Case-Control Association Testing

Introduction

- Identifying susceptibility variants for common/complex diseases has proven to be very difficult despite major advances in high-density genome scans.
- It is believed that most common disorders are influenced by numerous variants, with each variant contributing a relatively small effect (difficult to detect).
- Linkage Analysis Methods: identify regions that related affecteds share IBD in excess of what is expected under null hypothesis of no linkage (poor power for complex diseases)
- Alternatively association studies, also known as linkage disequilibrium studies, can be used to identify susceptibility variants.

Introduction

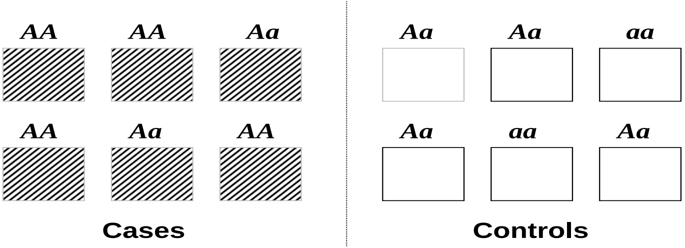
- Association mapping is now routinely being used to identify loci that are involved with complex traits.
- Technological advances have made it feasible to perform case-control association studies on a genome-wide basis with hundreds of thousands of markers in a single study.
- We consider testing a genetic marker for association with a disease in a sample of unrelated subjects.
- Case-control association methods essentially test for independence between trait and allele/genotype.

Case-Control Association Testing

- Allelic Association Tests
  - Allele is treated as the sampling unit
  - Typically make an assumption of Hardy-Weinberg equilibrium (HWE). Alleles within an individual are conditionally independent, given the trait value.
- Genotypic Association Tests
  - Individual is the sampling unit
  - Does not assume HWE

Case-Control Association Testing

- Below is a simple example to illustrate association testing at a genetic marker with two allelic types, **A** and **a**



Pearson's χ^2 Test for Allelic Association

- The classical Pearson's χ^2 test is often used for allelic association testing.
- This test looks for deviations from independence between the trait and allele.
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled "1" and "2"
- Let N_{ca} be the number of cases and N_{co} be the number of controls with genotype data at the marker.

Pearson's χ^2 Test for Allelic Association

- Below is a 2×2 contingency table for trait and allelic type

	Cases	Controls	Total
Allele 1	n_1^{ca}	n_1^{co}	n_1
Allele 2	n_2^{ca}	n_2^{co}	n_2
Total	$2N_{ca}$	$2N_{co}$	T

- n_1^{ca} is the number of type 1 alleles in the cases and $n_1^{ca} = 2 \times$ the number of homozygous $(1,1)$ cases $+$ the number of heterozygous (1,2) cases
- n_2^{co} is the number of type 2 alleles in the controls and $n_2^{co} = 2$ $\times$ the number of homozygous $(2,2)$ controls $+$ the number of heterozygous (1,2) controls
- Hypotheses
  - H_0: there is *no association* between the row variable and column variable
  - H_a: there *is* an association between the two variables

Pearson's χ^2 Test for Allelic Association

- Can use Pearson's χ^2 test for independence. The statistic is:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

- What is the the expected cell number under H_0? For each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1) = (2-1)(2-1) = 1$ degree of freedom

LHON Example: Pearson's χ^2 Test

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- Corresponding 2×2 contingency table for trait and allelic type

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Intuition for the test: Suppose H_0 is true, allelic type and case-control status are independent, then what counts would we expect to observe?
- Recall that under the independence assumption $P(A \text{ and } B) = P(A)P(B)$

LHON Example: Pearson's χ^2 Test

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Let n be the total number of alleles in the study. Assuming independence, the expected number of case alleles that are of type T is:

$$n \times P(\text{Allele is from a Case and Allelic type is T})$$

$$= nP(\text{Allele is from a Case})P(\text{Allelic type is T})$$

$$= 656\left(\frac{178}{656}\right)\left(\frac{550}{656}\right) = \frac{(178)(550)}{656} = 149.2378$$

LHON Example: Pearson's χ^2 Test

- Expected Counts

	Cases	Controls	Total
Allele T	149.2378	400.7622	550
Allele C	28.7622	77.2378	106
Total	178	478	656

- $$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + \cdots + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- The *p*-value is

$$P(\chi_1^2 \geq 4.369) = .037$$

The Armitage Trend Test for Genotypic Association

- The most common genotypic test for unrelated individuals is the Armitage trend test
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled “1” and “2”
- Let $Y_i = 2$ if individual i is homozygous (1,1), 1 if the i is heterozygous, and 0 if i is homozygous (2,2)
- Let $X_i = 1$ if i is a case and 0 if i is a control.
- A simple linear regression model of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

The Armitage Trend for Genotypic Association

- To test this hypothesis, the Armitage trend test statistic is

$$A_r = \frac{\hat{\beta}_1^2}{VAR(\hat{\beta}_1)} = Nr_{xy}^2$$

  where r_{xy}^2 is the squared correlation between genotype variable Y and phenotype variable X.
- Note that the variance estimate for Y that is used in the calculation of the Armitage trend test is the sum of the squared deviations of Y from the fitted values of Y for regression with only an intercept term.
- Under the null hypothesis, A_r will follow an approximate χ^2 distribution with 1 degree of freedom.
- The Armitage trend test can be shown to be valid when HWE does not hold.

LHON Example: Armitage Trend Test

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- The Armitage test statistic for this data is

$$A_r = Nr_{xy}^2 = 328(.0114) = 3.74$$

- The *p*-value is

$$P(\chi_1^2 \geq 3.743) = .053$$

Odds Ratios: Genetic Association

Odds Ratios (ORs) Allele Counting

	Cases	Controls
T	A	B
C	C	D

$$OR_T = \frac{\text{odds of disease with T allele}}{\text{odds of disease with C allele}}$$
$$= \frac{(A/B)}{(C/D)} = \frac{A \times D}{B \times C}$$

- Allele counting model essentially assumes an additive model
- Genotype TT has twice the risk (or protection) of heterozygous genotype CT.
- Same risk (or protection) for the comparison of heterozygous CT genotype and homozygous CC genotype.

Odds Ratios (ORs) Allele Counting

	Cases	Controls
T	A	B
C	C	D

- $OR_T = 1$ implies no association between genotype and disease
- $OR_T > 1$ implies that the T allele is associated with the disease
- $OR_T < 1$ implies that the T allele is protective

Confidence Intervals for Odds Ratios (ORs)

	Cases	Controls
T	A	B
C	C	D

$$OR = \frac{A \times D}{B \times C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{A \times D}{B \times C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

LHON Example: Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{158 \times 86}{392 \times 20} = 1.7332$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{158} + \frac{1}{392} + \frac{1}{20} + \frac{1}{86}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

$$= exp(log(1.7332) - 1.96 \times 0.2665) = 1.03$$

- Upper limit of 95% CI $= 2.92$

Odds Ratios (ORs) for Genotypes

	Cases	Controls
TT	A	B
CT	A'	B'
CC	C	D

- Typically choose a reference genotype. For this example we will let CC be the reference genotype.

$$OR_{TT} = \frac{\text{odds of disease in an individual with the TT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

$$OR_{CT} = \frac{\text{odds of disease in an individual with the CT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

Odds Ratios (ORs) for Genotypes

- To get odds ratios and confidence intervals for genotypes, logistic regression is used:

  $$\begin{aligned}&\text{log(odds of disease for individual } i\text{)}\\&= \beta_0 + \beta_{CT} I\{G_i = CT\} + \beta_{TT} I\{G_i = TT\} + \epsilon_i\end{aligned}$$

  where G_i is the genotype for individual i, and $I\{G_i = CT\}$ is 1 if $G_i = CT$ and 0 otherwise.
- The coefficient estimates for $\hat{\beta}_{CT}$ and $\hat{\beta}_{TT}$ can be used to calculate odds ratios:

  $$OR_{CT} = exp(\hat{\beta}_{CT})$$

  $$OR_{TT} = exp(\hat{\beta}_{TT})$$
- 95% CI for OR_{CT} is

  $$exp(\hat{\beta}_{CT} \pm 1.96 \times s.e.(\hat{\beta}_{CT}))$$

Odds Ratios (ORs) for Genotypes: LHON Example

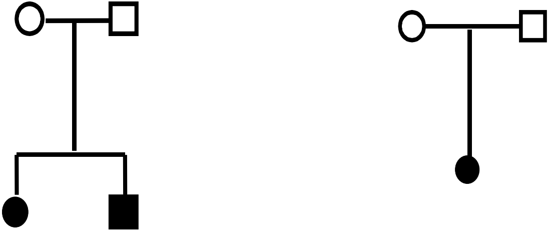
- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

Estimating Relatedness

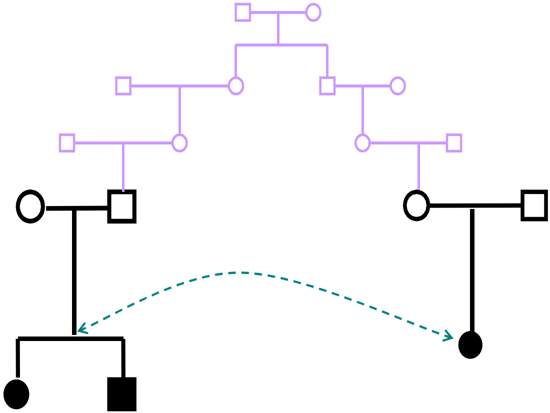
Incomplete Genealogy

- Many statistical methods for genetic data, e.g. linkage and association methods, are based on assumptions of independent samples or samples with known relationships.



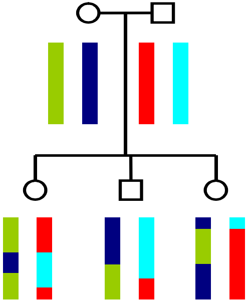
Incomplete Genealogy

- Misspecified and cryptic relationships can invalidate many of these methods.



Identifying Relative Pairs

- A chromosome inherited by an offspring from a parent is actually a mosaic (created by recombination) of the parent's two chromosomes.
- In the picture below, positions on the chromosomes that are the same color are identical by decent (IBD).



Identifying Relative Pairs

- In principle, could determine the relationship between two individuals by simply looking at the percentage of IBD sharing in the genome for the two
    - parent-offspring sharing: 50% of genome
    - sibs: 50% of genome (on average)
    - avuncular: 25% of genome (on average)
- However, we do not directly observe IBD sharing. We only observe DNA sequences.

Genome Screen Data to Identify Relative Pairs

- It is now common to have genome screen data on hundreds of thousands of genetic markers.
- Genome screen data can be used to infer genealogical relationships.
- Example: Suppose we are interested in identifying the relationship between two individuals and assume for now that haplotype phase is known.
- Observed sequence on a chromosome from individual 1:

  ...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...

- Observed sequence on a chromosome from from individual 2:

  ...GGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

- If haplotype phase is known, blocks of identical DNA sequences can be used to infer relationships.

Genome Screen Data to Identify Relative Pairs

- Stanley F Nelson (UCLA Department of Human Genetics): IBD sharing between relatives: rapid drop in number of blocks yet size drops asymptotically:
    - 1st cousins: n=20-30, average size∼20-30mb
    - 2nd cousins: n=5-8, average size∼20mb
    - 3rd cousins: n=1-3, average size ∼18mb
    - 4th cousin: n=0-1, average size ∼16mb
    - 5th cousins: n=0-1, average size ∼14mb
    - 6th cousins: n=0-1, average size∼12mb

Hidden Markov Model for Identifying Relative Pairs

- McPeek and Sun (2000) developed approximate likelihood method to identify relative pairs for close relationships
- Stankovich et al. (2005) extended method for more distantly related pairs (degree 13: 6th cousin). Software is GBIRP
- Uses a 2-state Hidden Markov model for IBD status (yes/no) to approximate the likelihood
- Likelihood is a function of the distance between genetic markers, frequency of alleles between the markers, and relationship of individuals

Hidden Markov Model for Identifying Relative Pairs

- Find pairwise relationship that maximizes the log likelihood ratio for the observed genome screen data (g_1, g_2) over various types of relationships (up to 6th cousins)

$$log \frac{P(g_1, g_2 | related)}{P(g_1, g_2 | unrelated)}$$

- High power to identify relationships up to degree eight (third cousins once removed)
- Typical error in degree for relationship $\leqslant$ eight is 1

GBIRP Results for Known Relationships

Table: GBIRP MS Pairs

ID1	ID2	Truth	Estimate
20001	30001	2	2
23908	24501	3	3
5809	3701	3	3
45101	45201	4	4
6807	9603	5	6
4801	3701	5	5
8201	42204	5	6
7202	7804	5	7
31001	7603	6	6
4801	5809	6	6
6802	21006	6	6
30602	20503	7	7
30603	9803	7	7
133505	30103	7	9
32204	1303	8	7
33404	4204	8	8
23804	1303	8	8
30501	7037	9	9
2901	602	9	∅
6202	602	9	∅
8003	1704	10	∅
4902	42204	10	∅
20503	1203	11	9
24001	32801	11	12
30501	7902	13	∅

IBD Sharing Probabilities

- IBD sharing probabilities are another measure of relatedness for pairs of individuals
- For any pair of outbred individuals i and j, let δ_k be the probability that i and j share k alleles IBD at a locus where k is 0, 1, or 2.

IBD Sharing Probabilites for Outbreds

Relationship	δ_2	δ_1	δ_0
Parent-Offspring	0	1	0
Full Siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Half Siblings	0	$\frac{1}{2}$	$\frac{1}{2}$
Uncle-Nephew	0	$\frac{1}{2}$	$\frac{1}{2}$
First Cousins	0	$\frac{1}{4}$	$\frac{3}{4}$
Double First Cousins	$\frac{1}{16}$	$\frac{6}{16}$	$\frac{9}{16}$
Second Cousins	0	$\frac{1}{16}$	$\frac{15}{16}$
Unrelated	0	0	1

- Note that $\sum_{k=0}^{2} \delta_k = 1$

Estimating IBD Sharing Probabilities: EM Algorithm

- It is often not be possible to determine exactly how many alleles a pair share IBD.
- Can estimate IBD sharing probabiliting wsing genetic marker data across the genome.
- Choi, Wijsman, and Weir (2009) proposed using an EM algorithm to estimate the IBD probabilities for this problem.

Estimating IBD Sharing Probabilities: EM Algorithm

- Suppose the data consists of N genetic markers accross the genome
- Assume for now that at we observe IBD sharing at each marker for individuals i and j in the sample
- Let X_k be the number of markers for which i and j share k alleles IBD, and let let δ_k be the probability that i and j share k alleles IBD at a merek where k is 0, 1, or 2..
- If the IBD sharing process at the markers is observed, what would the likelihood function be?

Estimating IBD Sharing Probabilities: EM Algorithm

- The likelihood function for the IBD sharing process would have the following multinomial distribution

$$L(X_0, X_1, X_2) = \frac{N!}{X_0!X_1!X_2!}\delta_0^{X_0}\delta_1^{X_1}\delta_2^{X_2}$$

  where $X_k = \sum_{r=1}^{N} I\{\ i$ and j share k alleles IBD at marker $r\}$
- Could estimate the δ_k's using the X_k's, which are the sufficient statistics: The MLE is $\hat{\delta}_k = \frac{X_k}{N}$ for $k = 0, 1, 2$.
- The IBD process, however is not observed.
- What is the complete data and what is the observed data?

Expectation Step of EM Algorithm

- The X_k values are the unobserved complete data.
- The observed data is the genotype data for individuals i and j at the N markers, and the X_k values are the missing data
- The E step of the EM algorithm calculates the expected value of X_k conditioned on the observed genotype data.
- Remember that initial values for the δ_k's need to be given for the EM algorithm.
- Let $\delta^0 = (\delta_0^0, \delta_1^0, \delta_2^0)$ be the initial values.
- Let $\mathbf{G} = (G_1, \ldots G_r, \ldots G_N)$, where $G_r = (G_{i_r}, G_{j_r})$ is the genotype data at marker r for i and j.

Expectation Step of EM Algorithm

- $X_2 = \sum_{r=1}^{N} I\left\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\right\}$
- $E\left[X_2 | \mathbf{G}, \delta^0\right] =$

$$\sum_{r=1}^{N} E\left[I\left\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\right\} | \mathbf{G}, \delta^0\right]$$

$$= \sum_{r=1}^{N} E\left[I\left\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\right\} | G_r, \delta^0\right]$$

$$= \sum_{r=1}^{N} P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r | G_r, \delta^0\right)$$

$$= \sum_{r=1}^{N} \frac{P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r | \delta^0\right)}{P\left(G_r | \delta^0\right)}$$

Expectation Step of EM Algorithm

- The numerator of the summand is $P\left(\text{ } i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r | \delta^0\right)$

$$= P\left(G_r | \text{ } i \text{ and } j \text{ share 2 alleles IBD at marker } r, \delta^0\right) \times$$

$$P\left(\text{ } i \text{ and } j \text{ share 2 alleles IBD at marker } r | \delta^0\right)$$

$$= P\left(G_r | \text{ } i \text{ and } j \text{ share 2 alleles IBD at marker } r, \delta^0\right) \delta_2^0$$

- $P\left(G_r | \text{ } i \text{ and } j \text{ share 2 alleles IBD at marker } r\right)$ will be based on the population allele frequency distribution at marker r.

Expectation Step of EM Algorithm

- For simplicity, assume that marker r is a SNP with the 2 allelic types labeled "0" and "1""
- Let p_r be the frequency of allelic type 1 in the population at marker k, where $0 < p_r < 1$.
- If the genotype of i is (1,1) and the genotype of j is (1,1) at marker r, then $P\left(G_r \middle| \, i \text{ and } j \text{ share 2 alleles IBD at marker } r\right) = p_r^2$ (if HWE is assumed).
- What is the probability if the genotype of i is (1,2) and the genotype of j is (2,2) at marker r?
- What is the probability if the genotype of i is (1,2) and the genotype of j is (1,2) at marker r?

Expectation Step of EM Algorithm

- From these probabilities, we can obtain $E\left[X_2|\mathbf{G},\delta^0\right] =$

$$\sum_{r=1}^{N} \frac{P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r|\delta^0\right)}{P\left(G_r|\delta^0\right)}$$

- Can similarly obtain $E\left[X_1|\mathbf{G},\delta^0\right]$ and $E\left[X_0|\mathbf{G},\delta^0\right]$, where

$$X_1 = \sum_{r=1}^{N} I\left\{\ i \text{ and } j \text{ share 1 alleles IBD at marker } r\right\}$$

and

$$X_0 = \sum_{r=1}^{N} I\left\{\ i \text{ and } j \text{ share 0 alleles IBD at marker } r\right\}$$

Maximization Step of EM Algorithm

- The M step involves maximizing the expected value of the log-likelihood (obtained in the E step) with respect to the δ_k parameters.
- The MLE is:
  - $\hat{\delta}_0 = \frac{E\left[X_0|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
  - $\hat{\delta}_1 = \frac{E\left[X_1|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
  - $\hat{\delta}_2 = \frac{E\left[X_2|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
- The next step is to set $\delta^1 = \hat{\delta}$ and then return to the E step of the algorithm.
- Continue iterating between the E and M step until the $\hat{\delta}^i$ values converge.

Estimating Kinship Coefficients

- Kinship coefficients can also be used to quantify relationships between two individuals.

Table: Kinship Coefficients

Relationship	ϕ
Parent-Offspring	1/4
Full Siblings	1/4
Half Siblings	1/8
Uncle-nephew	1/8
First Cousins	1/16
Double First Cousins	1/8
Second Cousins	1/64
unrelated	0

- Note that $\phi = \frac{1}{2}\delta_2 + \frac{1}{4}\delta_1$

Estimating Kinship Coefficients

- Thornton and McPeek (submitted) propose a method to estimate kinship coefficients using genetic marker data
- Consider once again a marker r with 2 allelic types labeled "0" and "1"
- Let p_r be the frequency of allelic type 1, where $0 < p_r < 1$.
- Consider two individuals i and j. For individual i, let $Y_{i_r} = \frac{1}{2}$ $\times$ (the number of alleles of type 1 in individual i at marker r). So the value of Y_{i_r} is 0, $\frac{1}{2}$, or 1. Similarly define Y_{j_r} for individual j.
- It can be shown that $Cov(Y_{i_r}, Y_{j_r}) = p_r(1 - p_r)\phi_{ij}$, where ϕ_{ij} is the kinship coefficient for i and j.
- Rearrange terms to see that $\phi_{ij} = \frac{Cov(Y_{i_r}, Y_{j_r})}{p_r(1-p_r)}$

Estimating Kinship Coefficients

- This relationship will hold for markers across the genome (with the allele frequency distribution changing for each marker).
- Can use data across the genome to estimate kinship coefficients for pairs of individuals
- Let N be the total number of markers in the data.
- For any pair of individuals i and j, can estimate ϕ_{ij} with

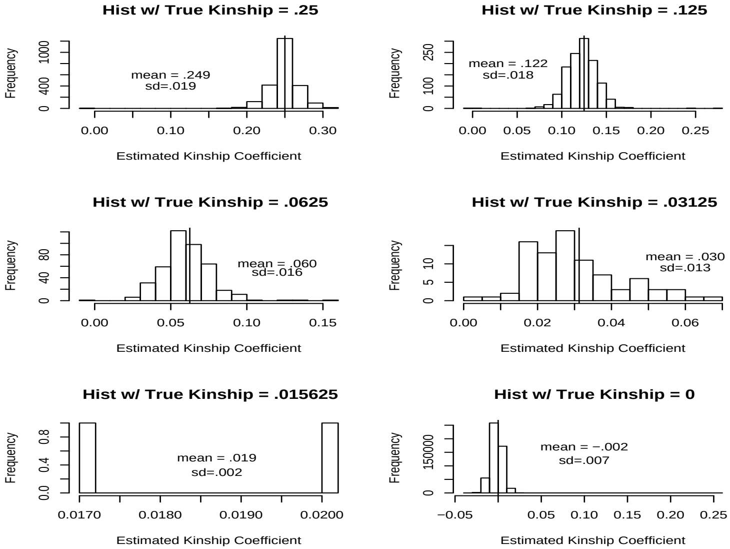
$$\hat{\phi}_{ij} = \frac{1}{N} \sum_{r=1}^{N} \frac{(Y_{i_r} - \hat{p}_r)(Y_{j_r} - \hat{p}_r)}{\hat{p}_r(1 - \hat{p}_r)}$$

where $\hat{p}_r$ is an allele frequency estimate for the type 1 allele at marker r

Estimating Kinships Using GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals. Individuals labeled as white, non-Hispanic were considered.
- 10K SNP array (10,081 SNPs) on 22 autosomal chromosomes
- Estimated kinship coefficients using genome-screen data

Estimating Kinships Using COGA Data


Hist w/ True Kinship = .25

Hist w/ True Kinship = .125

Hist w/ True Kinship = .0625

Hist w/ True Kinship = .03125

Hist w/ True Kinship = .015625

Hist w/ True Kinship = 0

Estimating Kinships Using COGA Data

- From the given pedigrees, two pairs of individuals that should have a kinship coefficient of .25 appear to be unrelated (estimated kinship coefficients of -0.006 and -0.003, respectively)
- Two pairs of individuals that should have a kinship coefficient of .125 appear to be unrelated (estimated kinship coefficients of -0.003 and 0.002, respectively)
- 9 pairs of "unrelated" individuals have a kinship coefficient around .125
- 2 pairs of "unrelated" individual have a kinship coefficient around .25

Population Structure

Nonrandom Mating

- HWE assumes that mating is random in the population
- Most natural populations deviate in some way from random mating
- There are various ways in which a species might deviate from random mating
- We will focus on the two most common departures from random mating:
  - inbreeding
  - population subdivision or substructure

Nonrandom Mating: Inbreeding

- Inbreeding occurs when individuals are more likely to mate with relatives than with randomly chosen individuals in the population
- Increases the probability that offspring are homozygous, and as a result the number of homozygous individuals at genetic markers in a population is increased
- Increase in homozygosity can lead to lower fitness in some species
- Increase in homozygosity can have a detrimental effect: For some species the decrease in fitness is dramatic with complete infertility or inviability after only a few generations of brother-sister mating

Nonrandom Mating: Population Subdivision

- For subdivided populations, individuals will appear to be inbred due to more homozygotes than expected under the assumption of random mating.
- Wahlund Effect: Reduction in observed heterozygosity (increased homozygosity) because of pooling discrete subpopulations with different allele frequencies that do not interbreed as a single randomly mating unit.

Wright's F Statistics

- Sewall Wright invented a set of measures called F statistics for departures from HWE for subdivided populations.
- F stands for fixation index, where fixation being increased homozygosity
- F_{IS} is also known as the inbreeding coefficient.
  - The correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (**I**ndividual within the **S**ubpopulation)
- F_{ST} is a measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations
  - Is defined as the correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (**S**ubpopulation within the **T**otal population).

Wright's F Statistics

- F_{IT} is not often used. It is the overall inbreeding coefficient of an individual relative to the total population (**I**ndividual within the **T**otal population).

Genotype Frequencies for Inbred Individuals

- Consider a bi-allelic genetic marker with alleles A and a. Let p be the frequency of allele A and $q = 1 - p$ the frequency of allele a in the population.
- Consider an individual with inbreeding coefficient F. What are the genotype frequencies for this individual at the marker?

Genotype	AA	Aa	aa
Frequency			

Generalized Hardy-Weinberg Deviations

- The table below gives genotype frequencies at a marker for when the HWE assumption does not hold:

Genotype	AA	Aa	aa
Frequency	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

  where $q = 1 - p$

- The F parameter describes the deviation of the genotype frequencies from the HWE frequencies.
- When $F = 0$, the genotype frequencies are in HWE.
- The parameters p and F are sufficient to describe genotype frequencies at a single locus with two alleles.

F_{st} for Subpopulations

- Example in Gillespie (2004)
- Consider a population with two equal sized subpopulations. Assume that there is random mating within each subpoulation.
- Let $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$
- Below is a table with genotype frequencies

Genotype	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$

- Are the subpopulations in HWE?
- What are the genotype frequencies for the entire population?
- What should the genotypic frequencies be if the population is in HWE at the marker?

F_{st} for Subpopulations

- From the table below it is clear that there are too many homozygotes in this population.

Genotype	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
Freq. Population	$\frac{1}{2}$	$\frac{5}{16}$	$\frac{3}{8}$	$\frac{5}{16}$
Hardy-Weinberg Frequencies	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

- To determine a measure of the excess in homozygosity from what we would expect under HWE, solve

$$2pq(1 - F_{ST}) = \frac{3}{8}$$

- What is F_{st}?

F_{st} for Subpopulations

- The excess homozygosity requires that $F_{ST} = \frac{1}{4}$
- For the previous example the allele frequency distribution for the two subpopulations is given.
- At the population level, it is often difficult to determine whether excess homozygosity in a population is due to inbreeding, to subpopulations, or other causes.
- European populations with relatively subtle population structure typically have an F_{st} value around .01 (e.g., ancestry from northwest and southeast Europe),
- F_{st} values that range from 0.1 to 0.3 have been observed for the most divergent populations (Cavalli-Sforza et al. 1994).

F_{st} for Subpopulations

- F_{st} can be generalized to populations with an arbitrary number of subpopulations.
- The idea is to find an expression for F_{st} in terms of the allele frequencies in the subpopulations and the relative sizes of the subpopulations.
- Consider a single population and let r be the number of subpopulations.
- Let p be the frequency of the A allele in the population, and let p_i be the frequency of A in subpopulation i, where $i = 1, \ldots, r$
- F_{st} is often defined as $F_{st} = \frac{\sigma_p^2}{p(1-p)}$, where σ_p^2 is the variance of the p_i's with $E(p_i) = p$.

F_{st} for Subpopulations

- Let the relative contribution of subpopulation i be c_i, where $$\sum_{i=1}^{r} c_i = 1.$$

Genotype	AA	Aa	aa
Freq. Subpop$_i$	p_i^2	$2p_iq_i$	q_i^2
Freq. Population	$\sum_{i=1}^r c_ip_i^2$	$\sum_{i=1}^r c_i2p_iq_i$	$\sum_{i=1}^r c_iq_i^2$

  where $q_i = 1 - p_i$

- In the population, we want to find the value F_{st} such that $2pq(1 - F_{st}) = \sum_{i=1}^r c_i2p_iq_i$
- Rearranging terms:

$$F_{st} = \frac{2pq - \sum_{i=1}^r c_i2p_iq_i}{2pq}$$

- Now $2pq = 1 - p^2 - q^2$ and $\sum_{i=1}^r c_i2p_iq_i = 1 - \sum_{i=1}^r c_i(p_i^2 + q_i^2)$

F_{st} for Subpopulations

- So can show that

$$
\begin{aligned}
F_{st} &= \frac{\sum_{i=1}^{r} c_i(p_i^2 + q_i^2) - p^2 - q^2}{2pq} \\
&= \frac{\left[\sum_{i=1}^{r} c_i p_i^2 - p^2\right] + \left[\sum_{i=1}^{r} c_i q_i^2 - q^2\right]}{2pq} \\
&= \frac{Var(p_i) + Var(q_i)}{2pq} \\
&= \frac{2Var(p_i)}{2p(1-p)} \\
&= \frac{Var(p_i)}{p(1-p)} \\
&= \frac{\sigma_p^2}{p(1-p)}
\end{aligned}
$$

Estimating F_{st}

- Let n be the total number of sampled individuals from the population and let n_i be the number of sampled individuals from subpopulation i
- Let $\hat{p}_i$ be the allele frequency estimate of the A allele for the sample from subpopulation i
- Let $\hat{p} = \sum_i \frac{n_i}{n} \hat{p}_i$
- A simple F_{st} estimate is $\hat{F}_{ST_1} = \frac{s^2}{\hat{p}(1-\hat{p})}$, where s^2 is the sample variance of the $\hat{p}_i$'s.

Estimating F_{st}

- Weir and Cockerman (1984) developed an estimate based on the method of moments.

$$MSA = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\hat{p}_i - \hat{p})^2$$

$$MSW = \frac{1}{\sum_i (n_i - 1)}\sum_{i=1}^{r} n_i \hat{p}_i (1 - \hat{p}_i)$$

- Their estimate is

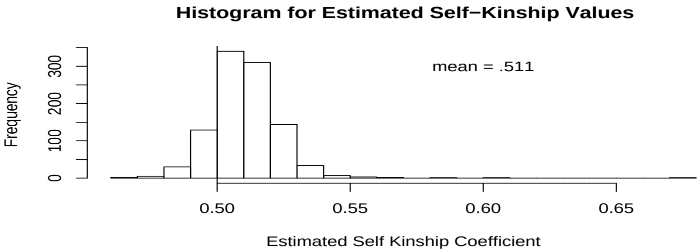
$$\hat{F}_{ST_2} = \frac{MSA - MSW}{MSA + (n_c - 1)MSW}$$

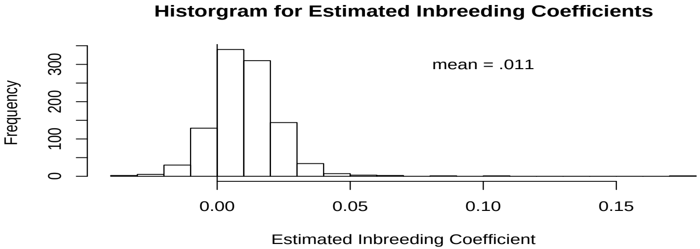
where $n_c = \sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$

GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals.
- Individuals labeled as white, non-Hispanic were considered.
- Estimated self-kinship and inbreeding coefficients using genome-screen data

COGA Data


Histogram for Estimated Self−Kinship Values
mean = .511
Frequency
0
100
200
300
0.50
0.55
0.60
0.65
Estimated Self Kinship Coefficient


Historgram for Estimated Inbreeding Coefficients
mean = .011
Frequency
0
100
200
300
0.00
0.05
0.10
0.15
Estimated Inbreeding Coefficient

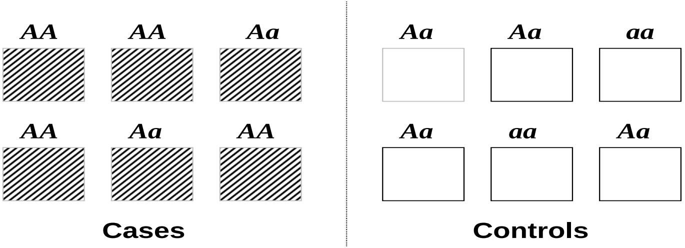
Association Testing with Cryptic Population Structure

Family Based Association Tests

- The popularity of family-based association tests, such as the TDT and FBAT, are largely due to fact that they are robust to population heterogeneity
- Can be used to protect against potential problems of unknown population substructure.
- What are some of the limitations of family based designs?
- Family-based tests are generally less powerful than case-control association methods

Case-Control Association Testing Review

- Consider testing for association between a disease and a genetic marker
- Idea is to look for an association by comparing allele/genotype frequencies between the cases (affected individuals) and the controls (unaffected individuals).



Population Structure and Association Testing

- The observations in genome-wide case-control association studies can have several sources of dependence.
- Population structure, the presence of subgroups in the population with ancestry differences, is a major concern for association studies
- Population structure is often cryptic.
- Neglecting such structure in the data can lead to seriously spurious associations.

Balding-Nichols Model

- A model that is often used for population structure is the Balding-Nichols model (Balding and Nichols, 1995).
- Consider unrelated outbred individuals that are sampled from a population with K subpopulations.
- Assume that an individual can be a member of only one subpopulation, i.e., there is no admixture.
- Under the Balding-Nichols model, the allele frequency for each subpopulation, $1, 2, \ldots, K$, is a random draw from a beta distribution with parameters $p(1 - F_{st})/F_{st}$ and $(1 - p)(1 - F_{st})/F_{st}$, where $0 < p < 1$
- The parameter p can be viewed as the ancestral allele frequency and F_{st} can be viewed as Wright's standardized measure of variation in the population

Balding-Nichols Model: Covariance Structure

- Consider a single bi-allelic marker (e.g. a SNP) with allele labels “0” and “1”
- Let N be the number of sampled individuals with genotype data at the marker.
- Let $Y = (Y_1, \ldots Y_N)$ where $Y_i =$the number of alleles of type 1 in individual i, so the value of Y_i is 0, 1, or 2.
- Under the Balding-Nichols model:
    - Individual i has inbreeding coefficient equal to F_{st}
    - If individuals i and j are are both from the same subpopulation then $Corr(Y_i, Y_j) = F_{st}$
    - If i and j are from different subpopulations then $Corr(Y_i, Y_j) = 0$
- F_{st}, the number of subpopulations K, and the subpopulation memberships for the sample individuals will be unknown when there is cryptic population structure.

If there is no structure then the covariance matrix of Y will be a function of the identity matrix:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

If there is structure then the covariance matrix of Y will be a function of :

$$\mathbf{\Sigma} = \begin{pmatrix} 1 + F_{st} & F_{st} & \dots & 0 \\ F_{st} & 1 + F_{st} & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 + F_{st} \end{pmatrix},$$

Methods for Population Structure

- There are three general approaches that have been proposed to correct for cryptic population structure in case-control
- Genomic Control
- Principal Components Analysis
- Structured Association

Observations from a Single Population: The Armitage Trend Test

- We previously introduced the Armitage Trend Test.
- It is the most common genotypic test for unrelated individuals
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled “1” and “2”
- Let $Y_i = 2$ if individual i is homozygous (1,1), 1 if the i is heterozygous, and 0 if i is homozygous (2,2)
- Let $X_i = 1$ if i is a case and 0 if i is a control.
- A simple linear regression model of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

The Armitage Trend for Genotypic Association

- To test this hypothesis, the Armitage trend test statistic is

  $$A_r = \frac{\hat{\beta}_1^2}{VAR(\hat{\beta}_1)} = N r_{xy}^2$$

  where r_{xy}^2 is the squared correlation between genotype variable Y and phenotype variable X.
- Under the null hypothesis, A_r will follow an approximate χ^2 distribution with 1 degree of freedom.

Genomic Control

- Devlin and Roeder (1999) proposed correcting for substructure via a method called "genomic control."
- The idea is to use data across the genome to correct for cryptic structure
- Let N be the number of individuals in the study.
- Let $\mathbf{X} = (X_1, \ldots X_N)$ be a phenotype indicator vector for case control status where $X_i = 1$ if i is a case and $X_i = 0$ if i is a control
- Let M be the number of bi-allelic markers (e.g. SNPs) in the data. Consider a marker s, where $1 \leqslant s \leqslant M$, and let $\mathbf{Y}_s = (Y_{1_s}, \ldots Y_{N_s})$ where Y_{i_s} =the number of alleles of type 1 in individual i at marker s.

Genomic Control

- For each marker s, the Armitage trend statistic is calculated
$$A_{r_s} = Nr^2_{XY_s}$$
where $r^2_{XY_s}$ is the squared correlation between the genotype variable $\mathbf{Y}_s$ for marker s and the binary phenotype variable $\mathbf{X}$.
- If there is no population structure, the distribution of A_{r_s} will approximately follow a χ^2 distribution with 1 degree of freedom.
- If there is population structure, the statistic will deviate from a χ^2_1 distribution due to an inflated variance.

Genomic Control

- Use $\lambda = \frac{median(A_{r_1},\ldots,A_{r_s},\ldots A_{r_M})}{.456}$ as a correction factor for cryptic structure, where .456 is the median of a χ^2_1 distribution.
- λ will be ≈ 1 if there is no population structure. $\lambda > 1$ indicates that there is population structure.
- The uniform inflation factor λ is then applied to the Armitage trend statistic values
$$\tilde{A}_{r_s} = \frac{A_{r_s}}{\lambda}$$
- $\tilde{A}_{r_s}$ will approximately follow a χ^2 distribution with 1 degree of freedom.
- For the Armitage statistic, the variance is calculated assuming individuals are unrelated (calculation based on the identity matrix).
- Genomic control inflates this variance to account for the cryptic structure (unknown F_{st} values)

Principal Components Analysis

- Price et al. (2006) proposed corrected for structure in association studies by using principal components analysis (PCA)
- They developed a method called EIGENSTRAT for association testing in structured populations.
- If there is cryptic structure then the covariance matrix of Y will be an unknown:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1+F_{st} & F_{st} & \dots & 0 \\ F_{st} & 1+F_{st} & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1+F_{st} \end{pmatrix},$$

EIGENSTRAT

- They propose estimating $\boldsymbol{\Sigma}$ by an empirical covariance matrix $\hat{\boldsymbol{\Sigma}}$ with components $\hat{\Sigma}_{ij}$:

  $$\hat{\Sigma}_{ij} = \frac{1}{M}\sum_{s=1}^{M}\frac{(Y_{is}-2\hat{p}_s)(Y_{js}-2\hat{p}_s)}{\hat{p}_s(1-\hat{p}_s)}$$

  where $\hat{p}_s$ is an allele frequency estimate for the type 1 allele at marker s
- Principal components (eigenvectors) for $\hat{\boldsymbol{\Sigma}}$ are obtained.
- For each eigenvector, and individual in the sample has a value
- The top principal components are viewed as continuous axes of variation that reflect subpopulation genetic variation in the sample.
- Individuals with "similar" values for a particular top principal component will have "similar" ancestry for that axes.

EIGENSTRAT

- The top principal components (highest eigenvalues) are used as covariates in a multi-linear regression.

$$Y_s = \beta_0 + \beta_1 X + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \cdots + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$