**BRIEFINGS IN BIOINFORMATICS.** VOL 9. NO 1. 1–13
Advance Access publication December 14, 2007
doi:10.1093/bib/bbm058

# Three lectures on case–control genetic association analysis

*Wentian Li*

Submitted: 5th June 2007; Received (in revised form): 1st November 2007

## Abstract

The purpose of this review is to focus on the three most important themes in genetic association studies using randomly selected patients (case, affected) and normal samples (control, unaffected), so that students and researchers alike who are new to this field may quickly grasp the key issues and command basic analysis methods. These three themes are: elementary categorical analysis; disease mutation as an unobserved entity; and the importance of homogeneity in genetic association analysis.

***Keywords:*** *genetic association; case-control; single-nucleotide-polymorphism (SNP); linkage disequilibrium; categorical data analysis; aggregation paradox*

## INTRODUCTION

Genetic association analyses, i.e. the examination of statistical correlation between a person's genetic marker genotype with his phenotype or disease status, has become a common task in human genetics and human disease studies. There are several reasons for this trend. First, it has been shown by statistical power analysis that genetic association studies require less samples than pedigree linkage analysis to map disease genes with low penetrance [1]. Second, with the genomic infrastructure in place, such as the complete DNA sequence of the human genome [2, 3], the location of single-nucleotide-polymorphism (SNP) genetic markers [4, 5], and the ever inexpensive high-throughput genotyping technologies [6, 7], it is more cost-effective and easier to carry out candidate-gene, regional, or whole-genome association studies. Third, it has been discovered that the issue of population stratification and spurious association signal can be manageable if better study design is used and if serious care is taken during analysis [8–11]. The publication of genome-wide association studies of seven common diseases using thousands of cases and controls from Wellcome Trust Case Control Consortium exemplified the current status of this field [12].

A consequence of the rapid developments in the field of genetic association study is the large number of publications. Figure 1 shows the number of methodology papers on association analyses per year listed in the online bibliography maintained by the author (URL: http://www.nslij-genetics.org/ld/): it can be seen that after Risch and Merikangas's seminal paper in 1996, the number of methodology papers jumped to around 50 per year for the next three years. The next jump occurred in 2000–2001 with the anticipation and completion of the human genome project, to 100–180 papers per year. One may wait to see whether a publication peak has been reached in the past 2–3 years. For researchers and students new to this field, it is impractical to read 1500 method-oriented research articles, not to mention even more application papers. Reading all review articles is no less overwhelming. Here is just a partial list: *Nature Reviews Genetics* carried nearly 30 review articles related to association analysis one way or another, with the following among them: [13–22]; *Lancet* published a series of review and

Corresponding author. Wentian Li. The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY 11030, USA. Tel: +1-516-562-1076; Fax: +1-516-562-1153; E-mail: wli@nslij-genetics.org

**Wentian Li** received his PhD in physics from Columbia University in 1989. He worked at Santa Fe Institute, Cold Spring Harbor Laboratory, Columbia University Medical Centre, and Rockefeller University, before taking the current position at North Shore LIJ Health System.

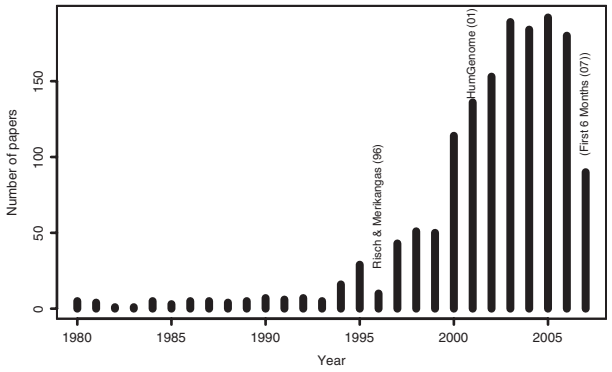


**Figure 1:** Annual number of methodology papers of genetic association analysis as listed in http://www.nslij-genetics.org/ld/.

introductory articles in 2005 on genetic epidemiology with association as the major component [23–29]; *Annual Review* journals published many reviews that can be linked to association studies such as [30–33]; *Current Opinion in Genetics & Development* carried 'Genomes and Evolution' volumes focused on human genetics topics [Vol.11 Issue 6 (2001), Vol.12 Issue 6 (2002), Vol.16, Issue 6 (2006)]; both *American Journal of Human Genetics* [34–38], and *Trends in Genetics* [39–46] occasionally publish reviews and perspectives on genetic association analysis.

To help a novice to digest this mountainous information, one solution is to limit the learning to only a few important topics or themes. The goal of this review is to focus on three such essential pieces of information that are required knowledge in genetic association: first, the necessary mathematical background; second, the fact that the disease-gene-mutation as well as the haplotype it resides is not directly observable; and third, the issue that heterogeneity of various forms is the most likely cause for false association signal. The following two review articles in [21, 47] are probably on a comparable level and readers can also consult the *Handbook on Statistical Genetics* in [48].

## LECTURE 1: STATISTICAL BACKGROUND—CATEGORICAL DATA ANALYSIS

One commonly asked question is: what kind of mathematical background is needed for genetic association analysis? Due to the involvement of professional statisticians in the field, many sophisticated statistical and mathematical techniques have been applied. But the most basic tools used in association studies belong to the 'categorical data analysis' [49], for the obvious reason that both the disease status and genotype are discrete and categorical.

The simplest scenario of genetic association analysis is to study a SNP's genotype frequency in a group of independent patients (cases) and a group of normal individuals (controls). This might be considered as the 'hydrogen atom' for genetic association, borrowing a term from physics to describe the minimum model system. Two types of questions can be asked concerning the genotype or allele frequency difference in the case and control groups. One is 'how big the genotype or allele frequency difference is?' Another is 'how likely that we see this genotype frequency difference by chance?' The first question is answered by 'estimation' and the second question by 'statistical testing.' The main difference between the two is that estimation is usually less affected by the sample size, whereas a test result is greatly influenced by how many samples are collected in the dataset. Later in this section, we will discuss the issue that precision of an estimation, the confidence interval, is also affected by the sample size.

After a routine test of Hardy–Weinberg equilibrium (which is explained in detail in both [47] and [21]) in control samples for the purpose of revealing genotype errors [50], allele and genotype frequency differences in the case and control groups can be tested for three tables: 2-by-3 genotype counts table, 2-by-2 allele counts table, and the 'better' of the two 2-by-2 genotype counts table by combining heterozygous genotype AB with either one of the homozygous genotype [Table 1(A), (B), (C)]. The last approach is essentially a model selection between the genetic dominant and recessive mode of inheritance: assuming allele $A$ is the mutant allele whose frequency is higher in the case group than the control group, then for the dominant model, $AA$ and $AB$ genotypes are equivalent in terms of their contribution to disease risk, and for the recessive model, $AB$ and $BB$ are equivalent.

Each test can be carried out by Pearson's chi-square ($\chi^2$) test, or Fisher's exact test which is usually used when some genotype counts are smaller than five (details can be found in [49]). For statistical tests, a 'test statistic' is first calculated. For example, for Pearson's test, the test statistic is a summary of discrepancy between the observed (O) and expected (E) genotype/allele count: $X^2 = \sum_{i=1}^{i_{\max}} (O_i - E_i)^2 / E_i$

**Table 1:** An example of a case-control data

| | *AA* | *AB* | *BB* | total |
|---|---|---|---|---|
| (A) Genotype counts | | | | |
| Case | a = 10 | b = 190 | c = 800 | a+b+c = 1000 |
| Control | d = 3 | e = 100 | f = 900 | d+e+f = 1003 |
| | ***A*** | ***B*** | | total |
| (B) Allele counts | | | | |
| Case | $x_{11}$ = 2a+b = 210 | $x_{12}$ = b+2c = 1790 | | 2(a+b+c) = 2000 |
| Control | $x_{21}$ = 2d+e = 106 | $x_{22}$ = d+2f = 1900 | | 2(d+e+f) = 2006 |
| | ***AA***+***AB*** | ***BB*** | ***AA*** | ***AB***+***BB*** | total |
| (C) Two ways of grouping heterozygotes with homozygotes | | | | | |
| Case | a+b = 200 | c = 800 | a = 10 | b+c = 990 | a+b+c = 1000 |
| Control | d+e = 103 | f = 900 | c = 3 | d+e = 1000 | d+e+f = 1003 |

There are 1000 case samples and 1003 control samples, whose genotype distribution is shown in the table (A); the number of A and B allele counts is in (B). The genotype counts in (C) are converted from (A) by combining AB with either AA or BB. Note that the total counts in (B) doubles the counts in (A), and the two tables in (C) correspond to the dominant and recessive models if allele *A* is considered as the risk allele.

($i_{\max} = 6$ and 4 for Table 1 (panel A) and (panel B), respectively). The test statistic calculated from the data is then compared to its 'null distribution', i.e. the distribution of the test statistic by chance alone. The probability of finding the observed test statistic or larger under the null distribution is called the *P*-value. A *P*-value in our case is the probability of finding the genotype/allele frequency difference as large as, or larger than what is observed in our data, *if in reality the case and the control have the same genotype/allele frequency*. When the *P*-value calculated from the data is less than 0.01—i.e. if there is no difference between the two groups, then our data is as unlikely as 1 out of 100 by chance—we say the test is significant at the 0.01 level.

The *R* script (for an introduction of the freely available statistical package *R*, from http://www.r-project.org, [51, 52]) for the three tests on the data in Table 1 can be found in Figure 2. Let's call the genotype-based test on Table 1 (panel A) GBT, allele-based test on Table 1 (panel B) ABT, and the maximum of the test statistics in two tables in Table 1 (panel C) MAX2. Then under the null hypothesis (i.e. genotype/allele frequencies are the same in case and in control group), the test statistic in GBT follows the $\chi^2$ distribution with two degrees of freedom, and the test statistic in ABT follows the $\chi^2$ distribution with one degree of freedom. For MAX2, the null distribution of the test statistic may not be known exactly, because it depends on the statistical correlation between two test statistics under dominant and recessive models, which in turn depends on model parameters such as disease penetrance and allele frequencies [53, 54].

For any $\chi^2$ distributed test statistic with *df* degrees of freedom, one can decompose it to two $\chi^2$ distributed test statistics with $df1$ and $df2$ degrees of freedom and their sum $df1 + df2$ is equal to *df* [49]. For example, the test statistic in GBT can be decomposed to two $\chi^2$ distributed values each with one degree of freedom. One of them is the test statistic in a commonly used test called Conchran–Armitage test (CAT) [21, 49]. CAT tests whether $\log(r)$, where $r$ is the (number of cases)/(number of cases + number of controls) ratio, changes linearly with the *AA*, *AB*, *BB* genotype with a non-zero slope. Note that since *AB* is positioned between *AA* and *BB* genotype, the genotype is not just a categorical variable, but an *ordered* categorical variable. Also note that although CAT is genotype-based, its value is closer to the allele-based ABT test statistic. This is confirmed by the numerical values in Figure 2.

From Figure 2, it can be seen that both the Pearson's chi-square test and Fisher's exact test lead to very similar results. The effect of sample size on test results can be easily demonstrated by multiplying all genotype counts by 2—even though the genotype frequencies as well as the difference between the two groups are unchanged, the test result becomes more significant (with much smaller *P*-values) when sample size is increased. The *P*-values obtained from GBT, ABT, MAX2, CAT tests are highly correlated, but not identical. Some differences between different tests were discussed in Ref. [55, 56], but according to Ref. [21], 'there is no generally accepted answer to the question of which SNP test to use'. The key issue seems to be that we do not know the true mode of inheritance of a disease with respect to a particular susceptibility gene and that information should help us to pick a particular test over other tests. The idea of combining two test statistics by a maximum operation as used in MAX2 test is a way to protect against uncertain disease model and this type of 'robust test' has been discussed in theoretical statistics [57, 58].

As for the 'estimation' part of the association analysis, besides the genotype/allele frequency difference (e.g. for the data in Table 1 (panel B), the allele *A* frequency in case group is 210/2000 = 10.5%,

```
gc <- c(10, 190, 800, 3, 100, 900)
ac <- c(2*gc[1]+gc[2], gc[2]+2*gc[3], 2*gc[4]+gc[5], gc[5]+2*gc[6])
gc1 <- c(gc[1]+gc[2], gc[3], gc[4]+gc[5], gc[6])
gc2 <- c(gc[1],gc[2]+gc[3], gc[4], gc[5]+gc[6])
pvg <- chisq.test( matrix(gc, ncol=3, byrow=T), corr=FALSE)$p.value
pva <- chisq.test( matrix(ac, ncol=2, byrow=T), corr=FALSE)$p.value
pvg1 <-  chisq.test( matrix(gc1, ncol=2, byrow=T), corr=FALSE)$p.value
pvg2 <-  chisq.test( matrix(gc2, ncol=2, byrow=T), corr=FALSE)$p.value
pvb <- min(pvg1, pvg2)

print( c(pvg, pva, pvb))  # 6.918239e-09 9.150309e-10 1.224003e-09
pvg.f <- fisher.test( matrix(gc, ncol=3, byrow=T))$p.value
pva.f <- fisher.test( matrix(ac, ncol=2, byrow=T))$p.value
pvg1.f <-  fisher.test( matrix(gc1, ncol=2, byrow=T))$p.value
pvg2.f <-  fisher.test( matrix(gc2, ncol=2, byrow=T))$p.value
pvb.f <- min(pvg1.f, pvg2.f)
print( c(pvg.f, pva.f, pvb.f))  # 2.412721e-09 8.047005e-10 1.132535e-09

pvcat <- prop.trend.test(gc[1:3], gc[1:3]+gc[4:6], score=c(0, 0.5, 1))$p.value
print( c(pvcat) ) # 9.820062e-10

gc <- gc*2
...                             # repeat the tests
print( c(pvg, pva, pvb))        # 4.786203e-17 4.716312e-18 8.379499e-18
print( c(pvg.f, pva.f, pvb.f))  # 1.231881e-17 3.485271e-18 6.810263e-18
print( c(pvcat) ) # 5.422705e-18
```

**Figure 2:** An *R* script for carrying out Pearson's $\chi^2$ test and Fisher's exact test on the data in Table 1. *gc*, *ac*, *gc*1, *gc*2 are the data in Table 1 (panel A), (panel B), (panel C). *pvg*, *pva*, *pvg*1, *pvg*2 are the resulting *P*-value from the Pearson's $\chi^2$ test. *pvb* is the minimum *P*-value of the dominant and the recessive model. *pvg.f*, *pva.f*, *pvg1.f*, *pvg2.f*, *pvb.f* are the similar *P*-values from Fisher's exact test. *pvcat* is the *P*-value for the Conchran-Armitage trend test. The numerical values after a # sign are the output from running the script. The last four lines show that when the sample size is doubled, the test results are even more significant (*P*-values are much smaller).

that in control group is $106/2006 = 5.3\%$, differs by 5.2%), another commonly used quantity is odds-ratio (OR). Odds itself is already a ratio: e.g. numbers of allele A ('for') versus numbers of allele that is not A ('against') can be used to estimate an odds . For the data in Table 1 (panel B), the odds for A in case group is estimated to be $210/1790 = 0.117$, and the odds for A in control group is estimated to be $106/1900 = 0.056$. If the allele A is rare, then odds is approximately equal to the allele frequency (e.g. $0.117 \approx 0.105$ and $0.056 \approx 0.053$). For the allele count data in Table 1 (panel B) the OR (case-over-control, for A) is 2.102878. Note that if we switch row and column, odds ('for' case, given a particular genotype) cannot be estimated, because in a case–control study design, the affection status is picked first and the genotype is determined later, so genotype is not given. For example, in the first table in Table 1 (panel A), we cannot say the odds 'for' case given $AA$ genotype is 10:3, as the case samples always consist of a very small percentage of the general population, and this odds should be $\leq 1$, not $>1$. However, a nice feature of case–control design is that even when odds cannot be defined when row and column are switched, the OR (A-over-B, 'for' case) can still be defined and unchanged by this operation.

There is one quantity that properly combines the 'estimation' and the 'testing' part of an analysis: the confidence interval (CI). We often see reports of the 95% confidence interval for odds-ratio (95% CI of OR), which means that if we repeat the same estimation of OR with the similarly generated data, 95% of the time the true value of OR would fall within that interval. The formula used for estimating CI of OR can be traced to Woolf [59, 60], which is restated here [for the notation used, please refer to Table 1 (panel B)]:

$$OR = \frac{x_{11}/x_{12}}{x_{12}/x_{22}} = \frac{x_{11}x_{22}}{x_{12}x_{21}}$$

$$\text{define} \qquad \sigma = \sqrt{\frac{1}{x_{11}} + \frac{1}{x_{12}} + \frac{1}{x_{21}} + \frac{1}{x_{22}}},$$

$$95\%\ \text{CI of OR} = [OR \cdot e^{-1.96\sigma},\ OR \cdot e^{1.96\sigma}].$$

```
ci.or <- function(counts, alpha){    # alpha=0.05 corresponds to 95%CI
 f <- qnorm(1- alpha/2)              # if alpha=0.05, f=1.96
 or <- counts[1]*counts[4]/(counts[2]*counts[3])
 sq <- sqrt(1/counts[1]+1/counts[2]+1/counts[3]+1/counts[4])
 upper <- exp( log(or) + f*sq)
 lower <- exp( log(or) - f*sq)
 res <- c(lower, or, upper)
 res
}

print( ci.or(ac, 0.05))             # 1.650411 2.102878 2.679390
print( ci.or(ac, 0.01))             # 1.529428 2.102878 2.891339

ac <- ac*2                          # double the sample size
print( ci.or(ac, 0.05))             # 1.771784 2.102878 2.495842
print( ci.or(ac, 0.01))             # 1.678927 2.102878 2.633882
```

**Figure 3:** An *R* function for estimating odds-ratio (OR) and its confidence interval (CI). The two input variables for the function are: *counts*: the 2-by-2 count table written as a 4-element array; *alpha*: one minus the percentage for the CI (e.g. *alpha* = 0.05 for 95% CI). The result for the data in Table 1 (panel A) is shown. The last three lines show the result when the sample size is doubled.

The number 1.96 comes from the normal distribution: the probability for sampling a standard normal distributed value within [−1.959964, 1.959964] is 95%. Similarly, there is a 99% chance to sample a standard normal random value within [−2.575829, 2.575829]. The estimation of confidence interval for OR is implemented in an *R* script in Figure 3. Note that as expected, the 99% CI is wider than the 95% CI; also note that when the sample size is doubled, both CIs are narrowed.

## LECTURE 2: DISEASE GENE MUTATION IS NOT DIRECTLY OBSERVABLE IN GENETIC ANALYSES

One of the important facts about genetic analysis is that we do not know directly the location of the disease gene, we only know the genetic markers nearby which might be in linkage disequilibrium with it; nor do we know directly the haplotype phase or the parental origin of the disease gene mutation. Some link between the observed and the hidden unobserved variable is needed.

Table 2 (panel A) shows the four possible haplotypes with the disease locus and a SNP marker nearby, with haplotype frequencies of $f_1 = p_{HA}$, $f_2 = p_{HB}$, $f_3 = p_{hA}$, $f_4 = p_{hB}$. The mutant frequency is $p_H = f_1 + f_2$, and wild-type allele frequency $p_h = f_3 + f_4$. Similarly, the marker allele frequencies are $p_A = f_1 + f_3$ and $p_B = f_2 + f_4$. The linkage

**Table 2:** Notations for haplotypes

| Disease locus | Freq | Marker | Freq | Hap-freq | Comments |
|---|---|---|---|---|---|
| (A) Four haplotypes | | | | | |
| H (mutant) | $f_1+f_2$ | A | $f_1+f_3$ | $f_1$ | Main mutant-carrying haplotype |
| H (mutant) | $f_1+f_2$ | B | $f_2+f_4$ | $f_2$ | Rare (zero freq in founders) |
| h (wild type) | $f_3+f_4$ | A | $f_1+f_3$ | $f_3$ | Common |
| h (wild type) | $f_3+f_4$ | B | $f_2+f_4$ | $f_4$ | Common |

| Disease locus | Freq | Marker | Freq | Hap-freq. | $D$ | $D'$ |
|---|---|---|---|---|---|---|
| (B) Three haplotypes | | | | | | |
| H | $f_1$ | A | $f_1+f_3$ | $f_1$ | $f_1-f_1(f_1+f_3)=f_1f_4$ | 1 |
| h | $f_3+f_4$ | A | $f_1+f_3$ | $f_3$ | $f_3-(f_3+f_4)(f_1+f_3)=-f_1f_4$ | 1 |
| h | $f_3+f_4$ | B | $f_4$ | $f_4$ | $f_4-(f_3+f_4)f_4=f_1f_4$ | 1 |

(A) The four haplotypes and their frequencies; (B) A special situation of three haplotypes ('three-haplotype scenario') where $D'$ is always equal to 1.

disequilibrium (LD) is measured by a difference between the haplotype frequency and the product of independent allele frequencies and an obvious choice is:

$$D = \sum_{i=H,h} \sum_{j=A,B} p_{ij} |p_{ij} - p_i \cdot p_j|$$

For Table 2 (panel A), the four $|p_{ij} - p_i p_j|$ terms in the definition of $D$ are actually identical, for example, $p_{HA} - p_H p_A = f_1 - (f_1 + f_2)(f_1 + f_3) = -f_2 + (f_1 + f_2)(f_2 + f_4) = -(p_{HB} - p_H p_B)$.

There is a problem in using $D$, however: it depends on allele frequencies. Considering the following scenario ('three-haplotype scenario'): when the mutant allele $H$ was first introduced into a population with the nearby marker allele $A$, most of haplotypes would be $h-A$ and $h-B$, very few were $H-A$, and none of the haplotypes were $H-B$. Using the formula in Table 2 (panel B), we have $D = f_1 f_4 = p_H p_B$, which is a function of allele frequencies at both two loci. More generally, since the haplotype frequency should always be smaller than the frequency of its constituent alleles, we have $p_{HA} = p_H p_A + D \le p_H$ and $p_{HA} = p_H p_A + D \le p_A$, which lead to $D \le \min(p_H p_B, p_h p_A)$ [61]. Let us call the minimum of multiple known upper bounds 'ceiling' – we can rewrite $D \le \text{ceiling}(D)$. Another measure of LD is to normalize $D$ by its ceiling:

$$D' = \frac{D}{\text{ceiling}\,(D)}$$

For the three-haplotype scenario in Table 2 (panel B), $D' = 1$ because $p_h p_A = (f_3 + f_4)(f_1 + f_3) = f_3 + f_1 f_4 > f_1 f_4 = p_H p_B$, so $\min(p_H p_B, p_h p_A) = p_H p_B$.

Is the maximum $D'$ in the three-haplotype scenario the best we can hope for between a marker and the disease locus? If we detect the allele $A$ in nearby marker, do we guarantee that there is a mutant allele in the disease locus? From Table 2 (panel B), the answer is no, because $A$ can also sit on a haplotype with the wild-type allele $h$. In other words, the correlation between $A$ and $H$ is not 'perfect'. The correlation between $A$ and $H$ is maximized when the odds-ratio $(p_H/p_h) : (p_A/p_B)$ is maximized, and the product of $D'$ (squared) and this odds-ratio leads to a third measure of LD:

$$r^2 = \frac{D^2}{p_H^2 p_B^2} \cdot \frac{p_H p_B}{p_h p_A} = \frac{D^2}{p_H p_h p_A p_B}.$$

$r^2$ is equal to 1 (maximized) only when $f_2 = f_3 = 0$, or, when there is a one-to-one correspondence between allele $A$ and mutant $H$.

If both the disease locus and its nearby marker are coded numerically with $H = 1$, $h = 0$, $A = 1$, $B = 0$, then the covariance of the two variables is $f_1 - (f_1 + f_2)(f_1 + f_3) = D$, variances are $p_H p_h$ and $p_A p_B$, and the correlation coefficient between the two numerically coded variable is:

$$C = \frac{\text{Cov}\,(\text{dis, marker})}{\sqrt{\text{Var(dis)}}\sqrt{\text{Var(marker)}}} = \frac{D}{\sqrt{p_H p_h p_A p_B}},$$

whose square is exactly $r^2$. The $r^2$ is also related to a quantity widely used in information theory, the mutual information [62]:

$$M = \sum_{i=H,h} \sum_{j=A,B} p_{ij} \log \frac{p_{ij}}{p_i \cdot p_j}.$$

This relationship ($r^2 = 2M$ assuming the natural log is used) can be proved by a Taylor expansion [49, 63–65]:

$$\begin{aligned} M &= \sum_{i=H,h} \sum_{j=A,B} (p_i p_j + D_{ij}) \log\left(1 + \frac{D_{ij}}{p_i p_j}\right) \\ &\approx \sum_{ij} (p_i p_j + D_{ij}) \left(1 + \frac{D_{ij}}{p_i p_j} - \frac{D_{ij}^2}{2 p_i^2 p_j^2}\right) \\ &= \sum_{ij} \left(\frac{D_{ij}^2}{p_i p_j} - \frac{D_{ij}^2}{2 p_i^2 p_j^2}\right) = \frac{D^2}{2} \sum_{ij} \frac{1}{p_i p_j} \\ &= \frac{D^2}{2} \left(\frac{p_A + p_B}{p_H p_A p_B} + \frac{p_A + p_B}{p_h p_A p_B}\right) = \frac{D^2}{2 p_H p_h p_A p_B} = \frac{r^2}{2}. \end{aligned}$$

For a marker to be effective in genetic association study, its allele frequency difference in case and control groups has to be highly correlated with that for the disease gene locus. Let us put superscripts to allele and haplotype frequencies to distinguish cases from controls and define $\Delta p_H = p_H^{\text{case}} - p_H^{\text{control}}$, $\Delta p_A = p_A^{\text{case}} - p_A^{\text{control}}$ for their differences. The notations $p_{A|H}, p_{A|h}$ are used for the $A$ allele frequency conditional on disease gene locus alleles. If we assume $p_{A|H} = p_{AH}/p_H = f_1/(f_1 + f_2)$ and $p_{A|h} = p_{Ah}/p_h = f_3/(f_3 + f_4)$ remain approximately the same in case and control group and the superscript is not needed, then following Pritchard and Przeworski [34]:

$$\begin{aligned} \Delta p_A &= p_A^{\text{case}} - p_A^{\text{control}} = (p_{A|H}^{\text{case}} p_H^{\text{case}} + p_{A|h}^{\text{case}} p_h^{\text{case}}) \\ &\quad - (p_{A|H}^{\text{control}} p_H^{\text{control}} + p_{A|h}^{\text{control}} p_h^{\text{control}}) \\ &\approx p_{A|H}(p_H^{\text{case}} - p_H^{\text{control}}) + p_{A|h}(p_h^{\text{case}} - p_h^{\text{control}}) \\ &= p_{A|H}(p_H^{\text{case}} - p_H^{\text{control}}) + p_{A|h}(1 - p_H^{\text{case}} - (1 - p_H^{\text{control}})) \\ &= (p_{A|H} - p_{A|h}) \Delta p_H \\ &= \left(\frac{f_1}{f_1 + f_2} - \frac{f_3}{f_3 + f_4}\right) \Delta p_H \\ &= \frac{f_1 f_4 - f_2 f_3}{p_H p_h} \Delta p_H = \frac{D}{p_H p_h} \Delta p_H. \end{aligned}$$

For the Pearson's chi-square test, the chi-square test statistic $X^2$ is proportional to the square of the allele frequency difference, and inversely proportional to the product of two allele

frequencies: $X^2 \sim (\Delta p)^2/(pq)$ [34, 56], so one implication of the above formula is:

$$X^2_{\text{marker}} \sim \frac{(\Delta p_A)^2}{p_A p_B} = \frac{D^2}{p_H^2 p_h^2 p_A p_B}(\Delta p_H)^2 = \frac{D^2}{p_H p_h p_A p_B}\frac{(\Delta p_H)^2}{p_H p_h} \sim \frac{D^2}{p_H p_h p_A p_B} X^2_{\text{disease-locus}},$$

i.e. the two chi-square test statistics, one at the disease locus and another at the SNP marker, are related by $r^2$. Because this equation makes a crucial claim that test statistic at a nearby marker can be translated to that at the disease locus itself [15, 21, 34], it can be called the 'fundamental formula for linkage disequilibrium mapping'. Other alternative versions of this fundamental formula can be found in Refs [66–68].

A careful reader might notice that during the derivation of this fundamental formula, if the roles of A and H are switched, it leads to the incorrect conclusion of $X^2_{\text{disease-locus}} = r^2 X^2_{\text{marker}}$. One explanation can be that it is more reasonable to assume $p^{\text{case}}_{A|H} = p^{\text{control}}_{A|H}$ than assuming $p^{\text{case}}_{H|A} = p^{\text{control}}_{H|A}$, because the first conditioning is to condition on the true cause of the disease status, whereas the second is only to condition on a 'hitchhiked' neutral marker, thus is unlikely to be true. However, how good the quality of the first approximation is may depend on whether there are other causes of the disease, i.e. other susceptibility loci or environmental factors unrelated to the region under study [67, 68]. Clearly more theoretical investigation of this issue is needed.

Besides the uncertainty of unobserved disease gene, the phase and parental origin of alleles are also not directly observable. However, methods exist to infer ('reconstruct') the haplotypes. One of the approaches, the EM (expectation and maximization) algorithm, will be introduced here, using the data in Table 3 (taken from Ref. [69]) for joint genotype counts of two loci, one with alleles H and h, and one with alleles A and B.

The only samples whose phase cannot be unambiguously resolved are the double-heterozygous samples with the Hh-AB joint genotype. Of these samples, there are two possible phases: (i) H-A and h-B; (ii) H-B and h-A. We assume $n_1$ samples are in phase 1, and $n_2$ samples are in phase 2 ($n_1 + n_2 = e$). If $n_1$ and $n_2$ are known, then the haplotype frequencies $f_i, f_2, f_3, f_4$ can be determined as in Table 3 (panel B). On the other hand, if we know $f_1, f_2, f_3, f_4$ haplotype frequencies, $n_1$ should be proportional to $f_1 f_4$, $n_2$ proportional to $f_2 f_3$, and due to the constraint $n_2 + n_3 = e$, their value can be determined. This 'chicken and egg' problem can be tackled by first assuming a value for $n_1$ and $n_2$ (e.g., $n_1 = n_2 = 2$), solving for haplotype frequencies, using these haplotype frequencies to re-estimate $n_1$, $n_2$, and repeating the iteration cycle until there is no improvement in the haplotype frequency estimation. An *R*-script of this process for the data in Table 3 (panel A) is included in Figure 4.

**Table 3:** An example of EM algorithm

| | *AA* | *AB* | *BB* |
|---|---|---|---|
| (A) raw data | | | |
| HH | $a = 14$ | $b = 0$ | $c = 0$ |
| hH | $d = 34$ | $e = 4$ | $f = 0$ |
| hh | $g = 109$ | $h = 50$ | $i = 10$ |
| N=221 | | | |

| | **A** | **B** |
|---|---|---|
| (B) haplotype frequency when the phase of double-heterozygous samples is known | | |
| H | $f_1$ $N = 2a + b + d + n_1 = 62 + n_1$ | $f_2$ $N = 2c + b + f + n_2 = n_2$ |
| h | $f_3$ $N = 2g + d + h + n_2 = 302 + n_2$ | $f_4$ $N = 2i + f + h + n_1 = 70 + n_1$ |

(A) Joint genotype counts originally used in Ref. [69]. (B) Two-locus haplotype counts if the double-heterozygous joint genotype counts (e) can be partitioned into $e = n_1 + n_2$, where $n_1$($n_2$) is the number of samples in the H-A, h-B phase (H-B, h-A phase).

This so-called EM algorithm [70] described above is also called 'gene counting' in genetics [71] and introductory material can be found in several textbooks [61, 72–74]. To recognize the situation where EM algorithm can be applied, remember the two key components in our example: missing data/information (the number of double heterozygous samples in each phase: $n_1$ and $n_2$), and unknown parameters (the four haplotype frequencies: $f_1$, $f_2$, $f_3$, $f_4$). If the missing data can be estimated (imputed) when the parameter value is known (the E-step), and if the unknown parameter value can be obtained by the data (the M-step), the EM method is applicable. One final note is on a limitation of EM method [75]: if all our samples are double heterozygous ($a = b = c = d = f = g = h = i = 0$, $e > 0$), the limiting solution is not unique and dependent on the initial choice of $n_1, n_2$. Intuitively, in this situation, we do not have any information to base on in order for deciding the phase of a double heterozygote, except to claim that each one of the four possible haplotypes has the same frequency. An alternative explanation of this limitation is that the application of EM requires Hardy–Weinberg

```
g <- c(14, 0, 0, 34, 4, 0, 109, 50, 10)
n1 <- g[5]/2
n2 <- g[5]/2
N <- 2*sum(g)

delta <- 1
f <- c(0.25, 0.25, 0.25, 0.25)

while( delta > 1e-10 ){
 fold <- f
 f <- c(2*g[1]+g[2]+g[4]+n1, 2*g[3]+g[2]+g[6]+n2, 2*g[7]+g[4]+g[8]+n2, 2*g[9]+g[6]+g[8]+n1)/N
 n1 <- g[5]*f[1]*f[4]/(f[1]*f[4]+f[2]*f[3])
 n2 <- g[5]*f[2]*f[3]/(f[1]*f[4]+f[2]*f[3])
 delta <- sum( abs(f-fold) )
 print( round(f, 4) ) # 0.1448 0.0045 0.6878 0.1629
                      # 0.1483 0.0011 0.6843 0.1664
                      # 0.1491 0.0003 0.6835 0.1672
                      # 0.1493 0.0001 0.6833 0.1674
                      # 0.1493 0.0000 0.6833 0.1674
}
```

**Figure 4:** An *R* script for estimating haplotype frequency for the data in Table 3(panel A). The joint genotype counts in Table 3(panel A) are stored in a 9-element array *g*. $n_1$, $n_2$ are the missing data (number of double heterozygotes in phase I and phase 2), and $f_1$, $f_2$, $f_3$, $f_4$ are the unknown parameter values (haplotype frequencies). Both the missing data and unknown parameter values are initialized before the while loop, and updated within the loop. The loop exit condition is when the iteration does not improve the parameter value estimation.

equilibrium, which is violated in the double-heterozygous situation.

## LECTURE 3: THE IMPORTANCE OF HOMOGENEITY

Concerns of population heterogeneity was the main reason that for many years the method of choice in genetic analysis was pedigree-based, i.e. linkage analysis and family-based association [76, 77], and not the case–control association analysis. Spurious association signals due to population stratification occur if cases (or controls) disproportionally represent a genetically distinct subgroup and as a result, any allele frequency difference between this subgroup and the general population leads to an allele frequency difference between the case and the control group (see box 4 of Ref. [21]). The effect of population stratification is an example of the 'aggregation paradox' [78] with Simpson's [79] (or Pearson-Yule-Kendall-Simpson's, to be historically correct [80]) paradox as a special case. Simply stated, Simpson/aggregation paradox is the reversal/inconsistency of a statistical relationship between two variables as observed in subpopulation, when subpopulations are combined into a whole population. In association analysis, the inconsistency caused by aggregation can be manifested in at least two different ways: (i) false positives—there is no observed association between a marker and the disease in subpopulations, whereas significant association is observed in the combined population; (ii) false negatives—marker-disease association exists in subpopulations but disappears in the combined population.

Instead of using tables to illustrate these two situations, here I am adopting a graphic technique previously used in the discussion of Simpson's paradox [81]. Figure 5 shows the allele frequencies in case and control groups in two subpopulations as well as in the combined population. The *x*-axis is the proportion of samples from subpopulation-1: left end corresponds to 100% subpopulation-2, and right end for 100% subpopulation-1. The *y*-axis is the allele frequency in case and control groups represented by two straight lines (here the case line is above the control line). Moving along the line changes the composition of samples from the two populations.

In Figure 5A, the control group has 40% of the samples from subpopulation-1, and 60% from subpopulation-2, but the case group has 90% subpopulation-1 samples and 10% subpopulation-2 samples. In each subpopulation, the allele frequency

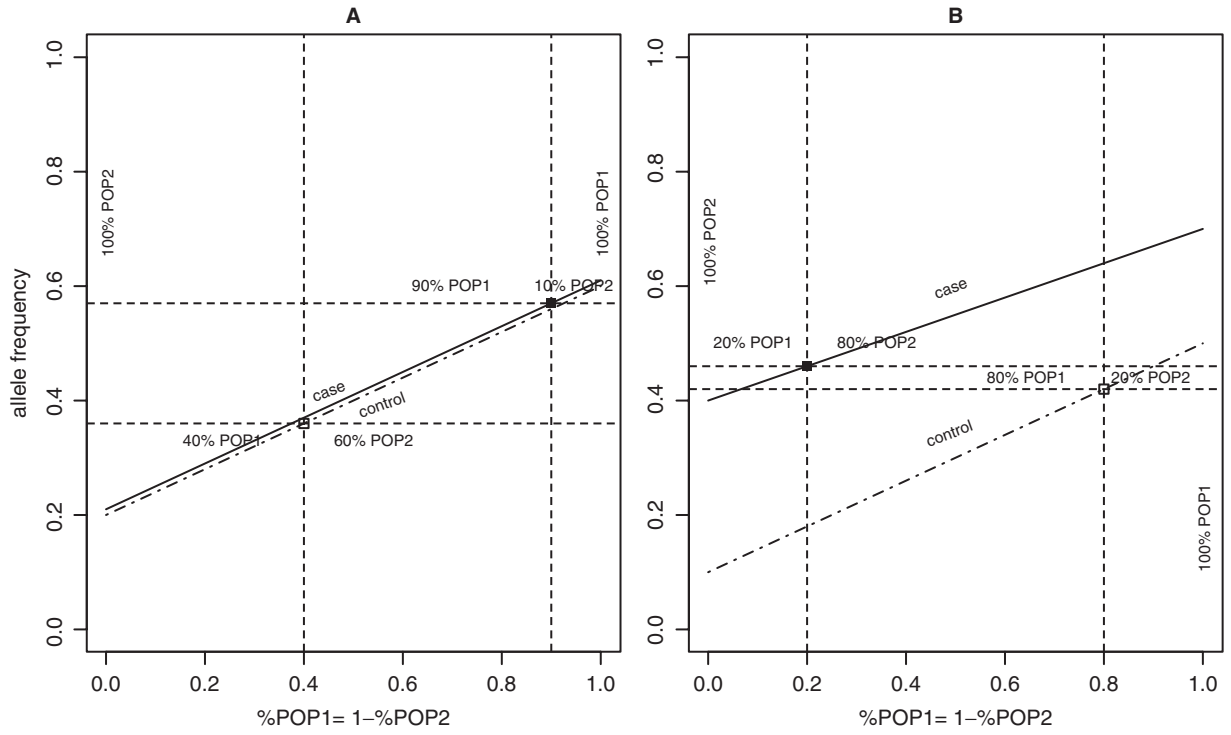


**Figure 5:** Graphic illustration of the population stratification effects: (A) false positive; (B) false negative.

difference between the case and control group is 0.01, whereas in the combined dataset, the difference is increased to 0.21. At a typical sample size, testing allele frequency difference of 0.01 in each subpopulation will not be significant, but testing the 0.21 difference in the combined population will be significant. In fact, if the allele frequency in case and control groups is shrunk from 0.01 to 0 testing allele frequency in subpopulations will never be significant, whereas it could be significant in the combined population.

Using the graphic representation, we can easily figure out situations where population stratification has no effect. For example, if both lines in Figure 5A are horizontal, any composition of the two subpopulations leads to the same allele frequency difference in the combined population. In fact, it is the situation where two subpopulations have no differential allele frequency for this SNP. In another example, when the two lines are parallel to each other and the subpopulation compositions in case and control group are the same, then the allele frequency difference in the combined population remains the same as those in the subpopulations. It is the situation when the case and the control group are well matched in their subpopulation compositions.

False negatives caused by population stratification are rarely discussed in the genetic association literature, but can be easily illustrated by Figure 5B. In this example, the allele frequency difference between case and control group in the two subpopulations is 0.3 and 0.2, respectively. But with different compositions of the two subpopulations (80% subpopulation-1 for control group and 20% subpopulation-1 for case group), the allele frequency difference is merely 0.04 in the combined group.

One might think that since the compositions of two subpopulations in case and control usually do not differ very much, population stratification is unlikely to cause any problem. However, Figure 6 shows a real example in genetic association study of type 2 diabetes in American Indians [82], where a 10% difference in subpopulation composition has lead to false positive signal. In Ref. [82], a haplotype derived from allotypic markers (called $Gm^{3:5,13,14}$) is examined in American Indians with various degrees of Indian heritage. The $Gm^{3:5,13,14}$-carrying frequency as a function of the Indian heritage (0 for European Caucasians, 7/8 for one non-Indian grand-grandparent, 1 for 100% Indian heritage, etc.) is shown in Figure 6 (crosses), which increases from 0.01 for 100% Indians to 0.66 for 100% Europeans),

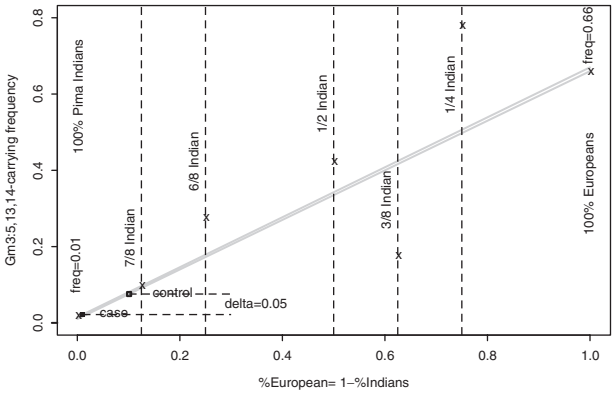


**Figure 6:** A graphic illustration of the effect of population stratification in studying the potential association between $Gm^{3:5,13,14}$ and type 2 diabetes using American Indian samples. The $Gm^{3:5,13,14}$-carrying probability as a function of non-Indian heritage is obtained from the data in Ref. [82] and shown in crosses. We also assume that the $Gm^{3:5,13,14}$-carrying probability is practically the same between the case (type 2 diabetes) and the control group, indicated by the two straight lines. The solid square indicates the case group with 99.9% Indian heritage, and the open square indicates the control group with 99% Indian heritage. An artifact of $Gm^{3:5,13,14}$-carrying frequency difference of 0.05 is produced by this unequal proportion of American Indian heritage in case and control group.

a dramatic change with the population proportion. Suppose the control group is 99% Indian and the diabetes group is 99.9% Indian, even if there is no difference between $Gm^{3:5,13,14}$-carrying frequency in cases and controls, that in the combined dataset is expected to be 0.05. The numbers of 99% and 99.9% Indian in the control and the case group are not unrealistic since the prevalence of type 2 diabetes is higher in Indian population than that in European population, and it is more likely to sample a 100% Indian heritage case sample than a 100% Indian control sample.

The example in Figure 6 points to a way of protecting against population stratification effect: avoiding SNPs with large allele frequency differences between ethnic groups. How large a difference in allele frequency is considered to be large? Wright's *F*-statistic can be used for the purpose of measuring the effect of admixing subpopulations [83]. Denote the allele frequencies of an allele in two subpopulations as $p_1$ and $p_2$ (and for another allele, the frequencies are $q_1 = 1 - p_1$, $q_2 = 1 - p_2$). The heterozygosity (frequency of the heterozygote) in the two subpopulations is $2p_1q_1$, $2p_2q_2$, and that of the combined population is $H_{\text{whole}} = 2\overline{p} \cdot \overline{q}$, where $\overline{p} = wp_1 + (1-w)p_2$, $\overline{q} = wq_1 + (1-w)q_2$ are the averaged allele frequency of the two alleles (with the mixing proportion of the two subpopulations being $w$ and $1 - w$).

If the heterozygosity is calculated within each subpopulation, and added with the appropriate weights, we have $H_{\text{sub}} = w2p_1q_1 + (1-w)2p_2q_2$. It turns out that heterozygosity always increases when subpopulations are combined, i.e. $H_{\text{whole}} \geq H_{\text{sub}}$. We can illustrate this inequality for $w = 1/2$:

$$\begin{aligned} H_{\text{whole}} &= 2 \cdot \frac{p_1 + p_2}{2}\frac{q_1 + q_2}{2} = \frac{p_1q_1 + p_2q_2 + p_1q_2 + p_2q_1}{2} \\ &= p_1q_1 + p_2q_2 + \frac{p_1q_2 + p_2q_1 - p_1q_1 - p_2q_2}{2} \\ &= p_1q_1 + p_2q_2 + \frac{(p_1 - p_2)^2}{2} \geq H_{\text{sub}} \end{aligned}$$

The larger the allele frequency difference $|p_1 - p_2|$, the bigger the inflation of heterozygosity in the combined population. The percentage increase of the heterozygosity by combining subpopulations is the Wright's *F*-statistic:

$$\begin{aligned} F \equiv \frac{H_{\text{whole}} - H_{\text{sub}}}{H_{\text{whole}}} &= \frac{2w(1-w)(p_1 - p_2)^2}{H_{\text{whole}}} \\ &\approx \frac{2w(1-w)(p_1 - p_2)^2}{H_{\text{sub}}} \\ &= \frac{(p_1 - p_2)^2}{(p_1q_1)/(1-w) + (p_2q_2)/(w)} \end{aligned} \quad (1)$$

For more than two subpopulations, the formula is more complicated. Other meanings of *F*-statistic are discussed in detail in Ref. [83].

The *F* value can be marker-specific as well as dataset-specific, but an $F = 10^{-1}$ seems to capture the typical variation between major ethnic groups (e.g. *F* for autosomal SNPs typed in HapMap between European and Chinese/Japanese is roughly 0.07, and that between Yoruba African and Chinese/Japanese is 0.12 [4]), and $F = 10^{-4} \sim 10^{-3}$ seem to describe variation between regions of an isolated population (e.g. 0.00137 for 40 microsatellite markers between regions in the Icelandic population [84]).

It is tempting to attribute the non-replication of association study results [85–87, 9] to population stratification, but it is not necessary to be the case [88, 89]. There are heterogeneities of other forms than that of the allele frequency. Locus heterogeneity refers to the situation in which the susceptibility gene can be located at different chromosomal locations for different patients. Allelic heterogeneity

refers to the situation where mutations can occur anywhere in the same disease gene: coding, non-coding or regulatory regions. For example, it was found that there are three functional variants in the IFN regulatory factor 5 conveying risk for the disease systemic lupus erythematosus [90]. It is not impossible that this type of allelic heterogeneity could weaken the association signal to such an extent that the disease gene fails to be detected by an association study. Yet two other sources of heterogeneity are the disease subtypes and uncertainty in disease diagnosis. What is considered to be a single disease may have different etiologies involving different genes. From this perspective, the contribution of environmental factors can also vary from patients to patients, thus be called heterogeneous. It is with these complications caused by heterogeneity that the authors of Ref. [91] cautioned that 'homogeneous populations are not a panacea'. Common sense and care are perhaps the best weapon against errors caused by heterogeneity.

## BEYOND THIS REVIEW

Even within the scope of the three themes covered in this review, not all topics are discussed. For Lecture 1, one might be interested in learning about power analysis (e.g. the power.prop.test subroutine in *R*) and multiple testing correction (Bonferroni correction for independent and correlated tests, false discovery rates, etc.). For Lecture 2, alternative ways for reconstructing haplotypes may be encountered [92–95], and one may also address the question of whether reconstructing haplotype step is truly necessary for LD mapping [21, 96, 97, 98]. For Lecture 3, one can study several proposals on correcting population stratifications using markers unlinked to the disease, such as the genomic control method [99,100] and the one implemented in the STRUCTURE program [101–103].

### Key Points

- Categorical data analysis provides all the tools needed for genetic association studies. There are many SNP tests with various advantages and disadvantages.
- Because the location, the haplotype phase and the mode of inheritance of a gene mutation are unknown, a variety of approaches focus on how to better infer the unobservable and on how to measure the correlation between observables and the unobservable.
- Population stratification is an example of the aggregation paradox, with Simpson's paradox as another example. The aggregation paradox can be easily illustrated and explained by simple graphs.

**Acknowledgements**

This review is based on the lectures I gave at the 'Theoretical and Practical Course: Pharmacogenomics – Genetic Epidemiology and Web-Based Tools' in March 2005 at Guadalajara, Mexico. I thank Prof. Luis Figuera for his invitation and acknowledge the financial support from the International Centre for Genetic Engineering and Biotechnology (ICGEB). I also thank Yaning Yang, Elena Kowalsky, Jose Luis Santiago-Alvarez for reading an early draft of the paper, and Lisa J. Mao for proofreading the revised version of the paper.

## *References*

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–7.
2. Venter JC, Adams MD, Meyers EW, *et al.* The sequence of the human genome. *Science* 2001;**291**:1304–51.
3. Lander ES, Linton LM, Birren B, *et al.* (International Human Genome Sequencing Consortium) Initial sequencing and analysis of the human genome. *Nature* 2001;**409**: 860–921.
4. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.
5. Barnes MR. Navigating the HapMap. *Brief Bioinformatics* 2006;**7**:211–24.
6. Matsuzaki H, Dong S, Loi H, *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Meth* 2004;**1**:109–111.
7. Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. *Nat Rev Genet* 2006;**7**:632–44.
8. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;**361**:598–604.
9. Freedman ML, Reich D, Penney KL, *et al.* Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;**36**:388–93.
10. Marchini J, Cardon LR, Phillips MS, *et al.* The effect of human population structure on large genetic association studies. *Nat Genet* 2004;**36**:512–17.
11. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 2007;**80**:921–30.
12. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**: 661–78.
13. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. *Nat Rev Genet* 2000;**1**:182–90.
14. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001;**2**:91–9.
15. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in human genome. *Nat Rev Genet* 2002;**2**:299–310.
16. Wall KD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003;**4**:587–97.
17. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004;**5**:89–100.

18. Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies *Nat Rev Genet* 2004;**5**:589–97.
19. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**:95–108.
20. Wang WYS, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6**:109–18.
21. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;**7**:781–91.
22. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet* 2006;**7**:885–91.
23. Burton PR, Tobin MD, Hopper JL. Genetic epidemiology 1: key concepts in genetic epidemiology. *Lancet* 2005;**366**:941–51.
24. Teare MD and Barrett JH. Genetic epidemiology 2: genetic linkage studies. *Lancet* 2005;**366**:1036–44.
25. Cordell HJ and Clayton DG. Genetic epidemiology 3: genetic association studies. *Lancet* 2005;**366**:1121–31.
26. Palmer LJ and Cardon LR. Genetic epidemiology 4: shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005;**366**:1223–34.
27. Hattersley AT and McCarthy MI. Genetic epidemiology 5: what makes a good genetic association study? *Lancet* 2005;**366**:1315–23.
28. Happer JL, Bishop DT, Easton DF. Genetic epidemiology 6: population-based family studies in genetic epidemiology. *Lancet* 2005;**366**:1397–1406.
29. Davey Smith G, Ebrahim S, Lewis S, *et al*. Genetic epidemiology 7: genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005;**366**:1484–98.
30. Zwick ME, Cutler DJ, Chakravarti A. Patterns of genetic variation in Mendelian and complex traits. *Ann Rev Genomics Hum Genet* 2000;**1**:387–407.
31. Weir BSm Hill WG. Estimating F-statistics. *Ann Rev Genet* 2002;**36**:721–50.
32. Crawford DC, Akey DT, Nickerson DA. The patterns of natural variation in human genes. *Ann Rev Genomics Hum Genet* 2005;**6**:287–312.
33. Serre D, Hudson TJ. Resources for genetic variation studies. *Ann Rev Genomics Hum Genet* 2006;**7**:443–57.
34. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001;**69**:1–14.
35. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *Am J Hum Genet* 2004;**75**:353–62.
36. McKaigue PM. Prospects for admixture mapping of complex traits. *Am J Hum Genet* 2005;**76**:1–7.
37. Risch N. The SNP endgame: a multidisciplinary approach. *Am J Hum Genet* 2005;**76**:221–6.
38. Thomas DC, Haile RW, Duggan D. Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet* 2005;**77**:337–45.
39. Reich DE, Lander ES. On the allelic spectrum of human diseases. *Trends Genet* 2001;**17**:502–10.
40. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002;**18**;19–24.
41. Lee C. Irresistible force meets immovable object: SNP mapping of complex diseases. *Trends Genet* 2002;**18**:67–9.
42. Nordborg M, Tavare S. Linkage disequilibrium: what history has to tell us. *Trends Genet* 2002;**18**:83–90.
43. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003;**19**:135–40.
44. Munafo MR, Flint J. Meta-analysis of genetic association studies. *Trends Genet* 2004;**20**:439–44.
45. Di Rienzo A, Hudson RR. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet* 2005;**21**:596–601.
46. Evans DM, Cardon LR. Genome-wide association: a promising start to a long race. *Trends Genet* 2006;**22**:350–54.
47. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinformatics* 2002;**3**:146–53.
48. Balding DJ, Bishop M, Cannings C., edited. *Handbook of Statistical Genetics*. 3rd edn. John Wiley & Sons, Chichester, UK, 2007.
49. Agresti A. *Categorical Data Analysis*. 2nd edn. Wiley 2002 2nd edn. Wiley, 2007.
50. Gomes I, Collins A, Lonjou C, *et al*. Hardy-Weinberg quality control. *Ann Hum Genet* 1999;**63**:535–8.
51. Venables WN. *An Introduction to R*. 2002; Network Theory, Bristol, UK.
52. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th edn. Springer, New York, 2002.
53. Freidlin B, Zheng G, Li Z, *et al*. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 2002;**53**:146–52.
54. Zheng G, Freidlin B, Gastwirth JL. Comparison of robust tests for genetic association using case-control studies. In: Rojo J (ed.). *Optimality: The Second Erich L. Lehmann Symposium – IMS Lecture Notes*, Vol. **49**, 2006: 253–65. Institute of Mathematical Statistics, Bethesda, MD.
55. Sasieni PD. From genotypes to genes: doubling the sample size, *Biometrics* 1997;**53**:1253–61.
56. Suh YJ, Li W. Genotype-based case-control analysis, violation of Hardy-Weinberg equilibrium, and phase diagrams. In: *Proceedings 5th Asia-Pacific Bioinformatics Conference* eds. D Sankoff, L Wang, F Chin, Imperial College Press, London, 2007: 185–94.
57. Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 1977;**64**:247–54.
58. Freidlin B, Podgor MJ, Gastwirth JL. Efficiency robust tests for survival or ordered categorical data. *Biometrics* 1999;**55**:883–6.
59. Woolf B. On estimating the relation between blood and disease. *Ann Hum Genet* 1955;**19**:251–3.
60. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999;**55**:597–602.
61. Weir BS. *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associations, 1996.
62. Li W. Mutual information functions versus correlation functions. *J Stat Phys* 1990;**60**:823–37.
63. Li W, Reich J. A complete enumeration and classification of two-locus models. *Hum Hered* 2000;**50**:334–49.
64. Nothnagel M. *The Definition of Multilocus Haplotype Blocks and Common Diseases*. PhD Thesis, Humboldt-Universität zu: Berlin, 2004.

65. Liu Z, Lin S. Multilocus LD measures and tagging selection with generalized mutual information. *Genet Epi* 2005;**29**:353–64.
66. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* 1999;**63**:1531–40.
67. Terwilliger JD, Hiekkalinna T. An utter refutation of the ''fundamental theorem of the HapMap''. *Euro J Hum Genet* 2006;**14**:426–37.
68. Moskvina V, O'Donovan MC. Detailed analysis of the relative power of direct and indirect studies and the implications for their interpretation. *Hum Hered*, 2007;**64**:63–73.
69. Hamilton DC, Cole DEC. Standardizing a composite measure of linkage disequilibrium. *Ann Hum Genet* 2004;**68**:234–39.
70. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977; **39**:1–38.
71. Smith CAB. Counting methods in genetical statistics. *Ann Hum Genet* 1957;**21**:254–76.
72. Sham P. *Statistics in Human Genetics*. London: Arnold, 1998.
73. Ott J. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD, 1997.
74. Lange K. *Mathematical and Statistical Methods for Genetic Analysis*. 2nd edn. Springer, New York, 2002.
75. Mano S, Yasuda N, Katoh T, *et al*. Notes on the maximum likelihood estimation of haplotype frequencies. *Ann Hum Genet* 2004;**68**:257–264.
76. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996;**59**:983–9.
77. Schaid DJ. Transmission disequilibrium, family controls, and great expectations. *Am J Hum Genet* 1998;**63**:935–41.
78. Stigler SM. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 2002.
79. Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B* 1951;**13**:238–41.
80. Good IJ, Mittal Y. The amalgamation and geometry of two-by-two contingency tables. *Ann Stat* 1987;**15**:694–711.
81. Tan A. A geometric interpretations of Simpson's paradox. *College Math J* 1986;**17**:340–1.
82. Knowler WC, Williams RC, Pettitt DJ, *et al*. $Gm^{3:5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988;**43**:520–26.
83. Hartl DL, Clark AG. *Principles of Population Genetics*. 4th edn. Sunderland, MA: Sinauer Associations, 2006.
84. Helgason A, Yngvadóttir B, Hrafnkelsson B, *et al*. An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005;**37**:90–5.
85. Ioannidis JP, Ntzani EE, Trikalinos TA, *et al*. Replication validity of genetic association studies. *Nat Genet* 2001;**29**:306–9.
86. Ioannidis JP, Trikalinos TA, Ntzani EE, *et al*. Genetic associations in large versus small studies: an empirical assessment *Lancet* 2003;**361**:567–71.
87. Lohmueller ME, Pearce CL, Pike M, *et al*. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;**33**:177–82.
88. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prevent* 2002;**11**:505–12.
89. Wacholder S, Rothman N, Caparaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prevent* 2002;**11**:513–20.
90. Graham RR, Kyogoku C, Sigurdsson S, *et al*. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci* 2007;**104**:6758–63.
91. Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotech* 1998;**9**:578–94.
92. Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990;**7**:111–22.
93. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;**68**:978–89.
94. Marchini J, Cutler D, Patterson N, *et al*. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006;**78**:437–50.
95. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006;**78**:629–44.
96. Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;**73**:1316–29.
97. Clayton D, Chapman J, Cooper J. The use of unphased multilocus genotype data in indirect association studies. *Genet Epi* 2004;**27**:415–28.
98. Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. *J Am Stat Assoc* 2006;**101**:89–104.
99. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;**55**:997–1004.
100. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theo Pop Biol* 2001;**60**:155–66.
101. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;**65**:220–8.
102. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genet* 2000;**155**:945–59.
103. Falush D, Stephens M, Pritchard JK. Inference of population structure: extensions to linked loci and correlated allele frequencies. *Genetics* 2003;**164**, 1567–87.