Chapter 3

Haplotypes and Linkage Disequilibrium

In the previous chapter we considered the use of Single Nucleotide Polymorphism (SNPs) as which one can use in the context of the case-random experimental design. These markers were aimed at detecting functional variations associated with an increase in the susceptibility to a disease. These variations were assumed to be SNPs themselves. In this chapter we extend the investigation the joint genetic information stored in a collection of markers. Usage of this joint information may potentially increase the sensitivity of detection.

Before going into the details, let us remind ourselves what is the form of the data we may expect to have. Again, the COMT data is useful. Recall that the initial genotyping was successfully applied to 7 markers, using quantitative genotyping of pools. This, by its own nature, provide us only with information regarding the marginal frequency of alleles. The same 7 SNPs were then individually genotyped over 70 DNA samples. Three SNPs were selected, based on their association with the disease, and based on the individual genotyping outcomes. The selected SNPs went through further individual genotyping over the entire samples of both cases and random controls, to provide us with the final results that were reported in the previous chapter. In this chapter we will attempt to investigate the ways by which more information can be extracted from individual genotype data; information that can help, for example, in selection of SNPs for genotyping.

In this investigation we will assume, as typically is the case in the case-random design, that no familial information is available for the determination of the phase of the genotype, i.e. the gamete to which each of the two alleles in each marker belong. One can partially overcome this difficulty using statistical estimation. The Expectation-Maximization (EM) algorithm is a procedure for the computation of the maximum-likelihood estimation of the unknown parameter in a statistical model. The algorithm resolves lack of phase data by treating it as an unknown parameter in an appropriate statistical model. In the first
section we describe the algorithm and write a Matlab code for its application.

In the section that follows we take a short break from the main topic of this section and describe some of the commonly used parameters of LD between a pair of SNPs. These parameters are important for the planning of a population-based genetic-association study. With them, one can estimate the expected power of a trial and make an educated decision regarding the required sample size and marker density. The values of these parameters in the population can be evaluated by applying the EM algorithm to the genotypes of a sample. Our main goal in this section is the assessment of the sampling error involved in such estimation procedure. This assessment will be conducted with the aid of simulations.

Returning to the main subject-matter, we will consider the use of the joint information over several markers for detection. We will propose a test based on the haplotype data. We will also look for for an extension of the results from the previous chapter, results that relate linkage disequilibrium to the detection power of an haplotype based test.

### 3.1 Haplotypes

Genotyping an individual over several markers provides us with pair of alleles at each of the markers. One thing it does not provide us with is the phase of these alleles. Given one of the gametes, only a single allele at each marker belongs to that gamete. Such allele is physically connected to the appropriate allele in the next marker and in the rest of the markers. They are said to have the same phase. Yet, nothing in the genotyped data of an single individual can guide us in the decision which allele in a given marker is physically related to an allele on the other marker by having the same phase. In some case, this lack of information is nonsignificant. For example, if the genotype in one of two markers is homozygous we can deduce the alleles of the markers along each of the two gametes. However, if both genotypes are heterozygous then two distinct configurations of haplotypes are possible.

When provided with information from other relatives, one can frequently infer the phase status by marking as unlikely assignment of alleles to gametes that force too many crossover events to have occurred. Unfortunately, in a trial situation like the we consider here it is typically not case that such information is at hand.

A remedy to the lack in phase data can be found by treating the problem as a statistical inference problem with partially missing data and applying statistical tools that are designed for the scenario of missing data. Such a tool is the Expectation-Maximization algorithm, or the EM in short. Given a sample of genotypes, the EM algorithm computes the maximum-likelihood estimate of the relative frequency of haplotypes in the population. Consider a collection of $m$ bi-allelic markers. The total number of different gamete alleles is $2^m$. One’s aim is to assign a distribution over these $2^m$ gametes. The problem is how to a assign the distribution in such a way which is most concordant with the
3.1. HAPLOTYPES

observed sample.

By construction, the EM algorithm uses the Hardy-Weinberg assumption of gamete independence. Given this assumption, the computation of the likelihood of the sample for a given distribution of gametes in the population is carried out iteratively. Given any distribution over gametes, an iteration of the algorithm produces an updated distribution which have higher likelihood value with respect to the sample. The iteration proceeds until the algorithm converge.

In order to demonstrate the principles of the EM algorithm we first apply it to the sample of genotypes over two bi-allelic markers. For illustration, consider the sample distribution over combination of genotypes for the SNPs rs737865 and rs165688 in the for the random controls (for the 2680 individuals who had genotype values for both markers) given in Table 3.1. We can clearly determine from this data that the haplotype C-A appears in the sample at least \(2 \times 1 + 54 + 75 = 131\) times. This follows from the fact that the haplotype is present in two copies in the double-homozygous individual in the first row of the table, and is presented in one copy in the 54 individuals in the second row and in the 75 individuals in the fourth row. However, this haplotype may also be present in one copy among some of the individuals in the fifth row of double-heterozygous. What we don’t known is the percentage of such individuals. The EM algorithm treats that percentage as a parameter and tries to estimate it form the data.

<table>
<thead>
<tr>
<th>rs737865</th>
<th>rs165688</th>
<th>frequency</th>
<th>haplotype 1</th>
<th>haplotype 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/C</td>
<td>A/A</td>
<td>1</td>
<td>C-A</td>
<td>C-A</td>
</tr>
<tr>
<td>C/C</td>
<td>G/A</td>
<td>54</td>
<td>C-A</td>
<td>C-G</td>
</tr>
<tr>
<td>C/C</td>
<td>G/G</td>
<td>364</td>
<td>C-G</td>
<td>C-G</td>
</tr>
<tr>
<td>C/T</td>
<td>A/A</td>
<td>75</td>
<td>C-A</td>
<td>T-A</td>
</tr>
<tr>
<td>C/T</td>
<td>G/A</td>
<td>978</td>
<td>C-A and T-G or C-G and T-A?</td>
<td></td>
</tr>
<tr>
<td>C/T</td>
<td>G/G</td>
<td>274</td>
<td>C-G</td>
<td>T-G</td>
</tr>
<tr>
<td>T/T</td>
<td>A/A</td>
<td>508</td>
<td>T-A</td>
<td>T-A</td>
</tr>
<tr>
<td>T/T</td>
<td>G/A</td>
<td>367</td>
<td>T-A</td>
<td>T-G</td>
</tr>
<tr>
<td>T/T</td>
<td>G/G</td>
<td>59</td>
<td>T-G</td>
<td>T-G</td>
</tr>
</tbody>
</table>

Table 3.1: Joint distribution of control genotypes for a pair of markers in the COMT gene

Assume \(0 \leq \theta \leq 1\) to be the ratio of double-heterozygous individuals which have the haplotype C-A (and also T-G). The frequency of the haplotype C-A would be then \(131 + \theta \times 978\). In a similar way we can carry out the computation of the relative frequency of the other 3 haplotypes. The results are presented in Table 3.2.

Having the frequencies of Table 3.2, we can turn back and reevaluate the expected percentage \(\theta\) of (C-A,T-G) double-heterozygote among the double-heterozygote. Assuming the Hardy-Weinberg Equilibrium, one gets that the
### Chapter 3. Haplotypes and Linkage Disequilibrium

The frequency in the sample of the four haplotypes, given the percentage $\theta$ of (C-A,T-G) double-heterozygote.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-A</td>
<td>$131 + \theta \times 978$</td>
</tr>
<tr>
<td>C-G</td>
<td>$1056 + (1 - \theta) \times 978$</td>
</tr>
<tr>
<td>T-A</td>
<td>$1458 + (1 - \theta) \times 978$</td>
</tr>
<tr>
<td>T-G</td>
<td>$759 + \theta \times 978$</td>
</tr>
</tbody>
</table>

Table 3.2: The frequency in the sample of the four haplotypes, given the percentage $\theta$ of (C-A,T-G) double-heterozygote.

The probability of obtaining a double-heterozygote is:

$$\frac{131 + \theta \cdot 978}{5360} \times \frac{759 + \theta \cdot 978}{5360} + \frac{1056 + (1 - \theta) \cdot 978}{5360} \times \frac{1458 + (1 - \theta) \cdot 978}{5360}.$$

The probability of obtaining a (C-A,T-G) double-heterozygote, on the other hand, is:

$$\frac{131 + \theta \cdot 978}{5360} \times \frac{759 + \theta \cdot 978}{5360}.$$

It turns out that the expected relative frequency of (C-A,T-G) double-heterozygote among the set of all double-heterozygote, is equal to:

$$\frac{(131 + \theta \cdot 978) \times (759 + \theta \cdot 978)}{(131 + \theta \cdot 978) \times (759 + \theta \cdot 978) + (1056 + (1 - \theta) \cdot 978) \times (1458 + (1 - \theta) \cdot 978)}.$$

Replacing the existing value of $\theta$ by this update completes an iteration. Applying the iteration several times will lead to a convergence of $\theta$ to a fixed-point. This fixed-point, once plugged into Table 3.2, provides the EM estimate of the haplotype sample frequencies.

The function `EM2.m` calculates the expected haplotype frequencies from genotype data. Since we intend to run simulations as part of the investigation of the statistical properties of the EM algorithm, this function was written so it can apply the algorithm in parallel:

```matlab
function H = EM2(G)
    MAXITER = 100;
    EPS = 0.001;
    % EM2 compute the haplotype frequency from the genotype data of 2 markers.
    % Input: G = a (9 by repeats) matrix of genotype frequencies, ordered by
    % the genotype of the first marker and sub-ordered by the by the second
    % marker
    % Output: H = a (4 by repeats) matrix of haplotype frequencies. The order
    % of the haplotypes is [1-1; 1-2; 2-1; 2-2].
```
3.2. PARAMETERS OF LINKAGE DISEQUILIBRIUM

\[
H_0 = \begin{bmatrix}
2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 2
\end{bmatrix};
\]

\[H_0 = H_0 \cdot G;\]
\[U = G(5,:);\]
\[th = 0.5 \cdot \text{ones}(1, \text{size}(G, 2));\]
\[th_0 = \text{zeros}(1, \text{size}(G, 2));\]
\[\text{iter} = 0;\]

```matlab
while (\text{iter} <= \text{MAXITER} & \text{max(abs(th-th0))} >= \text{EPS})
    th0 = th;
    th = (H0(1,:) + th.*U).*((H0(4,:) + th.*U)) ./ ...
        ((H0(1,:) + th.*U).*((H0(4,:) + th.*U)) + ...
        (H0(2,:) + (1-th).*U).*((H0(3,:) + (1-th).*U)));
    \text{iter} = \text{iter} + 1;
end
H = H0 + [th.*U; (1-th).*U; (1-th).*U; th.*U];
```

Applying the algorithm to the genotype data of the two SNPs under consideration yields:

```matlab
>> G = [1 54 364 75 978 274 508 367 59]'
G =
  1
  54
  364
  75
  978
  274
  508
  367
  59
>> EM2(G)
ans =
  1.0e+003 *
      0.1598
      2.0052
      2.0802
      0.7878
```

3.2 Parameters of linkage disequilibrium

3.3 QTL detection using haplotypes

For a haplotype, composed of \( m \) markers, consider the \( 2^m \times 2 \) table 3.3 describing the joint distribution of alleles of the functional polymorphism and the haplotypes in a random population:

\[
\begin{array}{ccc}
H_{111...11} & \pi_{111...11,1} & \pi_{111...11,2} & \pi_{111...11,..} \\
H_{111...12} & \pi_{111...12,1} & \pi_{111...12,2} & \pi_{111...12,..} \\
H_{111...21} & \pi_{111...21,1} & \pi_{111...21,2} & \pi_{111...21,..} \\
H_{111...22} & \pi_{111...22,1} & \pi_{111...22,2} & \pi_{111...22,..} \\
... & ... & ... & ... \\
H_{222...22} & \pi_{222...22,1} & \pi_{222...22,2} & \pi_{222...22,..} \\
\pi_1 & \pi_2 & 1
\end{array}
\]

Table 3.3: Joint distribution of alleles

Here \( M_1 \) and \( M_2 \) are the two alleles of the anonymous marker \( M \). The entries \( \pi_{ij} \) are the relative frequencies of gametes with allele \( M_i \) at locus \( M \) and allele \( D_j \) at locus \( D \). We use the dot notation for marginal frequencies. Hence, for example, \( \pi_1 = \pi_{11} + \pi_{12} \) is the marginal probability of sampling the allele \( M_1 \). The correlation parameter \( R^2 \) between the two loci is defined to be:

\[
R^2 = \frac{(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})^2}{\pi_1 \cdot \pi_2 \cdot \pi_1 \cdot \pi_2}.
\]

This parameter is the square of the standard Pearson correlation coefficient which one obtains by assigning the numerical values to the alleles. Note that the parameter \( R^2 \) takes values between zero and one. A value of one corresponds to perfect linkage. In that case the marker is equivalent to the functional locus from the statistical perspective. On the other extreme, a value of zero corresponds to perfect linkage disequilibrium (among the random controls). As we will see below, in this case the marker reflects none of the information projected from the functional site (at least to the level of the approximation of the power that we consider here). In other places we discuss the issue of the distribution of this parameter in various loci and in various populations. We also deal with the problem of estimation of its value from sampled genotype data. Here we regard only the effect of this parameter on the power to detect a susceptibility locus.

3.3.1 The non-centrality parameter of the allele test

Let us investigate the relation between the LD parameter \( R^2 \), computed between the functional polymorphism and the genotyped marker, and the statistical power of a test which evaluated the association between the phenotype and the genotype of that marker. In order to focus the investigation, we will carry it first in the context of the multiplicative model and the allelic test. We will
3.3. QTL DETECTION USING HAPLOTYPES

then proceed by extending the result to then genotype test and to more general models under local alternatives.

Under the Hardy-Weinberg equilibrium for the null distribution and the multiplicative model for penetrance, the distribution of the genotypes follows the Hardy-Weinberg assumption also under the alternative. It follows that the testing procedure may be reduced to the comparison of allelic frequency between two Bernoulli samples of gametes. We claim that the non-centrality parameter of the allelic test computed based on the genotyped data is equal in this case to the product of $R^2$ with the non-centrality parameter we would have obtained had we access to the genotype data at the functional locus. In other words, the non-centrality parameter at a marker is proportional to the non-centrality parameter at the susceptibility locus. The closer the LD parameter is to zero, the less likely it is that a functional polymorphism will be detected.

It is worthwhile to note the negating effects of the parameter $R^2$ and of the sample size $N$. Both change the value of the non-centrality parameter in a proportional manner, but they work in opposite directions. Thus, one may sustain the changes of detection by using a denser collection of markers (or a population with higher levels of linkage disequilibrium). Alternatively, one can increase the sample size. Indeed, which way to go — the way of increasing the number of markers to genotype, or the way of increasing the sample size — is a matter of availability of the different resources.

The claim of a linear relation between $R^2$ and the non-centrality parameter can be justified using the following mathematical theorem. In this theorem the difference between the frequency of an allele at the marker $M$ in the cases and the difference in the controls is given is shown to be proportional to difference of the allele frequency in the functional polymorphism $D$. This theorem is based on the notion of statistical sufficiency. Specifically, it makes use of the assumption that conditional on the allele at locus $D$, the distribution of the alleles at locus $M$ is the same for cases and random controls. In other words, all the relevant information for making inference at the given region is concentrated at locus $D$. The information at locus $M$ is merely a reflection of that information. Had we had the information at $D$, we wouldn’t have bothered to acquire the genotypic status at locus $M$, which adds no extra data.

**Lemma 3.3.1.1.** Assume that the locus $D$ is sufficient for statistical inference. Namely, the conditional distribution of the alleles at locus $M$ for the cases coincide with that of the random sample. Then

$$
P^e(M_1) - P^r(M_1) = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_1\pi_2} \times [P^e(D_1) - P^r(D_1)],$$

where $P^e(\cdot)$ ($P^r(\cdot)$) is the distribution among cases (resp. random controls).

**Proof:** Conditioning on the allele at $D$, which is a sufficient statistic, yields:

$$P^e(M_1) = P^e(M_1 | D_1)P^e(D_1) + P^e(M_1 | D_2)P^e(D_2).$$

Also,

$$P^r(M_1) = P^r(M_1 | D_1)P^r(D_1) + P^r(M_1 | D_2)P^r(D_2).$$
Consequently,

\[ P_a(M_1) - P_r(M_1) = [P_r(M_1 | D_1) - P_r(M_1 | D_2)] [P_a(D_1) - P_r(D_1)] \]

However,

\[ P_r(M_1 | D_1) - P_r(M_1 | D_2) = \frac{\pi_{11} - \pi_{12}}{\pi_1} = \frac{\pi_{11} \pi_{2} - \pi_{12} \pi_{1}}{\pi_1 \pi_2} = \frac{\pi_{11} \pi_{22} - \pi_{12} \pi_{21}}{\pi_1 \pi_2}. \]

The non-centrality parameter of the allelic statistic \( Z^2 \) at the locus \( D \) is given by:

\[ \mu^2(D) = \frac{[P_a(D_1) - P_r(D_1)]^2}{\pi_1 \pi_2} \times \frac{2 N^a N^r}{N}. \]

Similarly, the parameter of non-centrality for the test statistic computed for the marker \( M \) is given by:

\[ \mu^2(M) = \frac{[P_a(M_1) - P_r(M_1)]^2}{\pi_1 \pi_2} \times \frac{2 N^a N^r}{N}. \]

It follows directly from Lemma 3.3.1.1 that:

**Theorem 3.3.1.1.** Under the conditions of Lemma 3.3.1.1:

\[ \mu^2(M) = R^2 \times \mu^2(D). \]

The multiplicative assumption. The test statistic. Haplotypes versus joint genotypes. Parameter of non-centrality and LD. Haplotypes versus SNPs.

### 3.4 Fine mapping with haplotypes