

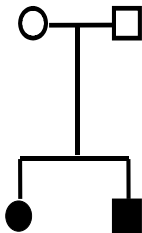
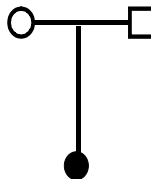
Estimating Relatedness

Timothy Thornton

Summer Institute in Statistical Genetics 2011
Module 10
Lecture 6

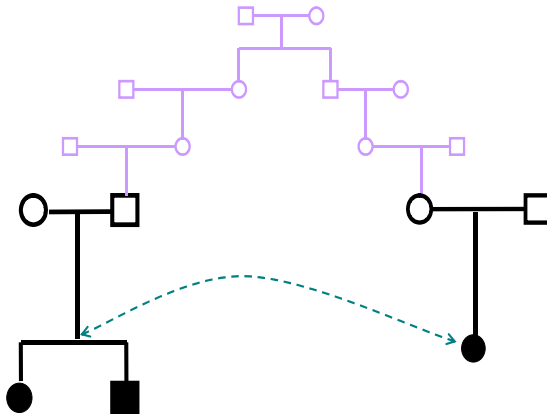
Incomplete Genealogy

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



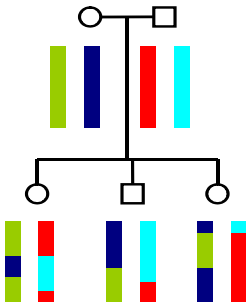
Incomplete Genealogy

- 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 descent (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:
...TATACGTGCACCTG**GATTACAGATTACAGATTACAGATTACA**TTGCATCGATCGAA...
- Observed sequence on a chromosome from individual 2:
...GGATCCTGAACCTA**GATTACAGATTACAGATTACAGATTACA**AATGCTTCGATGGAC...
- 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 $\sim 20-30\text{mb}$
 - 2nd cousins: $n=5-8$, average size $\sim 20\text{mb}$
 - 3rd cousins: $n=1-3$, average size $\sim 18\text{mb}$
 - 4th cousin: $n=0-1$, average size $\sim 16\text{mb}$
 - 5th cousins: $n=0-1$, average size $\sim 14\text{mb}$
 - 6th cousins: $n=0-1$, average size $\sim 12\text{mb}$

Hidden Markov Model for Identifying Relative Pairs

- McPeck 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)

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

- High power to identify relationships up to degree eight (third cousins once removed)
- Typical error in degree for relationship $\leq$ 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	$\emptyset$
6202	602	9	$\emptyset$
8003	1704	10	$\emptyset$
4902	42204	10	$\emptyset$
20503	1203	11	9
24001	32801	11	12
30501	7902	13	$\emptyset$

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 Probabilities 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 possible to determine exactly how many alleles a pair share IBD.
- Can estimate IBD sharing probabilities using 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 across the genome
- Assume for now that 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 δ_k be the probability that i and j share k alleles IBD at a marker 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 $X_k = \sum_{r=1}^N I \{ i \text{ and } j \text{ share } k \text{ alleles IBD at marker } r \}$

- 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 $\mathbf{G} = (G_1, \dots, G_r, \dots, G_N)$, where $G_r = (G_{ir}, G_{jr})$ is the genotype data at marker r for i and j .

Expectation Step of EM Algorithm

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

$$\begin{aligned} & \sum_{r=1}^N E [I \{ i \text{ and } j \text{ share 2 alleles IBD at marker } r \} | \mathbf{G}, \delta^0] \\ &= \sum_{r=1}^N E [I \{ i \text{ and } j \text{ share 2 alleles IBD at marker } r \} | G_r, \delta^0] \\ &= \sum_{r=1}^N P (i \text{ and } j \text{ share 2 alleles IBD at marker } r | G_r, \delta^0) \\ &= \sum_{r=1}^N \frac{P (i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r | \delta^0)}{P (G_r | \delta^0)} \end{aligned}$$

Expectation Step of EM Algorithm

- The numerator of the summand is

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

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

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

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

- $P(G_r | i \text{ and } j \text{ share 2 alleles IBD at marker } r)$ 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
$$P(G_r | i \text{ and } j \text{ share 2 alleles IBD at marker } r) = p_r^2 \text{ (if HWE is assumed).}$$
- 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 $E[X_2|\mathbf{G}, \delta^0] =$

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

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

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

and

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

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:
 - $\hat{\delta}_0 = \frac{E[X_0|\mathbf{G}, \delta^0]}{E[X_0|\mathbf{G}, \delta^0] + E[X_1|\mathbf{G}, \delta^0] + E[X_2|\mathbf{G}, \delta^0]}$
 - $\hat{\delta}_1 = \frac{E[X_1|\mathbf{G}, \delta^0]}{E[X_0|\mathbf{G}, \delta^0] + E[X_1|\mathbf{G}, \delta^0] + E[X_2|\mathbf{G}, \delta^0]}$
 - $\hat{\delta}_2 = \frac{E[X_2|\mathbf{G}, \delta^0]}{E[X_0|\mathbf{G}, \delta^0] + E[X_1|\mathbf{G}, \delta^0] + E[X_2|\mathbf{G}, \delta^0]}$
- 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 IBD Sharing Probabilities: Method of Moments

- Purcell et al. (2007) proposed a method of moments estimator for IBD sharing probabilities
- Estimate IBD sharing probabilities based on IBS sharing for pairs of individuals
- Implements the IBD sharing method of moments estimator in their software package PLINK

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 McPeck (2010) 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 $Y_{i_r} = \frac{1}{2} \times (\text{the number of alleles of type 1 in individual } i \text{ at marker } r)$. 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 $\text{Cov}(Y_{i_r}, Y_{j_r}) = p_r(1 - p_r)\phi_{ij}$, where ϕ_{ij} is the kinship coefficient for i and j .
- Rearrange terms to see that $\phi_{ij} = \frac{\text{Cov}(Y_{i_r}, Y_{j_r})}{p_r(1 - p_r)}$

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

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

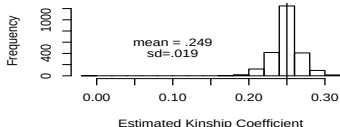
where $\hat{p}_r$ is an allele frequency estimate for the type 1 allele at marker r

Estimating Kinships Using GAW 14 COGA Data

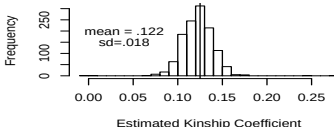
- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals. Individuals 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

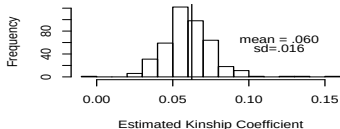
Hist w/ True Kinship = .25



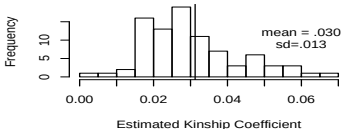
Hist w/ True Kinship = .125



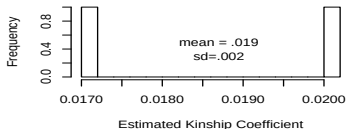
Hist w/ True Kinship = .0625



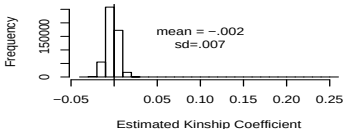
Hist w/ True Kinship = .03125



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

Kinship Estimation in Structured Populations

- Manichaikul A et al. (2010) propose a method to estimate kinship coefficients in the presence of population structure.
- Applicable to structured samples for which there is no admixture
- Gives accurate kinship coefficient estimates for individuals who are from the same population for relationships up to degree 3 (first-cousins)
- Gives negative estimates for unrelated individuals from different populations
- Two advantages of this method are
 - does not require allele frequency estimates at the marker: is based on the amount of IBS sharing for individuals
 - very fast computationally!

References

- Choi Y, Wijsman EM, Weir BS (2009). Case-control association testing in the presence of unknown relationships. *Genet. Epi.* **33**, 668-678.
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873.
- McPeck MS and Sun L (2000). Statistical Tests for Detection of Misspecified Relationships by Use of Genome-Screen Data, *Am. J. Hum. Genet.* **66**, 1076-1094.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575.

- Stankovich J, Bahlo M, Rubio JP, Wilkinson CR, Thomson R, Banks A, Ring M, Foote SJ, Speed TP (2005). Identifying nineteenth century genealogical links from genotypes. *Hum. Genet.* **117**, 188-199
- Thornton T, McPeck MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.