Methods for Cryptic Structure

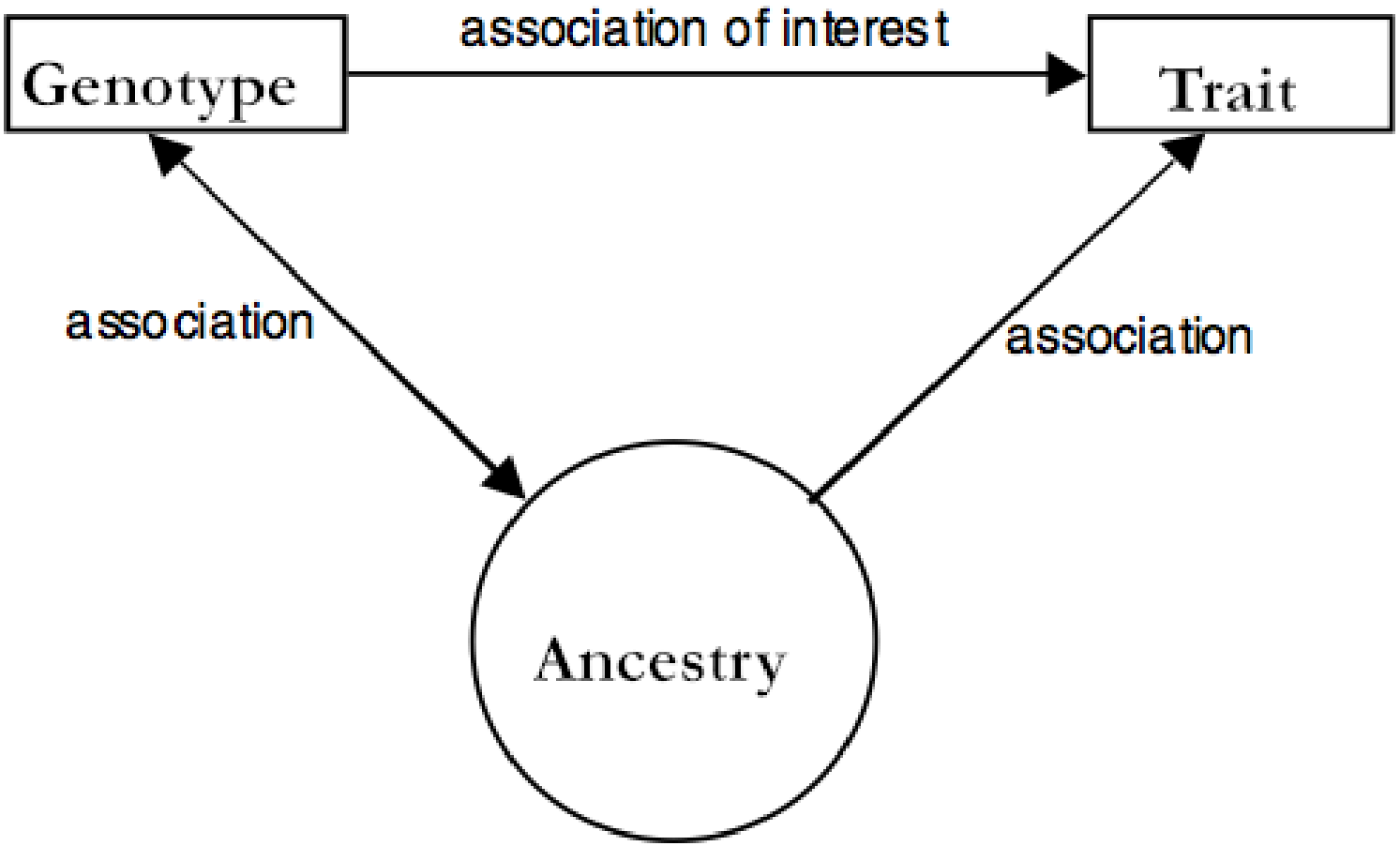
Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 10

Population Structure and Association Testing

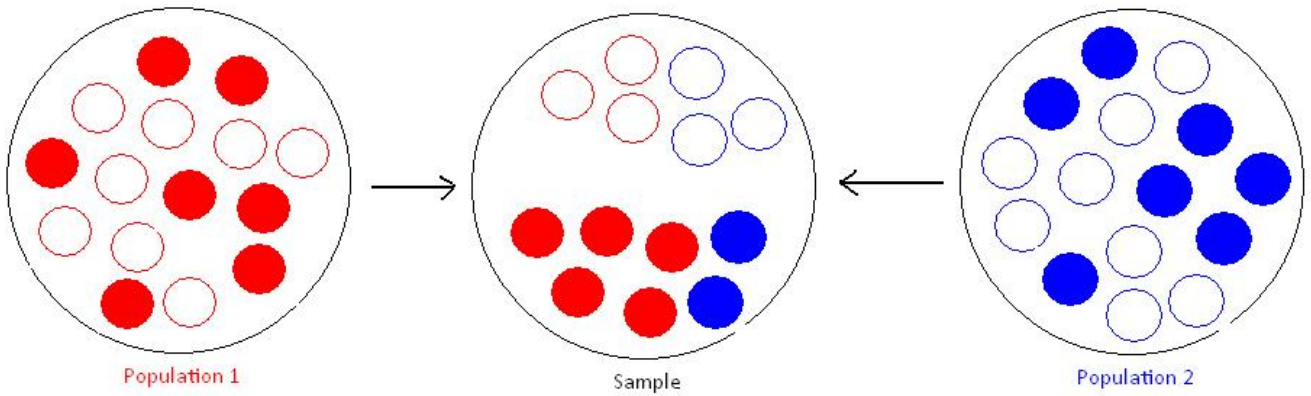
- The observations in genome-wide association studies can have several sources of dependence.
- Population structure, the presence of subgroups in the population with ancestry differences or admixed ancestry, is a major concern for association studies
- Population structure has long been recognized as a confounding factor in genetic association studies.
- Heterogeneous genomes of sample individuals can lead to both spurious association and reduced power if not properly accounted for

Confounding due to Ancestry



Spurious Association

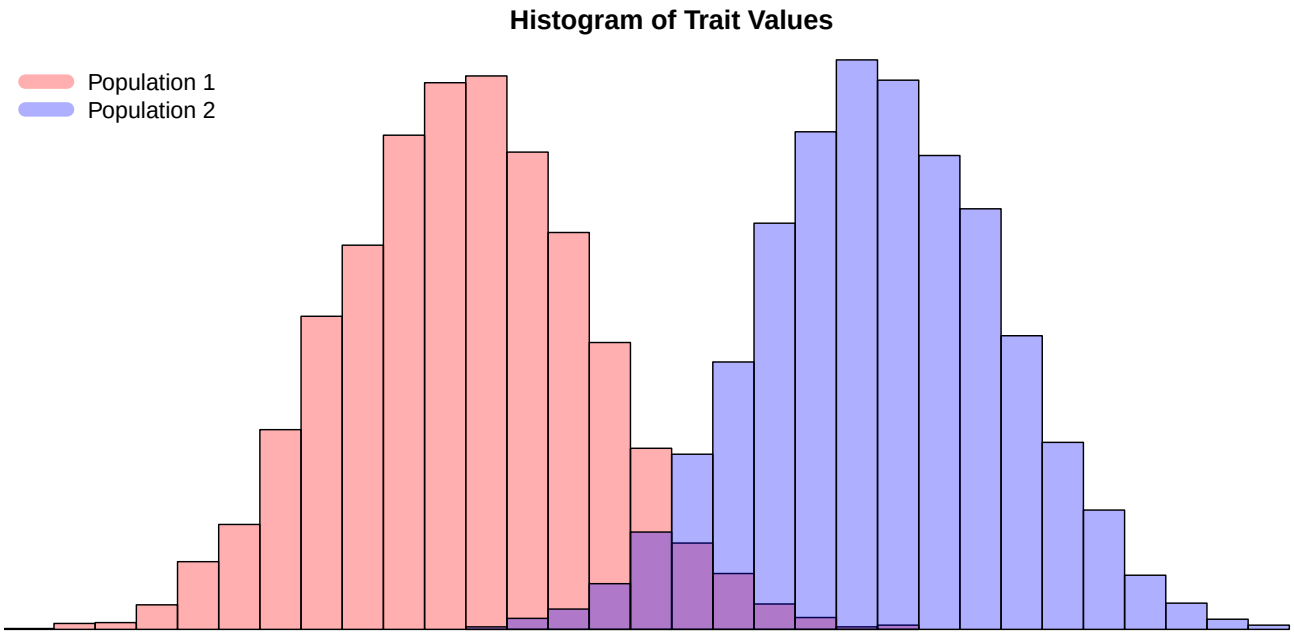
- Case/Control association test
  - Comparison of allele frequency between cases and controls.
- Consider a sample from 2 populations:



  - Red population overrepresented among cases in the sample.
  - Genetic markers that are not influencing the disease but with significant differences in allele frequencies between the populations
    $\Longrightarrow$ spurious association between disease and genetic marker

Spurious Association

- Quantitative trait association test
  - Test for association between genotype and trait value
- Consider sampling from 2 populations:



- Blue population has higher trait values.
- Different allele frequency in each population
  $\Longrightarrow$ spurious association between trait and genetic marker for samples containing individuals from both populations

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, i.e., the subpopulations are $1, 2, \ldots, K$.
- 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 subpopulation k, where $1 \leq k \leq K$, is a random draw from a beta distribution with parameters $p(1 - F_{st_k})/F_{st_k}$ and $(1 - p)(1 - F_{st_k})/F_{st_k}$, where $0 < p < 1$
- The parameter p can be viewed as the ancestral allele frequency and F_{st_k} can be viewed as Wright's standardized measure of variation for subpopulation k.

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 $X = (X_1, \ldots X_N)$ where X_i =the number of alleles of type 1 in individual i, so the value of X_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 k, then $Corr(X_i, X_j) = F_{st_k}$
  - If i and j are from different subpopulations then $Corr(X_i, X_j) = 0$
- The F_{st_k} values, the number of subpopulations K, and the subpopulation memberships for the sample individuals will generally be unknown when there is cryptic population structure.

Balding-Nichols Model: Covariance Structure

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

$$\mathbf{I_0} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

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

$$\mathbf{\Sigma_0} = \begin{pmatrix} 1 + F_{st_1} & F_{st_1} & \dots & 0 \\ F_{st_1} & 1 + F_{st_1} & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 + F_{st_K} \end{pmatrix},$$

Case-Control Design

- Methods have been proposed to correct for cryptic population structure in case-control studies by using data across the genome.
- Let N be the number of individuals in the study.
- Let $\mathbf{Y} = (Y_1, \ldots Y_N)$ be a phenotype indicator vector for case control status where $Y_i = 1$ if i is a case and $Y_i = 0$ 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 $1 \leq s \leq M$, and let $\mathbf{X}_s = (X_{1_s}, \ldots X_{N_s})$ where X_{i_s} =the number of alleles of type 1 in individual i at marker s.

Genomic Control

- Devlin and Roeder (1999) proposed correcting for substructure via a method called "genomic control."
- For each marker s, the Armitage trend statistic is calculated

$$A_{r_s} = N r^2_{X_s Y}$$

  where $r^2_{X_s Y}$ is the squared correlation between the genotype variable $\mathbf{X}_s$ for marker s and the binary phenotype variable $\mathbf{Y}$.
- 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 $\lambda = \frac{median(A_{r_1},...,A_{r_s},...A_{r_M})}{.456}$ as a correction factor for cryptic structure, where .456 is the median of a χ^2_1 distribution.
- 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.

Genomic Control

- Another way to view genomic control is as follows:

$$A_{r_s} = \frac{T^2}{Var_0(T)}$$

where T is a measure of allele frequency differences between cases and controls and $Var_0(T)$ is the variance of T under the null hypothesis

- For the Armitage statistic, $Var_0(T)$ is calculated assuming individuals are unrelated (calculation based on the identity matrix).
- Genomic control inflates this variance to account for the structure (unknown F_{st} values)

$$\tilde{A}_{r_s} = \frac{T^2}{\lambda Var_0(T)}$$

Identifying Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA is method for calculating the **principal components** that explain differences among the sample individuals in the genetic data
- To perform a PCA for population structure inference, first calculate an empirical covariance matrix $\hat{\mathbf{\Psi}}$ with components $\hat{\psi}_{ij}$:

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

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

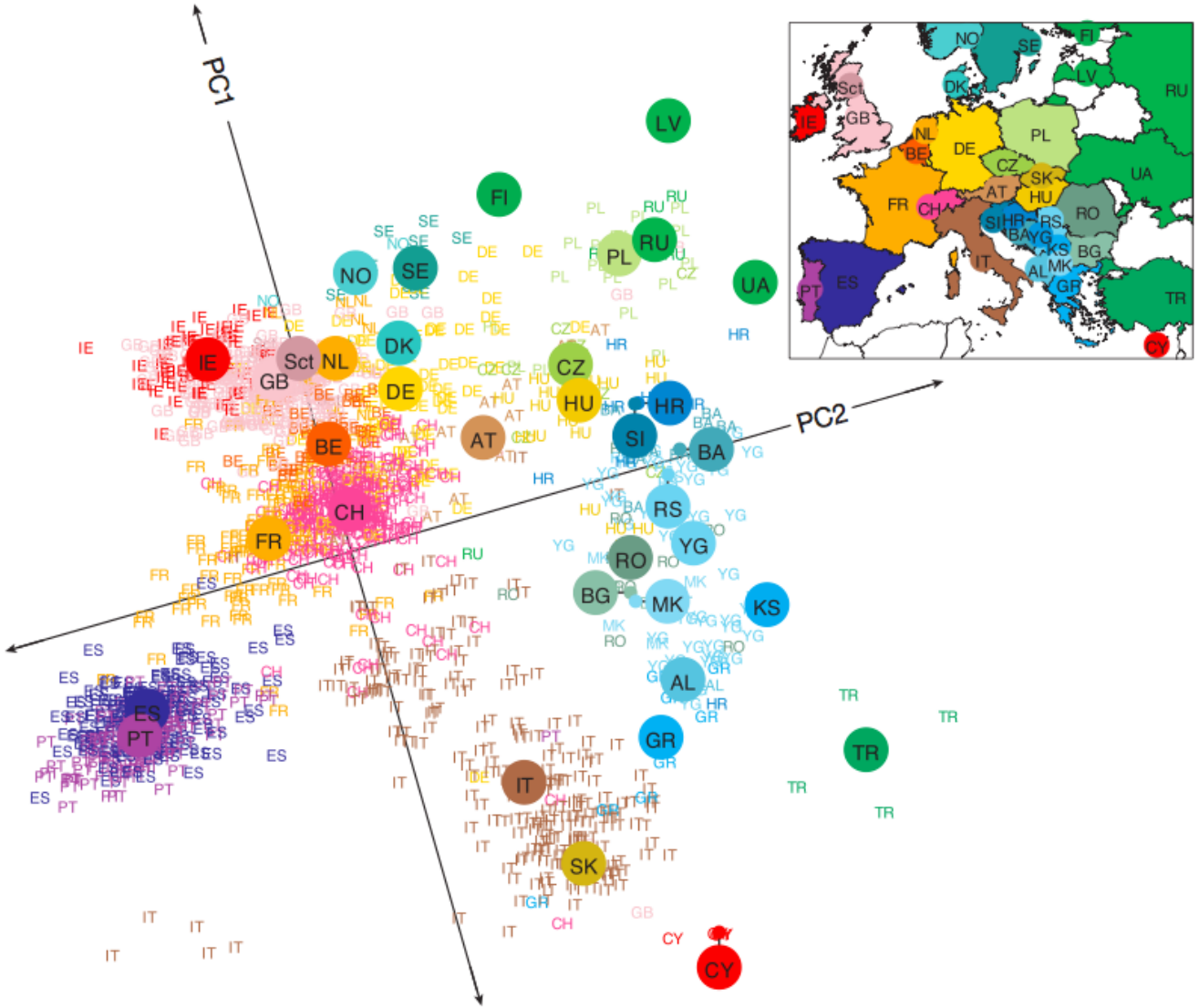
Identifying Population Structure with PCA

- Principal components (eigenvectors) for $\hat{\mathbf{\Psi}}$ can then be obtained via a singular value decomposition (SVD).
- 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.
- Does PCA actually work in practice?

PCA of Europeans

- An application of principal components to genetic data (Novembre et al. 2008) showed that among Europeans for whom all four grandparents originated in the same country, the first two principal components computed using 200,000 SNPs could map their country of origin quite accurately in the plane

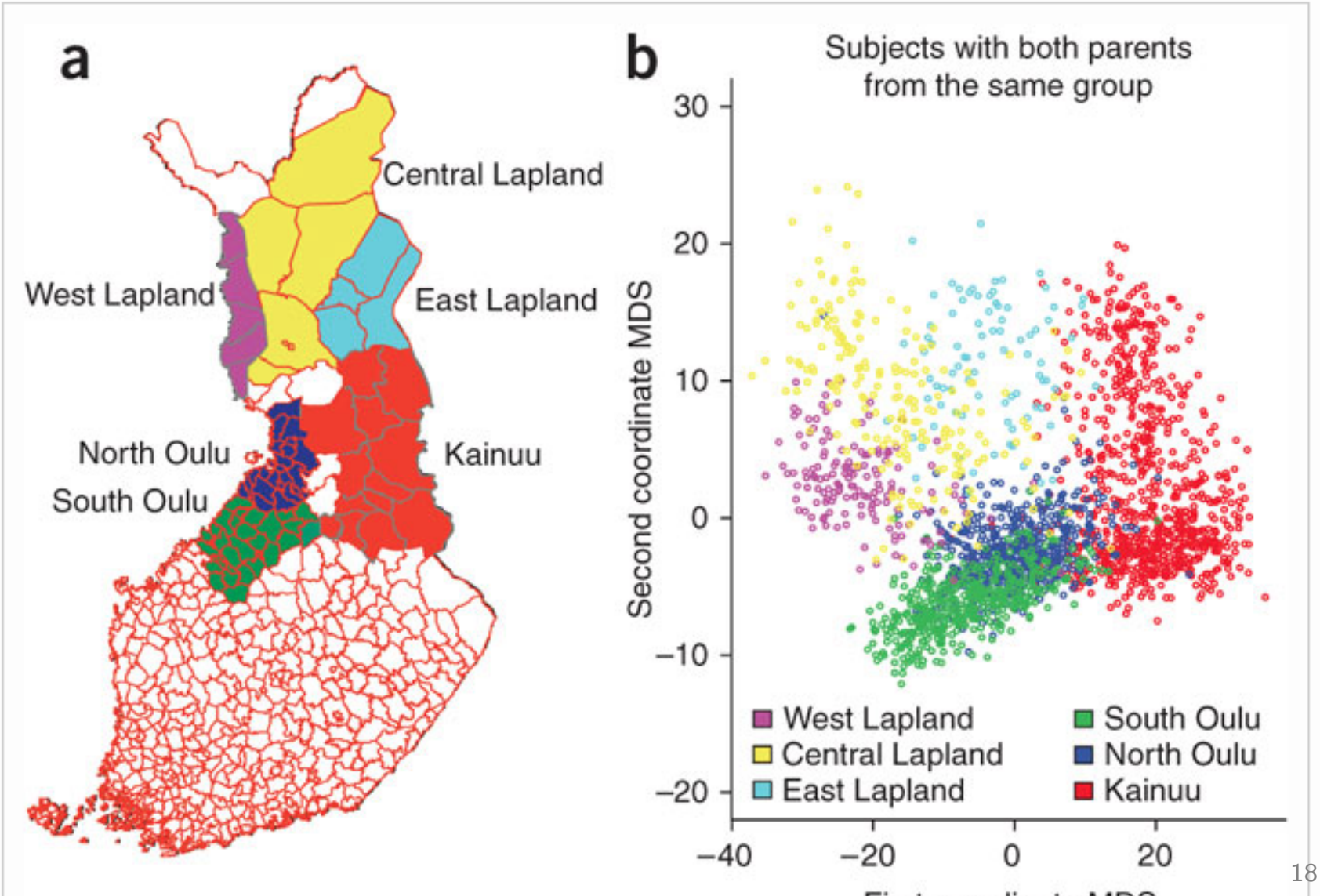
PCA of Europeans



PCA in Finland

- There can be population structure in all populations, even those that appear to be relatively "homogenous"
- An application of principal components to genetic data from Finland samples (Sabatti et al., 2008) identified population structure that corresponded very well to geographic regions in this country.

PCA in Finland



Principal Components Analysis: Association Testing

- Price et al. (2006) proposed corrected for structure in genetic association studies by using PCA
- They developed a method called EIGENSTRAT for association testing in structured populations where the top principal components (highest eigenvalues) from a PCA can be used as covariates in a multi-linear regression.

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

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

Computation of Eigenstrat

- The kth axis of variation is defined to be the kth eigenvector of $\hat{\Psi}$ (that is, the eigenvector with kth largest eigenvalue).
- Thus, the ancestry a_{jk} of individual j along the kth axis of variation equals coordinate j of the kth eigenvector.
- Adjustments of genotypes for ancestry
- Let X_{js} be the genotype of individual j ($X_{js} = 0$, 1 or 2) at marker s, and let a_j be the ancestry of individual j along a given axis of variation.
- $X_{js}^{adjusted} = X_{js} - \gamma_s a_j$, where $\gamma_s = \sum_i a_i X_{is} / \sum_i a_i^2$
- γ_s is a regression coefficient for ancestry predicting genotype across individuals j with valid genotypes at marker s.
- A similar adjustment is performed for each axis of variation.

Computation of Eigenstrat

- Adjustments of phenotypes for ancestry
- $Y_j = 1$ if j is a case and $Y_j = 0$ if j is a control
- $Y_j^{adjusted} = Y_j - \delta a_i$, where $\delta = \sum_i a_i Y_i / \sum_i a_i^2$
- The EIGENSTRAT statistic adjusted for ancestry at marker s is

$$EIG_s = (N - K - 1) \times \left[corr \left(Y_s^{adjusted}, X^{adjusted} \right) \right]^2$$

where N is the number of subjects in the sample and K is the number of axes of variation used to adjust for ancestry.

- EIG_s will approximately follow a χ^2 distribution with 1 degree of freedom.

Structured Samples with Related Individuals

- The methods discussed so far have been developed for samples with unrelated individuals
- Methods may not be valid in samples with related individuals (known and/or unknown)
- Many genetic studies have samples with related individuals
- Consider testing for association between a trait and a single genetic marker in a sample that includes related individuals.
- Assume that there are some related individuals in the sample from a structured population.
- If individuals i and j are from population k and they are related, the correlation, however, is no longer F_{st_k}!
- The correlation for the pair is now ψ_{ij}, where ψ_{ij} is a function of both the kinship coefficient for individuals i and j and F_{st_k}.

ROADTRIPS

- The ROADTRIPS approach of Thornton and McPeek (2010) incorporates an empirical covariance matrix $\hat{\boldsymbol{\Psi}}$.
- The top principal components in EIGENSTRAT are not able to capture the complicated covariance structure due to the related individuals in the samples.
- Instead of taking the top principal components of the matrix $\hat{\boldsymbol{\Psi}}$, ROADTRIPS uses the entire matrix to correct the variance of a general class of statistics for unknown structure
- ROADTRIPS extensions have been developed for a number of association tests including Pearson χ^2 test, the Armitage trend test, the corrected χ^2 test, the W_{QLS} test, and the M_{QLS} test.
- ROADTRIPS is a valid association method for partially or completely unknown population and pedigree structure.

EMMAX: Linear Mixed Effects Models For Structure

- Kang et al. [Nature Genet, 2010] proposed the EMMAX variance components method:

$$Y = \beta_0 + \beta_1 X + \mathbf{g} + \boldsymbol{\epsilon},$$

$$\text{with } \mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \hat{\boldsymbol{\Psi}}) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$$

- The variance components, σ_g^2 and σ_e^2, are then estimated based on this model using either a maximum likelihood or restricted maximum likelihood (REML) approach.
- The null hypothesis of no association is $H_0 : \beta_1 = 0$, and the alternative hypothesis is : $H_A : \beta_1 \neq 0$
- The EMMAX statistic is the score statistic for testing the null hypothesis of $\beta_1 = 0$.
- Similar mixed effects methods proposed by Lippert et al. (2011, Nat Methods) and Stephens et al. (2012, Nat Genet).

Structured Association Method

- Pritchard et al. (2000) developed the program STRUCTURE
- This method uses MCMC to cluster genetically similar individuals into "subpopulations."
- Association testing is performed within each of the subpopulations
- Assignments of individuals are highly sensitive to the number of subpopulations K in the sample, which is often unknown when there is cryptic structure
- Not computationally feasible for genome-wide association studies

Structured Association Method

- Let K be the number of populations, Let N be the number of individuals in the sample. We first consider the case when K is known.
- We assume that each individual originates in one of the K populations, i.e., there is no admixture.
- Let L denote the number of loci and let the vector X denote the genotypes of the loci for the individuals.
- Let the vector Z be the populations of origin for the individuals. Let P be the allele frequencies for the populations. The vectors Z and P are both unknown

Structured Association Method

- The elements of the vectors are as follows:

$$
\begin{aligned}
(x_l^{(i,1)}, x_l^{(i,2)}) &= \text{genotype of ith individual at the lth locus,} \\
&\quad \text{where } i = 1, 2, ..., N \text{ and } l = 1, 2, ..., L \\
z_i &= \text{population from which ith individual originated} \\
p_{klj} &= \text{frequency of allele j at lth locus in populaion k} \\
&\quad \text{where } k = 1, 2, ..., K \text{ and } j = 1, 2, ..., J_l
\end{aligned}
$$

- Now, given the allele frequencies and the population of origins, we have that $x_l^{(i,a)}$ are independent and $Pr(x_l^{(i,a)} = j | Z, P) = p_{z_i lj}$.

Prior Distributions

- For the prior distribution of Z, we have that the z_i's are independent and each one follows a discrete uniform distribution where $Pr(z_i = k) = \frac{1}{K}$ for $k = 1, 2, ..., K$.
- This prior seems reasonable since we have no information on the population of origin before observing any genotype information
- For population k at locus l we have have that the prior distribution of $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l}) \sim D(\lambda_1, \lambda_1, ..., \lambda_{J_l})$ where D is a Dirichlet distribution.
- Before observing any genotype data we have no information about the allele frequency distributions. To account for our ignorance of this, we make $\lambda_1 = \lambda_2 = ... = \lambda_{J_l} = 1$.

The Gibbs Sampling MCMC Algorithm

- Let x^i denote the observed genotype data for individual i for all loci. Now

$$
\begin{aligned}
Pr(z_i = k|X, P) &= \frac{Pr(x^i|z_i = k, P)Pr(z_i = k|P)Pr(P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)Pr(z_i = k'|P)Pr(P)} \\
&= \frac{Pr(x^i|z_i = k, P)\frac{1}{K}Pr(P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)\frac{1}{K}Pr(P)} \\
&= \frac{Pr(x^i|z_i = k, P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)}
\end{aligned}
$$

Note that $Pr(x_l^{(i,a)} = j|Z, P) = p_{z_i l j}$, for individual i, population z_i, locus l, and allele j. So we can easily obtain likelihood of the genotype data of individual i since $Pr(x^i|z_i = k, P) = \prod_{l=1}^{L} p_{klx_l^{(i,1)}} p_{klx_l^{(i,2)}}$

The Gibbs Sampling MCMC Algorithm

- We will now give the full condition for P. We have that $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l})$ and

$$Pr(p_{kl}|X, Z) \quad \propto \quad p_{kl1}^{\lambda_1 + n_{kl1}} \cdot p_{kl2}^{\lambda_2 + n_{kl2}} \cdots p_{klJ_l}^{\lambda_{J_l} + n_{klJ_l}}$$

where n_{klj} is the number of copies of allele j found at locus l in individuals that are assigned to population k. So, $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l})$ is a $D(\lambda_1 + n_{kl1}, \lambda_1 + n_{kl2}, ..., \lambda_{J_l} + n_{klJ_l})$

The Gibbs Sampling MCMC Algorithm

- Now that we have the full conditionals, we can do implement the Gibbs sampler.
- Step 0: Start with an initial value of Z by randomly drawing from the discrete uniform distribution $\frac{1}{k}$ for each component. Denote this initial value as Z^0.
- Step 1: Draw P^n from $Pr(P|X, Z^{n-1})$
- Step 2: Draw Z^n from $Pr(Z|X, P^n)$
- For n=1,2,..., repeat Step 1 and Step 2.

The Gibbs Sampling MCMC Algorithm: Example

- Two random mating populations were simulated. Each individual in the sample comes are from one of the two populations.
- Prichard et al. (2000) performed simulations under this model for 5 loci and for 15 loci.
  - The number of individuals in their sample was is over 100.
  - Algorithm performs very well in clustering individuals for the 5 loci case and is almost perfect for the 15 loci case
- Assigning individuals to originating populations depend on a number of factors which include the number of individuals in the sample which in turn affects the accuracy of P, the number of loci, and the allele frequency within the populations.
- Interest in obtaining clustering results if the number of loci is decreased to 4 and the sample size is reduced to 20 (9 from one population 1 and 11 from population 2).

The Gibbs Sampling MCMC Algorithm: Example

Simulated from the following two populations:

Allele Frequencies at Each Locus for The Two Populations

	Locus 1	Locus 2	Locus 3	Locus 4
Population 1	Allele 1: .90 Allele 2: .10	Allele 1: .20 Allele 2: .80	Allele 1: .10 Allele 2: .20 Allele 3: .70	Allele 1: .50 Allele 2: .30 Allele 3: .20
Population 2	Allele 1: .10 Allele 2: .90	Allele 1: .80 Allele 2: .20	Allele 1: .70 Allele 2: .20 Allele 3: .10	Allele 1: .20 Allele 2: .10 Allele 3: .70

The Gibbs Sampling MCMC Algorithm: Example

Genotypes of Sample Individuals at Each Locus

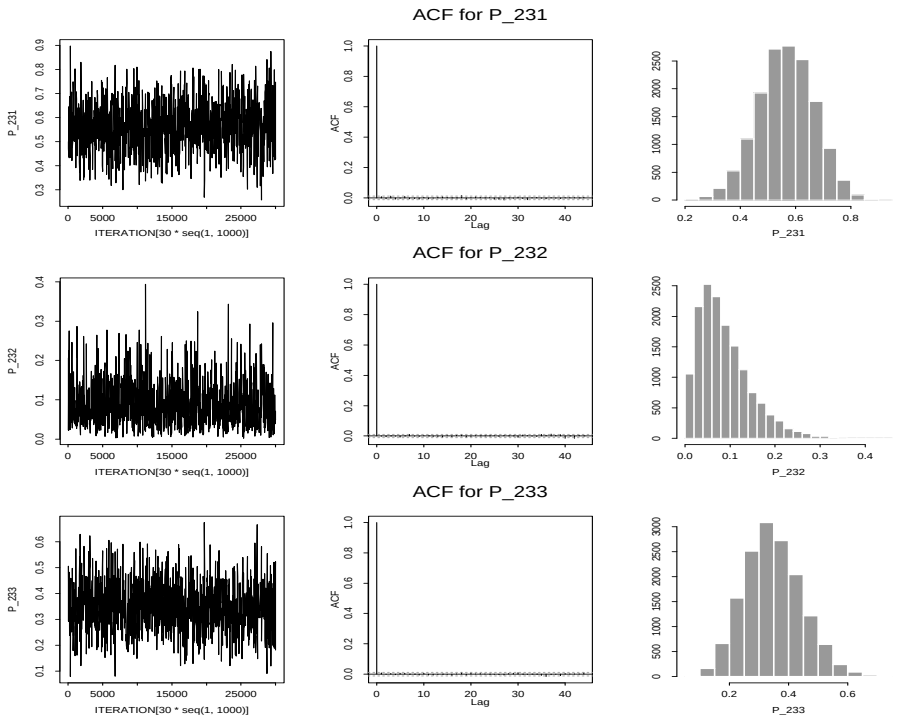
	Locus 1	Locus 2	Locus 3	Locus 4
Person 1	2, 2	1, 2	1, 1	3, 3
Person 2	2, 2	1, 1	1, 1	3, 3
Person 3	2, 2	2, 2	2, 3	1, 1
Person 4	2, 1	1, 1	1, 1	2, 3
Person 5	2, 2	1, 2	1, 1	1, 1
Person 6	2, 2	2, 1	1, 1	3, 1
Person 7	1, 1	1, 2	2, 3	3, 1
Person 8	1, 1	2, 2	1, 3	3, 1
Person 9	1, 1	2, 2	3, 1	3, 1
Person 10	1, 1	2, 1	2, 2	2, 3
Person 11	2, 2	1, 2	1, 1	3, 2
Person 12	2, 2	1, 1	2, 1	1, 3
Person 13	1, 1	2, 1	3, 3	3, 1
Person 14	2, 1	1, 2	1, 1	3, 1
Person 15	1, 1	2, 2	3, 3	3, 2
Person 16	1, 1	2, 2	3, 3	1, 1
Person 17	1, 2	1, 1	1, 1	1, 1
Person 18	2, 2	1, 1	1, 3	3, 3
Person 19	1, 1	2, 2	3, 2	2, 1
Person 20	2, 2	1, 1	3, 2	2, 3

Individuals from population 1: 3, 7, 8, 9, 10, 13, 15, 16, and 19.

Individuals from population 2: 1, 2, 4, 5, 6, 11, 12, 14, 17, 18, and 20.

The Gibbs Sampling MCMC Algorithm: Example

Below are plots for the allele frequency parameters for population 2 at locus 3. The histograms are from samples following the 15,000th iteration (30,000 total iterations).



The Gibbs Sampling MCMC Algorithm: Example

Marginal Distribution of Z

	TRUE POPULATION	Pr(Z=1)	Pr(Z=2)	Correct Infer.
Person 1	2	0.1202667	0.8797333	*YES*
Person 2	2	0.1287333	0.8712667	*YES*
Person 3	1	0.3936667	0.6063333	*NO*
Person 4	2	0.4176667	0.5823333	*YES*
Person 5	2	0.06886667	0.9311333	*YES*
Person 6	2	0.0774	0.9226	*YES*
Person 7	1	0.7792667	0.2207333	*YES*
Person 8	1	0.5047333	0.4952667	*UNSURE*
Person 9	1	0.4913333	0.5086667	*UNSURE*
Person 10	1	0.9305333	0.06946667	*YES*
Person 11	2	0.2615333	0.7384667	*YES*
Person 12	2	0.2529333	0.7470667	*YES*
Person 13	1	0.7080667	0.2919333	*YES*
Person 14	2	0.07993333	0.9200667	*YES*
Person 15	1	0.8038667	0.1961333	*YES*
Person 16	1	0.63453337	0.3654667	*YES*
Person 17	2	0.1385333	0.8614667	*YES*
Person 18	2	0.1385333	0.746	*YES*
Person 19	1	0.8957333	0.1042667	*YES*
Person 20	2	0.7332667	0.2667333	*NO*

The posterior marginal distribution of Z was obtained and used to determine whether the individuals were properly classified. An individual is considered to be properly classified if the mode of the

References

- Balding DJ, Nichols RA (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. *Genetica* **96**, 3-12.
- Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909.

References

- Pritchard J, Stephens M, Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.Y., Freimer, N. B., Sabatti, C. Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-354.
- 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.