Chapter 8. Association Analysis: Multiple Marker Methods

The association methods we discussed till now are single marker methods. It is likely that multiple markers contain more genetic information. The multiple marker methods are either considering multiple tightly linked markers in one candidate gene or considering multiple markers in different genes.

When we perform candidate gene studies, several tightly linked markers are usually typed within a candidate genes. There is strong evidence that several mutations within a single gene can interact to create a super allele that has a large effect on the observed phenotype (Schaid et al. 2002). Some examples in humans include a gene that influences intestinal lactase activity (Hollox et al. 2001); a gene responsible for human lipoprotein lipase (Clark et al. 1998); the HPC2/ELAC2 gene, which increases the risk for prostate cancer (Tavtigian et al. 2001); and a gene that influences actions of catecholamines, which influence bronchodilation and hence, asthma (Drysdale 2000). The biologic explanation for these haplotype effects is that several mutations in a gene cause several amino acid changes in the ultimate protein product, and the joint effect of these amino acid changes can have a much larger influence on the function of the protein product than any single amino acid change. This emphasizes the importance of examining candidate genes by SNP haplotype.

The methods by considering different markers in different genes are motivated by gene-gene interaction. As an example of gene-gene interaction, let's consider two biallelic markers. Table 8.1 gives the penetrances of two-locus genotypes and the penetrances of marginal one-locus genotypes for each of the markers. From the table, we can see that the genotypes aaBB, AaBb, and AAbb are high risk genotypes by considering the two loci jointly, however, we can see nothing by examining each one of the two markers.

Table 8.1 The penetrances of the two-locus genotypes and marginal genotypes (the frequency of each of the four alleles is 0.5)

	AA	Aa	aa	Marginal
BB	0	0	0.2	0.05
Bb	0	0.1	0	0.05
bb	0.2	0	0	0.05
Marginal	0.05	0.05	0.05	

Because complex traits presumably arise from multiple interacting genes located throughout the genome, it would be appropriate to search for and to analyze sets of marker loci in different genes jointly rather than to test each marker in isolation.

§ 8.1 Methods for Candidate Genes Based on Unrelated Individuals

8.1.1 Likelihood Ratio Test for Qualitative trait

Assume that we have sampled n_1 cases and n_2 controls ($n = n_1 + n_2$). Each sampled individual has genotype at m tightly linked markers within a candidate gene. Once the haplotype frequencies are estimated with the EM algorithm they can be compared between

affected cases and normal controls by use of a likelihood ratio statistic (Fallin, D. 2000, Fallin, et al. 2001, Fallin D. and Schork, N.,2000).

$$HLR = 2(\log L_{\text{cases}} + \log L_{\text{controls}} - \log L_{\text{pooled}}),$$

which has an asymptotic χ^2 distribution with degrees of freedom equal to the number of haplotypes compatible with the sample minus one. Assume that there are H possible haplotypes denoted by $h_1, \ldots, h_H$. Let $p = (p_1, \ldots, p_H)$ and $q = (q_1, \ldots, q_h)$ denote the haplotype frequencies in cases and controls. The null hypothesis of no association is $H_0 : p = q$. Let N_i, M_i, and K_i denote the number of haplotype h_i in cases, controls, and pooled sample, respectively, where N_i, M_i, and K_i need to be estimated from cases, controls, and pooled sample, respectively. The log-likelihood under null is given by

$$l(p) = \sum_{i=1}^{H} K_i \log p_i$$

and the maximum log-likelihood is given by $\log L_{pooled} = \max_{H_0} l(p) = \sum_{i=1}^{H} K_i \log(\hat{p}_i) = \sum_{i=1}^{H} 2n\hat{p}_i \log(\hat{p}_i)$, where $\hat{p}_i = \frac{K_i}{2n}$ is the MLE of p_i usually by EM algorithm. The maximum of log-likelihood under whole parameter space $\log L_{cases} + \log L_{controls}$ can be obtained in similar way.

This estimation-based likelihood ratio test is sensitive to any departure from the equality of the haplotype frequencies in cases and controls, including the possibility that more than one haplotype is associated with the disease (Fallin, D., 2000, Fallin, et al., 2001). For sparse data, empirical p-values may be more reliable than these based on the asymptotic distribution. The empirical p-values can be obtained through reshuffling the case/control status among the individuals and recalculating the haplotype frequencies and the corresponding log-likelihoods. The main disadvantage of this test is that the number of haplotypes with estimated frequencies different from zero tends to be large, which increases the number of degrees of freedom, thereby limiting the power of the test. This may happen even when linkage disequilibrium exists among markers, so the actual number of haplotypes is not large. Furthermore, this test does not provide a way of making inference on individual haplotypes and is also not easy to generalize to deal with quantitative traits.

8.1.2. Haplotype Trend Regression (Zaykin et al. 2002)

This test is applicable to both qualitative trait and quantitative trait. Consider a sample of n unrelated individuals where each individual has genotypes at several SNPs on a candidate gene. Let $H + 1$ be the number of possible haplotypes formed by the typed SNPs, and $h_1, \ldots, h_{H+1}$ denote all haplotypes. Let y_i and g_i denote the trait value and multi-marker genotype of individual i, respectively. For qualitative trait, $y_i = 1$ if the ith individual is diseased; otherwise $y_i = 0$. We code the genotype g_i through the haplotypes that are compatible with g_i and denote the numerical code of g_i by $X_i = (x_{i1}, \ldots, x_{iH})^T$. If haplotypic phases are available or in case of no ambiguity,

$$x_{ij} = \begin{cases} 2 \text{ if } g_i \text{ is homozygote for haplotype } h_j \\ 1 \text{ if } g_i \text{ is heterozygous and include haplotype } h_j \\ 0 \text{ otherwise.} \end{cases} .$$

We can use a linear model

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_H x_{iH} + \epsilon_i, \tag{8.1}$$

for a quantitative trait, and a logistic linear model

$$\log \frac{p_i}{1-p_i} = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_H x_{iH} \tag{8.2}$$

for a qualitative trait, where $p_i = \Pr(y_i = 1|g_i)$.

If the phases are unknown and can not be unambiguous reconstructed, haplotype frequencies are estimated via the EM algorithm. In this case, g_i must be heterozygous and we define the genotype score x_{ij} as the posterior probability that individual i has haplotype h_j given its genotype g_i, and x_{ij} can be written as

$$x_{ij} = p(h_j|g_i) = \delta_{ij} \frac{\Pr(h_j h_j^{com})}{\sum_k \Pr(h_{kj} h_{kj}^{com})}$$

where $\sum_k$ is over all the compatible pairs of haplotypes $h_{kj}h_{kj}^{com}$ that are compatible with g_i.where ϵ_i is a random error. The null hypothesis is $H_0 : \alpha_1 = \cdots = \alpha_H = 0$.

For a qualitative trait, we can use LRT or score test (same as that given in section 6.4.3). For quantitative trait, the F statistic can be used to test the null hypothesis, where

$$F = \frac{SSR/H}{SSE/(n-H-1)},$$

$SSR = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$ and $SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$. If y_i follows a normal distribution, then F has a F distribution with degrees of freedom $(H, n-H-1)$. The normal assumption as well as the asymptotic distribution for qualitative trait may be not appropriate due to the large number of haplotypes. Using permutation procedure to get the empirical p-value is an alternative.

8.1.3 Generalized T^2 test

Recently Xiong et al. (2002) proposed the generalized T^2 statistic for case-control association studies of complex traits that simultaneously utilizes multiple biallelic (SNP) markers. This statistic is a corollary to that originally developed for multivariate analysis and is known in this context as two-sample Hotelling's T^2 statistic, see for example Anderson (1984). Consider a case-control with n_1 cases and n_2 controls. Suppose that m biallelic markers (e.g. SNPs) have been typed in the sample of cases and controls. The jth marker has alleles B_j and b_j, respectively. The genotype of the jth marker for the ith individual from cases is coded by the indicator variable

$$X_{ij} = \begin{cases} 1, & \text{if the genotype is } B_jB_j \\ 0, & \text{if the genotype is } B_jb_j \\ -1, & \text{if the genotype is } b_jb_j \end{cases}.$$

Similarly defined an indicator variable, Y_{ij}, for an individual from controls. Let

$$X_i = (X_{i1}, \ldots, X_{im})^T, Y_i = (Y_{i1}, \ldots, Y_{im})^T;$$
$$\bar{X}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ij}, \bar{Y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{ij};$$
$$\bar{X} = (\bar{X}_1, \ldots, \bar{X}_m)^T, \bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_m)^T,$$

where $\bar{X}$ and $\bar{Y}$ are the mean vectors for cases and controls, respectively.

The pooled-sample variance-covariance matrix of the indicator variables for the marker genotypes is

$$S = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(X_i - \bar{X})(X_i - \bar{X})^T + \sum_{i=1}^{n_2}(Y_i - \bar{Y})(Y_i - \bar{Y})^T\right].$$

The two-sample Hotelling's T^2 statistics is then defined as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})^T S^{-1}(\bar{X} - \bar{Y}).$$

Under the null hypothesis of no linkage disequilibrium between any marker in the set and a disease locus, the covariance matrix of the indicator variables for the marker genotypes of the cases, $\sum_1 = Cov(X_i, X_i)$, and the covariance matrix of indicator variables for the controls, $\sum_2 = Cov(Y_i, Y_i)$, are equal. Hence, when the sample size is large enough, under the null hypothesis, T^2 is asymptotically distributed as a χ^2 distribution with m degrees of freedom (Anderson 1984). Note that m is the rank of matrix S. When S is not of full rank, one uses its generalized inverse.

8.1.4. Similarity based Method

The basic idea behind the similarity-based methods is that the haplotypes with disease mutation are the descendents of one or few haplotypes, and thus have larger similarity around disease mutation.

Consider a case-control design with n cases and m controls. Suppose that k biallelic markers (e.g. SNPs) have been typed in the sample of cases and controls. For two haplotypes $h_1 = (h_{11}, \ldots, h_{1k})$ and $h_2 = (h_{21}, \ldots, h_{2k})$, where h_{ij} denotes the allele of haplotype i at jth marker, there are many way to define the similarity of the two haplotypes. We summarize some of the definitions in the literature.

1. Counting Similarity (CS): number of markers at which the two haplotypes have the same allele, that is, $AS = \sum_{i=1}^{k} I(h_{1i} = h_{2i})$, $I(\cdot)$ is a indicator function.

2. Length Similarity (LS): The the length spanned by the longest continuous interval of matching alleles.

3. Local Length Similarity (LLS): For marker $l(l = 1, \ldots, m)$, LLS is the length spanned by the longest continuous interval of matching alleles around marker l.

The case of known phases

In this case, we know the two haplotypes of each of the individuals. Suppose there are totally H haplotypes $h_1, \ldots, h_H$, and there are n_i and m_i haplotype h_i in the cases and in the controls, respectively. We denote the $2n$ and $2m$ haplotypes in cases and in controls by $h_1^{ca}, \ldots, h_{2n}^{ca}$ and $h_1^{co}, \ldots, h_{2m}^{co}$, respectively. Let D denote the average similarity for all pairs of haplotypes in cases minus the average similarity for all pairs of haplotypes in controls, that is,

$$\begin{aligned} D &= \frac{1}{4n^2}\sum_{i=1}^{2n}\sum_{j=1}^{2n} S(h_i^{ca}, h_j^{ca}) - \frac{1}{4m^2}\sum_{i=1}^{2m}\sum_{j=1}^{2m} S(h_i^{co}, h_j^{co}) \\ &= \frac{1}{4n^2}\sum_{i=1}^{H}\sum_{j=1}^{H} n_i n_j S(h_i, h_j) - \frac{1}{4m^2}\sum_{i=1}^{H}\sum_{j=1}^{H} m_i m_j S(h_i, h_j) \\ &= \sum_{i=1}^{H}\sum_{j=1}^{H} [\hat{p}_i\hat{p}_j - \hat{q}_i\hat{q}_j] S(h_i, h_j), \end{aligned}$$

where $\hat{p}_i = \frac{n_i}{2n}$ and $\hat{q}_i = \frac{m_i}{2n}$ are the MLE of p_i, q_i, frequencies of haplotype h_i in cases and in controls, respectively. If we use S to denote the $H \times H$ matrix of the similarities with $(i, j)th$ element $S(h_i, h_j)$. Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_H)^T$ and $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_H)$. Then $D = \hat{p}^T S\hat{p} - \hat{q}^T S\hat{q}$. Tzeng et al. (2003) have calculated $Var(D) = V(p, q)$ which is a complicated function of S, p and q (see Appendix B of Tzeng et al. (2003).). Under null $p = q$, $E(D) = 0$ and $\hat{p}^{(0)} = \hat{q}^{(0)} = \frac{n_i + m_i}{4n}$, Then

$$T = \frac{D}{\sqrt{V(\hat{p}^{(0)}, \hat{q}^{(0)})}} \tag{8.1}$$

asymptotically has standard normal distribution.

If we use the LLS, the statistic T has relation with the location of the marker. For each of the marker l, we have a statistic $T(l)$, we can use statistic

$$T = \max_{1 \le l \le k} T(l)$$

as statistic to test the association between the chromosome region and the trait.

The case of unknown phases

For the CS, we do not estimate haplotypes, we can use the similarity of the genotypes. For two genotypes $g_1 = (g_{11}, \ldots, g_{1m})$ and $g_2 = (g_{21}, \ldots, g_{2m})$, where $g_{ij} = 0, 1$, or 2 as usual, the AS of the two genotypes is $\sum_{i=1}^{m} |g_{1i} - g_{2i}|$.

For other similarity measures, we first need to estimate the haplotype frequencies in cases, controls, and the pooled sample. Denote the frequencies of haplotype h_i in cases, controls, and pooled sample by $\hat{p}_i, \hat{q}_i$, and $\hat{p}_i^{(0)}$, respectively.

The statistic is still given by (8.1). However, since there are addition uncertainty introduced by using EM to estimate haplotype frequencies, the standard normal distribution will not a very good approximation to the distribution. In this case, we may use permutation procedure to evaluate the p-value of the test.

References

Anderson TW (1984). *An Introduction to Multivariate Statistical Analysis.* Second Edition, John Wiley & Sons, Inc..

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 63:595–612.

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB. (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci USA* 97(19):10483–10488.

Fallin D (2000) *Haplotype-Based Approaches for Genetic Case-Control Studies.* dissertation, Western Reserve University, Cleveland.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Shork N (2001) Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: application to APOE Locus Variation and Alzhaimer's Disease. *Genome Research*, 11:143-151.

Fallin D, Schork N (2000) Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. *Am. J. of Hum. Genet.*, 67:947-959.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. (2001) *Am J Hum Genet* 68(1):160-172.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score test for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70:425–434.

Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, et al. (2001) A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 27(2):172-180.

Tzeng JY, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 72:891–902.

Xiong M, Zhao J, Boerrwinkle, E (2002) Generalized T^2 Test for Genome Association Studies. *Am J Hum Genet,* 70: 1257-1268.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotype with discrete and continuous traits in samples of unrelated individuals. *Hum Heredity* 53:79–91