Population-Based Genetic Association Studies

- **Allelic Association**

Two loci on a chromosome A and B have alleles $\{A_i\}_1^m$ and $\left\{B_j\right\}_1^n$ occurring at frequencies $\{p_i\}_1^m$ and $\left\{q_j\right\}_1^n$ in the population. The genotypes of each subject consists of two haplotypes, one from the paternal gamete and the other from the maternal gamate. Each haplotype consists of two alleles, one at locus A and the other at locus

B. There are a total of *mn* possible haplotypes, which can be denoted as $\left\{A_iB_j\right\}_{i=1,\ldots m;j=1,\ldots n}$, and whose frequencies can be denoted as $h_{11}, h_{12}\ldots, h_{mn}$.

Def: Two alleles are said to be associated if the frequency of the joint occurrence of alleles A_i and B_j in a gamete (the haplotype freq. of A_iB_j) is not equal to the product of the allele frequencies of A_i and B_j, i.e. $h_{ij} \neq p_iq_j$.

When $h_{ij} > (<) p_iq_j$, we say that A_i and B_j are positively (negatively) associated.

- **Maintenance of allelic associations:**

  linkage disequilibrium

Q: If the freq. of the haplotype A_iB_j in the current generation is h_{ij}^0, what will be the frequency of the same haplotype in the next generation, assuming random mating within the population?

A: Let θ be the recombination fraction. Then

$$h_{ij}^{1} = (1-\theta)h_{ij}^{0} + \theta p_i q_j$$

so

$$h_{ij}^{1} - h_{ij}^{0} = \theta(p_i q_j - h_{ij}^{0}) \tag{1}$$

The haplotype freq. will not change if $h_{ij}^{0} = p_i q_j$, i.e. if there is no allelic association. If $h_{ij}^{0} = p_i q_j$,$\forall i, j$,there will be no change in haplotype freq. from generation to generation, and we say that the two loci are in linkage equilibrium. Otherwise the two loci are in linkage disequilibrium (LD) (allelic association).

From (1) we have, after *k* generations,

$$h_{ij}^{k} - p_i q_j = (1-\theta)^k (h_{ij}^{0} - p_i q_j)$$

HW: Suppose that the recombination fraction between two loci is 0.01. How many generations would it take to halve the magnitude of allelic associations (as measured by the discrepancies of the haplotype frequencies from their equilibrium values) between two loci, assuming random mating?

- **Reasons for LD** (allelic association)
  - Physical Proximity
  - Random Genetic Drift (random changes in allele or haplotype freq. from one generation to the next)
  - Mutation
  - Selection
  - Founder Effect

– Population Admixture/ Stratification(spurious association)

- **Spurious Association Caused by Population Admixture / Stratification:**

  A special case of Simpson's paradox: When population subgroups have varied allele frequencies and there is a positive correlation in the frequencies of two alleles across the population subgroups, then the two alleles will appear to be positively associated in the population as a whole, even if there exists no such association within each population subgroup.

HW Consider three populations that have reached linkage equilibrium with respect to the allele A and B. Suppose the population size (N), the frequencies of allele A and allele B, and the frequency of AB are as follows:

N	A	B	AB
1000	0.3	0.5	0.15
2000	0.2	0.4	0.08
10000	0.05	0.1	0.005

Do these alleles A and B show linkage equilibrium in the combined population?

- **Association Analysis As a Tool for Fine Mapping**

  Although allelic association (LD) is a result of the complex interplay between many aspects of the evolutionary history of the population, it is generally accepted that substantial LD is only likely to occur between loci with a recombination fraction of less than 1%.

*Note that, although linkage analysis is a powerful tool for detecting the presence of a disease locus in a chromosome region, it is not efficient for fine mapping, since the discrimination between small differences in recombination frequency requires data on a large number of families.

- **Study Designs for Association Studies**
  - Population-Based Case-Control Studies

- Family-Based Case-Control Studies

- **Case-control Studies:**

  Studying genetic associations between a disease and a set of genetic markers on a group of diseased subjects ('cases') and a group of non-diseased subjects ('controls')

*Note:

- A case-control sample is a biased sample from the target population since the ’cases’ are usually over-represented.
- In case-control samples, since the sample is biasedly sampled, it is not possible to estimate the disease prevalence in each ’exposure’ group directly. However, we can show that the ratio of ’odds’ of exposure in the diseased to that in the controls is identical to the ratio of the odds of disease in the exposed to that in the unexposed.

-Odds Ratio (exposed vs. non-exposed) is estimable in a case-control sample!

*Note:

$$Odds \quad of \quad exposure \equiv \frac{Pr(exposed)}{Pr(nonexposed)}$$

similarly,

$$Odds \quad of \quad disease \equiv \frac{Pr(diseased)}{Pr(nondiseased)}$$

*Note: When the disease is rare, odds of disease$\approx$ Pr(disease)

- Odds-Ratio (OR) Estimation in Genetic Association Study

-Contingency Table (Categorical Data)

ex: cervical cancer and DQ gene (DQ3 allele)

	Negative	Heterozygous	Homozygous	Total
Case	40	45	28	113
Control	273	100	43	416
Total	313	145	71	529

Q: Is the DQ3 gene associated with the cervical cancer?

Q: If yes, is the effect of DQ3 allele dominant, codominant, or recessive?

*Note : the effect of an allele is

– 'dominant' if,

$$OR(heterozygous) = OR(homozygous)$$

– 'codominant' if

$$OR(heterozygous)^2 = OR(homozygous)$$

– 'recessive' if

$$OR(hetrozygous) = 1 \ \& \ OR(homozygous) > 1$$

-general

	# of	DQ3	alleles	
	0	1	2	total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

$$\widehat{OR}(hetero) = \frac{r_1 s_0}{r_0 s_1}$$

$$\widehat{OR}(homo) = \frac{r_2 s_0}{r_0 s_2}$$

-When the effect of DQ3 is codominant, i.e.

$$OR(hetero)^2 = OR(homo)$$

we can estimate this common OR by maximum likelihood estimation in the logistic model.

-logistic model:

$$\log \frac{Pr(D=1|x)}{Pr(D=0|x)} = \beta_0 + \beta x$$

where $x = \#$of DQ3 alleles, $\beta = \log OR$.

-likelihood:

$$L = \prod_{j=1}^{N} Pr(D_j|x_j) = \prod_{i=1}^{N} \frac{\exp\{D_i(\beta_0 + \beta x_j)\}}{1 + \exp(\beta_0 + \beta x_j)}$$

-We can then obtain the MLE of β (and hence $OR = e^{\beta}$) by maximizing $\log L$ over (β_0, β_1)

-Iterative numerical algorithm (such as Newton method) is required to obtain the MLE of OR. [SAS: PROC LOGISTIC; R: glm(family=binomial)]

HW: Using any software you are familiar with to obtain the MLE of the common OR (assuming codominant effect) in the cervical cancer example. (Together with the standard error)

Hint: Delta method:

$S.E.(\widehat{OR}) = \widehat{OR} \times S.E.(\widehat{\beta})$ where $\widehat{OR} = e^{\widehat{\beta}}$

- When the allelic effect is dominant, polling heterozygous and homozygous subjects to form a 2×2 table can lead to more efficient estimate of OR.

  (This is equivalent to a logistic regression using $X = 1$ for both heterozygous and homozygous, and $X = 0$ for negative)

- **Testing for Association**

H_0: the disease is not associated with the gene (i.e., H_0: $\beta = 0$ in the logistic model)

-Armitage's trend test:

$$X_G^2 = \frac{N(N\sum r_i x_i - R\sum n_i x_i)^2}{R(N-R)\left\{N\sum n_i x_i^2 - (\sum n_i x_i)^2\right\}}$$

where $x_i = i$, $i = 0, 1, 2$.

Under H_0, $X_G^2 \sim \chi^2(df = 1)$

*Note: Armitage's trend test is equivalent to the score test in the logistic regression model assuming codominant effect.

*Score Test:
Recall that the logistic model for the association between disease and the gene:

$$Pr(D = d|x_i; \beta_0, \beta) = \frac{\exp d(\beta_0 + \beta x_i)}{1 + \exp(\beta_0 + \beta x_i)}$$

where $d = 0, 1$, $i = 0, 1, 2$

So the loglikelihood is

$$l(\beta_0, \beta) = \sum_{d=0}^{1} \sum_{i=0}^{2} n_{di} \log Pr(D = d|x_i; \beta_0, \beta)$$

where $n_{di} = \#$ of subjects with $D = d$ and $x = x_i$.

The score function for β is

$$U(\beta_0, \beta) \equiv \frac{\partial}{\partial \beta} l(\beta_0, \beta)$$

Let $\widehat{\beta}_0$ be the solution to $0 = \frac{\partial}{\partial \beta_0} l(\beta_0, \beta)|_{\beta=0}$.

Then under H_0 ($\beta = 0$), $U(\widehat{\beta}_0, \beta)|_{\beta=0} \sim N(0, I_{\beta|\beta_0})$

Where

$$I_{\beta|\beta_0} = (I_{\beta\beta} - I_{\beta\beta_0} I_{\beta_0\beta_0}^{-1} I_{\beta_0\beta})|_{\beta=0,\beta_0=\widehat{\beta}_0}$$

and

$$I_{\beta\beta} = -\frac{\partial^2}{\partial \beta^2} l(\beta_0, \beta)$$

$$I_{\beta\beta_0} = -\frac{\partial^2}{\partial\beta\partial\beta_0} l(\beta_0, \beta) = I_{\beta_0\beta}$$

$$I_{\beta_0\beta_0} = -\frac{\partial^2}{\partial\beta_0^2} l(\beta_0, \beta)$$

so under H_0, the score test for $\beta = 0$ is based on

$$\frac{U^2(\beta_0, \beta)}{I_{\beta|\beta_0}}\Big|_{\beta=0,\beta_0=\hat{\beta}_0} \sim \chi^2(df = 1)$$

<u>HW</u> Show that the score test is equivalent to the Armitage's trend test.

- **Potential Drawback for Population-based Case-Control Studies**

  Spurious association may arise through 'population stratification'; that is, when the population at large consists of several subpopulations with different disease prevalence and different allele frequencies, a false positive finding may result even if there exists no association between disease and gene in each subpopulation. (Confounding effect!)

- **How to Avoid Spurious Associations Caused by Population Stratification?**

-Matching:
cases and controls are carefully matched with respect to some potential confounders such as race, ethnicity, nationality, ancestry and birthplace.

-Genotyping additional 'unlinked' markers (in linkage equilibrium with the original candidate markers), known as

'Genomic Control'.

- **Idea:**

  Using genome itself to induce controls, similar to family-based studies.

- **Variance Inflation Factor:**

Recall that the Armitage-trend-test statistics $X_G^2 \sim \chi^2(1)$ under H_0. When there exists population stratification, X_G^2 will no longer follow a chi-squared distribution under H_0, but instead follow a scaled chi-squared distribution, i.e., $X_G^2 \sim \lambda\chi^2(1)$ under H_0, where λ is a constant termed 'variance inflation factor'.

(You can imagine that this λ arises due to the correlation between genotypes among subjects, which is caused by the population stratification. For detailed theoretical derivation, see Devlin and Roeder(1999 Biometrics))

- **Estimation of Variance Inflation Factor**

Genotype a number of markers chosen at random throughout the whole genome so that it is unlikely that any one is tightly linked to a disease-susceptibility gene, and then the average of the Armitage-trend-test statistics across these unlinked markers can serve as an estimate of λ. (Recall that $E\left\{\chi^2(1)\right\} = 1$)

- **Adjusting the Test Statistic**

Let $\widehat{\lambda}$ be the estimate of λ using genomic control. We adjust the Armitage-trend-test statistic by $X_G^2/\widehat{\lambda}$, which will follow $\chi^2(1)$(approximately) under H_0.

Structure Association Method:

Modelling and estimating the population structure, the test for association incorporates the estimated population structure.