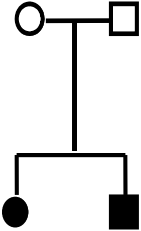
Estimating Relatedness in Homogenous Populations

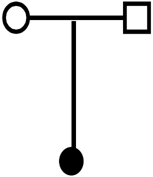
Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 7: Part I

Incomplete Genealogy

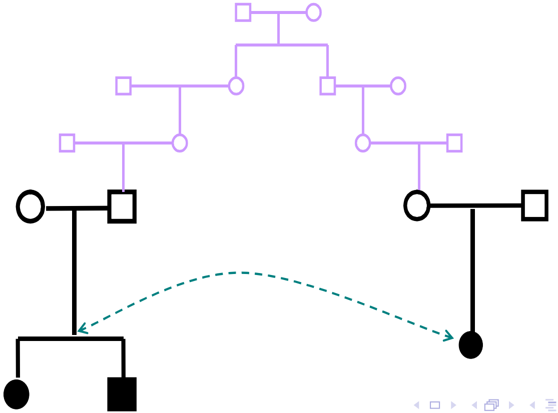
- Many statistical methods for genetic data, e.g. linkage and association methods, are based on assumptions of independent samples or samples with known relationships.





Incomplete Genealogy

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



Identifying Relative Pairs

- 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 $\leq$ 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 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

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 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 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 $E\left[X_2|\mathbf{G},\delta^0\right] =$

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

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

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

and

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

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 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 McPeek (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 $Y_{i_r} = \frac{1}{2}$ $\times$ (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

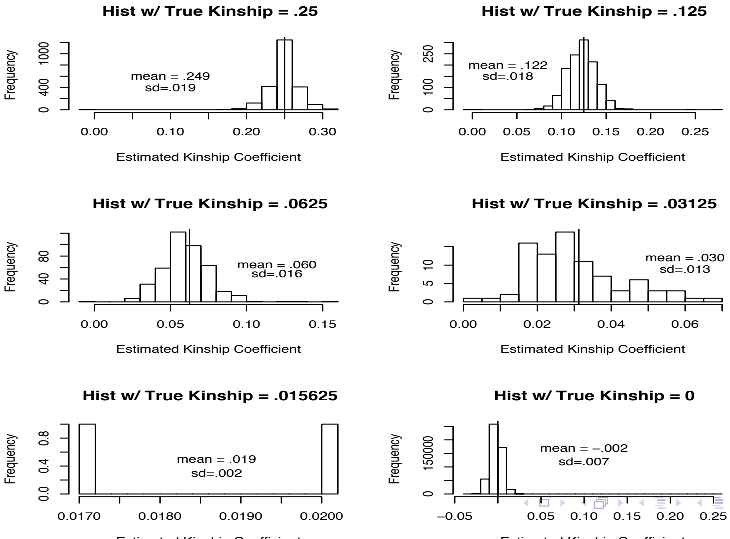
$$\hat{\phi}_{ij} = \frac{1}{N} \sum_{r=1}^{N} \frac{(Y_{i_r} - \hat{p}_r)(Y_{j_r} - \hat{p}_r)}{\hat{p}_r(1 - \hat{p}_r)}$$

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

References

- Choi Y, Wijsman EM, Weir BS (2009). Case-control association testing in the presence of unknown relationships. *Genet. Epi.* **33**, 668-678.
- McPeek 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.

References

- 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, McPeek MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.

Estimating Relatedness in Populations with Admixed Ancestry

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 7: Part II

Relatedness Inference in Structured Populations

- Popular algorithms for relationship inference are based on a strong assumption of population homogeneity
- This assumption is often untenable. GWAS often have cryptic population structure (or ancestry differences among the sample individuals)
- In samples with population structure, relationship estimation methods that assume homogeneity can give extremely biased results
- The degree of relatedness among related and unrelated sample individuals with similar ancestry are systematically inflated

Structured Populations with Distinct Ancestral Subpopulations

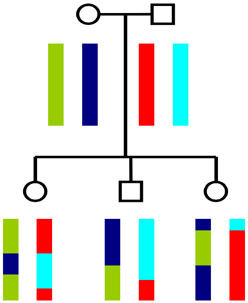
- Manichaikul A et al. (2010) propose an estimator, KING-robust, which stands for Kinship-based INference for Genome-wide association studies
- Estimates kinship coefficients in for individuals from ancestrally distinct subpopulations
- KING-robust estimates kinship coefficients for a pair of individuals by using the shared genotype counts as a measure of the genetic distance between the pair.
- Method does not require allele frequency estimates at the marker: is based on allele sharing counts for individuals
- Gives biased kinship estimates for individuals with different ancestry

Admixed Populations

- Genetic models used to identify related individuals from large scale genetic data often make simplifying assumptions about population structure – either random mating or simple structures.
- In reality, human populations do not mate at random nor are there simple endogamous subgroups.
- While GWAS have primarily examined populations of European ancestry, more recent studies involve admixed populations.
- A number of populations, including the two largest minority populations in the United States, Hispanics and African Americans, are known to have ancestral admixture of chromosomes from different continents.

Ancestry Admixture

- Consider two admixed parents, where each are admixed from different ancestral populations.
- In the picture below, positions on the chromosomes that are the same color are from the same ancestral population.



Relatedness Inference in Admixed Samples

- Thornton et al. (2012) proposed REAP (Relatedness Estimation in Admixed Populations) for relatedness inference in samples from populations with admixed ancestry
- Consider the problem of estimating relatedness in a set N of outbred individuals who are sampled from a population with admixture from K subpopulations
- Let $\mathbf{q}^s = (q_1^s, \ldots, q_K^s)^T$ denote the vector of subpopulation-specific allele frequencies at SNP s, where q_k^s is the allele frequency of SNP s in subpopulation k, $1 \leq k \leq K$.
- Define $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})^T$ to be the genome-wide ancestry vector for $i \in N$, where a_{ik} is the proportion of ancestry from subpopulation k for i, $a_{ik} \geq 0$ for all k, and $\sum_{k=1}^K a_{ik} = 1$.

Estimating Relatedness in an Admixed Population

- Let Y_i^s be the genotype variable for individual i, where $Y_i^s = \frac{1}{2} \times$ (the number of alleles of type 1 at SNP s in individual i). Similarly define Y_j^s for individual j.
- Conditional on $\mathbf{q}^s$, we assume alleles of an outbred individual i are independent, identically-distributed (i.i.d.) Bernoulli random variables , a modeling assumption made by other commonly-used models of population structure (Balding-Nichols model with admixture).
- We denote $\mu_i^s = E[Y_i^s | \mathbf{a}_i, \mathbf{q}^s]$ to be the expected value of Y_i^s conditional on $\mathbf{q}^s$ and $\mathbf{a}_i$ where

$$\mu_i^s = \mathbf{a}_i^T \mathbf{q}^s = \sum_{k=1}^{K} a_{ik} q_k^s,$$

- The variance of Y_i^s conditional on $\mathbf{q}^s$ and $\mathbf{a}_i$ is $.5\mu_i^s(1 - \mu_i^s)$.

Estimating Kinship Coefficients: Admixed Population

- For i and j from a homogenous populations, it can be shown that $\phi_{ij} = \frac{1}{2}\rho_{Y_i Y_j}$ for i and j , where $\rho_{Y_i Y_j}$ is the correlation of Y_i^s and Y_j^s.
- For estimating ϕ_{ij} in structured populations with admixture, we propose to similarly calculate the correlation of Y_i^s and Y_j^s
- Propose using a correlation that is calculated conditional on the admixture ancestry proportions of i and j as well as the subpopulation allele frequencies.

Estimating Kinship Coefficients: Admixed Population

- The conditional correlation that we estimate for inference on ϕ_{ij} is $\rho_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s}$, which is the correlation of Y_i^s and Y_j^s conditional on $\mathbf{a}_i$, $\mathbf{a}_j$, and $\mathbf{q}^s$.
- When genome-screen data is available for i and j we estimate ϕ_{ij} in the presence of population structure with admixture with the REAP estimator

$$\hat{\phi}_{ij}^{A} = \frac{1}{2}\hat{\rho}_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s}$$

where

$$\hat{\rho}_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s} = \frac{1}{|\mathcal{S}_{ij}|}\sum_{s\in\mathcal{S}_{ij}} \frac{(Y_i^s - \hat{\mu}_i^s)(Y_j^s - \hat{\mu}_j^s)}{\sqrt{.5\hat{\mu}_i^s(1-\hat{\mu}_i^s)}\sqrt{.5\hat{\mu}_j^s(1-\hat{\mu}_j^s)}},$$

Estimating IBD Sharing Probabilities: Admixed Populations

- Can also extend estimating IBD sharing probabilities in admixed populations.
- Define Z_{ij}^{s} as before to be an indicator for i and j sharing 0 alleles IBD at SNP s
- Can use the conditional expectation of Z_{ij}^{s} given $\mathbf{a}_i, \mathbf{a}_j, \mathbf{q}^s$ to obtain a method of moments estimator for δ_{ij}^{0} in the the presence of admixture.
- For any pair of individuals i and j from an admixed population,we have that

$$E(Z_{ij}^{s}|\mathbf{a}_i, \mathbf{a}_j, \mathbf{q}^s) = \left[(\mu_i^s)^2(1-\mu_j^s)^2 + (1-\mu_i^s)^2(\mu_j^s)^2\right]\delta_{ij}^{0}$$

Estimating IBD Sharing Probabilities: Admixed Populations

- Let $\mathcal{S}_{ij}$ be the set of markers in the genome screen for which both i and j have nonmissing genotype data.
- Our REAP method of moments for δ_{ij}^0 in the presence of admixture is

$$\hat{\delta}_{ij}^{0^A} = \frac{\sum_{s \in \mathcal{S}_{ij}} Z_{ij}^s}{\sum_{s \in \mathcal{S}_{ij}} \left[(\hat{\mu}_i^s)^2 (1 - \hat{\mu}_j^s)^2 + (1 - \hat{\mu}_i^s)^2 (\hat{\mu}_i^s)^2 \right]}$$

Estimating IBD Sharing Probabilities: Admixed Populations

- The remaining two IBD sharing probabilities, δ_{ij}^{1} and δ_{ij}^{2}, can be written as a function of δ_{ij}^{0} and ϕ_{ij}
- Estimate $\delta_{ij}^{1^A}$ with $\hat{\delta}_{ij}^{1^A} = 2 - 2\hat{\delta}_{ij}^{0^A} - 4\hat{\phi}_{ij}^{A}$
- Estimate $\delta_{ij}^{2^A}$ with $\hat{\delta}_{ij}^{2^A} = \hat{\delta}_{ij}^{0^A} + 4\hat{\phi}_{ij}^{A} - 1$.

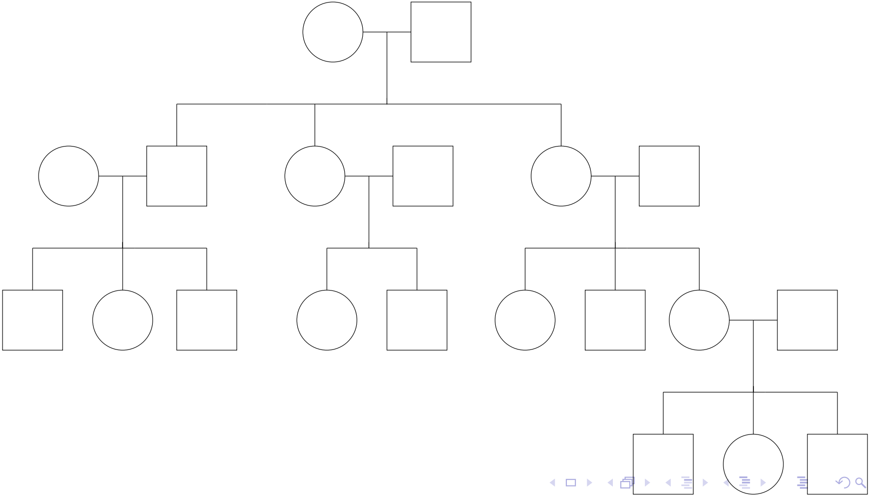
Simulation Studies: Relatedness and Population Structure

- Perform simulation studies, in which population structure and related individuals are simultaneously present
- The population structure settings used in the simulation studies are based on the Balding-Nichols model.
- For each SNP, an ancestral population allele frequency p was drawn from the uniform distribution on [0.1,0.9].
- We set $F_{ST} = .2$ in the Balding-Nichols model to simulate two highly divergent subpopulations.

Simulation Studies: Relatedness and Population Structure

- We consider population structure settings where individuals from an admixed population formed from two divergent subpopulations.
- Population structure setting 1 has individuals sampled from an admixed population formed from ancestral populations and where there is assortative mating.
- Population structure setting 2 has individuals sampled from an admixed population formed from ancestral populations where there is random mating
- We sample 400 individuals from 20 outbred pedigrees containing 1st, 2nd, 3rd, and 4th-degree relationships.

Pedigree Configuration

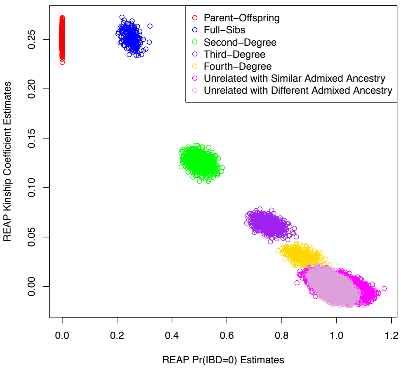


Simulation Studies: Relatedness and Population Structure

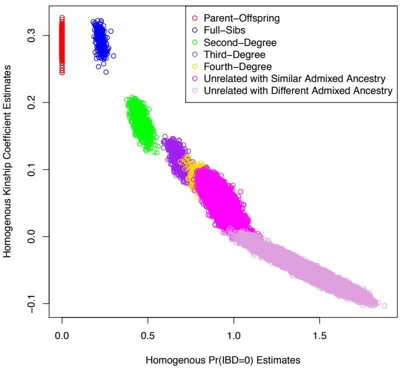
- For each of the two population structure settings we generate genotype data for 10,000 random SNPs.
- Genome-wide ancestry estimates used by REAP for the sample individuals were obtained by the *frappe* software program
- *frappe* implements an EM algorithm for simultaneously inferring each individuals ancestry proportion and allele frequencies in the ancestral populations.

Setting 1: Admixture from Two Ancestral Populations and Assortative Mating

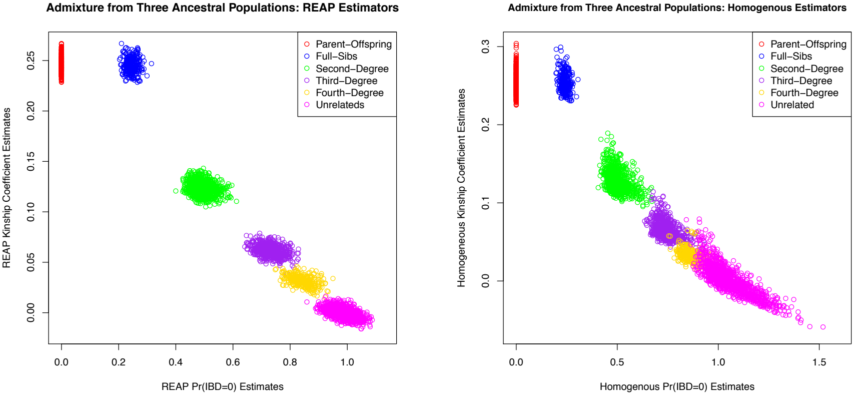
Admixture from Two Ancestral Populations: REAP Estimators



Admixture from Two Ancestral Populations : Homogenous Estimators

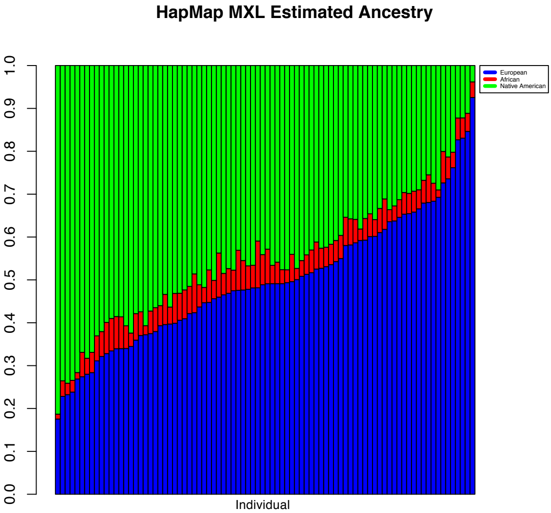


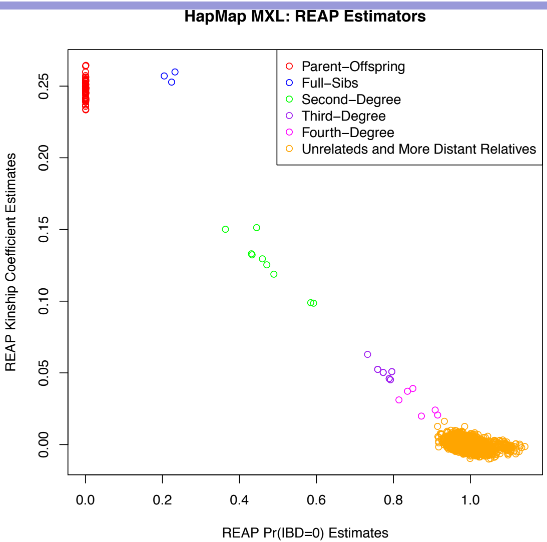
Setting 2: Admixture from Three Ancestral Populations and Random Mating



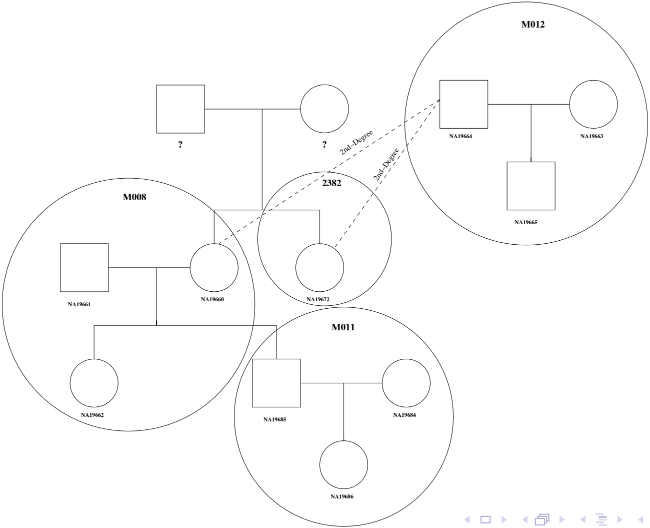
Estimating Kinship: HapMap Mex Sample

- Estimate estimating kinship coefficients and IBD sharing probabilities in the HapMap Mexicans in Los Angeles (MXL) sample of release 3 of phase III..
- Used *frappe* to estimate genome-wide ancestry for the 86 individuals in the sample
- We set the number of ancestral populations $K = 3$
  - HapMaP YRI for African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.


HapMap MXL Estimated Ancestry
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Individual
European
African
Native American


HapMap MXL: REAP Estimators
REAP Kinship Coefficient Estimates
REAP Pr(IBD=0) Estimates
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
○ Parent–Offspring
○ Full–Sibs
○ Second–Degree
○ Third–Degree
○ Fourth–Degree
○ Unrelateds and More Distant Relatives

Reconstructed HapMap MXL Extended Pedigree



Women's Health Initiative

- The Womens Health Initiative (WHI) is a national health study focusing on strategies for preventing chronic diseases in postmenopausal women.
- A total of 161,808 women aged 50-79 yrs. old were recruited from 40 clinical centers in the US between 1993 and 1998.
- The WHI cohort included
  - Two clinical trials of postmenopausal hormone therapy (estrogen alone and estrogen plus progestin)
  - A clinical trial of calcium and vitamin D supplements, and a dietary modification trial.

Genetic analysis of WHI-SHARe Minority Cohort

- Minority populations have largely been underrepresented in genetic studies despite bearing a disproportionately high burden for disease.
- WHI study opens up tremendous new possibilities for the identification of genetic risk factors associated with a number of clinical outcomes in the two largest minority populations in the U.S.
- The WHI SNP Health Association Resource (SHARe) minority cohort includes 8421 self-identified African American women from and 3587 self-identified Hispanic women
- 909,622 single nucleotide polymorphisms (SNPs) across the genome

Ancestry Estimation: WHI-SHARe data

- Used *frappe* to estimate genome-wide ancestry of every individual in the sample
- We set the number of ancestral populations $K = 4$
  - HapMaP YRI for African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.
  - HGDP East Asian samples for East Asian Ancestry

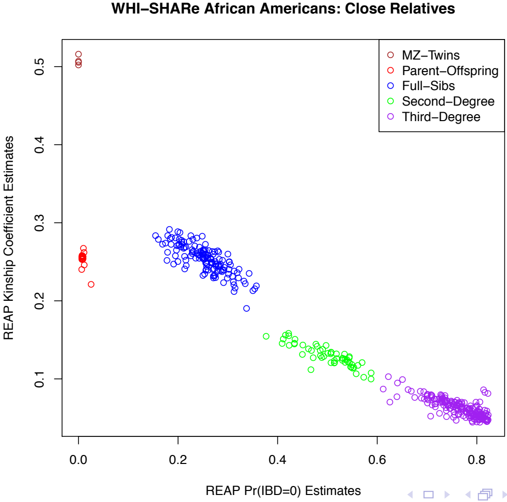
Relatedness Inference in WHI-SHARe

- No available genealogical information for the WHI-SHARe sample
- Used REAP to estimated relationships for all possible pairs:

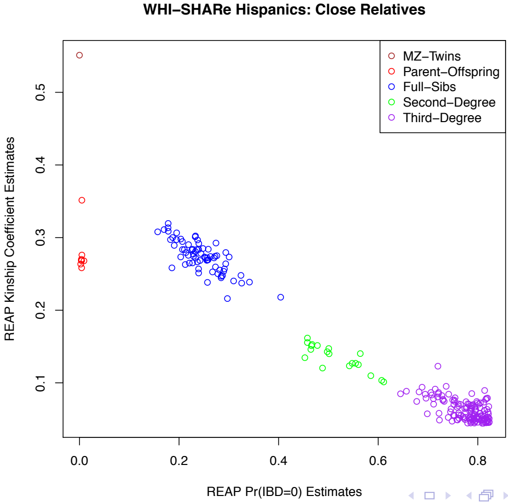
$$\binom{12008}{2} = 7,209,028$$

- Obtained estimates for kinship coefficients and IBD sharing probabilities

WHI-SHARe African Americans



WHI-SHARe Hispanics



Relatedness Inference in WHI-SHARe

- Also used the PLINK software (Purcell et al., 2007) method of moments kinship coefficient estimator: 8,932 pairs are identified to be either first or second degree relatives
- Our REAP kinship estimator that adjusts for individual specific ancestry identifies 344 individuals with kinship coefficients that are consistent with either first or second degree relatives

Relatedness Inference in WHI-SHARe

- Interestingly, there are individuals who are identified as second- and third-degree relative pairs by REAP but who have a different self-reported race/ethnicity, e.g. one individual is a self-report African American and the other is a self-report Hispanic.
- An advantage of the REAP approach is that robust relatedness estimates can be obtained for all individuals, even for individuals who have different admixed ancestry distributions and self-identify in different ethnic or nationality groups.

References

- 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.
- 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.
- Thornton T, Tang H, Hoffman TJ, Ochs-Balcom HM, Baan BJ, and Risch N (2012) Estimating Kinship in Admixed Populations *Am. J. Hum. Genet.* **91**