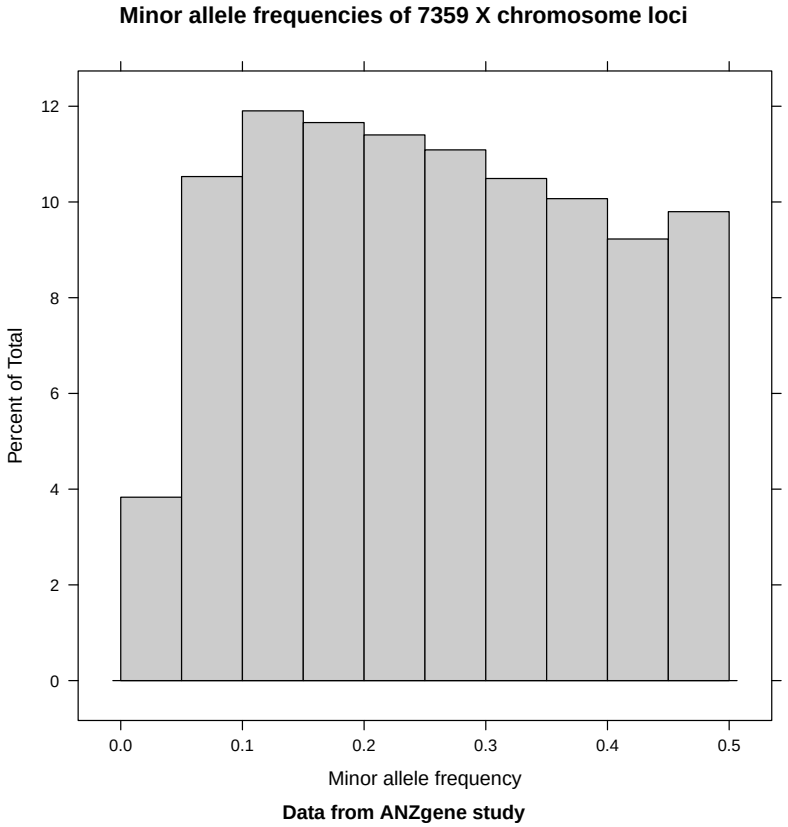

Figure 4.1: A histogram of minor allele frequencies (MAFs) for all 7359 X chromosome SNPs that passed quality-control procedures in the ANZgene GWAS. We see that the MAFs are non-uniformly distributed across the range of possible values $[0.01, 0.5]$ (Kolmogorov-Smirnov statistic $= 0.0451$, p-value $= 1.896 \times 10^{-13}$). The minimum MAF ≥ 0.01 is due to quality-control procedures — see chapter 5 for further details on the ANZgene data

and

$$\Pr(A/-|\text{Control}) = \Pr(A/A|\text{Control}) + \Pr(A/B|\text{Control})/2$$
$$\Pr(B/-|\text{Control}) = \Pr(B/B|\text{Control}) + \Pr(A/B|\text{Control})/2.$$

An implication of this, to be discussed further, is that the distribution of the male genotypes will depend on the genetic model used in the simulation. This in turn implies that the power we have to detect associations in males depends on the genetic model specified. This is at odds with our understanding that the power to detect associations in males should be identical regardless of the genetic model (see section 3.8).

To highlight this problem I wrote a simulation in R that uses Zheng's method to simulate genotype data in 2000 cases and 2000 controls at a single X chromosome SNP. Half the cases and half the controls are male and the data are simulated under an additive, a dominant, and a recessive genetic model with 1000 replicates for each.

The minor allele frequency is arbitrarily set at 0.3 and the GRR for B/B females is set at $r = 3$. Recalling table 3.10 and table 3.11, for females we have $\boldsymbol{\lambda}_f = (1, 2, 3)$ under an additive model, $\boldsymbol{\lambda}_f = (1, 3, 3)$ under the dominant model, and $\boldsymbol{\lambda}_f = (1, 1, 3)$ under the recessive model. For males, the GRRs for the $A/-$ and $B/-$ genotypes are given by $\boldsymbol{\lambda}_m = (1, 3)$ regardless of the genetic model. The variable of interest in this simulation is the number of male cases with the $B/-$ genotype — this should remain constant regardless of the genetic model according to section 3.8.

The results of the simulation are presented in figure 4.2 and we see that the distribution of male genotypes clearly depends on the underlying genetic model when using Zheng's simulation method. I attempt to correct this using my own simulation method which I now describe.

4.2 Methods

My simulation methodology is adapted from the *autosomal* simulation methods of Slager and Schaid (2001).

Simulation methodology

We consider a single locus on the X chromosome under the assumption of Hardy-Weinberg equilibrium. Recall that the distribution of genotypes in the wider population is then given by

$$\begin{aligned} g^{(f)} &:= \left(g_0^{(f)}, g_1^{(f)}, g_2^{(f)}\right) \\ &= \big(\Pr(A/A), \Pr(A/B), \Pr(B/B)\big) \\ &= \big((1-p_{MAF})^2, 2p_{MAF}(1-p_{MAF}), p_{MAF}^2\big) \end{aligned}$$

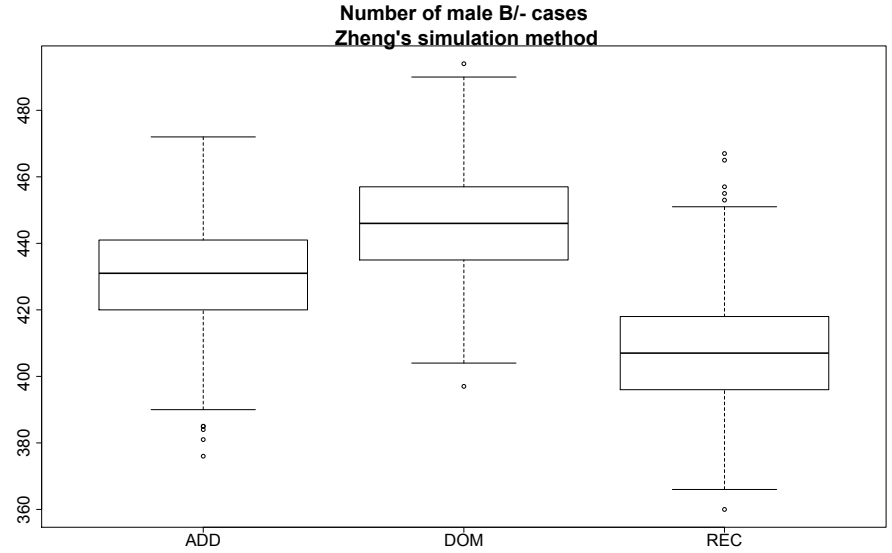

Figure 4.2: Boxplots showing the differing distributions of male $B/-$ cases under an additive, dominant, or recessive disease model when using Zheng's simulation method. The boxplots show the results of 1000 replicates for each genetic model.

in females, and

$$\begin{aligned} g^{(m)} &:= \left(g_0^{(m)}, g_2^{(m)}\right) \\ &= \left(\Pr(A/-), \Pr(B/-)\right) \\ &= \left(1 - p_{MAF}, p_{MAF}\right) \end{aligned}$$

in males, where p_{MAF} is the frequency of the minor allele in the wider population of males and females (see section 1.4.1). We assume that the minor allele, B, with population frequency, p_{MAF}, is the causal allele in all that follows.

Assume a random sample of R cases and S controls ($R + S = N$) and consider 2 possible alleles at the loci, A and B, where B is the risk allele. Given disease status and sex, the distribution of genotypes is multinomial with parameter vector $p^{(f)} = (p_0^{(f)}, p_1^{(f)}, p_2^{(f)})$ for female cases and $q^{(f)} = (q_0^{(f)}, q_1^{(f)}, q_2^{(f)})$ for female controls.

For males, the distribution of genotypes given disease status is binomial with $p^{(m)} = (p_0^{(m)}, p_2^{(m)})$ in cases and $q^{(m)} = (q_0^{(m)}, q_2^{(m)})$ in controls. The superscripts f, m are to denote females and males respectively while the subscripts $0, 1, 2$ are to remind us that we consider the male hemizygotes to be equivalent to the corresponding female homozygotes. The values of

$p^{(f)}, p^{(m)}, q^{(f)}, q^{(m)}$ are related to the genetic model through the following set of equations.

Firstly, consider the distribution of female genotype. We define

$$p^{(f)} = \left(\frac{f_0^{(f)} g_0^{(f)}}{\sum_{i=0}^{2} f_i^{(f)} g_i^{(f)}}, \frac{f_1^{(f)} g_1^{(f)}}{\sum_{i=0}^{2} f_i^{(f)} g_i^{(f)}}, \frac{f_2^{(f)} g_2^{(f)}}{\sum_{i=0}^{2} f_i^{(f)} g_i^{(f)}} \right)$$
$$q^{(f)} = \left(\frac{(1-f_0^{(f)}) g_0^{(f)}}{\sum_{i=0}^{2} (1-f_i^{(f)}) g_i^{(f)}}, \frac{(1-f_1^{(f)}) g_1^{(f)}}{\sum_{i=0}^{2} (1-f_i^{(f)}) g_i^{(f)}}, \frac{(1-f_2^{(f)}) g_2^{(f)}}{\sum_{i=0}^{2} (1-f_i^{(f)}) g_i^{(f)}} \right).$$

Here $f_0^{(f)}, f_1^{(f)}, f_2^{(f)}$ are the *penetrances* for the female genotypes A/A, A/B, B/B respectively. The higher the penetrance of a disease for a particular genotype, the more likely a person with the given genotype will be affected by the disease. The values of the $f^{(f)}$ are not usually known but can be estimated from the disease prevalence K and the genotypic relative risks $\boldsymbol{\lambda}_f = (\lambda_0, \lambda_1, \lambda_2) = (1, \lambda_1, \lambda_2)$ using[1]

$$\begin{aligned} f_0^{(f)} &= \frac{K}{g_2^{(f)} \lambda_2 + g_1^{(f)} \lambda_1 + g_0^{(f)}}, \\ f_1^{(f)} &= f_0^{(f)} \lambda_1, \\ f_2^{(f)} &= f_0^{(f)} \lambda_2. \end{aligned}$$

Thus the distribution of genotypes depends on the GRRs, the disease prevalence and the population allele frequencies. Note that this is also how Zheng et al. define the distribution of female genotypes.

However, as a point of difference to Zheng et al., I define the distribution of male genotypes by

$$p_i^{(m)} = \left(\frac{f_0^{(m)} g_0^{(m)}}{f_0^{(m)} g_0^{(m)} + f_2^{(m)} g_2^{(m)}}, \frac{f_2^{(m)} g_2^{(m)}}{f_0^{(m)} g_0^{(m)} + f_2^{(m)} g_2^{(m)}} \right)$$
$$q_i^{(m)} = \left(\frac{(1-f_0^{(m)}) g_0^{(m)}}{(1-f_0^{(m)}) g_0^{(m)} + (1-f_2^{(m)}) g_2^{(m)}}, \frac{(1-f_2^{(m)}) g_2^{(m)}}{(1-f_0^{(m)}) g_0^{(m)} + (1-f_2^{(m)}) g_2^{(m)}} \right)$$

where $f_0^{(m)}, f_2^{(m)}$ are the penetrances for the male genotypes $A/-, B/-$ respectively. Similarly to the female definitions, the values of the $f^{(m)}$ can be calculated from the disease prevalence K and the genotypic relative risks $\boldsymbol{\lambda}_m = (\lambda_0, \lambda_2) = (1, \lambda_2)$ using

$$\begin{aligned} f_0^{(m)} &= \frac{K}{g_2^{(m)} \lambda_2 + g_0^{(m)}} \\ f_2^{(m)} &= f_0^{(m)} \lambda_2. \end{aligned}$$

[1]There is an error in Slager and Schaid's definition of f_0 (Susan Slager, personal communication). This is the corrected version.

We have seen in section 4.1 that by applying Zheng et al.'s methodology the distribution of male genotypes depends on the genetic model. More specifically, we can now see that this is caused by the female heterozygous GRR, λ_1 (see Remark A.2 in the appendix).

To confirm that my simulation method does not suffer from the same flaw as that of Zheng et al., I repeat the simple simulation of section 4.1, but this time using my simulation methodology. We see in figure 4.3 that the distribution of male $B/-$ cases is now identical, regardless of the genetic model used in the simulation.

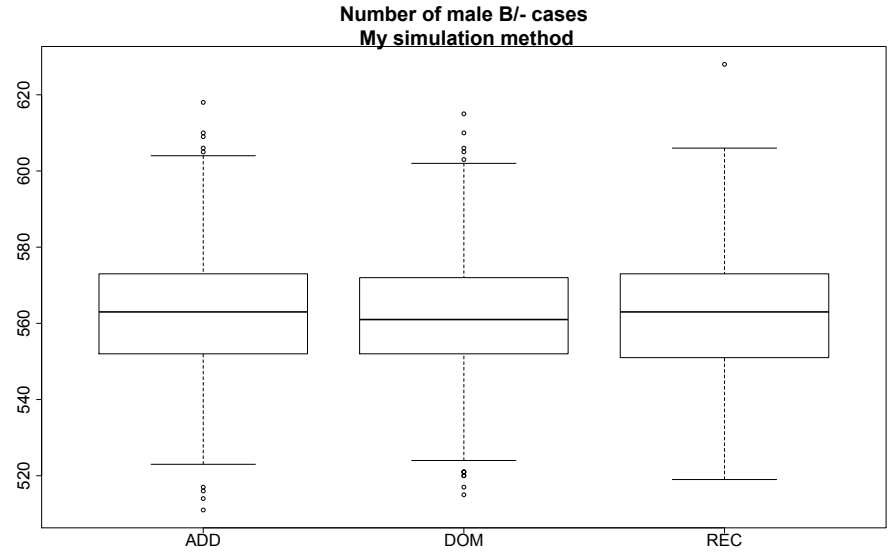

Figure 4.3: Boxplots showing the identical distributions of male $B/-$ cases under an additive, dominant, or recessive disease model when using my simulation method. The boxplots show the results of 1000 replicates for each genetic model.

Numerical example

We now work through a short example, using the equations of section 4.2, to give a sense of how the genotype frequencies will vary between the case and control cohorts in males and females. Consider a disease with prevalence rate $K = 1/1000$ under an additive genetic model with $r = 3$, i.e. $\boldsymbol{\lambda}_f = (1, 2, 3)$ and $\boldsymbol{\lambda}_m = (1, 3)$. Choosing $p_{MAF} = 0.05$ we have the population genotype frequencies

$$g^{(f)} = (0.9025, 0.095, 0.0025)$$
$$g^{(m)} = (0.95, 0.05).$$

Having defined the population level parameters we can now calculate the genotype frequencies in the case and control cohorts using the equations of section 4.2. These are

$$\begin{aligned} p^{(f)} &= (0.820, 0.173, 0.007) \\ q^{(f)} &= (0.903, 0.095, 0.002) \\ p^{(m)} &= (0.905, 0.095) \\ q^{(m)} &= (0.950, 0.050) \end{aligned}$$

to 3 decimal places.

Sensibly, the case cohort's genotype frequencies, p, now differ noticeably from the population genotypes frequencies, g, while the control cohort's genotype frequencies, q, remain close to the population genotype frequencies, g.

Simulation parameters and test statistics

Having established the validity of my simulation method, we now define the various test statistics whose performance we compare under different experimental designs. The eight tests we compare in my study, and the reasons for their inclusion, are:

- Clayton's 1 and 2 degree of freedom tests defined in section 3.7; denoted $S^{(1)}$ and $S^{(2)}$ respectively in the simulations. These two tests were proposed in the paper that motivated the topic of my thesis.

- Z_A^2, Z_C^2, Z_{mfA}^2 and Z_{mfG}^2 defined in section 3.6.1 and previously studied by Zheng et al. We investigate these here under a broader set of parameters and uncouple the dependence between the distribution of male genotypes and genetic models. The allele based test, Z_A^2, is the simplest test and should perform well given that we are simulating under HWE. The remaining three tests are novel ways of combining results across sexes.

- The χ^2 test on 2 degrees of freedom for the *female* genotype table or Cochran-Armitage trend test using only the *female* samples; denoted by χ^2_{female} and CA_{female} respectively in the simulations. When using PLINK these are the default association tests for X loci if the user performs a "naïve" analysis. Indeed, PLINK's CA_{female} test is used by the ANZgene consortium to test for association on the X chromosome in their GWAS on multiple sclerosis (see Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009, Methods).

There are obviously many parameters that can be altered in such simulations. To keep things as simple as possible we consider a sample consisting

of 2000 cases and 2000 controls, in line with recommended GWAS sample sizes. The genome-wide significance level is fixed at $\alpha = 10^{-8}$, in line with the consensus Type I error rate discussed in section 3.3. The prevalence of the disease is set at $K = \frac{1}{1000}$ which is the estimated prevalence of multiple sclerosis in people of northern European ancestry[2] (Oksenberg et al., 2008) and is a more realistic prevalence rate for complex genetic diseases than $K = 0.1$.

These parameters are kept fixed while we explore the effects of altering the parameters in table 4.2 on the performance of the various test statistics.

Parameter	Levels
Proportion of cases female	$1, 0.9, 0.75, 0.6, 0.5, 0.4, 0.25, 0.1, 0$
Proportion of controls female	Fixed at 0.5 *or* matched to the proportion of female cases
Genetic Model	Additive (ADD), dominant (DOM) *or* recessive (REC)
GRR	$\boldsymbol{\lambda}_f = (1, \lambda_1, \lambda_2), \boldsymbol{\lambda}_m = (1, r)$ defined by the genetic model with $\lambda_2 = r = 1.5$, 2.5 *or* 3
MAF in wider population	$0.01, 0.02, \ldots, 0.49, 0.50$

Table 4.2: Parameters varied in the simulation study

If we consider each level of the minor allele frequency separately, this gives a total of $9 \times 2 \times 3 \times 3 \times 50 = 8100$ experimental designs to assess the performance of the 8 test statistics on. We also report the results under the null hypothesis of no association, corresponding to $\boldsymbol{\lambda}_f = (1, 1, 1), \boldsymbol{\lambda}_m = (1, 1)$, to investigate the Type I error rates of the 8 tests.

All 8 tests are similar in computational complexity and so the criteria on which we rank them is empirical power. The empirical power of a test is defined as the number of times the test gives a genome-wide significant result divided by the number of replications for each set of parameter levels.

The tests implemented in my simulation study do not form an exhaustive list of association tests for X chromosome loci. However, our particular interest with this study is to compare Clayton's proposed test statistics to those methods previously studied or currently in use. Some test procedures (such as PLINK's logistic regression approach and the proposed GLM approach with sandwich estimates of the covariance matrix from Clayton (2008)) could not be included in my simulation due to computational constraints. Both of these methods require the fitting of at least one GLM per simulation replicate which greatly increases the simulation run-time.

[2]All the samples used in the ANZgene multiple sclerosis GWAS are all off European descent hence my choice of K

As it is, my simulation takes approximately 150 hours when run as a single process on a unix server running 4 $\times$ Quad Core CPUs @ 2.93GHz with 128GB of RAM. This run-time can be reduced by splitting the simulation across multiple processors but at the expense of increased programming complexity. I satisfied myself with an overall runtime of approximately 50 hours for the simulation with a common 50 : 50 control cohort, and a similar amount of time for the simulation with the case and control numbers matched by sex.

4.3 Results

These simulations clearly produce a large amount of data and so it requires some thought in order to present the results as clearly as possible. I will present and interpret plots of the power achieved by each test statistic for each "experimental design", e.g. an additive model with $r = 2.5$ using a common 50 : 50 control cohort. My aim is to explain the significant trends *across* experimental designs rather than focus on the smaller details of individual plots. That said, I will begin with a discussion of a single plot; each plot is of identical format and so this example will explain the key features shared by all the graphs.

Figure 4.4 is our example plot. It shows the empirical power for all 8 test statistics under a recessive disease model with $r = 2.5$ and both the case and control cohorts with a 50 : 50 sex ratio. We define the sex ratio by $females : males$ with 100 : 0 and 0 : 100 defined as all female and all male respectively. The key features of this example plot are described in the caption accompanying figure 4.4.

In what follows we do not include the results for the 100 : 0 all-female or 0 : 100 all-male cohorts. The test statistics are designed for mixed-sex samples and are not appropriate when either cohort, or indeed both, are made up of a single sex. The single-sex simulations were designed as control datasets to ensure the simulation behaved as intended and the results are not of interest to us here.

There are far too many experimental designs to give a commentary on each resulting plot individually. Indeed, to do so would be to ignore the more interesting and relevant "bigger picture". For this reason we present graphs together in blocks of 7 where each block is a single experimental design. We also exclude the individual points of each power curve and instead use a simple coloured line to display each test's power. The individual marks simply clutter the figure and make the distinctions between test statistics less clear.

There is some "wriggly-ness" in the line graphs as a result of simulation variance — applying a smoothing spline to remove this wriggly-ness was investigated but was found to distort the true results and was thus more

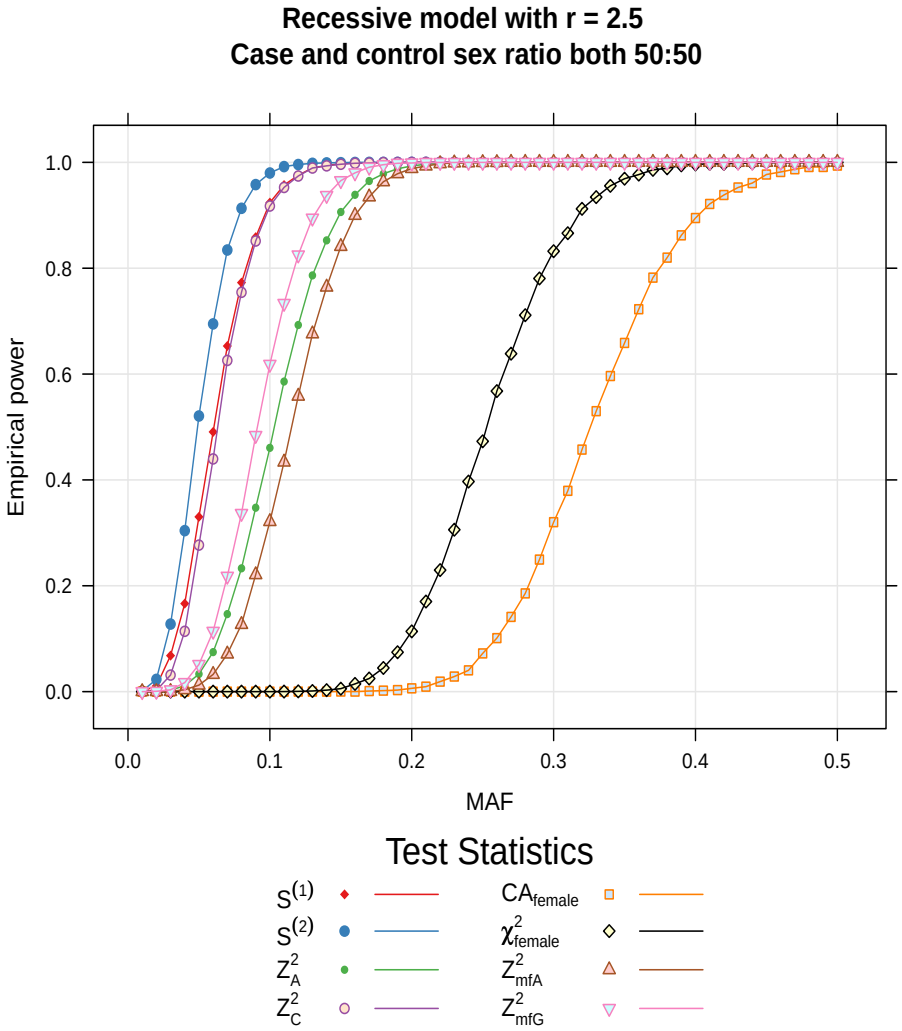

Figure 4.4: For each test statistic we can read off its power for a given minor allele frequency (MAF). In this example we see that $S^{(2)}$ is uniformly most powerful across the range of MAFs. Here the risk associated with the recessive genotype is so large that most tests achieve maximum power for a low MAF. Note that both female only tests, CA_{female} and χ^2_{female}, perform considerably worse than the 6 tests that use both male and female samples. This is to be expected and is a consistent result in all except the most extreme simulation set-ups.

hindrance than help.

We first discuss the results of the simulation study using a common $50:50$ control cohort. There is little difference between the results for the $r = 2.5$ and $r = 3$ simulations, and so the $r = 3$ results are not shown here. In the $r = 3$ case the power curves are simply "compressed" versions of the $r = 2.5$ curves that have been shifted to the left, i.e. each test has more power for lower MAFs.

4.3.1 $50:50$ control cohort

Ideally the cases and controls in a GWAS should be matched by sex but it is increasingly common to acquire control samples from a "genome-bank" due to the high costs of collecting and genotyping samples (McCarthy et al., 2008). These genome-bank cohorts typically consist of a roughly $50:50$ split of males and females. It is therefore of interest to analyse how these tests would perform in a GWAS using the common $50:50$ control cohort.

Type I error rates of tests

The Type I error rate of a test is the proportion of times the test reports a result that would lead us to falsely reject the null hypothesis. It is also known as the size of the test. To analyse the size of each test we simulate data under the null hypothesis of no association between genotype and phenotype. For each test we then calculate the proportion of times the test reports a significant result, where significance is given by the nominal level of $\alpha = 0.05$.

This simulation is done over the range of minor allele frequencies and case cohort sex-ratios, with each experimental design replicated 10,000 times. A well behaved test should therefore return significant results only 500 times under each experimental design for a Type I error rate of 0.05.

The results for each test, across the range of case cohort sex-ratios, are summarised in figure 4.5. Each box plot represents the distribution of Type I errors for the range of minor allele frequencies, and so ideally each plot should be tightly centred around the nominal value of 0.05. In each plot the number in the header, highlighted in orange, refers to the proportion of cases that are female. The figures should thus be read from left to right, top to bottom, to see how the increase in proportion of females in the case cohort effects the results.

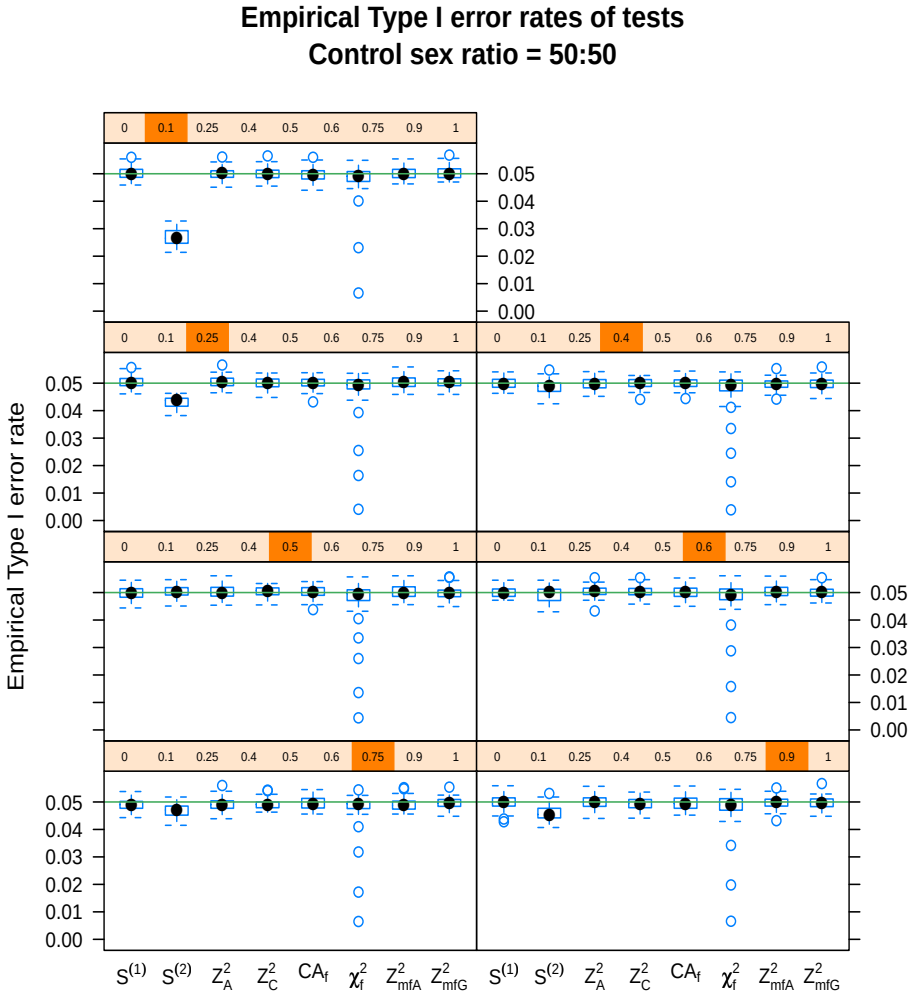

Figure 4.5: We see that the Type I error rates of all but the $S^{(2)}$ test are centred around the nominal 0.05 value (green line). For strongly male-biased case cohorts the $S^{(2)}$ test falsely rejects the null less often than its counterparts, i.e. $S^{(2)}$ is a conservative test. The $S^{(2)}$ test is also slightly conservative for the most highly female-biased case cohort. These results do raise questions about the appropriateness of the approximate χ^2 distribution for $S^{(2)}$. Note that for graphical purposes test names have been abbreviated: $CA_f \equiv CA_{female}$ and $\chi^2_f \equiv \chi^2_{female}$

Further to the discussion of figure 4.5, R's in-built function `chisq.test` (which is used in my simulation for some tests) will produce warnings when the appropriateness of the approximate χ^2 distribution is in doubt.The way the simulation is currently coded these warnings are noted but no action is taken. The results of this can be seen in the larger variation of the Type I error rate for the χ^2_{female} test. This is due to a low number of counts of a particular genotype/phenotype combination. This is typically only observed in the simulations with a minor allele frequency $0.01 - 0.05$ (results not shown).

In summary, we now know that for a strongly sex-biased case cohorts, in conjunction with a common $50:50$ control cohort, that the asymptotic distribution of the $S^{(2)}$ statistic may not be appropriate. We also note that the $S^{(2)}$ statistic is a conservative test, though this of much less concern. We therefore suggest further study of the approximate distribution of the $S^{(2)}$ statistic. An alternative approach to relying on the approximate distribution of the test would be to use permutation testing to get empirical p-values, but this is not pursued here. All other tests are of the correct size.

Additive models

Recall that under the additive model the female genotypic relative risks are $\boldsymbol{\lambda}_f = (1, \frac{r+1}{2}, r)$ and the male genotypic relative risks are $\boldsymbol{\lambda}_m = (1, r)$.

Figure 4.6 and figure 4.7 are the plots for the $r = 1.5$ and $r = 2.5$ simulations. Again the figures should be read from left to right, top to bottom, to see how the distribution of sexes in the case cohort effects the power to detect associations for X chromosome SNPs.

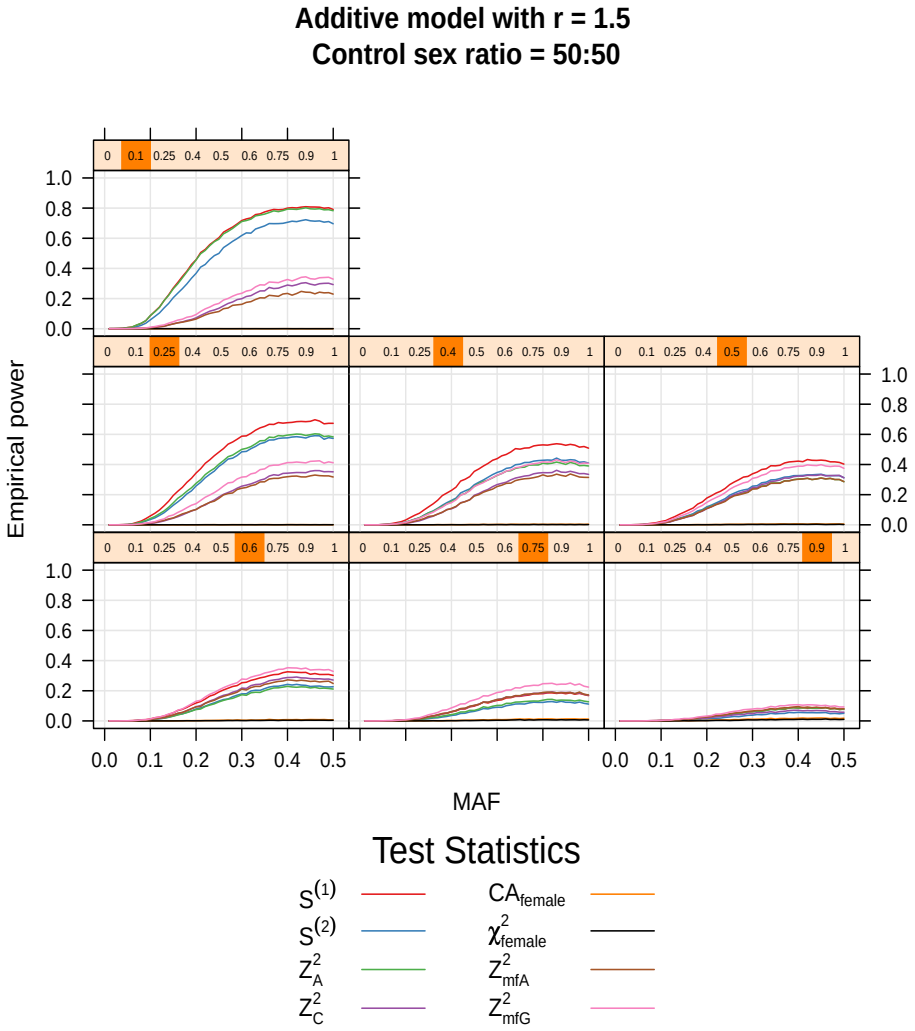

Figure 4.6: Recall that the number highlighted in orange in the header of each plot corresponds to the proportion of the case cohort that is female. The immediately apparent trend is that the power is greater for case cohorts that have a higher proportion of males. This is a common theme and will be discussed further. Clayton's 1 degree of freedom test, $S^{(1)}$, is consistently among the top 2 tests. We have less than 50% power to detect associations in most designs and particularly poor performance for low MAFs. The female only tests, CA_{female} and χ^2_{female}, have power $\leq 3\%$, regardless of the proportion of cases that are female. We note that for the female-biased case cohorts that the Z^2_{mfG} test is most powerful. This result will be relevant in the case study.

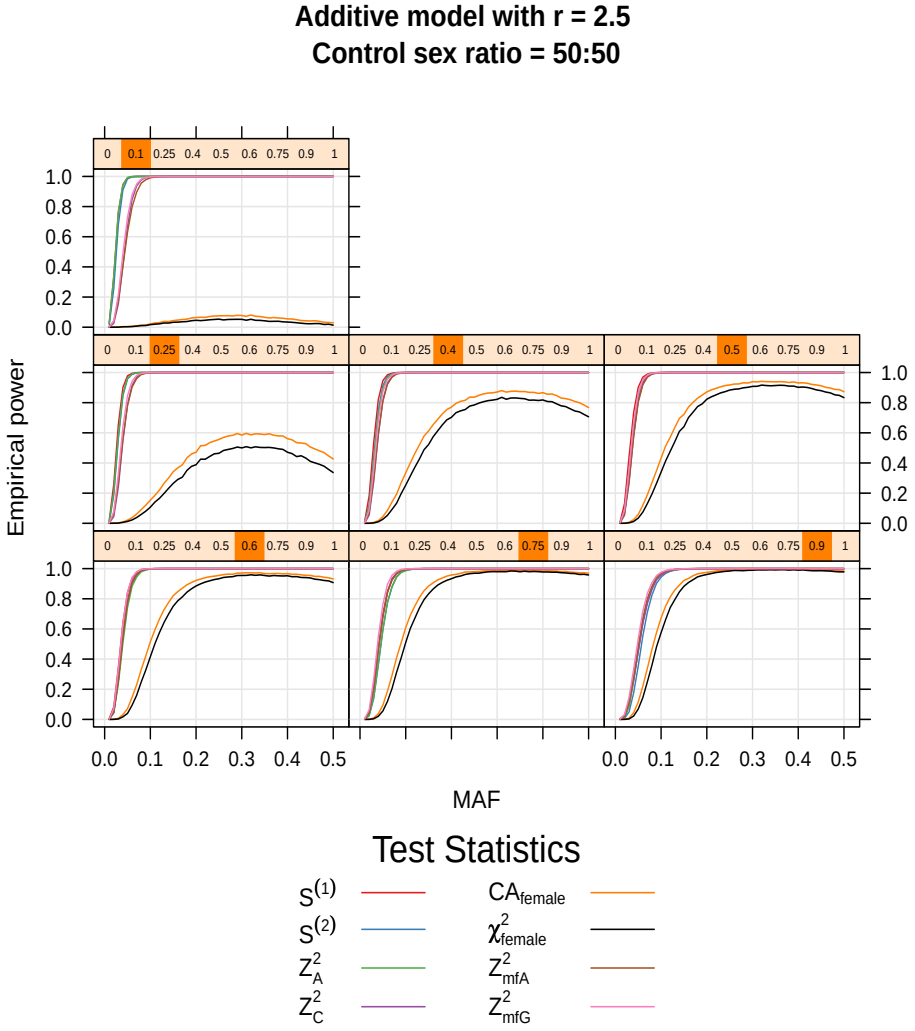

Figure 4.7: Increasing the genetic risks to $r = 2.5$ produces an immediate increase in the power to detect associations. In all but the most extremely female-biased case cohorts, maximum power is achieved for a MAF ≤ 0.1 and maintained thereafter. Clayton's tests are again amongst the best, but all the tests have high power and are fairly indistinguishable, aside from the less powerful CA_{female} and χ^2_{female} tests. The female-to-male ratio of the case cohort has far less impact on the results than for the $r = 1.5$ design in figure 4.6. Note the decay in power for the female-only tests, CA_{female} and χ^2_{female}, for the higher MAFs.

We see from figures 4.6 and 4.7 that Clayton's 1 degree of freedom test, $S^{(1)}$, is amongst the most powerful tests for all additive simulation set-ups. It therefore suggests itself as the best test for X chromosome loci when the true genetic model is additive. Indeed, by using Clayton's choice of $a = (0, 1, 2)$, $S^{(1)}$ is designed to test for additive effects so this is an expected result.

The simplest test, the allele based test Z_A^2, also performs quite well across additive simulations. However, recalling that the data were simulated assuming HWE this is not surprising and we must bear in mind the usual caveat that Z_A^2 is not reliable when HWE fails to hold. The decay in power for the female-only tests with high minor allele frequencies is an interesting result. This trend will be repeated, even more dramatically, in later results for the dominant model.

Recessive model

Recall that under the recessive model the female genotypic relative risks are $\boldsymbol{\lambda}_f = (1, 1, r)$ and the male genotypic relative risks are $\boldsymbol{\lambda}_m = (1, r)$. Figures 4.8 and 4.9 are the blocks of plots for the $r = 1.5$ and $r = 2.5$ simulations.

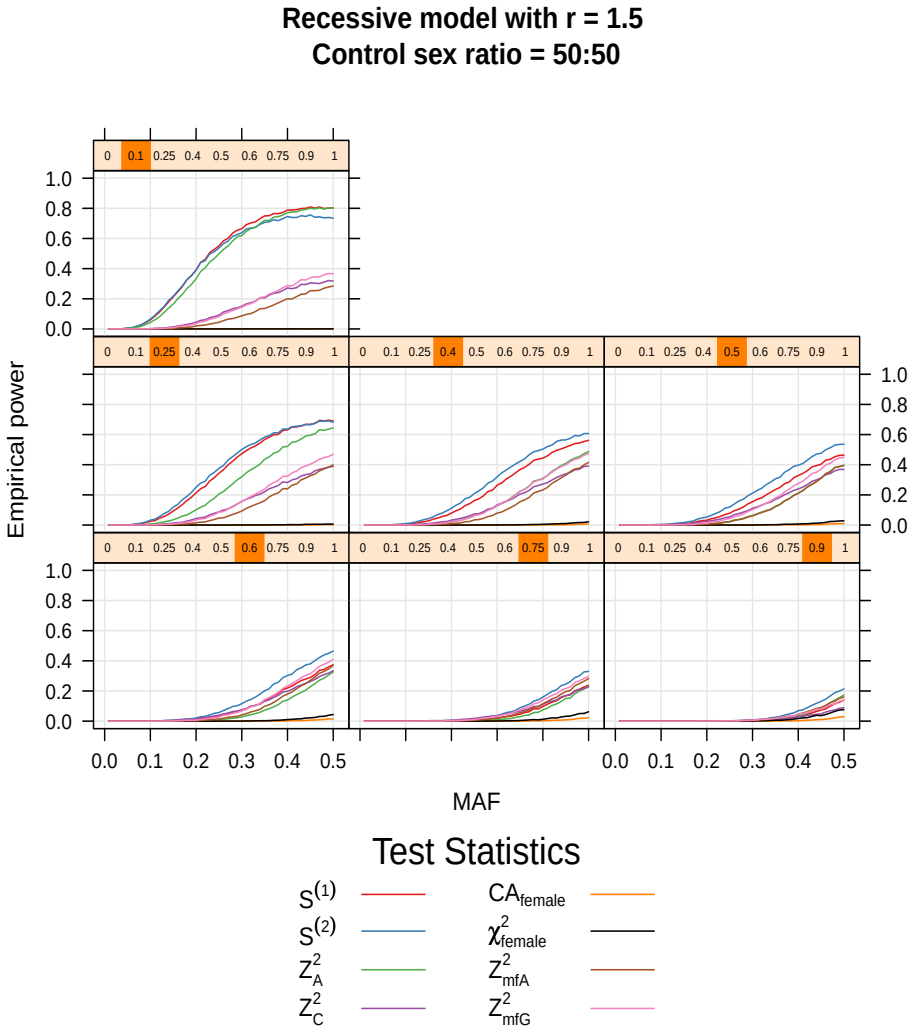

Figure 4.8: Again we see that the power increases with the proportion of the case cohort that is male. The power curves are not dissimilar to those for the additive model with $r = 1.5$, shown in figure 4.6. However, we see that $S^{(2)}$ achieves the highest power, except in strongly male-biased case cohorts. Zheng et al.'s Z^2_C, Z^2_{mfA} and Z^2_{mfG} tests perform noticeably worse than Clayton's tests. Except in the strongly male-biased case cohorts, we have little power to detect associations for all but the highest MAFs, and even then the maximum power is only $35\% - 60\%$.

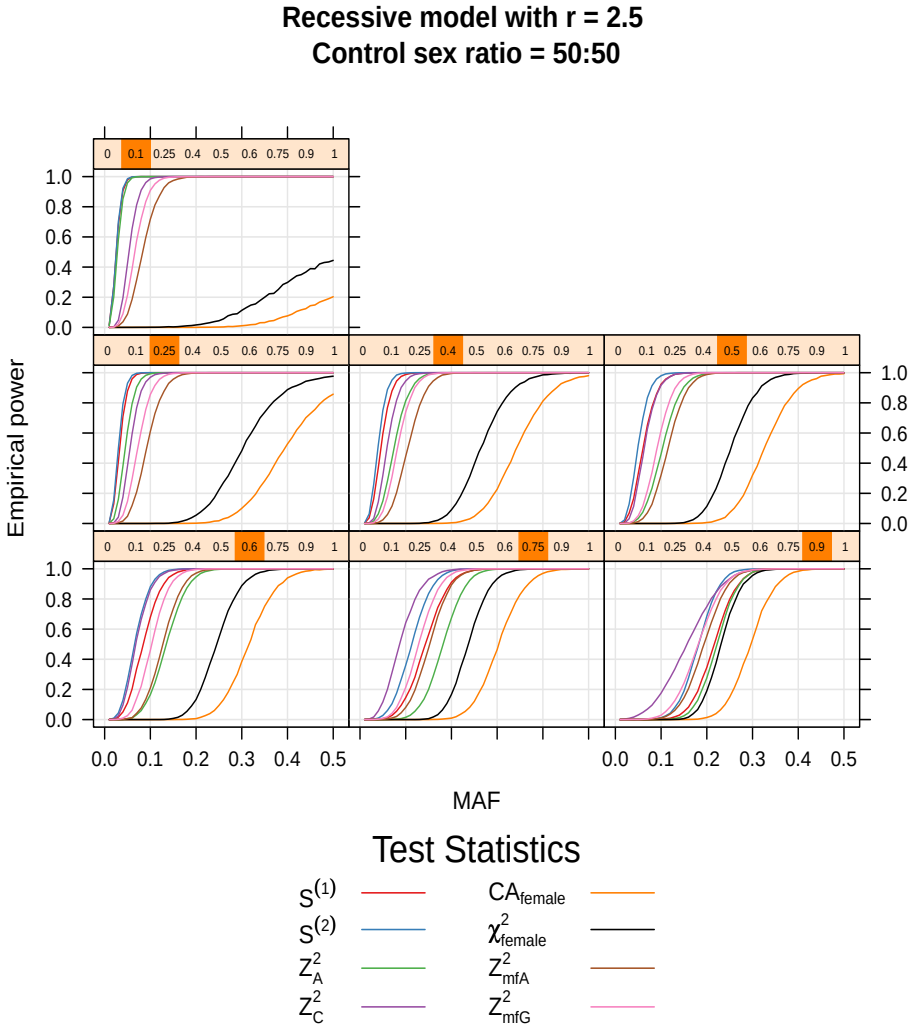

Figure 4.9: Increasing the genetic risk to $r = 2.5$ again has the obvious effect of increasing our power to detect associations. What is interesting here is the narrow range of MAFs between having "little power" and "maximum power". We generally either have very good prospects of detecting the association, or very little hope, with not much in between. This is in contrast to the much more gradual increase in power for the recessive model with $r = 1.5$, shown in figure 4.8. $S^{(2)}$ is again amongst the top 2 most powerful tests. However, for strongly female-biased case cohorts we see an advantage in using Z^2_C, particularly for low MAFs.

The results for the recessive model share a few points in common with the results for the additive model, such as:

- An increase in power with an increase in the proportion of cases that are male
- An increase in power as the genetic risk r increases
- Clayton's tests consistently amongst the best performing test statistics.

$S^{(1)}$ does perform slightly worse for the recessive model than for the additive model. This is to be expected since $S^{(1)}$ is designed to detect additive genetic effects, not recessive genetic effects. However, $S^{(1)}$ still performs better than most other tests for the recessive simulations. Clayton's 2 degree of freedom test, $S^{(2)}$, is designed to be robust to the underlying genetic model and it shows its strength here outperforming $S^{(1)}$ for all 7 plots.

Dominant model

Recall that under the dominant model the female genotypic relative risks are $\boldsymbol{\lambda}_f = (1, r, r)$ and the male genotypic relative risks are $\boldsymbol{\lambda}_m = (1, r)$. The test statistics behave quite differently for the dominant model when compared with the additive and recessive models. Rather than the power for each test increasing more-or-less monotonically across the range of MAFs, under the dominant model there is a distinct decay in the tails for the power curves. This will be discussed further after we examine the plots for the $r = 1.5$ and $r = 2.5$ dominant simulations in figures 4.10 and 4.11.

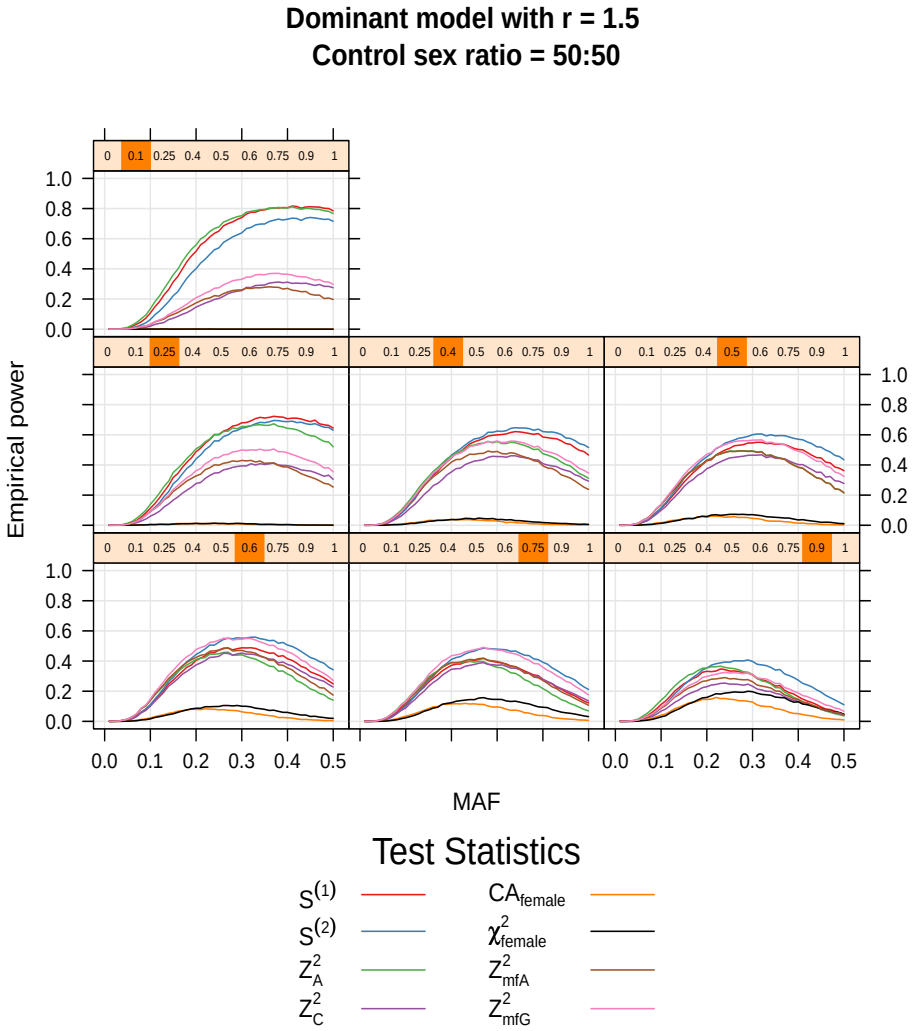

Figure 4.10: In all but the most male-biased case cohorts we see maximum power is achieved for a MAF $\approx 0.25 - 0.35$, with power then decaying for higher MAFs. We see that $S^{(1)}, S^{(2)}, Z_A^2$ typically outperform the remaining tests. Maximum power is limited, $\leq 60\%$ in most cases, with particularly poor performance for low MAFs.

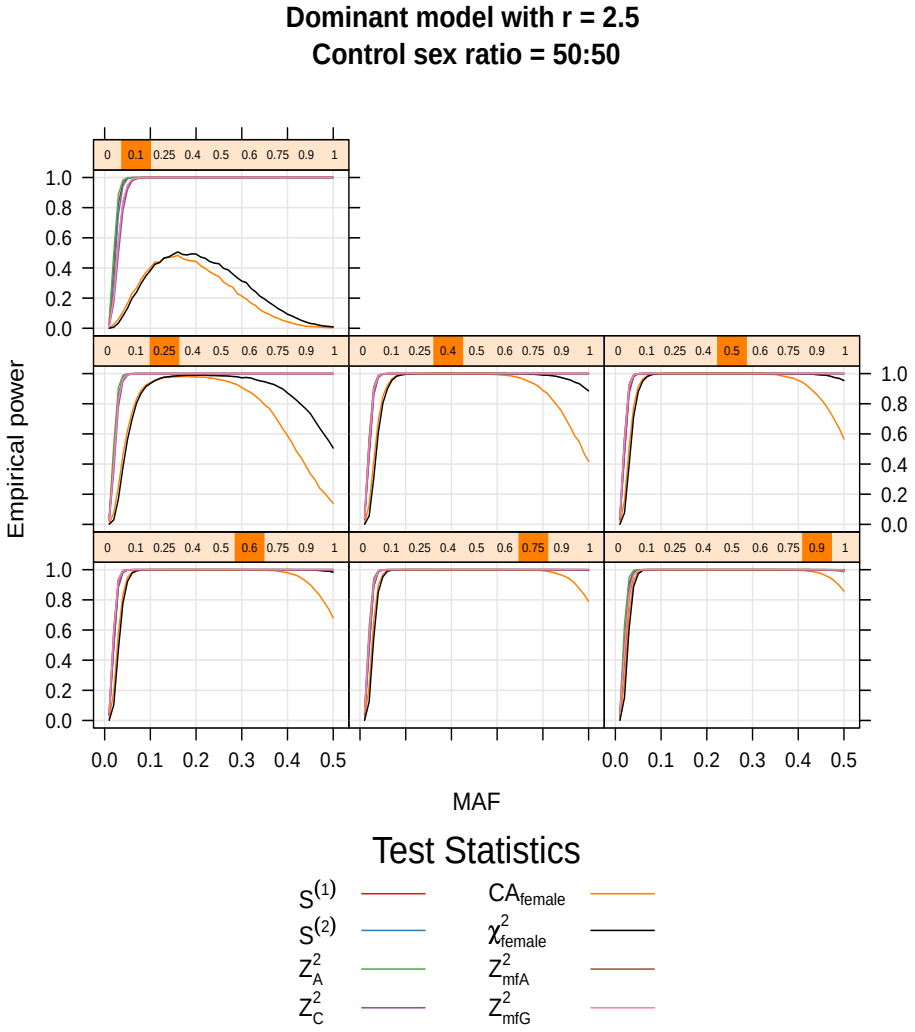

Figure 4.11: With the r-value increased to 2.5, the effect size is large enough to mostly "overcome" the decay in the tails of the power curves seen in figure 4.10. Indeed the effect size is so large that even the female-only tests achieve maximum power for 6 out of the 7 plots. The power curves for the remaining tests are indistinguishable — all achieving maximum power for very low MAFs — meaning that the choice of test for this genetic model is less important than in others since we have very high power using any of $S^{(1)}, S^{(2)}, Z^2_A, Z^2_C, Z^2_{mfA}$ or Z^2_{mfG}.

The lack of monotonicity in the power curves for the dominant model, most pronounced for small r-values, is a somewhat surprising result. My first reaction was to return to my code to check for any errors but none were found.

I have not been able to entirely satisfy myself with a suitable explanation for this decay in power. However, this decrease in power for higher minor allele frequencies is not entirely without precedent. An example of this can be seen in the work of Lettre et al. (2007) on autosomal association tests, though it is admittedly a less dramatic result. There are a number of differences between the parameters in effect in Lettre et al.'s simulation study and mine, however I believe the point remains. Figure 4.12 is a plot from Lettre et al. (2007) showing a decrease in power for the additive version of the Cochran-Armitage trend test when the MAF is high and the true genetic model is dominant.

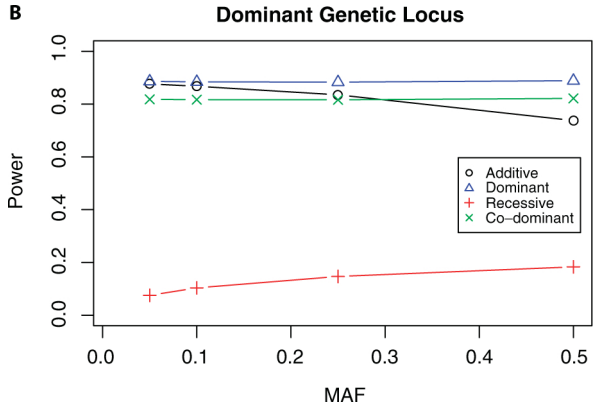

Figure 4.12: For an autosomal dominant genetic locus we see the additive version of the Cochran-Armitage has its lowest power for the highest minor allele frequency (MAF). We have seen a similar decrease in power in my simulations of X chromosome dominant genetic loci. This figure is adapted from figure 1B of Lettre et al. (2007).

To properly understand this result requires further work, however I believe it is due in part to the fact that for higher MAFs the "risk" genotypes are very common even in those unaffected by the disease. For example, under a dominant model with $r = 1.5$, $p_{MAF} = 0.45$ and $K = 1/1000$, the

frequencies of genotypes are given by

$$\begin{aligned} p^{(f)} &= (0.2243, 0.5505, 0.2252) \\ p^{(m)} &= (0.4490, 0.5510) \\ q^{(f)} &= (0.3026, 0.4949, 0.2025) \\ q^{(m)} &= (0.5501, 0.4499). \end{aligned}$$

The female "risk" genotypes $A/B, B/B$ make up 70% of the female control genotypes (compared to 77% of female cases) and the male risk genotype, $B/-$, makes up 45% of the male control genotypes (compared to 55% of male cases). The people in the case cohort *do not* have the disease yet there is a similar percentage of people with the risk genotype to the case cohort. It is therefore harder to separate cases and controls using the information of their genotype at this locus alone and so we have less power to detect associations.

For larger r-values the distribution of genotypes in the case cohort becomes highly skewed in favour of the "risk" genotypes $A/B, B/B, B/-$. For example, if we use the exact same p_{MAF} and K as before, but increase the r-value to $r = 3$ we get

$$\begin{aligned} p^{(f)} &= (0.1263, 0.6200, 0.2537) \\ p^{(m)} &= (0.2895, 0.7105) \\ q^{(f)} &= (0.3027, 0.4949, 0.2024) \\ q^{(m)} &= (0.5503, 0.4497). \end{aligned}$$

We see that while the distribution of genotypes in the control cohort remains almost identical, there are now 87% of female cases and 71% of male cases with the risk genotypes. It is thus far easier to separate the cases and controls based on their genotype at this loci than it was in the previous example and so we have more power to detect associations when the r-value is large.

4.3.2 Matching case and control cohorts by sex

The current implementation of some of the tests will at times cause the simulation to crash for the highly sex-biased cohorts, e.g. the $90:10$ case and control design. Unfortunately this means I have not been able to study the sex-matched experimental designs in as greater detail as the $50:50$ control cohort design. The simulations of the recessive genetic model are particularly problematic and never ran to completion in my study. I believe this is due to difficulties inverting the $\widehat{V}$ matrix in Clayton's tests — $\widehat{V}$ is singular, or near singular (ill-conditioned), for the strongly sex-biased cohorts where case and control numbers are matched by sex. Consequently I only present

simulation results for the additive and dominant genetic models under the sex-matched experimental design. There exist computational methods to handle the situations where $\widehat{V}$ is ill-conditioned, or singular, but these are not pursued here.

So that I could produce and analyse simulations where case and control numbers are matched by sex, the number of replications was reduced to 1000, down from 10000. The estimated power of each test is of course more variable as a result, and this can be seen in the more "wiggly" power curves in the following plots.

In the header of each plot, highlighted in orange, is now the proportion of cases *and* controls that are female. Again, the $r = 3$ results are very similar to the $r = 2.5$ results and so are not presented here.

Type I error rates of tests

As we did for the common 50 : 50 control cohort, we analyse the Type I error rate of each test across minor allele frequencies. These results are summarised in the box plots of figure 4.13. The key point from these plots is that the Type I error rates are all centred around the nominal level of 0.05. We therefore feel more confident in the appropriateness of these tests when the cases and controls are matched by sex, rather than when we use a common 50 : 50 control cohort.

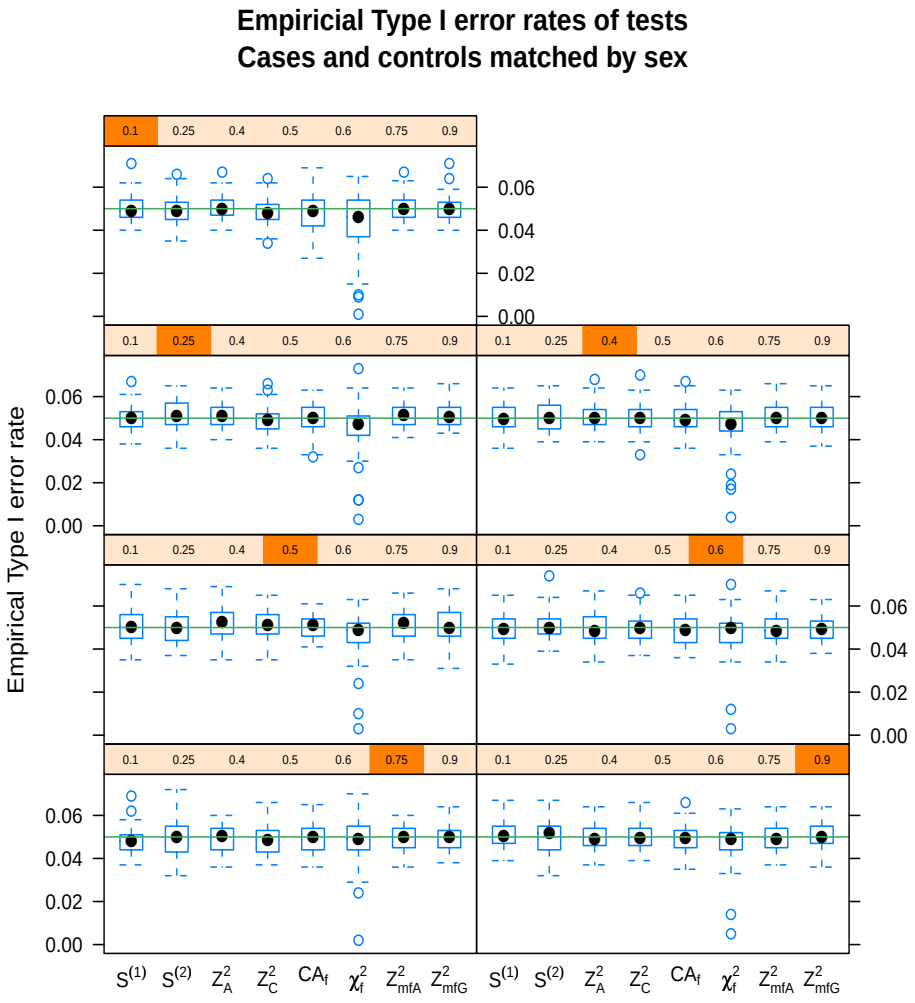

Figure 4.13: Each box plot represents the distribution of Type I error rates of a test across the range of MAFs $0.01-0.5$. All these box plots are correctly centred around the nominal level of 0.05 (green line). Further investigation reveals that the outliers in these plots correspond to simulations with low MAFs (results not shown).

As before, the loci with the lowest minor allele frequencies require some extra care to ensure there are sufficient counts in each genotype/phenotype category. Without this, a test statistic's approximate distribution may not hold, particularly if using the χ^2_{female} test.

Additive model

The simulation parameters are as for the additive model in the 50 : 50 control cohort simulations described in section 4.3.1. The results for the $r = 1.5$ and $r = 2.5$ simulations are shown in figures 4.14 and 4.15 respectively.

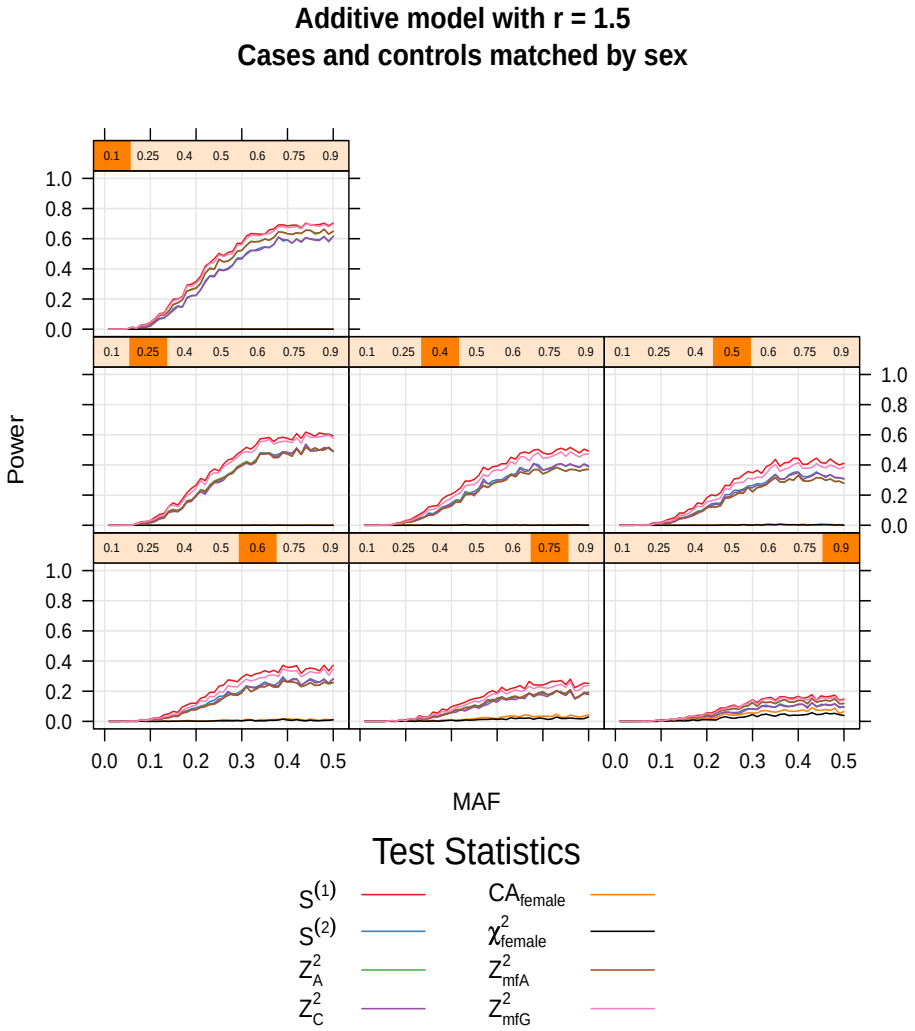

Figure 4.14: As the proportion of males in the whole sample increases we have more power to detect associations. There is again poor performance for low MAFs and the female-only tests have negligible power to detect associations. The two best performing tests are $S^{(1)}$ and Z^2_{mfG}, though even these have less than 50% power for the majority of loci.

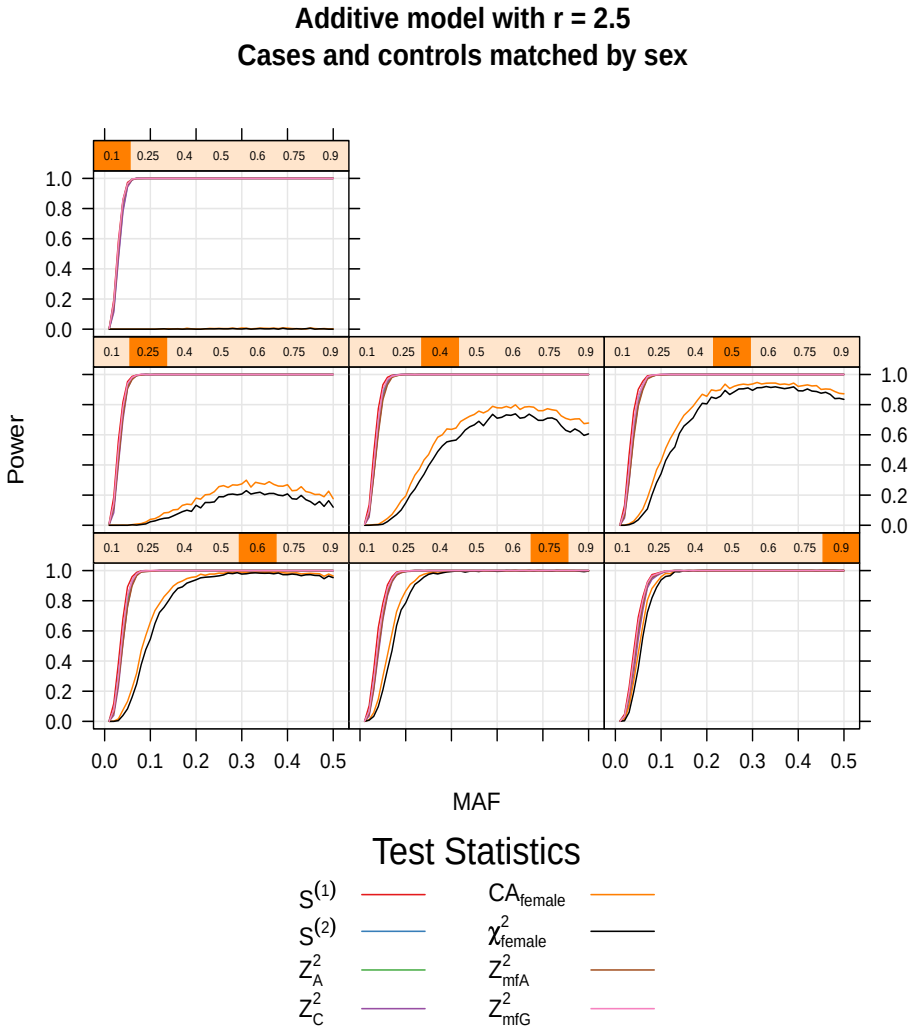

Figure 4.15: These results are very similar to those for the additive model with $r = 2.5$ and using a common $50 : 50$ control cohort (see figure 4.6). We have "maximum power" for low MAFs in each plot when using $S^{(1)}, S^{(2)}, Z^2_A, Z^2_C, Z^2_{mfA}$ or Z^2_{mfG}. The female-only tests simply cannot match these other tests even when up to 90% of samples are female.

What is interesting from these simulations is that the results are very similar to those for when a common 50 : 50 control cohort is used. This suggests that the matching of case and control numbers by sex may be less important than it first appears. The $S^{(1)}$ statistic is again the top performing test here, as to be expected for additive simulations.

Dominant model

The simulation parameters are as for the dominant model in the 50 : 50 control cohort simulations described in section 4.3.1. The results for the $r = 1.5$ and $r = 2.5$ simulations are shown in figures 4.16 and 4.17 respectively.

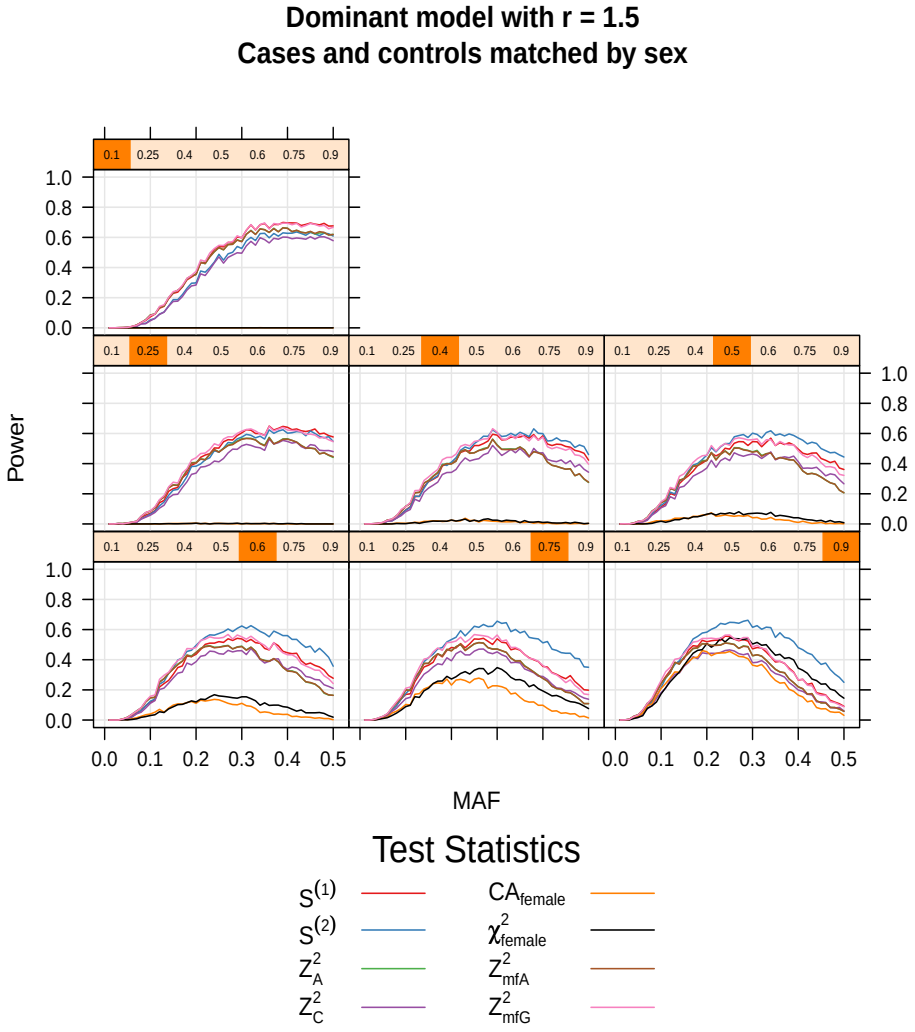

Figure 4.16: The difference between using one of $S^{(1)}, S^{(2)}, Z_A^2$ or one of $Z_C^2, Z_{mfA}^2, Z_{mfG}^2$ is far less than when using the common $50:50$ control cohort. This suggests that Zheng et al.'s tests perform better when case and control numbers are matched by sex. We again see the decay in power for all tests at higher MAFs. The maximum power for the best test, $S^{(2)}$, is consistently around 60% regardless of the proportion of samples that are female.

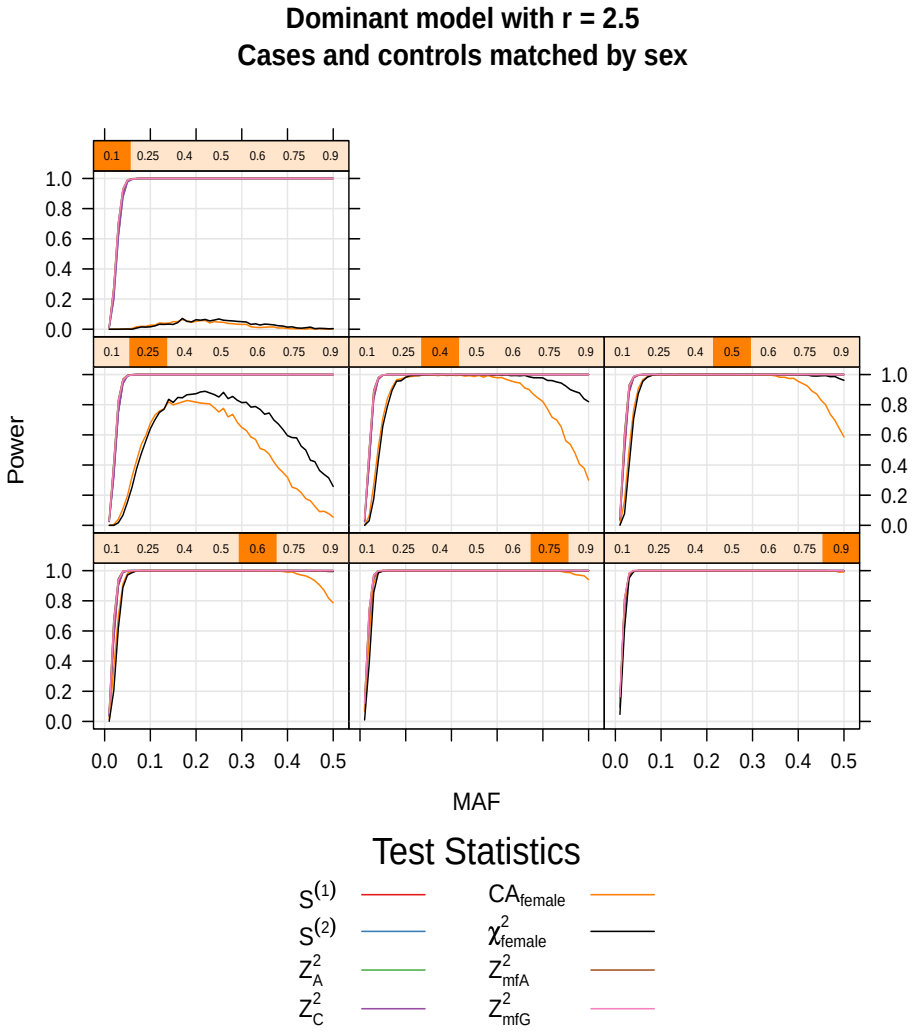

Figure 4.17: These results are also quite similar to the dominant model with $r = 2.5$ where a common $50 : 50$ control cohort is used (see figure 4.11). We have "maximum power" for very low MAFs and so we should be able detect any high-risk dominant X chromosome loci in real data with ease.

The power curves for each test (excluding the female-only tests) remain quite similar for the dominant model across the proportion of samples that are female. This is in contrast to when a common 50 : 50 control cohort is used; in that situation the power steadily decreases for each test as the proportion of male cases increases. These results show that we typically have more power to detect a dominant genetic model when the case and control cohorts are matched by sex but that this is only noticeable for low r-values. As the r-value increases the dominant genetic effect becomes so large that it can be detected by most tests regardless of the female-to-male ratio.

4.4 Discussion and conclusions of the simulation study

The results of the simulation study can be used to help choose the best statistic for testing genotype/phenotype associations in X chromosome GWAS data. Several key points have been have been illustrated in the simulation study, some that are more intuitively clear and obvious than others.

Firstly, discarding males from the analysis will lead to a large drop in power and is a wasteful strategy. We have more power to detect X chromosome genotype/phenotype associations in males as there are only 2 genotypes for male X chromosome loci, compared to the 3 genotypes for females. However, in testing males we have no power to discriminate between genetic models; this requires female samples. This suggests using all samples to detect associations, and when an association is found, to use the female samples to discriminate between genetic models.

It is clear that PLINK's "default" female-only tests are a poor choice to analyse X chromosome GWAS data — in every simulation considered here we have more power when using a test that combines male and female samples. PLINK includes other methods that do make use of both male and female samples to analyse X chromosome data, and these should be favoured when using this software. However, these methods are not the default option in PLINK and so users can easily apply the female-only tests unwittingly.

Secondly, we have little power to detect associations for the lowest minor allele frequencies, particularly when the allele confers only a small risk. Therefore it will be more difficult to identify these "rare" alleles using any of the methods considered. Indeed, the current crop of GWA studies are not the optimal experiments to identify these rare variants, though progress will be made as technology advances (see Altshuler et al., 2008, for further discussion on the difficulties in identifying rare variants involved in human diseases).

Finally, we discuss some practical matters. Clayton's tests are consistently among the best overall methods of those considered here. Clayton's

1 degree of freedom test, $S^{(1)}$, performs surprisingly well even when the true genetic model is not additive. Clayton's 2 degree of freedom test, $S^{(2)}$, is even more robust to misspecification of the true genetic model and has consistently high power.

The allele based test, Z_A^2, also has good power to detect associations for all genetic models, but is of course only applicable when HWE holds. A standard quality-control procedure in GWA studies is to remove SNPs that depart too far from the HWE assumption, so this may not be the constraint it first seems. Zheng et al.'s tests have more power when the case and control numbers are matched by sex but there are few situations when one of these is the optimal choice of test.

If we are to promote an overall "best test" from the simulations then it would be one of Clayton's $S^{(1)}$ or $S^{(2)}$. Given the additional complexities of specifying a sensible biological model for the X chromosome, discussed in section 3.8, the robustness of $S^{(2)}$ makes it quite appealing. However, we have seen that there may be problems with the asymptotic distribution of $S^{(2)}$ when using a highly-male biased case cohort in conjunction with a common $50:50$ control cohort. We suggest further study of the distribution of $S^{(2)}$ and the investigation of alternative evaluation approaches such as permutation testing to assuage these concerns.

A typical GWAS will have a reasonable balance of males and females in both case and control cohorts, unless the disease has a sex bias (e.g. multiple sclerosis). Clayton's $S^{(1)}$ or $S^{(2)}$ statistics will allow us to confidently detect associations, across a range of genetic models, for such a balanced study.

The plots in section 4.3 can also assist in choosing which test to use for a particular experimental design. However, the best way to choose which test statistic to analyse X chromosome GWAS data with is to enter the study's design into a simulation program such as mine. The simulation code I have written can easily handle a range of experimental designs (such as an unbalanced number of cases and controls), different genetic models, and the addition of new test statistics. A researcher can then use the simulation results to select the best test statistic for their particular experiment and estimate the power they will have to detect associations in their data.

4.4.1 Extensions

Clayton (2008) only considers an additive version of $S^{(1)}$, that is with $a = (0, 1, 2)$. There appears to be no reason not to consider different choices of a_i, corresponding to a recessive or dominant version of the 1 degree of freedom test, but these have not been investigated here due to time constraints. In line with the autosomal simulation results of Lettre et al. (2007), we would expect the dominant and recessive tests to perform well for their respective genetic models but to perform poorly when the test does not reflect the true underlying genetic model.

Other tests that have not been included due to computational and time constraints include the Cochran-Mantel-Haenszel test (implemented in PLINK), the logistic regression of PLINK, and similar GLM type approaches proposed in Clayton (2008). These methods all warrant further study. Following Zheng et al. (2007) it would be nice to repeat this simulations with a variety of departures from HWE to see how this assumption effects the performance of the various tests.

The simulation study allows us to analyse the various tests in a controlled environment. In this setting the data are not subject to the noise that real data displays and so we have a fairer background to compare the tests on. While this is obviously advantageous, the simulation study has its limits. For one, the simulation cannot capture the various nuances of true biological data. The simulation makes assumptions such as HWE and the equivalence of certain genotypes that may not be true in reality.

Ultimately the best evaluation of these methods would be to compare their performance on a real data set that contains an X chromosome locus that is truly associated with a phenotype. This sort of "gold standard" data set is not available to me, however I do have access to data from the ANZgene GWAS of multiple sclerosis that may contain interesting X chromosome loci. We discuss the reasons why the X chromosome is a good candidate region for multiple sclerosis risk loci in the following case study.

Chapter 5

Case study

5.1 Multiple sclerosis and GWA studies

The data for my case study comes from the Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) who published a GWAS of multiple sclerosis (MS) in July 2009. I begin with an explanation of MS and what is currently known about the underlying genetics of it, a description of the ANZgene study and its results, and then carry out an analysis of the X chromosome data from the ANZgene study.

What is multiple sclerosis?

Multiple sclerosis (MS) is a disease that affects the central nervous system and can, to varying degrees, interfere with the transmission of nerve impulses throughout the brain, spinal cord and optic nerves. It affects an estimated 18,000 Australians and some 2,500,000 worldwide, yet we do not understand why some people are susceptible, and others are not, nor why there is such large variation in the severity of the symptoms for those affected.

Having a first-degree relative such as a parent or sibling with MS increases an individual's risk of developing the disease several-fold above the risk for the general population (source: MS Australia). Unlike "pure" genetic diseases such as cystic fibrosis and Huntington's disease, multiple sclerosis is a complex disease with both genetic and environmental risk factors. Hence there are genetic factors at play, but how many and their associated risks are still unknown.

MS can be loosely divided into two forms: relapsing remitting (RR) and primary progressive (PP), which are the mild and extreme ends of the spectrum respectively. Interestingly there is an increased risk in females (Oksenberg et al., 2008) which suggests a possible role for the X chromosome in determining susceptibility. The increased risk in females is particularly evident for the more common relapsing-remitting form, with an estimated $3.65 : 1$ female-to-male ratio for the RR form (Source: ANZ-

gene supplementary material[1] `http://www.nature.com/ng/journal/v41/n7/suppinfo/ng.396_S1.html`). For the primary-progressive form this sex ratio is reversed with PPMS being more common in males than in females. However, to date no risk-assocated loci for MS have been identified on the X chromosome. A comprehensive review of the genetics of MS can be found in Oksenberg et al. (2008).

The recent GWAS published by the Australia and New Zealand Multiple Sclerosis Genetics Consortium identified several new risk-associated loci (all located on the autosomes), as well as replicating the well-known human leukocyte antigen (HLA chromosome 6p22) association and several, more recently discovered MS loci (International Multiple Sclerosis Genetics Consortium (IMSGC), 2007). We hope that by using refined analysis methods that we may discover in the ANZgene data new MS risk loci on the X chromosome.

5.1.1 Description of the ANZgene study

The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) is a team of more than 40 investigators from 11 institutions in Australia and New Zealand. In the discovery phase of the study the group genotyped 2,000 people with MS of European-ancestry from Australia and New Zealand. These 2000 cases were genotyped on the Illumina Infinium Hap370CNV array, described in section 1.5.1. DNA for genotyping was derived from a blood or saliva sample. As a part of the study the quality of the two DNA sources were shown to be equally good (Bahlo et al., 2009).

It was decided to acquire controls from two genome-banks to maximise the acquisition of genotypes for MS cases. The control samples are all of European-ancestry and were obtained from two genome-banks in the United Kingdom and United States. For a full description of the cohorts and the quality control procedures we refer the reader to the original paper.

Since the control cohort genotypes are from genome-banks it is clear that the cases and controls have not been matched by sex. Indeed, there is quite a female-bias in the case cohort owing to the greater prevalence of MS in females than males. After quality control (QC) procedures there remain $R = 1618$ cases (407 with primary-progressive MS, 1211 with relapsing-remitting MS) and $S = 3413$ controls. Of the cases, $R_m = 445$ are male and $R_f = 1173$ are female, giving an approximate $2.64 : 1$ female-to-male ratio in the case cohort. There is a $1.27 : 1$ female-to-male ratio of primary progressive MS cases and a $3.55 : 1$ female-to-male ratio of relapsing remitting MS cases in the discovery phase of the GWAS. Of the controls, $S_m = 1478$ are female and $S_f = 1935$ are female, giving an approximate $1.31 : 1$ female-to-male ratio in the control cohort. There is a significant difference between the

[1] NB: this estimate likely suffers from a female sampling-bias but is still reflective of the true increased prevalence of RRMS in females.

proportion of samples that are female in the case and control cohorts (two sample proportion test, $p \leq 2.2 \times 10^{-16}$).

Unlike in the simulation, when using real data there is the possibility of a sample being assigned a "no-call" (NC) at a SNP. That is, we will not have genotypes for every sample at every SNP in the analysis. We simply ignore these NCs for the time being and so the test of association at each SNP will not necessarily use all $N = R+S = 5031$ samples. The "missingness" rate at a SNP is defined as the number of samples a NC divided by the total number of samples. This missingness rate is significantly higher in the X chromosome data than in the autosomal data for the ANZgene study (mean missingness rate for X chromosome = 0.0016, mean missingness rate for autosomes = 0.0012, Welch t-test of the hypothesis that the mean missingness rate is higher for the X chromosome data than for the autosomal data: $p < 2.2 \times 10^{-16}$)

We have seen in figure 4.1 that the minor allele frequencies (MAFs) of the X chromosome SNPs used in the analysis of the ANZgene data are non-uniformly distributed on the interval $[0.01, 0.5]$. The lower bound of this interval is 0.01 as all SNPs with a lower MAF are excluded from analysis during QC. The genotyping data for SNPs with a low MAF (≤ 0.01 is a common threshold) tend to be of a poorer quality and the analytical tools are not as accurate for these SNPs. Note that this is not the only reason for why a SNP may fail QC and be excluded from the final analysis.

The X chromosome analysis is conducted on a set of 7359 SNPs that passed QC out of a possible 12917 SNPs genotyped by the Illumina Hap370CNV array. The Hap370CNV array contains 361 probes in the pseudo-autosomal regions of the X chromosome but none of these passed QC in the ANZgene study, thus there is no need for a separate analysis of the PAR probes.

Like the methods for association testing on the X chromosome, the QC procedures for X chromosome data are not as refined as for their autosomal counterparts. It is therefore likely that more X chromosome SNPs than necessary are excluded from the final analysis due to "failure" in QC. Refinement of the quality-control procedures for X chromosome data would also likely lead to an increase in power to detect genotype/phenotype associations.

The original analysis of the ANZgene data used PLINK v1.02 (Purcell et al., 2007) to implement the additive version of the Cochran-Armitage trend test. This of course means that only the female samples were used in the original analysis of the X chromosome ANZgene data (see section 3.6.1). We will discuss the results of the X chromosome analysis that featured in the original paper once I have presented my own analysis, to compare the two analyses.

Other MS GWA studies

Prior to the ANZgene study there had been one other GWAS of multiple sclerosis, published by the International Multiple Sclerosis Genetics Consortium (IMSGC) (2007). The IMSGC study used quite a different experimental design to the ANZgene study, indeed quite different to most GWA studies. The samples in the IMSGC came from 931 family trios, consisting of an affected child and both parents. Many of the the findings in the IMSGC study were validated by the ANZgene study.

5.2 Analysis of the ANZgene X chromosome data

To decide which is the best test statistic to analyse the X chromosome data from the ANZgene study, I ran my simulation code with the parameters chosen to reflect the experimental design of the ANZgene study. These parameters are summarised in table 5.1.

Parameter	Value
N_R	1618
N_S	3413
Proportion of cases female	1173/1618
Proportion of controls female	1935/3413
K	1/1000
Genetic model	ADD, DOM, REC
r-value	1.5, 2.5
Minor allele frequency	$0.01 - 0.5$

Table 5.1: Parameters in effect for MS simulation

There are 3 genetic models and 2 r-values giving a total of 6 experimental designs over the range of MAFs. In each experimental design we simulate 10000 replicates for every minor allele frequency. Note that this is a female-biased study, particularly in the case cohort, and so according to the simulations in chapter 4 we expect to have somewhat reduced power to detect associations.

Type I error rates of tests

Our first step is to estimate the size of each test. As before, this is done by simulating data under the null hypothesis $\big(\boldsymbol{\lambda}_f = (1,1,1), \boldsymbol{\lambda}_m = (1,1)\ \big)$ and recording the proportion of times each test returns a p-value less than the nominal level of $\alpha = 0.05$ for each minor allele frequency. The results are presented in the box plots of figure 5.1.

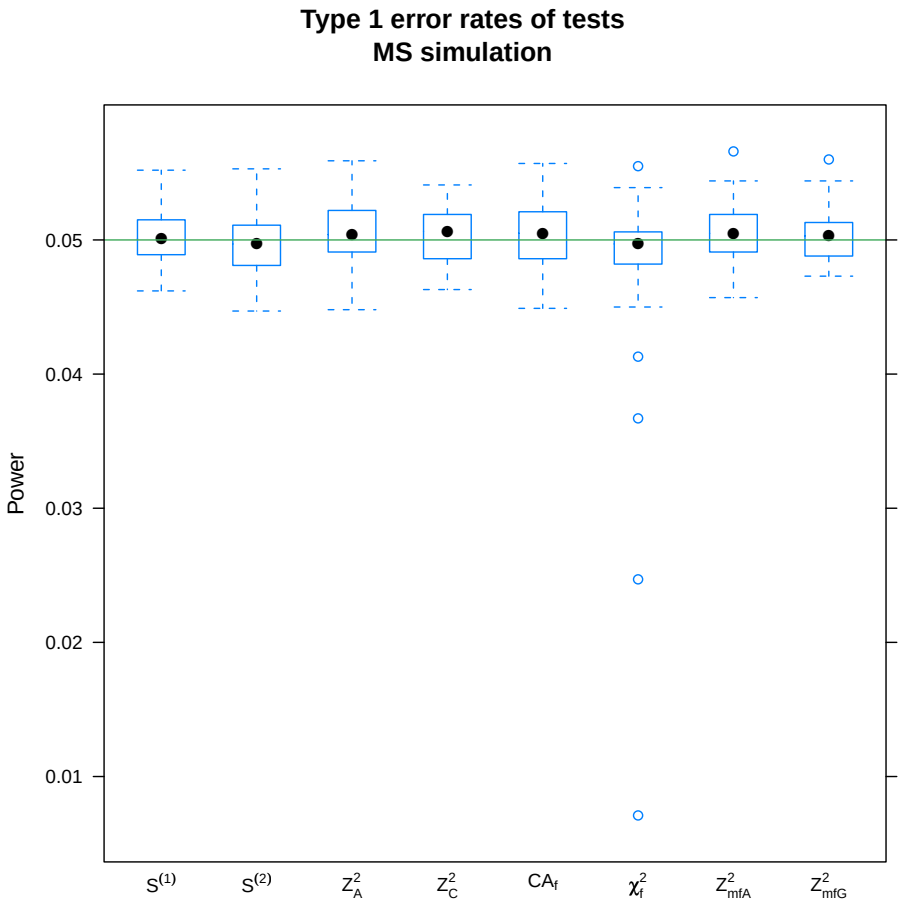

Figure 5.1: The Type I error rates of all tests are centred around the nominal level of 0.05 (green line). This gives us confidence that any of these methods will be appropriate to analyse the MS data.

Choice of test for the ANZgene X chromosome data

Since all tests are of the correct size we now examine the empirical power of each test for each genetic model. Of course we do not know the true genetic model for MS and so our choice of test will not only depend on the empirical power but also the robustness of the test to misspecification of the genetic model.

The simulation results for the ANZgene GWAS are shown in figure 5.2. In each plot the genetic model is highlighted in orange and the r-value is highlighted in green. Unlike in the previous simulation plots, the female-to-male ratios of both cohorts are now fixed. Each row corresponds to an additive, dominant or recessive model, and each column to an r-value of 1.5 or 2.5.

The simulation results do not highlight a single test statistic as being most powerful for the analysis of the ANZgene X chromosome data. The "best" test is quite dependent on the underlying genetic model. Clayton's $S^{(2)}$ is the most robust of the tests and so has an advantage over its nearest competitor, Z^2_{mfG}, since we do not know the true genetic model. There is little power, regardless of the test, to detect additive or recessive associations with low r-values.

Having studied these results I have chosen to apply the $S^{(2)}$ test to the ANZgene X chromosome data. I favour this test owing to the aforementioned robustness to the underlying genetic model. We can see from figure 5.2 that we will have the most power to detect a a genotype/phenotype association at a dominant X locus, as expected.

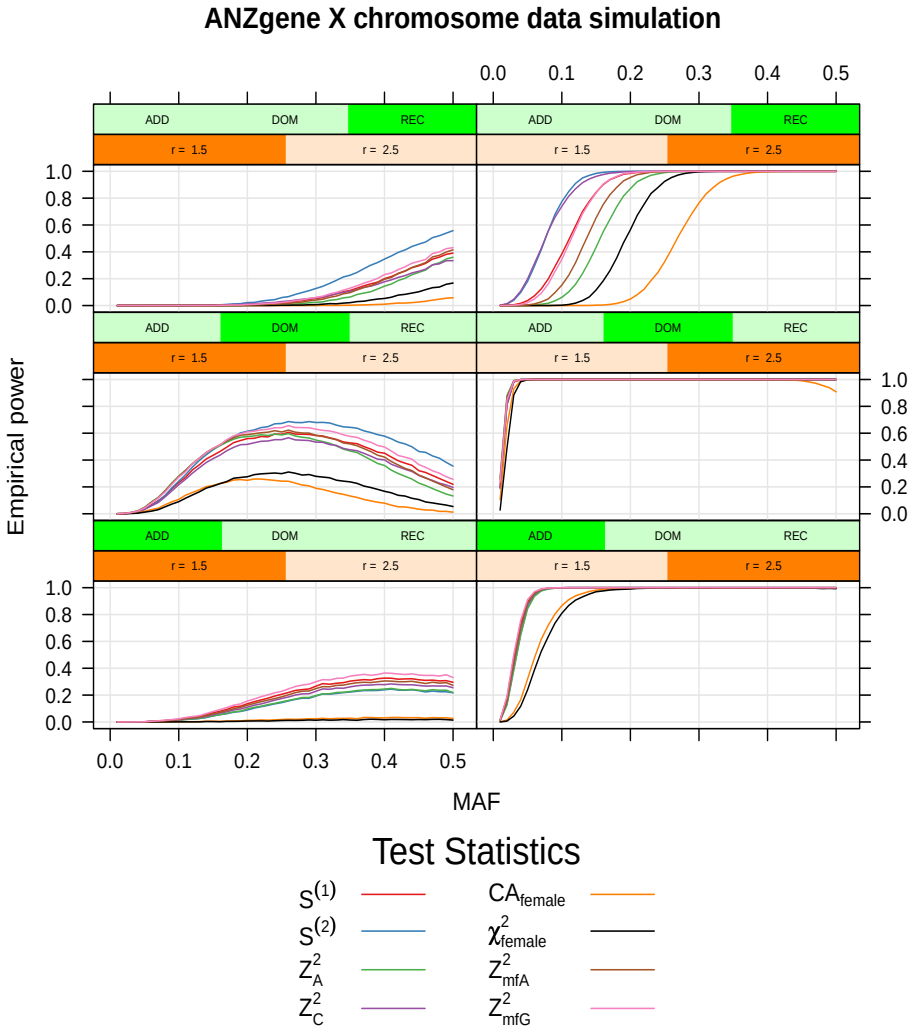

Figure 5.2: Recall that each cohort in the ANZgene study is female-biased, particularly the case cohort. As noted in figure 4.6, the Z^2_{mfG} is the most powerful test for the additive model with low r-value due to this female sampling-bias. In fact in this study Z^2_{mfG} slightly outperforms the usually favoured 1 degree of freedom test, $S^{(1)}$, across all genetic models. The 2 degree of freedom test, $S^{(2)}$, is the best test across the dominant or recessive models. This again highlights the advantage of $S^{(2)}$ as it does not rely on the specification of a genetic model like the 1 degree of freedom tests do.

5.2.1 Results

Almost all the required functions are already written as a part of my simulation study. It is therefore quite a simple process to analyse the real data. The data is in the form of a large tab-separated text file, known as a pedigree file, with each row corresponding to a sample.

The data I will use for the analysis has been through all the pre-processing and quality-control procedures, however there is still one anomaly I must remove from the data. For reasons that are not entirely clear there are 814 male genotypes that have been called as heterozygous by GenCall (the Illumina proprietary software that produces the discrete genotype calls from the raw data generated by the SNP chips) and these were not excluded in the quality control stage. Since this data does include any SNPs from the pseudo-autosomal regions these are clearly genotyping errors and must be set to "no-calls" (i.e. missing) before we can begin the analysis. The only other processing of the pedigree file that I am required to do is to convert the genotype calls to the format appropriate for my computer code.

The first six columns of the pedigree file relate to sample identification, such as sample ID, sample sex and sample phenotype. The remaining 7359 columns are each sample's genotype for the X chromosome SNPs that passed QC. An additional *map* file gives the location of each of these SNPs on the X chromosome.

The p-values reported in the association tests of a GWAS can be very small and so the p-values are typically transformed to the $-\log_{10}$ scale. These transformed p-values can be graphed as a *Manhattan plot*. A Manhattan plot consists of the set of SNPs along the x-axis, ordered by location in the genome, and for each SNP the transformed p-value, $-\log_{10}(p)$, is plotted on the y-axis. Thus, small p corresponds to large $-\log_{10}(p)$ and the plot looks somewhat like the skyscraper silhouette of a big city, hence its name.

Figure 5.3 is the Manhattan plot for the ANZgene X chromosome data when analysed with Clayton's $S^{(2)}$ statistic. The dashed-line in orange corresponds to the consensus Type I error rate of $\alpha = 10^{-8}$(NB: the original analysis used a slightly larger Type I error rate of $\alpha = 5 \times 10^{-7}$). Furthermore, the ANZgene study prioritised SNPs for the replication phase if they ranked in the top 500 SNPs by significance. This "replication threshold" for the top 500 SNPs was $p < 9.2 \times 10^{-4}$ and so we also include this line in green on the plot as well.

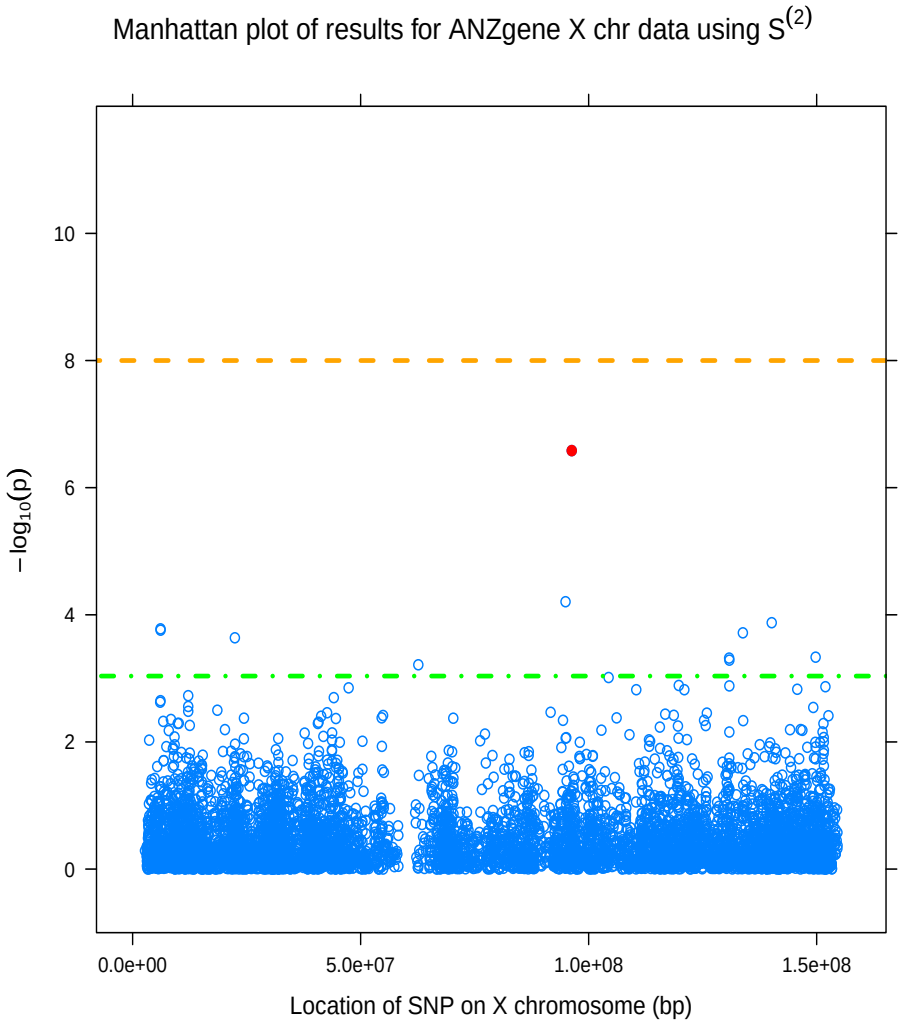

Figure 5.3: We see that no SNP achieves genome-wide significance (orange line). However, there are eleven SNPs with p-values less than the replication threshold of 9.2×10^{-4} (i.e. points above the green line) which would place these SNPs in the top 500 of the original ANZgene study. Not all these 11 SNPs are distinguishable in the plot as several lie very close to one another on X chromosome. One standout SNP (coloured in red) has $p \approx 3 \times 10^{-7}$ and will be discussed further.

The SNP highlighted in red in figure 5.3 has a strong association with case/control status. This SNP has a low minor allele frequency (MAF = 0.03) which may be causing a spurious association. We need to return to the raw genotyping data for this SNP to check the genotyping quality as well as examine the results for SNPs in linkage disequilibrium with this SNP and look at haplotypes containing this SNP to confirm whether this is a true positive. A further 10 SNPs with p-values less than the replication threshold also warrant further investigation, however we do not go into the details of this here.

The original analysis of the X chromosome data from the ANZgene study discovered 4 SNPs with p-values less than the replication threshold. Recall that this analysis was performed with the additive version of the Cochran-Armitage trend test using only the female samples. We have seen from simulations that we have greater power to detect true associations when we include the male samples in the testing procedure. This is reflected in the 11 putatively associated SNPs we have discovered using the $S^{(2)}$ statistic as opposed to the original 4 associated SNPs identified using PLINK's female-only test. Interestingly, these 4 "significant" X chromosome SNPs from the original analysis are no longer significant when we apply Clayton's $S^{(2)}$ test.

A quantile-quantile plot (q-q plot) is a nice way to visualise the increase in power when using the $S^{(2)}$, as opposed to PLINK's female-only Cochran-Armitage trend test, to analyse the ANZgene X chromosome data. In figure 5.4 are the qq-plots of the $-\log_{10}$(p-values) for the two statistics overlaid on a common figure.

Under the null hypothesis we would expect the p-values to be uniformly distributed on the set $\mathcal{U} = \{\frac{1}{n}, \ldots, \frac{n}{n}\}$, where $n = 7359$ is the number of X chromosome SNPs in the analysis. This null distribution is represented in figure 5.4 on the $-\log_{10}$ scale by the line $y = x$. In this figure, points above the $y = x$ line correspond to observed p-values that are smaller than we would expect by chance.

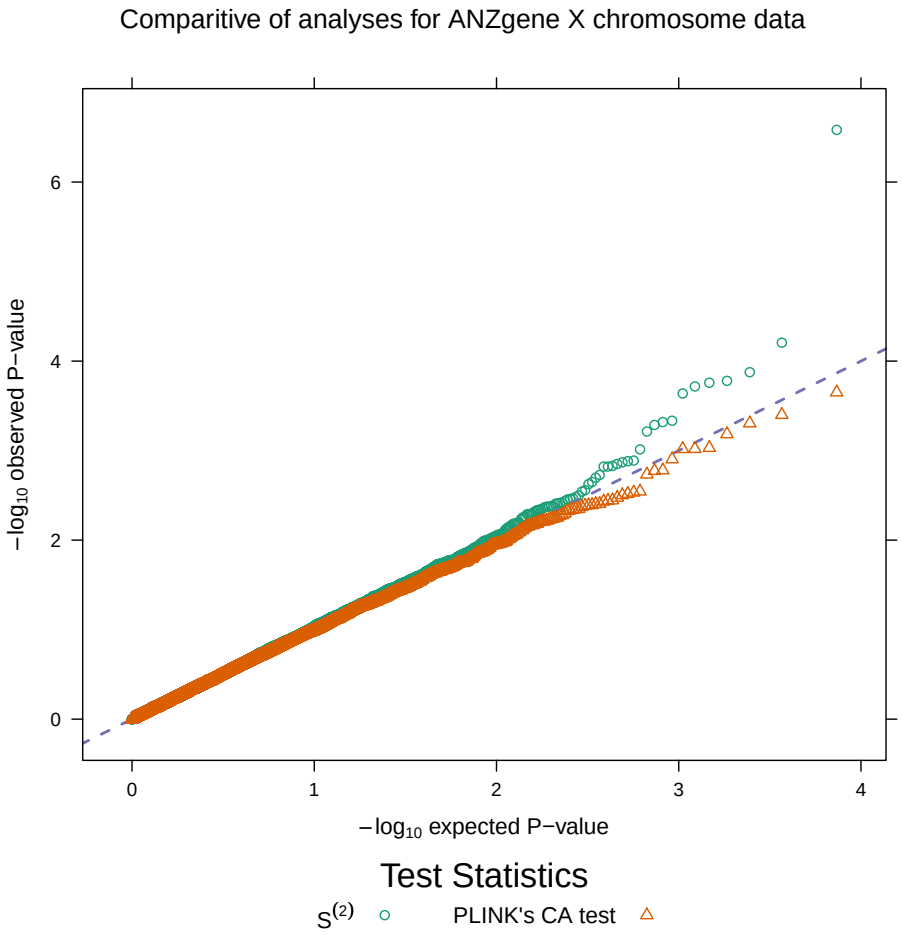

Figure 5.4: This q-q plot is indicative of the higher power we have to detect associations when using the $S^{(2)}$ test as opposed to PLINK's female-only test. We see that the smallest observed p-values of the $S^{(2)}$ statistic lie well above the line $y = x$, representing the expected p-values under the null distribution. This means that we would not expect these observed values by chance alone, and so suspect these results to be underlying true associations. In contrast, the smallest observed p-values of the Cochran-Armitage statistic in fact lie below the $y = x$ line. This means that even the smallest observed p-values from PLINK's test are not more extreme than what we would expect by chance, and so we suspect these results of being false positives.

Discussion

The computer program written for the simulation study, and the results of chapter 4, have helped us to select the best test to analyse the ANZgene X chromosome data. The chosen test, $S^{(2)}$, has identified 11 SNPs with p-values that, under the conditions of the original study, would have promoted these SNPs to the replication phase of the GWAS.

The results for these 11 SNPs will be studied in greater detail to determine their appropriateness for follow-up biological studies, but this work is beyond the scope of my thesis. For now, we note that 3 of these 11 SNPs were also identified as having p-values less than the replication threshold when the data was analysed with the Z^2_{mfG} statistic (results not shown). We have seen from the simulations of section 5.2 that the Z^2_{mfG} test is typically less powerful than the $S^{(2)}$ test for this data, and so this gives us confidence that these 3 associations in particular are worthy of further study.

We saw in chapter 4 that by discarding male samples we are losing considerable power to identify risk-associated X chromosome loci. The results of the case study reiterate this point — by discarding males from the analysis we may well be missing out on identifying X chromosome loci of interest.

Chapter 6

Discussion

Genome-wide association studies have improved our understanding of many complex diseases such as type 1 and type 2 diabetes, inflammatory bowel disease, prostate cancer and breast cancer (McCarthy et al., 2008). These studies require the investment of substantial time, effort and money — each SNP chip costs approximately $350 and then there are the associated costs of subject recruitment, sample collection and analysis. We would like to extract as much information as possible from this precious data. It is therefore surprising that the considerable amount of data from the X chromosome is often analysed in an ad-hoc and less than ideal manner.

It frequently occurs that separate, "rival" GWA studies of the same disease are running concurrently and so the pressure to publish first is high. As a result of this competitiveness, most studies aim to identify an exciting result quickly (which will normally be located on the autosomes — the so-called "low hanging fruit") and to publish their result. The more subtle associations, located in regions that are harder to analyse such as the X chromosome, are pushed aside for later analysis which often never happens. An example of this neglect can be seen in Klein et al. (2005) who only analysed the autosomal regions, despite having genotyped 2334 X chromosome SNPs[1].

This neglect is self-perpetuating with "the problem of testing for genotype-phenotype association loci on the X chromosome ... receiving surprisingly little attention" (Clayton, 2008). That is to say, research has not always made use of X chromosome data and so there is little development in the methodology for such an analysis, while simultaneously *because* there is little developed methodology, research continues to not make full use of X chromosome data.

The challenges of X chromosome association testing are due to the many biological features and processes that make the X chromosome unique in

[1]Klein et al. identified a SNP on chromosome 1 that is highly associated with age-related macular degeneration (odds ratio = 7.4)

the human genome. The autosomes, which make up 22 of the 24 human chromosomes, can all be analysed in the same manner regardless of whether the sample is male or female. The intricacies of the X chromosome also make it challenging to construct statistical models that are interpretable and meaningful in a biological context.

Despite these challenges I have shown in my thesis that we can do considerably better than many of the current widely-used methods. In particular we have seen that the discarding of male samples (à la PLINK) is a wasteful strategy that leads to a large drop in power to detect genotype/phenotype associations. In fact, the higher the proportion of cases that are male the more power we will have. The loss in power for using female-only tests will be particularly evident for male-biased diseases such as autism, but is still noticeable even in a female-biased disease such as multiple sclerosis.

We have shown via the simulation study of chapter 4 that the methods proposed in Clayton (2008) are amongst the best for X chromosome analysis. My results supplement Clayton's work as I am not aware of any published data that has compared Clayton's proposed test statistics with others that are in wide-spread use. The results of the simulation study also allow us to estimate the power we have to detect X chromosome associations under a wide variety of experimental designs and genetic models, as well as estimate the Type I error rates of each test. We have seen that for strongly sex-biased case cohorts that the $S^{(2)}$ test is conservative, i.e. the χ^2 distribution on 2 degrees of freedom is not as good an approximation for these cohorts. If the user is overly worried by this then an alternative evaluation procedure such as permutation testing could be applied.

All the testing methods considered here rely on large sample approximations, and this is standard practice for GWA studies. For the most part these approximations will be good, but there are certainly situations when so-called exact methods would be an improvement. There is a large amount of literature on exact-methods for contingency tables (see Agresti, 1992, for a comprehensive review) yet these methods have not taken hold in GWA studies. This is likely to be because these methods are both computationally and theoretically more demanding, particularly when compared to something as simple as Pearson's χ^2 test. Nevertheless, there may well be benefits to using these exact-methods and I would expect these to become more popular as computational power increases. One possible approach would be to apply the approximate methods when the data is known to satisfy the χ^2 approximation (typically for SNPs with a MAF ≥ 0.05), and to apply exact-tests for the situations where the approximate methods are known to perform poorly (e.g. for SNPs with a MAF ≤ 0.05).

A somewhat simpler addition to the testing procedure would be the development of a MAX test (Podgot et al., 1996; Freidlin et al., 2002) for X chromosome data. A MAX test takes the most significant test at each SNP (e.g. the maximum value of the additive test statistic, the dominant test

statistic and the recessive test statistic) and then adjusts the p-value for the multiple models tested. This method is rapidly gaining favour for autosomal association testing but, like most GWAS techniques, an equivalent X chromosome procedure lags behind.

Further instances where the X chromosome "lags" behind the autosomes are in the quality control procedures and genotype calling. We saw for the ANZgene data that the missingness rate for the X chromosome was significantly higher than for the autosomes ($p < 2.2 \times 10^{-16}$) and only 57% of X chromosome SNPs passed QC. Development of more specific quality control and genotyping procedures for X chromosome data will lead to an immediate increase in power since we will have a greater number of higher quality observations.

Somewhat surprisingly the simulation results have shown us that the power to detect associations is not greatly affected by whether a common 50 : 50 control cohort is used or whether the case and control numbers are matched by sex. This further validates genome-bank control cohorts as a viable alternative to the costly process of genotyping thousands of control samples. This allows researchers to focus their attention on obtaining as many case samples as possible which will increase the power of the study.

Chapters 4 and 5 highlight the benefits of using a simulation study to choose the most powerful test statistic to analyse real data. By using the simulation program we were able to conclude that Clayton's $S^{(2)}$ statistic was the best choice of test for analysing the X chromosome data from the ANZgene multiple sclerosis GWAS. The subsequent analysis identified 11 SNPs that in the original study would have appeared in the top 500 associated SNPs and thus been prioritised for the replication phase of the study. These 11 SNPs will now be studied in further detail to determine their appropriateness for follow-up biological studies to investigate their role, if any, in multiple sclerosis.

A further analysis of the ANZgene data will be to try to identify loci associated with the severity of MS. The ANZgene data contains both primary-progressive MS and relapsing-remitting MS cases, and given the different sex-ratios between the two forms (see section 5.1) the X chromosome is again an obvious candidate region for SNPs that influence the severity of the disease.

There are a number of possible extensions to my work aside from those discussed in section 4.4.1 of the simulation study. The most challenging extension is to develop both simulation and analytical methods that more accurately reflect the complex biological processes of the X chromosome. The current standard assumptions, such as homogeneity of the X chromosome inactivation process, are likely to be overly simplistic. There are also further biological intricacies of the X chromosome that are not addressed by the current analysis methods, such as imprinted genes (imprinted genes are not inherited in the classical Mendelian manner).

Current autosomal GWAS analyses typically also do not incorporate any existing biological knowledge but for the X chromosome this type of information is likely to make a much larger difference in the power of certain statistics. Individualised tests that make use of existing knowledge of the biology underlying each X chromosome SNP would be the most powerful approach, but such an approach is very difficult and some time away from being implemented.

All the methods considered here are single SNP analyses. One could also consider haplotype tests and multi-locus methods to investigate interactions between loci, though these are obviously more challenging than single marker analyses.

In this thesis we have considered many of the challenges of GWA studies, in particular those provided by the X chromosome. This work contains a number of results that show that we can do better than we currently are when it comes to X chromosome association testing. Furthermore, we have given several specific implementations of these improvements and shown in a simulation study that these significantly increase our power to detect genotype/phenotype associations. By applying these methods to real X chromosome data from the ANZgene multiple sclerosis GWAS we have identified 11 additional association signals that merit follow-up, above what had been previously identified.

Appendix A

Appendix

Remark A.1. *The allele based test, X_A^2, is simply the Pearson's χ^2 test, X^2, of the 2×2 allele table.*

Proof. We use the special form of X^2 for 2×2 tables,

$$X^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

For the allele table

$$\begin{aligned} n &= 2N, \\ a &= 2r_0 + r_1, \\ b &= 2r_2 + r_1, \\ c &= 2s_0 + s_1, \\ d &= 2s_2 + s_1, \end{aligned}$$

and

$$\begin{aligned} a + b &= 2R, \\ c + d &= 2S, \\ a + c &= 2n_0 + n_1, \\ b + d &= 2n_2 + n_1 \end{aligned}$$

giving

$$X^2 = \frac{2N\Big[(2r_0 + r_1)(2s_2 + s_1) - (2s_0 + s_1)(2r_2 + r_1)\Big]^2}{2R \cdot 2S \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)}. \tag{A.1}$$

Now

$$\begin{aligned} (2n_0 + n_1)(2n_2 + n_1) &= \big(2N - [2n_2 + n_1]\big)\big(2n_2 + n_1\big) \\ &= 2N(2n_2 + n_1) - (2n_2 + n_1)^2 \end{aligned} \tag{A.2}$$

and similarly the term

$$\Big[(2r_0+r_1)(2s_2+s_1)-(2s_0+s_1)(2r_2+r_1)\Big]^2$$

can be rewritten as

$$\begin{aligned}
&= \Big[\big(2R-\{2r_2+r_1\}\big)\big(2s_2+s_1\big)-\big(2S-\{2s_2+s_1\}\big)\big(2r_2+r_1\big)\Big]^2\\
&= \Big[2R(2s_2+s_1)-(2r_2+r_1)(2s_2+2s_1)-2S(2r_2+r_1)+(2s_2+s_1)(2r_2+r_1)\Big]^2\\
&= \Big[2R(2s_2+s_1)-2S(2r_2+r_1)\Big]^2\\
&= \Big[2R\big(2n_2+n_1-\{2r_2+r_2\}\big)-2S\big(2r_2+r_1\big)\Big]^2\\
&= \Big(2r_2+r_1)(-2R-2S)+2R(2n_2+n_1)\Big]^2\\
&= \Big[-2N(2r_2+r_1)+2R(2n_2+n_1)\Big]^2\\
&= \Big[2N(2r_2+r_1)-2R(2n_2+n_1)\Big]^2 \qquad \text{(A.3)}
\end{aligned}$$

Substituting (A.2) and (A.3) into (A.1) gives

$$\begin{aligned}
X^2 &= \frac{2N\{2N(r_1+2r_2)-2R(n_1+2n_2)\}^2}{(2R)2(N-R)\{2N(n_1+2n_2)-(n_1+2n_2)^2\}}\\
&= X_A^2.
\end{aligned}$$

Thus the ABT, X_A^2, has an approximate χ^2 distribution on 1 degree of freedom. □

Remark A.2. *Under Zheng et al.'s simulation methodology the distribution of male genotypes depends on the genetic model through the female heterozygous relative risk, λ_1.*

Proof. Table A.1 recalls the X chromosome genotypic relative risks (GRRs) for males and females under the additive, dominant, and recessive genetic models. We see that the only difference between the GRRs for each genetic model is the value of λ_1. Therefore, for fixed r, the specification of a λ_1 is equivalent to the specification of the genetic model.

The female genotype probabilities are defined by

$$\begin{aligned}
p^{(f)} &= \left(\frac{f_0^{(f)}g_0^{(f)}}{\sum_{i=0}^2 f_i^{(f)}g_i^{(f)}}, \frac{f_1^{(f)}g_1^{(f)}}{\sum_{i=0}^2 f_i^{(f)}g_i^{(f)}}, \frac{f_2^{(f)}g_2^{(f)}}{\sum_{i=0}^2 f_i^{(f)}g_i^{(f)}}\right) \qquad \text{(A.4)}\\
q^{(f)} &= \left(\frac{(1-f_0^{(f)})g_0^{(f)}}{\sum_{i=0}^2 (1-f_i^{(f)})g_i^{(f)}}, \frac{(1-f_1^{(f)})g_1^{(f)}}{\sum_{i=0}^2 (1-f_i^{(f)})g_i^{(f)}}, \frac{(1-f_2^{(f)})g_2^{(f)}}{\sum_{i=0}^2 (1-f_i^{(f)})g_i^{(f)}}\right)
\end{aligned}$$

Model	$\boldsymbol{\lambda}_f = (\lambda_0, \lambda_1, \lambda_2)$	$\boldsymbol{\lambda}_m = (\lambda_0, \lambda_2)$
Dominant	$(1, r, r)$	$(1, r)$
Recessive	$(1, 1, r)$	$(1, r)$
Additive	$(1, \frac{r+1}{2}, r)$	$(1, r)$

Table A.1: Genotypic relative risks for the X chromosome in females and males under the three classical genetic models.

with

$$\begin{aligned} f_0^{(f)} &= \frac{K}{g_2^{(f)}\lambda_2 + g_1^{(f)}\lambda_1 + g_0^{(f)}}, \\ f_1^{(f)} &= f_0^{(f)}\lambda_1, \\ f_2^{(f)} &= f_0^{(f)}\lambda_2. \end{aligned} \tag{A.5}$$

From (A.4) we see that the female heterozygous genotype probabilities, $p_1^{(f)}$ and $q_1^{(f)}$, depend on λ_1 through the definition of $f_1^{(f)}$ in (A.5).

Under Zheng et al.'s simulation methodology the male genotype probabilities are then defined relative to the female genotype probabilities by

$$\begin{aligned} p^{(m)} &= \left(p_0^{(f)} + \frac{p_1^{(f)}}{2}, p_2^{(f)} + \frac{p_1^{(f)}}{2}\right) \\ q^{(m)} &= \left(q_0^{(f)} + \frac{q_1^{(f)}}{2}, q_2^{(f)} + \frac{q_1^{(f)}}{2}\right). \end{aligned}$$

The male genotype probabilities depend on the female heterozygous genotype probabilities, i.e. the male genotype probabilities depend on λ_1. Since the specification of λ_1 is equivalent to the specification of the genetic model this explains why the distribution of male genotypes depends on the genetic model in Zheng et al.'s simulation. □

Bibliography

Agresti, A. (1992). A survey of exact inference for contingency tables, *Statist. Sci.* **7**(1): 131–153.

Altshuler, D., Daly, M. J. and Lander, E. S. (2008). Genetic mapping in human disease, *Science* **322**(5903): 881–8.

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

Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20, *Nat Genet* **41**(7): 824–8.

Bahlo, M., Stankovich, J., Danoy, P., Hickey, P. F., Taylor, B. V., Browning, S. R., The Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Brown, M. A. and Rubio, J. P. (2009). Saliva derived DNA performs well in large-scale, high-density SNP microarray studies, *Manuscript in preparation* .

Balding, D. J. (2006). A tutorial on statistical methods for population association studies, *Nat Rev Genet* **7**(10): 781–91.

Carrel, L. and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females, *Nature* **434**(7031): 400–4.

Carvalho, B., Bengtsson, H., Speed, T. P. and Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data, *Biostatistics (Oxford, England)* **8**(2): 485–99.

Chow, J. and Heard, E. (2009). X inactivation and the complexities of silencing a sex chromosome, *Curr Opin Cell Biol* **21**(3): 359–66.

Clayton, D. (2008). Testing for association on the X chromosome, *Biostatistics* **9**(4): 593–600.

Clayton, D. and Leung, H.-T. (2009). *snpMatrix: The snp.matrix and X.snp.matrix classes*. R package version 1.8.0.
URL: *http://www-gene.cimr.cam.ac.uk/clayton/software/*

Cochran, W. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**(4): 417–451.

Cordell, H. J. and Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes, *Am J Hum Genet* **70**(1): 124–41.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, second edn, Chapman and Hall.

Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans, *Genetic Epidemiology* **32**(3): 227–34.

Flaquer, A., Rappold, G., Wienker, T. and Fischer, C. (2008). The human pseudoautosomal regions: a review for genetic epidemiologists, *European Journal of Human Genetics* .

Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Hum Hered* **53**(3): 146–52.

Guedj, M., Nuel, G. and Prum, B. (2008). A note on allelic tests in case-control association studies, *Ann Hum Genet* **72**(3): 407–409.

International HapMap Consortium (2003). The International HapMap Project, *Nature* **426**(6968): 789–96.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome, *Nature* **409**(6822): 860–921.

International Multiple Sclerosis Genetics Consortium (IMSGC) (2007). Risk alleles for multiple sclerosis identified by a genomewide study, *N Engl J Med* **357**(9): 851–62.

Joo, J., Kwak, M. and Zheng, G. (2009). Improving power for testing genetic association in case-control studies by reducing the alternative space, *Biometrics* (Epub ahead of print, 13 Apr 2009).

Kahvejian, A., Quackenbush, J. and Thompson, J. F. (2008). What would you do if you could sequence everything?, *Nat Biotech* **26**(10): 1125–33.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor h polymorphism in age-related macular degeneration, *Science* **308**(5720): 385–9.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*, 2nd edn, Springer.

Lettre, G., Lange, C. and Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits, *Genetic Epidemiology* **31**(4): 358–62.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease, *Nature Genetics* **39**(7 Suppl): S37–42.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat Rev Genet* **9**(5): 356–69.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, second edn, Chapman & Hall/CRC.

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

Oksenberg, J. R., Baranzini, S. E., Sawcer, S. and Hauser, S. L. (2008). The genetics of multiple sclerosis: Snps to pathways to pathogenesis, *Nat Rev Genet* **9**(7): 516–26.

Podgot, M., Gastwirth, J. and Mehta, C. (1996). Efficiency robust tests of independence in contingency tables with ordered classifications, *Statistics in Medicine* **15**(19): 2095–2105.

Purcell, S. (2009). *PLINK V1.05 manual*.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet* **81**(3): 559–75.
URL: *http://pngu.mgh.harvard.edu/purcell/plink/*

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Ritchie, M., Carvalho, B., Hetrick, K., Tavare, S. and Irizarry, R. (2009). R/bioconductor software for illumina's infinium whole-genome genotyping beadchips, *Bioinformatics* **25**(19): 2621.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Frankish, A., Lovell, F. L., Howe, K. L., Ashurst, J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., Andrews, T. D., Scott, C. E., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N. P., Hunt, S. E., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M. L., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D. A., Worley, K. C., Ainscough, R., Ambrose, K. D., Ansari-Lari, M. A., Aradhya, S., Ashwell, R. I. S., Babbage, A. K., Bagguley, C. L., Ballabio, A., Banerjee, R., Barker, G. E., Barlow, K. F., Barrett, I. P., Bates, K. N., Beare, D. M., Beasley, H., Beasley, O., Beck, A., Bethel, G., Blechschmidt, K., Brady, N., Bray-Allen, S., Bridgeman, A. M., Brown, A. J., Brown, M. J., Bonnin, D., Bruford, E. A., Buhay, C., Burch, P., Burford, D., Burgess, J., Burrill, W., Burton, J., Bye, J. M., Carder, C., Carrel, L., Chako, J., Chapman, J. C., Chavez, D., Chen, E., Chen, G., Chen, Y., Chen, Z., Chinault, C., Ciccodicola, A., Clark, S. Y., Clarke, G., Clee, C. M., Clegg, S., Clerc-Blankenburg, K., Clifford, K., Cobley, V., Cole, C. G., Conquer, J. S., Corby, N., Connor, R. E., David, R., Davies, J., Davis, C., Davis, J., Delgado, O., Deshazo, D., Dhami, P., Ding, Y., Dinh, H., Dodsworth, S., Draper, H., Dugan-Rocha, S., Dunham, A., Dunn, M., Durbin, K. J., Dutta, I., Eades, T., Ellwood, M., Emery-Cohen, A., Errington, H., Evans, K. L., Faulkner, L., Francis, F., Frankland, J., Fraser, A. E., Galgoczy, P., Gilbert, J., Gill, R., Glöckner, G., Gregory, S. G., Gribble, S., Griffiths, C., Grocock, R., Gu, Y., Gwilliam, R., Hamilton, C., Hart, E. A., Hawes, A., Heath, P. D., Heitmann, K., Hennig, S., Hernandez, J., Hinzmann, B., Ho, S., Hoffs, M., Howden, P. J., Huckle, E. J., Hume, J., Hunt, P. J., Hunt, A. R., Isherwood, J., Jacob, L., Johnson, D., Jones, S., de Jong, P. J., Joseph, S. S., Keenan, S., Kelly, S., Kershaw, J. K., Khan, Z., Kioschis, P., Klages, S., Knights, A. J., Kosiura, A., Kovar-Smith, C., Laird, G. K., Langford, C., Lawlor, S., Leversha, M., Lewis, L., Liu, W., Lloyd, C., Lloyd, D. M., Loulseged, H., Loveland, J. E., Lovell, J. D., Lozado, R., Lu, J., Lyne, R., Ma, J., Maheshwari, M., Matthews, L. H., McDowall, J., McLaren, S., McMurray, A., Meidl, P., Meitinger, T., Milne, S., Miner, G., Mistry, S. L., Morgan, M., Morris, S., Müller, I., Mullikin, J. C., Nguyen, N., Nordsiek, G., Nyakatura, G., O'Dell, C. N., Okwuonu, G., Palmer, S., Pandian, R., Parker, D., Parrish, J., Pasternak, S., Patel, D., Pearce, A. V., Pearson, D. M., Pelan, S. E., Perez, L., Porter, K. M., Ramsey, Y., Reichwald, K., Rhodes, S., Ridler, K. A., Schlessinger, D., Schueler, M. G., Sehra, H. K., Shaw-Smith, C., Shen, H., Sheridan, E. M., Shownkeen, R.,

Skuce, C. D., Smith, M. L., Sotheran, E. C., Steingruber, H. E., Steward, C. A., Storey, R., Swann, R. M., Swarbreck, D., Tabor, P. E., Taudien, S., Taylor, T., Teague, B., Thomas, K., Thorpe, A., Timms, K., Tracey, A., Trevanion, S., Tromans, A. C., d'Urso, M., Verduzco, D., Villasana, D., Waldron, L., Wall, M., Wang, Q., Warren, J., Warry, G. L., Wei, X., West, A., Whitehead, S. L., Whiteley, M. N., Wilkinson, J. E., Willey, D. L., Williams, G., Williams, L., Williamson, A., Williamson, H., Wilming, L., Woodmansey, R. L., Wray, P. W., Yen, J., Zhang, J., Zhou, J., Zoghbi, H., Zorilla, S., Buck, D., Reinhardt, R., Poustka, A., Rosenthal, A., Lehrach, H., Meindl, A., Minx, P. J., Hillier, L. W., Willard, H. F., Wilson, R. K., Waterston, R. H., Rice, C. M., Vaudin, M., Coulson, A., Nelson, D. L., Weinstock, G., Sulston, J. E., Durbin, R., Hubbard, T., Gibbs, R. A., Beck, S., Rogers, J. and Bentley, D. R. (2005). The DNA sequence of the human X chromosome, *Nature* **434**(7031): 325–37.

Sarkar, D. (2009). *lattice: Lattice Graphics v.17-26*. R package version 0.17-22.
URL: *http://CRAN.R-project.org/package=lattice*

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size, *Biometrics* **53**(4): 1253–61.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature* **445**(7130): 881–5.

Slager, S. L. and Schaid, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend, *Hum Hered* **52**(3): 149–53.

Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic, *in* D. R. Goldstein (ed.), *Science and Statistics: A Festschrift for Terry Speed*, Vol. 40 of *IMS Lecture Notes—Monograph Series*, Institute of Mathematical Statistics, Beachwood, OH, pp. 115–126.

Suzuki, D., Griffiths, A., Miller, J. and Lewontin, R. (1989). *An Introduction to Genetic Analysis*, fourth edn, W.H. Freeman and Company.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature* **447**(7145): 661–78.

Zheng, G. (2008). Can the allelic test be retired from analysis of case-control association studies?, *Ann Hum Genet* **72**(6): 848–851.

Zheng, G., Joo, J. and Yang, Y. (2009). Pearson's test, trend test, and MAX are all trend tests with different types of scores, *Ann Hum Genet* **73**(2): 133–140.

Zheng, G., Joo, J., Zhang, C. and Geller, N. L. (2007). Testing association for markers on the X chromosome, *Genetic Epidemiology* **31**(8): 834–43.