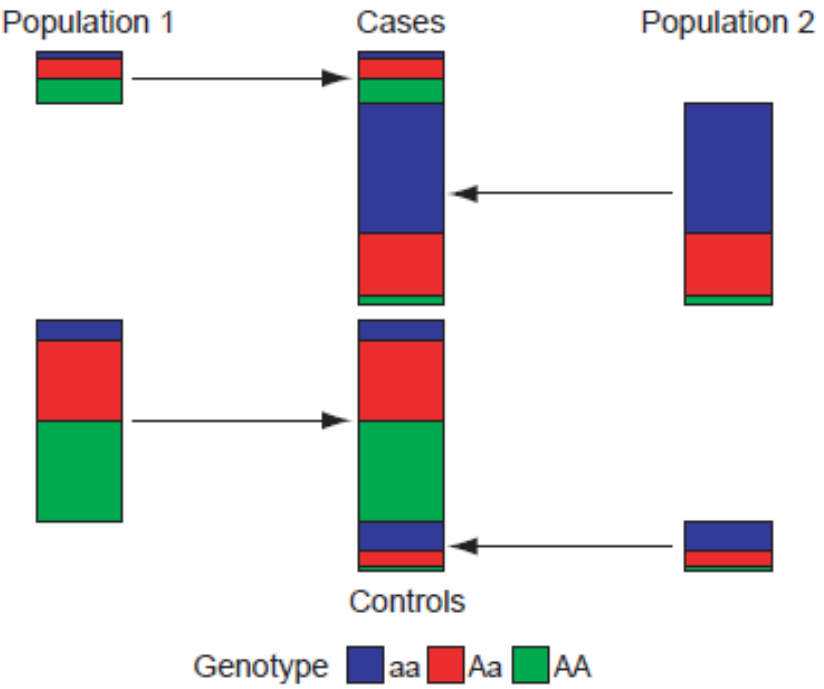
Population Structure and Association Analysis

02-715 Advanced Topics in Computational Genomics

Population Structure and Association Analysis

- Population structure in data causes false positives
  - Samples in the case population are usually more related
  - Any SNPs more prevalent in the case population will be found significantly associated with the trait.



Accounting for Population Structure in Association Analysis

- Needs to account for population structure in association mapping.

- Careful study design with each population represented in case/control groups in a balanced way.
  - Can be hard to control
  - The effect of cryptic population structure

Family-based Design vs. Population-based Design

- Family-based studies
  - The effect of population structure can be controlled by the use of parents' genotypes.
  - In practice, collecting genotypes from multiple individuals in a family can be hard. (e.g., late-onset diseases)

- Population-based design
  - Data collection is easier for a large number of unrelated individuals than a large number of families.
  - The control samples can be reused in different studies.

Accounting for Population Structure in Association Analysis

- Family-based method
  - Transmission disequilibrium test (TDT)
- Population-based method
  - Genomic control (Devlin & Roeder, Biometrics 1999)
  - Structured association (Pritchard et al., AJHG 2000)
  - EigenStrat: principal component analysis (Price et al., Nature Genetics 2006)

Transmission Disequilibrium Test (TDT)

- Genotype affected individuals and their parents (trio)

	Non-transmitted alleles		
Transmitted alleles	*M*	*m*	total
M	a	b	a+b
m	c	d	c+d
Total	a+c	b+d	2N

- Null hypothesis: (b/(b+c), c/(b+c)) is compatible with (0.5, 0.5)
  - Test statistic is given as $(b-c)^2/(b+c)$
- The non-transmitted alleles play the role of controls

Genomic Control (GC)

- Idea: Use the SNPs that are not associated with the trait to remove the effect of population stratification

- Genotype data consist of
  - Candidate genes to be tested
  - *L* supplementary loci (null loci) for estimating the inflation factor λ

- GC uses the inflation factor λ to correct the association statistic of the SNP in the candidate gene

- Limitation: the inflation factor λ is assumed to be the same across the genome, ignoring population admixture

Devlin & Roeder, Biometrics 1999

STRAT: Structured Association

(Pritchard et al., AJHG 2000)

- Idea: Within each subpopulation, an association between a genetic marker and the trait is a true association.

- Two-stage method
  - Step 1: Using Structure (Pritchard et al., Genetics 2000) and unlinked genetic markers,
    - estimate the population structure
    - assign sampled individuals to putative subpopulations
  - Step 2:
    - Test for association within the subpopulations inferred in Step 1

- Limitation
  - Running Structure is computationally demanding

Pritchard et al., AJHG 2000

STRAT: Step 2

- Given ancestry proportions $q_k^{(i)}$ for population *k*, individual *i* estimated by STRUCTURE
- H_0: The probability model for genotypes *c*'s under the null hypothesis of no association

$$\Pr_0[c^{(i,a)} = j \mid Q, P_0, \Phi] = \sum_k q_k^{(i)} p_{kj}$$

- H_1: The probability model for genotypes *c*'s the alternative hypothesis of association

$$\Pr_1[c^{(i,a)} = j \mid Q, P_1, \Phi] = \sum_k q_k^{(i)} p_{kj}^{[\phi^{(i)}]}$$

STRAT: Step 2

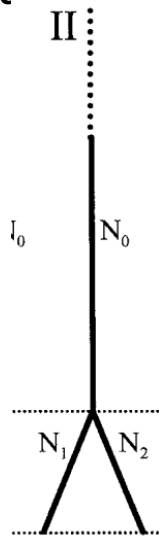
- Likelihood ratio test:

$$\Lambda = \frac{\Pr_1(C; \hat{P}_1, \hat{Q})}{\Pr_0(C; \hat{P}_0, \hat{Q})}$$

  - Large values indicate that the alternative hypothesis explains the data better.

Simulation Studies: No Admixture

- Assume two discrete populations
- Simulate genotypes of 150 affected and 150 control individuals at 100 unlinked loci
  - With sample size N, we have 2N chromosomes
  - Assume two populations have split 0.05N generations ago without migration
  - Controls: half of the controls came from each of the two subpopulations
  - Affected group: 100 from population 1, 50 from population 2



STRAT: Simulation Results

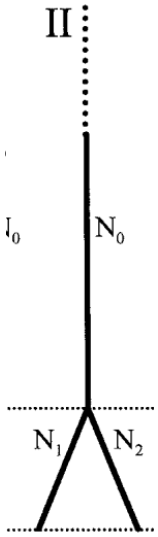
1A. Discrete Populations

p_1, p_2	STRAT	TDT	χ^2
	$P = .05$		
.5, .5	.050	.050	.049
.1, .1	.048	.049	.049
.5, .1	.050	.050	.437
.9, .1	.049	.050	.769
	$P = .01$		
.5, .5	.011	.010	.011
.1, .1	.009	.010	.009
.5, .1	.009	.010	.260
.9, .1	.010	.009	.649
	$P = 10^{-3}$		
.5, .1	0.83×10^{-3}	0.98×10^{-3}	.112
.9, .1	0.83×10^{-3}	0.90×10^{-3}	.506
	$P = 10^{-4}$		
.5, .1	0.97×10^{-4}	0.95×10^{-4}	.046
.9, .1	0.80×10^{-4}	0.76×10^{-4}	.370

- Rejection rates under the null hypothesis of no association
- p_1,p_2: allele frequencies for populations 1 and 2 at the given locus

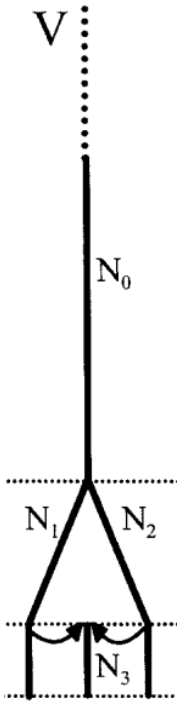
Simulation Studies: With Admixture

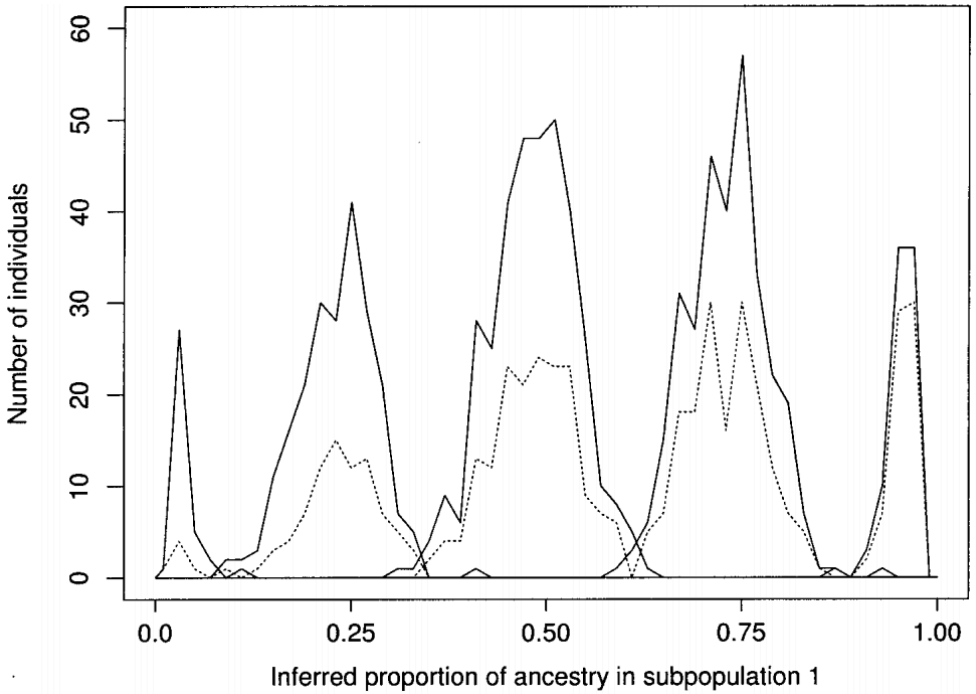
- Assume two discrete populations
- Simulate genotypes of 500 affected and 500 control individuals at 150 unlinked microsatellite loci
  - With sample size N, we have 2N chromosomes
  - Assume two populations have split 0.15N generations ago, followed by two generations of admixing
  - Controls: random draws from the whole population
  - Affected group: random draws from the whole population assuming disease risk mode for grand parents



Structure: Simulation Results

- Learning population structure using genotypes from two recently admixed populations
  - Dashed line – case group





STRAT: Simulation Results

p_1, p_2	STRAT	TDT	χ^2
	$P = .05$		
.5, .5	.048	.049	.054
.1, .1	.050	.049	.051
.5, .1	.044	.050	.623
.9, .1	.033	.050	.998
	$P = .01$		
.5, .5	.010	.010	.010
.1, .1	.010	.009	.010
.5, .1	.008	.010	.370
.9, .1	.005	.010	.979
	$P = 10^{-3}$		
.5, .1	0.83×10^{-3}	0.86×10^{-3}	.156
.9, .1	0.45×10^{-3}	1.06×10^{-3}	.874
	$P = 10^{-4}$		
.5, .1	0.85×10^{-4}	0.38×10^{-4}	.050
.9, .1	0.61×10^{-4}	0.92×10^{-4}	.662

- Rejection rates under the null hypothesis
- p_1,p_2: allele frequencies for populations 1 and 2 at the given locus

TDT vs. STRAT

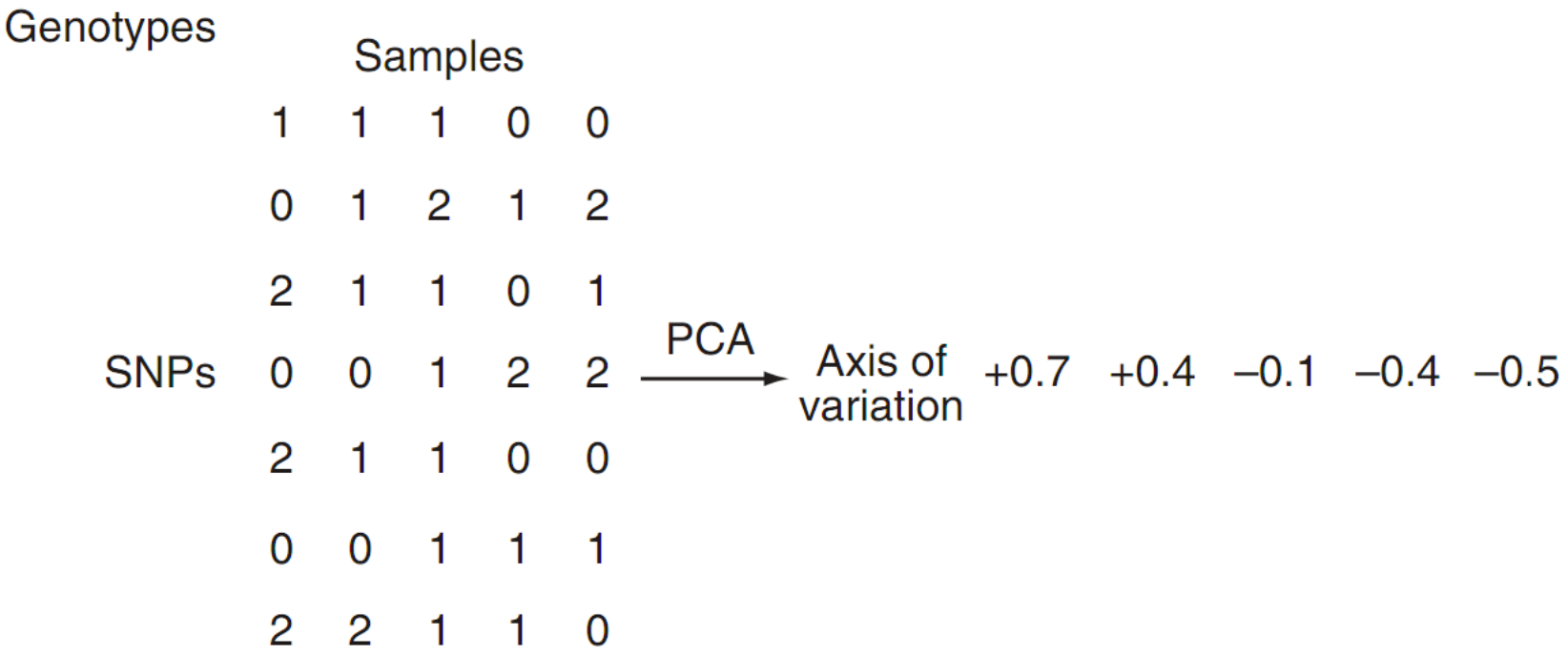
- TDT
  - Requires genotyping parents of the affected offspring
- STRAT
  - Requires genotypes for additional loci to infer population structure with *STRUCTURE*

EigenStrat

- Structured association approach
- Step 1: Run PCA on genotype data to infer the population structure
- Step 2: Perform association analysis after correcting for the population effects in genotype/phenotype data
- Advantages: low computational cost compared to STRAT

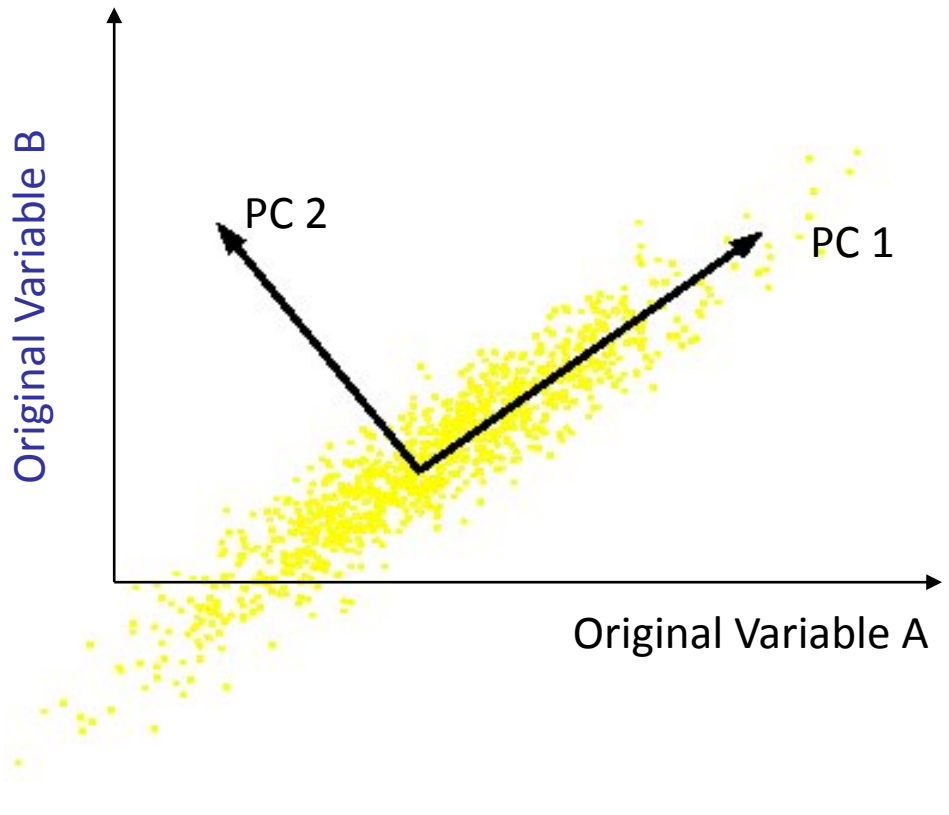
EigenStrat: Structured Association with PCA

- Step 1: (Inferring Ancestry) PCA is applied to genotype data to infer continuous axes of genetic variation



Price et al., Nature Genetics 2006

What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

EigenStrat: Structured Association with PCA

- Step 2: (Removing Ancestry Effects) Genotype at a candidate SNP and phenotype are continuously adjusted by amounts attributable to ancestry along each axis

Candidate SNP	2	2	1	1	0	→	1.0	1.4	1.1	1.6	0.8
Phenotype	1	1	0	0	0	→	0.3	0.6	0.1	0.4	0.5

- Step 3: (Association test) $\chi^2 = 0.07 \Longrightarrow$ no association

Simulation Procedure

- Given F_{ST}, For each SNP
  - Draw an ancestral population allele frequency p from uniform distribution [0.1 0.9]
  - Allele frequencies for populations 1 and 2, p_1 and p_2, are drawn from Beta($p(1-F_{ST})/F_{ST}$, $(1-p)(1-F_{ST})/F_{ST}$)
  - Draw SNPs using population allele frequencies p_1 and p_2

Simulation Study

- Discrete populations vs. admixed populations
- Moderate vs. extreme ancestry differences in the ancestry between cases/controls
  - Moderate: control (40% population 1, 60% population 2), case (60% population 1, 40% population 2)
  - Extreme: control (0% population 1, 100% population 2), case (50% population 1, 50% population 2)
- Datasets with candidate loci selected as follows
  - Random SNPs (no associations)
  - Differentiated SNPs (a large difference in allele frequencies between populations, but no associations)
    - Allele frequence 0.8 for population 1, 0.2 for population 2
  - Causal SNPs

Simulation Results

	χ^2	Genomic control	EIGENSTRAT
Discrete subpopulations with moderate ancestry differences between cases and controls			
Random SNPs	0.0008	0.0001	0.0001
Differentiated SNPs	0.8520	0.5007	0.0001
Causal SNPs	0.5117	0.2980	0.4860
Discrete subpopulations with more extreme ancestry differences between cases and controls			
Random SNPs	0.0365	0.0001	0.0001
Differentiated SNPs	1.0000	1.0000	0.0001
Causal SNPs	0.5073	0.0342	0.2666
Admixed population with ancestry differences between cases and controls based on ancestry risk *r*			
$r = 2$			
Random SNPs	0.0002	0.0001	0.0001
Differentiated SNPs	0.1600	0.1004	0.0001
Causal SNPs	0.5180	0.4367	0.4863
$r = 3$			
Random SNPs	0.0007	0.0001	0.0001
Differentiated SNPs	0.7757	0.5553	0.0001
Causal SNPs	0.5158	0.3328	0.4442

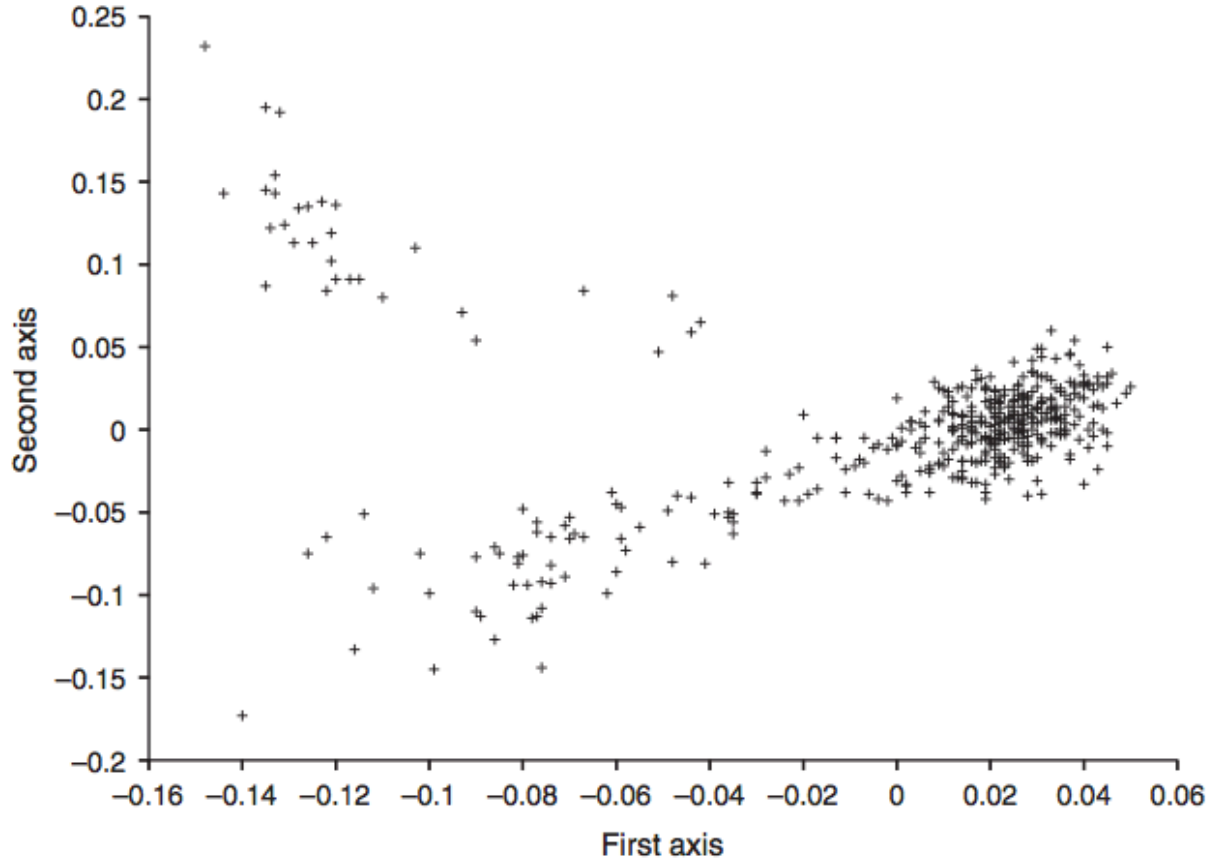
Simulation Results

- To correct for population stratification, a greater number of SNPs are required for less differentiated populations

# SNPs required	F_{ST}
20,000	0.005
50,000	0.002
100,000	0.001

PCA for Population Structure Discovery

Genetic variation between two southeast European populations



Genetic variation between northwest and southeast Europe

European American Dataset

- 488 European Americans genotyped at 116,204 SNPs
- A mutation in *LCT* gene is 100% associated with lactase persistence phenotype
  - This mutation was not included in this dataset
  - Look for an indirect association between a nearby SNP rs3769005, which is in 90% LD with the *LCT* mutation based on HapMap data, and the phenotype
- The region in chromosome 2 surrounding *LCT* gene is highly associated with the phenotype due to the the strong selective sweep in that region.

Association Results for SNPs Outside of Chromosome 2 (*LCT* gene)

SNP	χ^2	Genomic control	EIGENSTRAT
rs10511418	**45.11 (0.0000022)**	**31.55 (0.0023)**	11.57 (1.00)
rs2493880	**26.12 (0.037)**	18.27 (0.89)	8.17 (1.00)
rs4306808	**26.04 (0.039)**	18.21 (0.90)	8.83 (1.00)
rs2243133	**25.60 (0.049)**	17.90 (0.93)	5.88 (1.00)

Summary

- Genomic Control
  - Cannot handle the effect of admixed populations
- STRAT: structured association with *STRUCTURE*
  - Uses a generative model that explicitly models admixture
  - Computationally demanding
- EigenStrat
  - Does not provide intuition behind the admixing process
  - Significantly low computational cost than STRAT