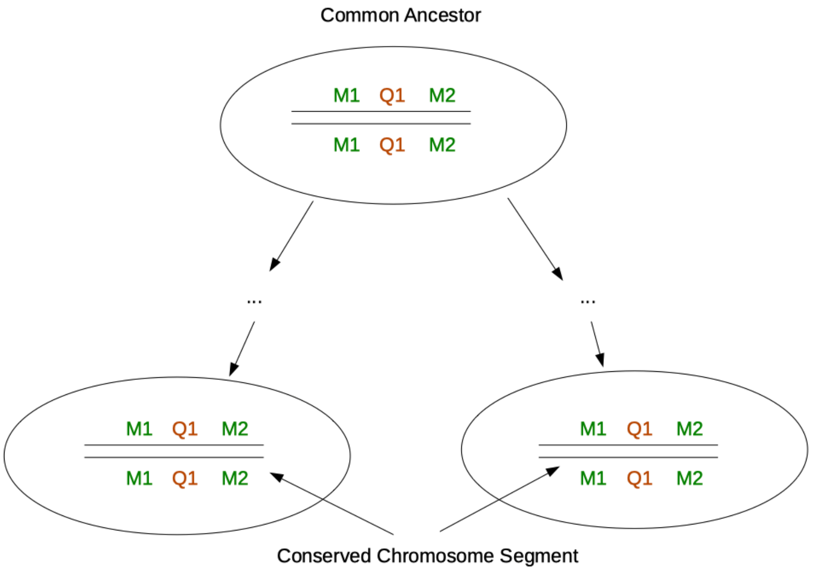
Chapter 11

Genome-Wide Association Studies (GWAS)

This chapter is based on chapter 6 of (Gondro et al., 2013). As such it provides a summary of some of the statistical methods used for genome-wide association studies (GWAS).

11.1 Single Marker Regression Tests

GWAS use linkage disequilibrium which correspond to associations of markers to causative mutations of quantiative trait loci. These associations are only expected to hold at the population level. They arise from small chromosomal segments that are inherited from a common ancestor. These chromosome segments which trace back to a common ancestor without any intervening recombination will carry identical marker alleles or marker haplotypes. If there is a QTL somewhere inside of such marker segments, they will also carry the same QTL allele. There are a number of statistical methods that use these associations to find locations of interesting QTL. A simple method is the single marker regression test.



In a random mating population without population substructures, the association between a marker and a QTL that is relevant for the expression of a phenotypic value of an economically important trait can be tested with a single marker regression as

$$y = Wb + Xg + e \tag{11.1}$$

where y is a vector of phenotypes, b is a vector of fixed effects, g is the marker effect and e is a vector of random error terms. These error terms are all identically and independently distributed with $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ where σ_e^2 corresponds to the error variance. The design matrix W links observations to fixed effects and the matrix X allocates records to the marker effect.

In this model the marker effect is treated as fixed and the model is additive which means that two copies of the same allele have twice the effect of a single marker allele and zero alleles have no effect at all. The underlying assumption is that a given marker will only affect the phenotypic observation of a trait if it is linked to an unobservable QTL.

The null hypothesis (H_0) is that the marker does not have an effect on the trait while the alternative hypothesis (H_A) is that the marker does have an effect on the trait. The null hypothesis is rejected if the test statistic F satisfies the condition $F > F_{\alpha,\nu1,\nu2}$ where $F_{\alpha,\nu1,\nu2}$ is the value of the F-distribution at significance level α and $\nu1$ and $\nu2$ degrees of freedom.

11.1.1 Example

Consider the following example dataset.

Table 11.1: Phenotypic and genotypic data for ten animals and one marker locus

Animal	Phenotype	SNP Allele 1	SNP Allele 2
1	2.03	1	1
2	3.54	1	2
3	3.83	1	2
4	4.87	2	2
5	3.41	1	2
6	2.34	1	1
7	2.65	1	1
8	3.76	1	2
9	3.69	1	2
10	3.69	1	2

We need a design matrix X to allocate both the mean and SNP alleles to phenotypes. In this case we will use an X matrix with number of rows equal to the number of observations and one column for the SNP effect. We will set the effect of the "1" allele to 0 which means that allele "2" is the allele with the positive effect on the phenotype. So the SNP effect column is the number of copies of the "2" allele. We assume a common mean μ as the only fixed effect. Hence the matrices X and W have the following structure.

$$W = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$X = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

The general mean and the SNP effect can be estimated as

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} W^TW & W^TX \\ X^TW & X^TX \end{bmatrix}^{-1} \begin{bmatrix} W^Ty \\ X^Ty \end{bmatrix} = \begin{bmatrix} 2.35 \\ 1.28 \end{bmatrix}$$

The F-value can be computed as

$$F = \frac{(n-1)(\hat{g}X^Ty - 1/ny^Ty)}{y^Ty - \hat{g}X^Ty - \hat{u}1_n^Ty} = 4.56$$

The tabulated value for $F_{0.05,1,9} = 5.12$ for a significance level $\alpha = 0.05$ and $\nu1 = 1$ and $\nu2 = 9$ degrees of freedom. Hence for this small dataset the null hypothesis of the SNP having no effect on the trait cannot be rejected.

11.2 Genome-Wide Association Experiments Using Haplotypes

Instead of using single markers, haplotypes of markers could be used in genome-wide associations. In this context, the term “haplotype” stands for a group of consecutive markers on the same chromosome. The effect of haplotypes in windows across the genome would be tested for their association with phenotype. The justification for using haplotypes is that marker haplotypes may be in greater linkage disequilibrium with the QTL alleles than single markers. If this is true, then the r^2[1] between the QTL and the haplotypes is increased, thereby increas- ing the power of the experiment.

[1] Note r^2 is defined as $r^2 = (f(A1B1)f(A2B2) - f(A1B2)f(A2B1)^2/(f(A1)f(A2)f(B1)f(B2))$ and measures how closely the two loci A and B are linked.

11.3 Fitting All Markers Simultaneously

There are two disadvantages of the approaches described above that fit single SNPs, haplotypes, or single genome regions in the analysis. One of these is the multiple testing problem, that is many thousands of tests are run, so the significance level must be very stringent to take this into account. Further, the setting of a significance threshold combined with the testing of so many marker effects means that the markers most likely to exceed the threshold are those with favorable error terms, so that the significant markers have overestimated effects. The second disadvantage, particularly of the single SNP approach, is that a region containing the true mutation can be hard to define, as a large number of SNP can be in LD with the QTL, such that significant SNP span a wide region. This is particularly problematic in livestock (and likely some plant species), as low, but non zero, LD extends for Mb. While a partial solution to this second problem is to jointly fit SNP in multiple or conditional regression, an even better solution to both these issues is to fit all SNP simultaneously. This involves fitting the same models that have been proposed for genomic prediction.

This can be achieved by fitting the SNPs as random effects (e.g., derived from a distribution), with different prior assumptions on the distribution of possible SNP effects (e.g., a Bayesian approach). The model is:

$$y = 1^T\mu + Xg + e$$

where g is now a vector of random SNP-effects. Because the above equation consists of a linear mixed-effect model, the solutions can be obtained by the well-known mixed-model equations.