Population Stratification

Benjamin Neale

Leuven August 2008

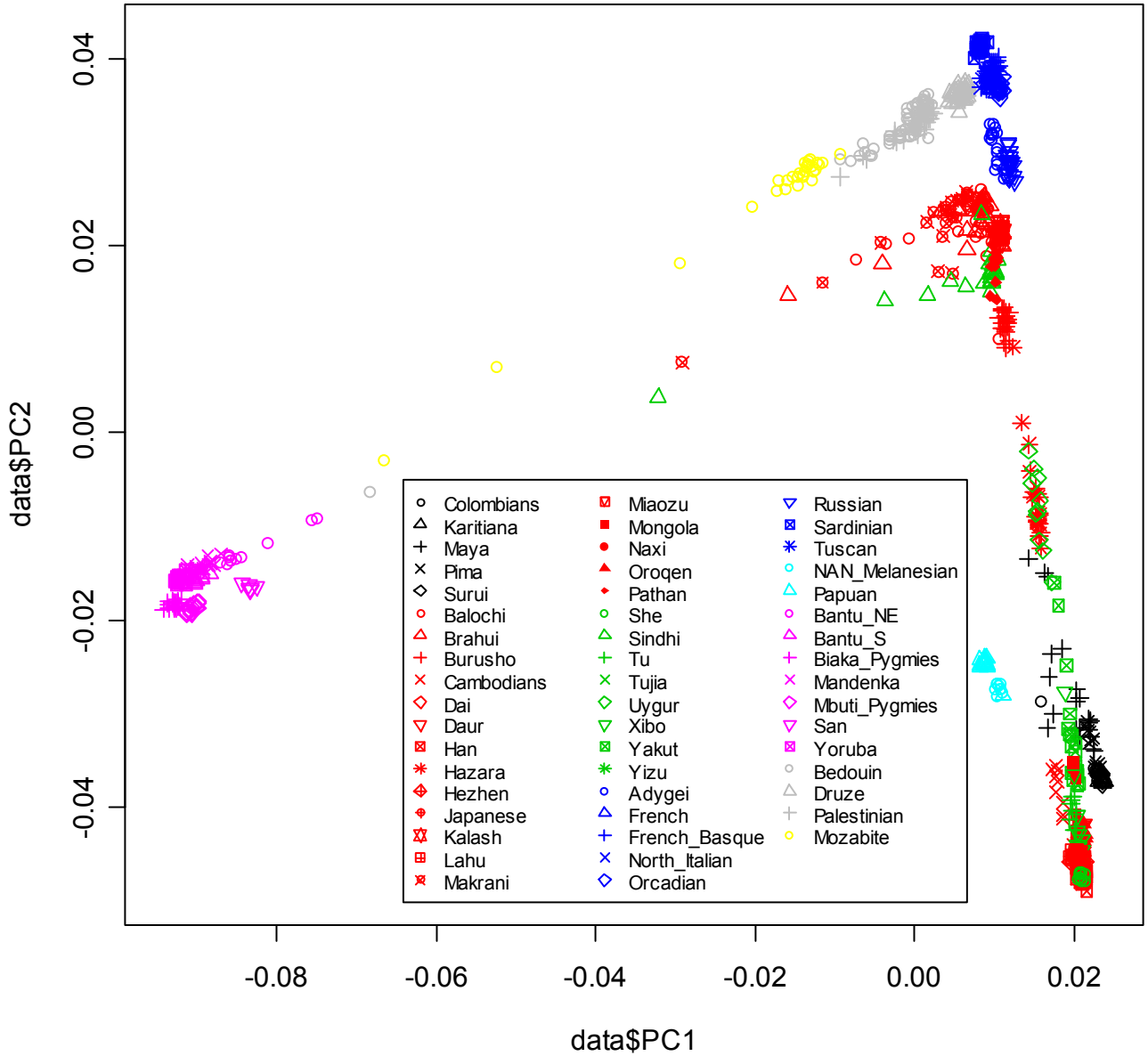
Objectives

- Population Stratification – What & Why?
- Dealing with PS in association studies
  - Revisiting Genomic Control (small studies)
  - EIGENSTRAT
  - PLINK practical
  - Other methods

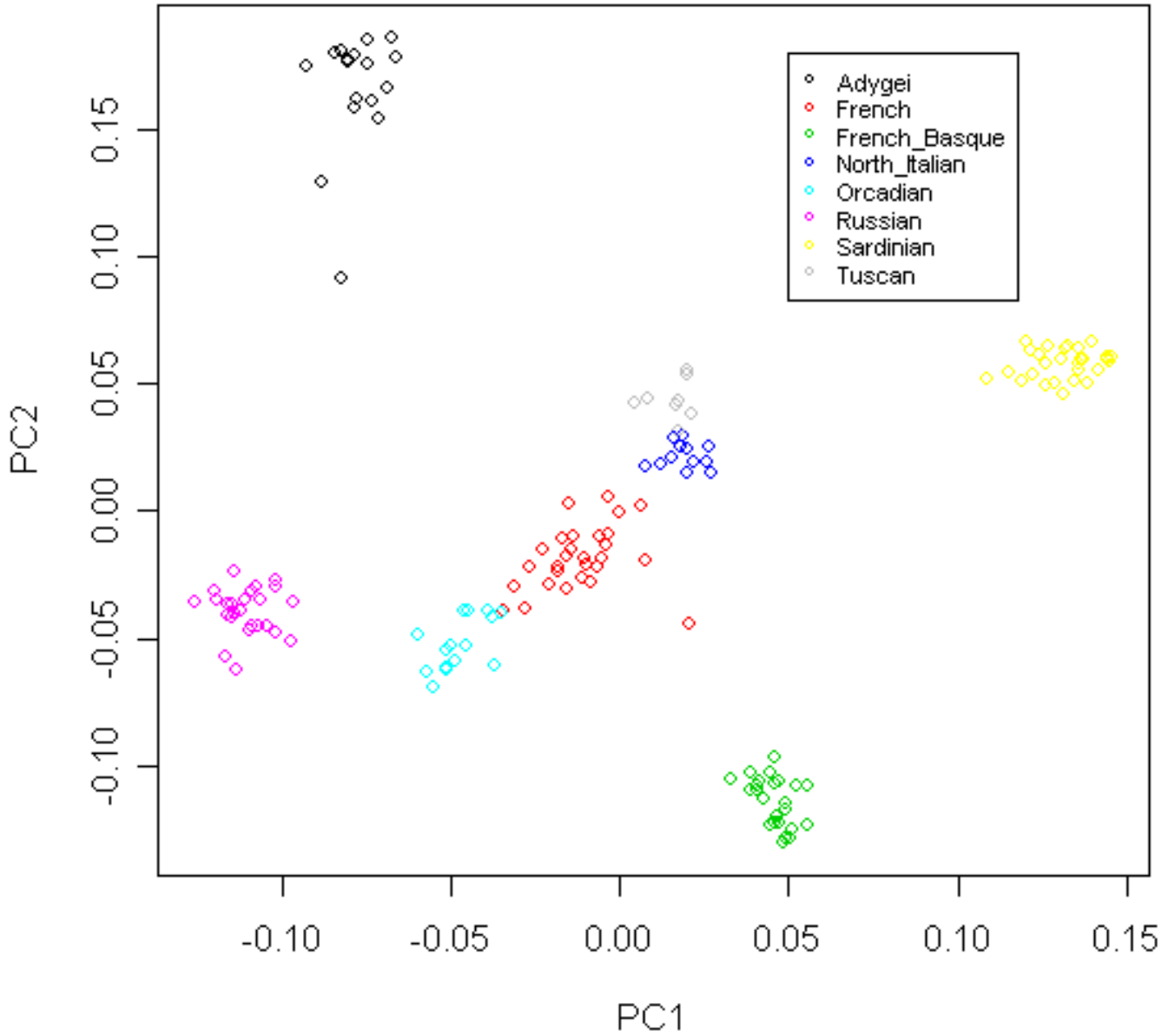
What is population stratification/structure (PS)?

- This just in! Human beings don't mate at random
  - Physical barriers
  - Political barriers
  - Socio-cultural barriers
  - Isolation by distance
- None of these barriers are absolute, and in fact by primate standards we are remarkably homogeneous
  - Most human variation is 'within population'
  - Reflects recent common ancestry (Out of Africa)
- Between population variation still exists, even though the vast majority of human variation is shared

Human Genetic Diversity Panel, Illumina 650Y SNP chip (Li et al. 2008, Science 319: 1100)



Human Genetic Diversity Panel, Europeans only



Why is **hidden** PS a problem for association studies?

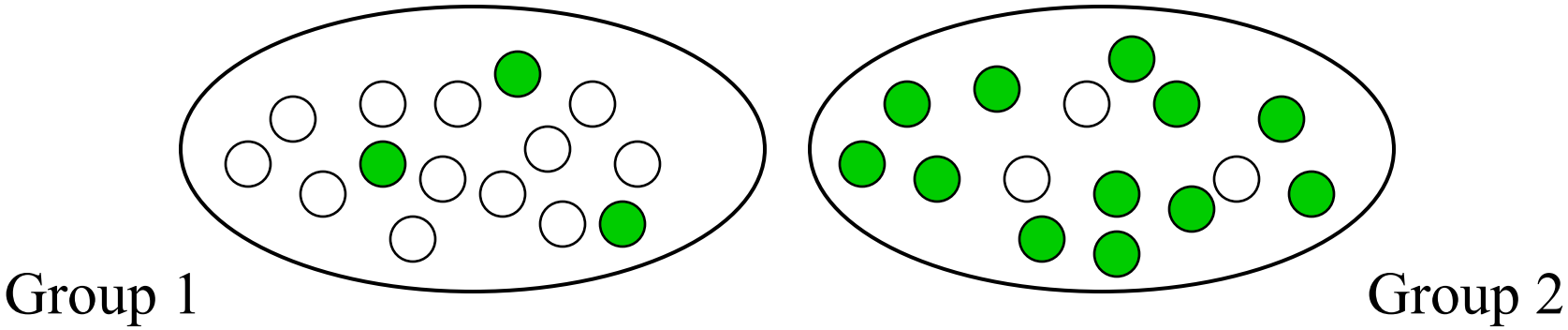
- Reduced Power
  - Lower chance of detecting true effects
- Confounding
  - Higher chance of spurious association finding

Requirements of stratification

- Both conditions necessary for stratification
  - –Variation in disease rates across groups
  - –Variation in allele frequencies

Visualization of stratification conditions

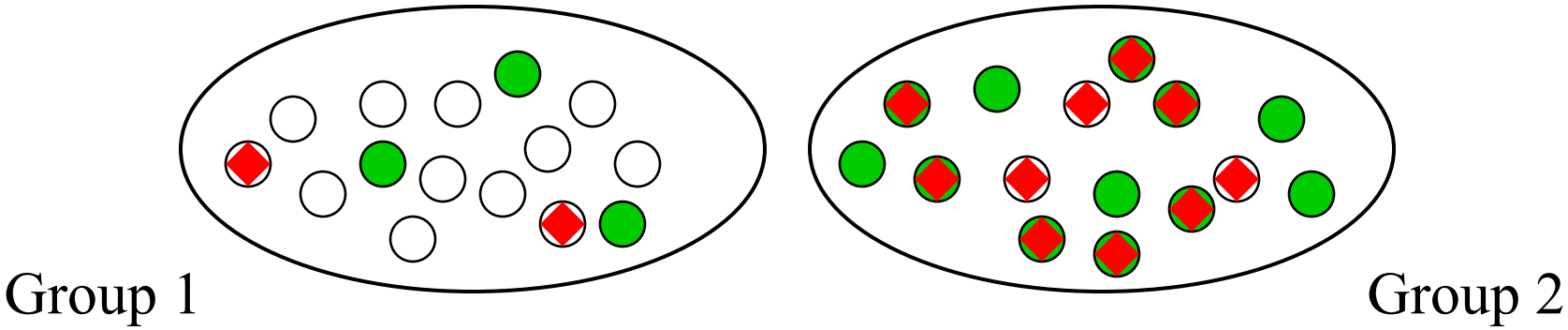
- Suppose that a disease is more common in one subgroup than in another…



- …then the cases will tend to be over-sampled from that group, relative to controls.

…and this can lead to false positive associations

- Any allele that is more common in Group 2 will *appear* to be associated with the disease.



- This will happen if Group 1 & 2 are "hidden" – if they are known then they can be accounted for.
- Discrete groups are not required – admixture yields same problem.

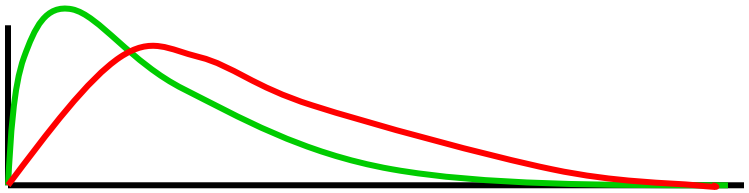
Dealing with PS in association studies

Family-based association studies

- Transmission conditional on known parental ('founder') genotypes
  - E.g. TDT
  - Recent review: Tiwari et al. (2008, Hum. Hered. 66: 67)
- Pros
  - Cast-iron PS protection
- Cons
  - 50% more genotyping needed (if using trios)
  - Not all trios are informative
  - Families more difficult to collect

Genomic Control (GC)

- Devlin and Roeder (1999) used theoretical arguments to propose that with population structure, the distribution of Chi-square tests is inflated by a constant multiplicative factor λ.



- To estimate λ, add a separate "GC" set of neutral loci to genotype, and calculate chi-square tests for association in these

- Now perform an adjusted test of association that takes account of any mismatching of cases/controls:

$$\chi^2_{GC} = \chi^2_{Raw}/\lambda$$

Genomic Control (GC)

- Correct χ^2 test statistic by inflation factor λ
- Pros
  - Easy to use
  - Doesn't need many SNPs
  - Can handle highly mismatched Case/Control design
- Cons
  - Less powerful than other methods when many SNPs available
  - Can't handle 'lactase-type' false positives
  - λ-scaling assumption breaks down for large λ

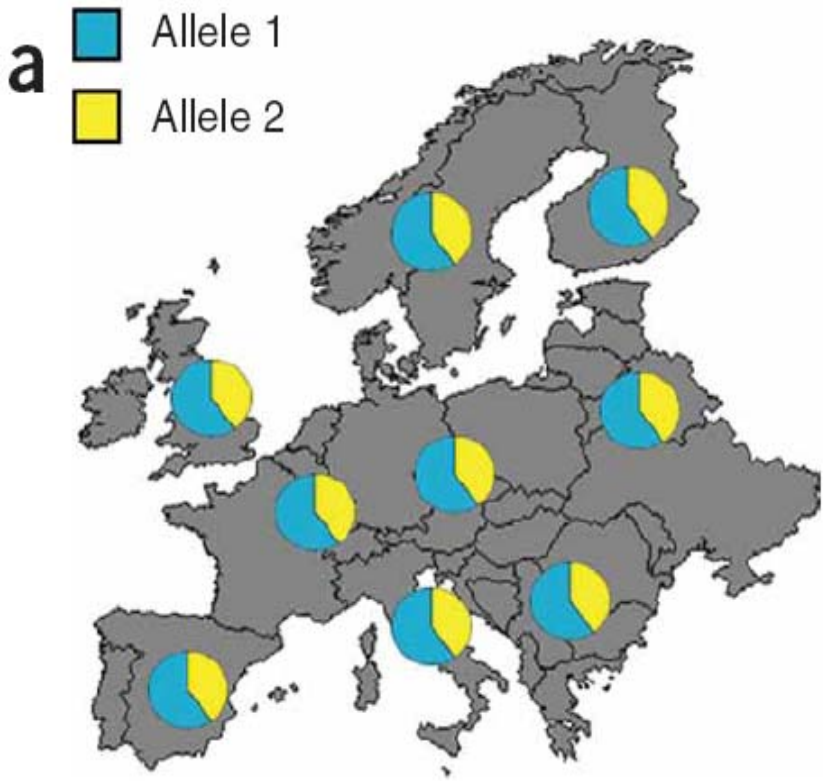
Genomic Control variants

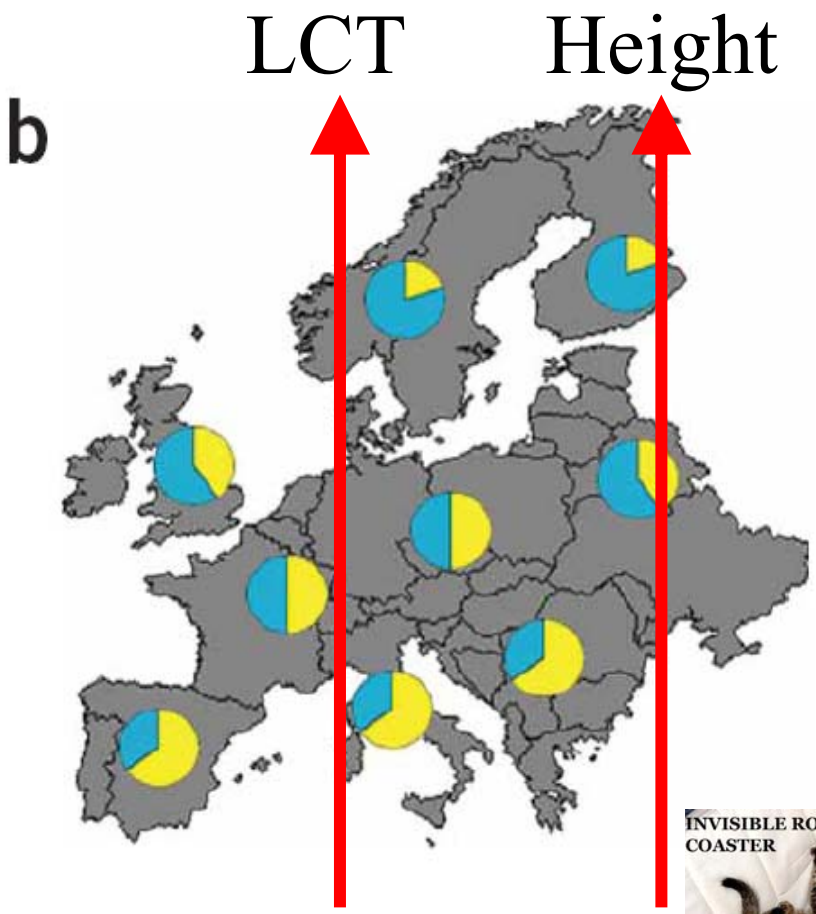
- GC_{med} (Devlin & Roeder 1999, Biometrics 55: 997)
  - $\lambda = \text{median}(\chi^2_{GC})/0.455$
- GC_{mean} (Reich & Goldstein 2001, Gen Epi 20: 4)
  - $\lambda = \text{mean}(\chi^2_{GC})$
  - Upper 95% CI of λ used as conservative measure
- GCF (Devlin et al. 2004, Nat Genet 36: 1129)
  - Test χ^2_{Raw}/λ as F-statistic
  - Recent work (Dadd, Weale & Lewis, submitted) confirms GCF as the best choice
- More variants on the theme
  - Use Q-Q plot to remove GC-SNP outliers (Clayton et al. 2005, Nat Genet 37: 1243)
  - Ancestry Informative Markers (Review: Barnholtz-Sloan et al. 2008, Cancer Epi Bio Prev 17: 471)
  - Frequency matching (Reich & Goldstein 2001, Gen Epi 20: 4)

Other methods

- Structured Association
  - E.g. *strat* (Pritchard et al. 2000, Am J Hum Genet 67: 170)
  - Fits explicit model of discrete ancestral sub-populations
  - Breaks down for small datasets, too computationally costly for large datasets
- Mixed modelling
  - Fits error structure based on matrix of estimated pairwise relatedness among all individuals (e.g. Yu et al. 2006, Nat Genet 38: 203)
  - Requires many SNPs to estimate relatedness well
  - Can't handle binary phenotypes (e.g. Ca/Co) well
- Still an active area of methodological development
  - Delta-centralization (Gorrochurn et al. 2006, Gen Epi 30: 277)
  - Logistic Regression (Setakis et al. 2006, Genome Res 16: 290)
  - Stratification Score (Epstein et al. 2007, Am J Hum Genet 80: 921)
  - Review: Barnholtz-Sloan et al. (2008, Cancer Epi Bio Prev 17: 471)

Genomic Control fails if stratification affects certain SNPs more than the average





Campbell et al. (2005, Nat Genet 37: 868)

EIGENSTRAT

An example: height associates with lactase persistence SNP in US-European sample

N	Total	2,179
	Tall	1,123
	Short	1,056
LCT −13910 genotype counts[c]	Total	392:918:869
	Tall	161:474:489
	Short	231:444:380
Hardy-Weinberg *P*	Total	5.6×10^{-7}
	Tall	0.03
	Short	2.5×10^{-5}
Association *P*		3.6×10^{-7}
OR (95% c.i.)[d]		1.37 (1.22–1.54)

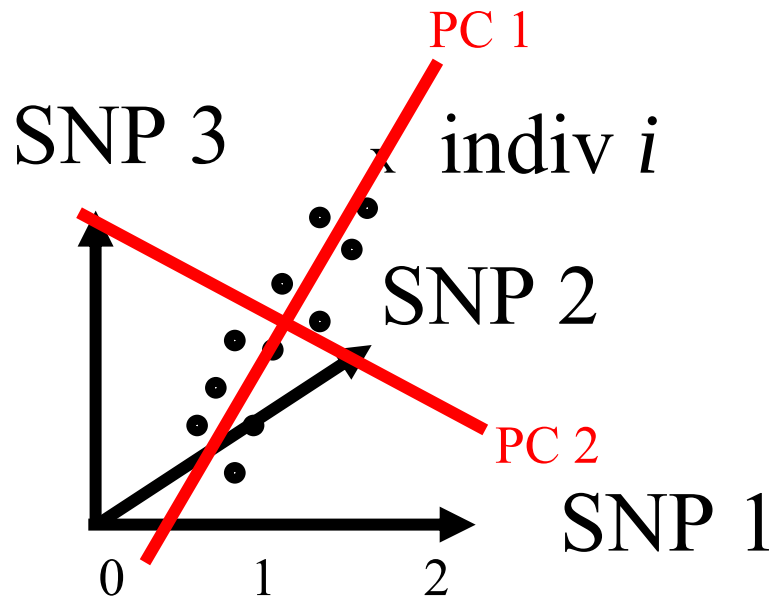
EIGENSTRAT

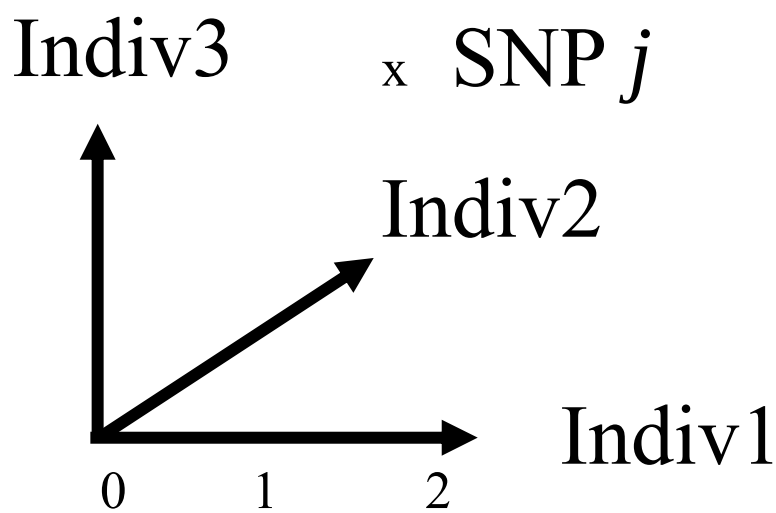
False Positive

The EIGENSTRAT solution

PCA for SNP data ("EIGENSTRAT")

$$\mathbf{X} = \begin{array}{cc} & m \text{ SNPs} \\ n \text{ indivs} & \begin{bmatrix} x_{11} & \cdots & \\ & x_{ij} & \cdots \\ & & x_{nm} \end{bmatrix} \end{array}$$





EIGENSTRAT

PCA properties

- Each axis is a linear equation, defining individual "scores" or SNP "loadings"

$$Z_i = a_1x_{i1} + .. + a_jx_{ij} + .. + a_mx_{nm}$$
$$Z'_j = b_1x_{1j} + .. + b_ix_{ij} + .. + b_nx_{nm}$$

- Axes can be created in either projection
- Max N^O axes = min(n-1,m-1)
- Each axis is at right angles to all others ("orthogonal")
- Eigenvectors define the axes, and eigenvalues define the "variance explained" by each axis

PC axis types

- PCA dissects and ranks the correlation structure of multivariate data
- Stratification is *one* way that correlations in SNPs can be set up
  - Stratification
  - Systematic genotyping artefacts
  - Local LD
  - (Theoretical) Many high-effect causal SNPs in a case-control study
- Inspection of PC axis properties can determine which type of effect is at work for each axis

Original EIGENSTRAT procedure

1) Code all SNP data {0,1,2}, where 1=het
2) Normalize by subtracting mean and dividing by

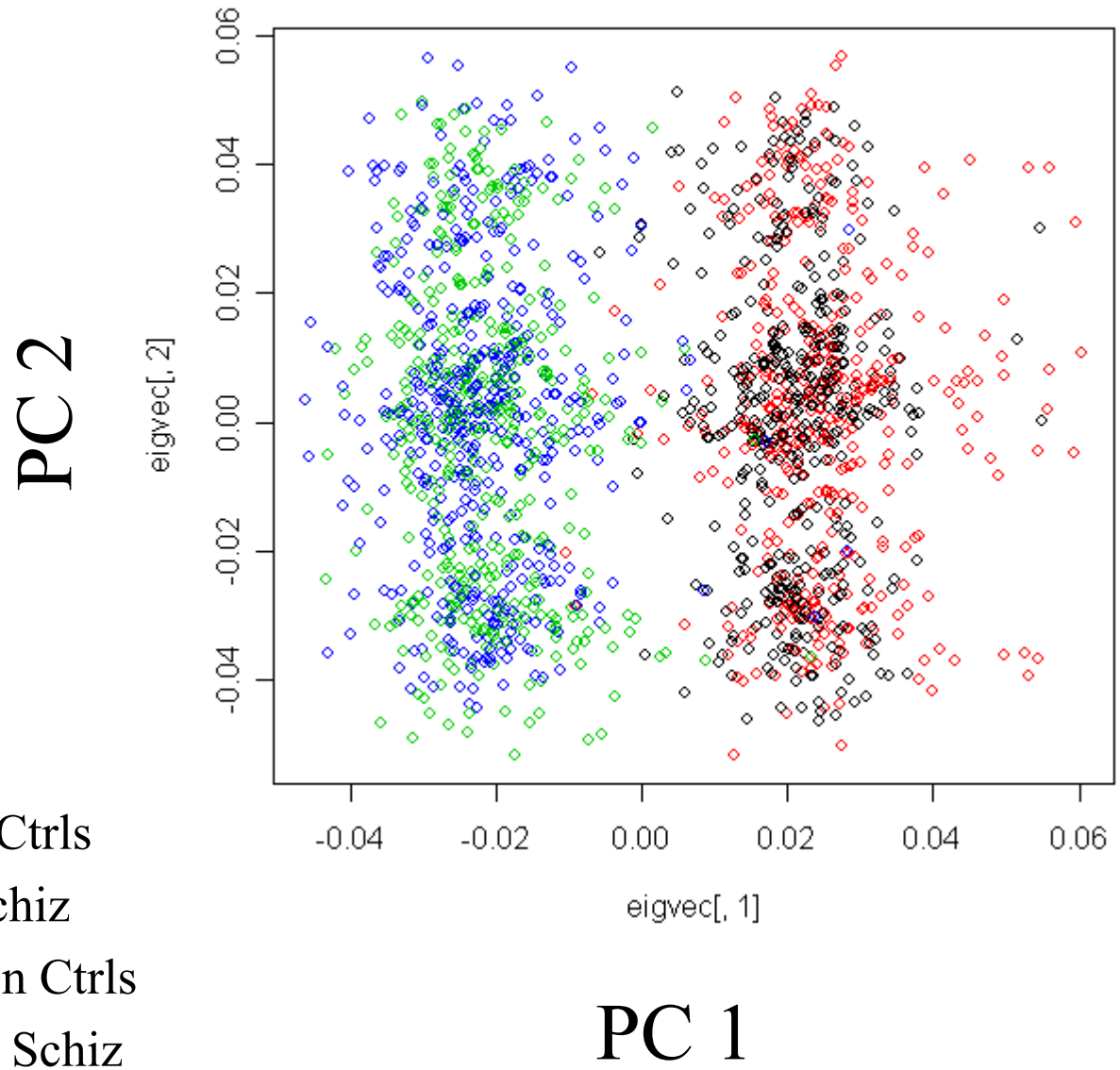
$$\sqrt{p(1-p)}$$

3) Recode missing genotype as 0
4) Apply PCA to matrix of coded SNP data
5) Extract scores for 1st 10 PC axes
6) Calculate modified Armitage Trend statistic using 1st 10 PC scores as covariates

Price et al. (2006, Nat Genet 38: 904)
Patterson et al. (2006, PLoS Genet 2: e190)
Earlier more general structure: Zhang et al. (2003, Gen Epi 24: 44)

Identifying PC axis types

EIGENSTRAT applied to genomewide SNP data typed in two populations



Black = Munich Ctrls

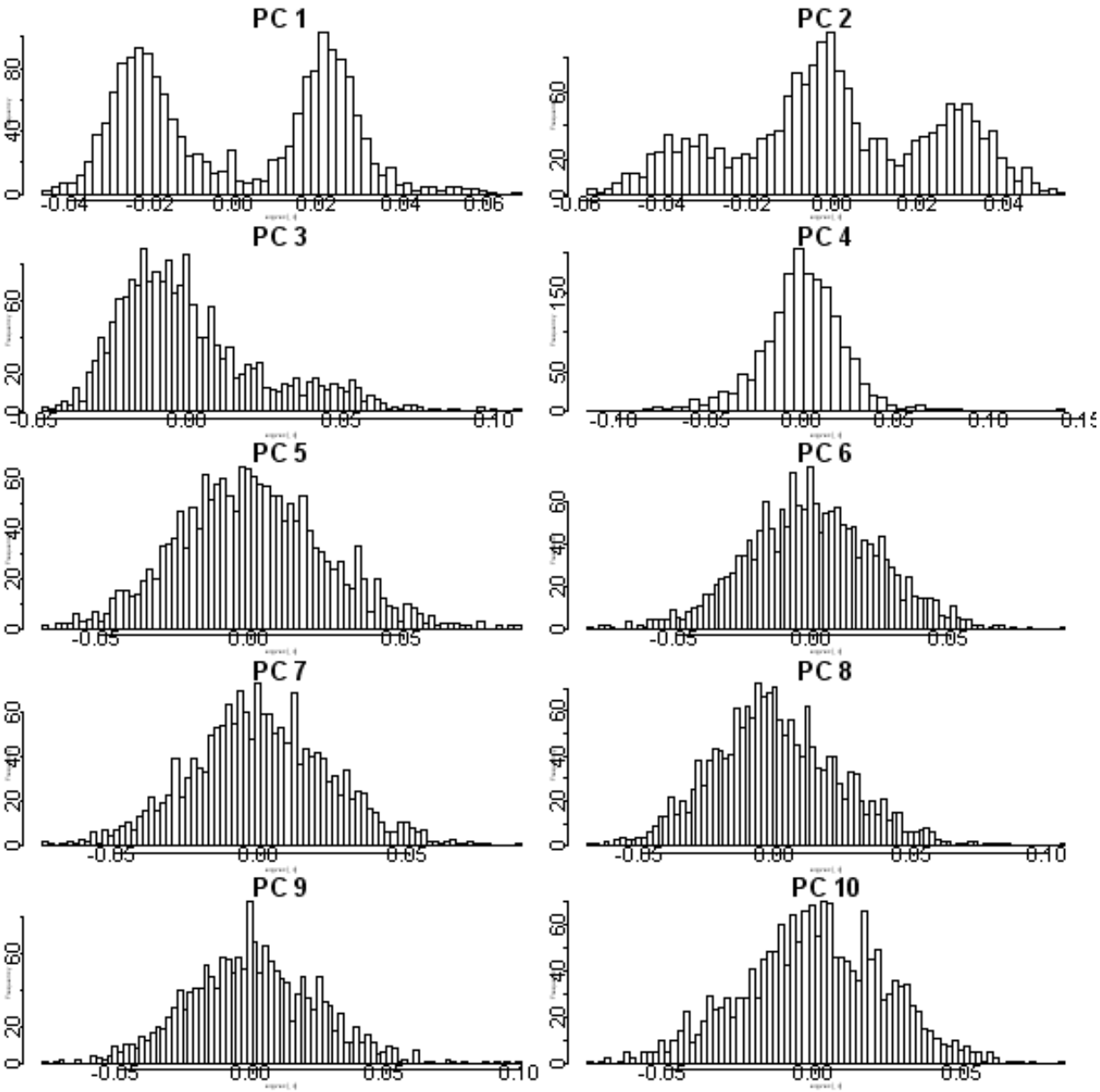
Red = Munich Schiz

Green = Aberdeen Ctrls

Blue = Aberdeen Schiz

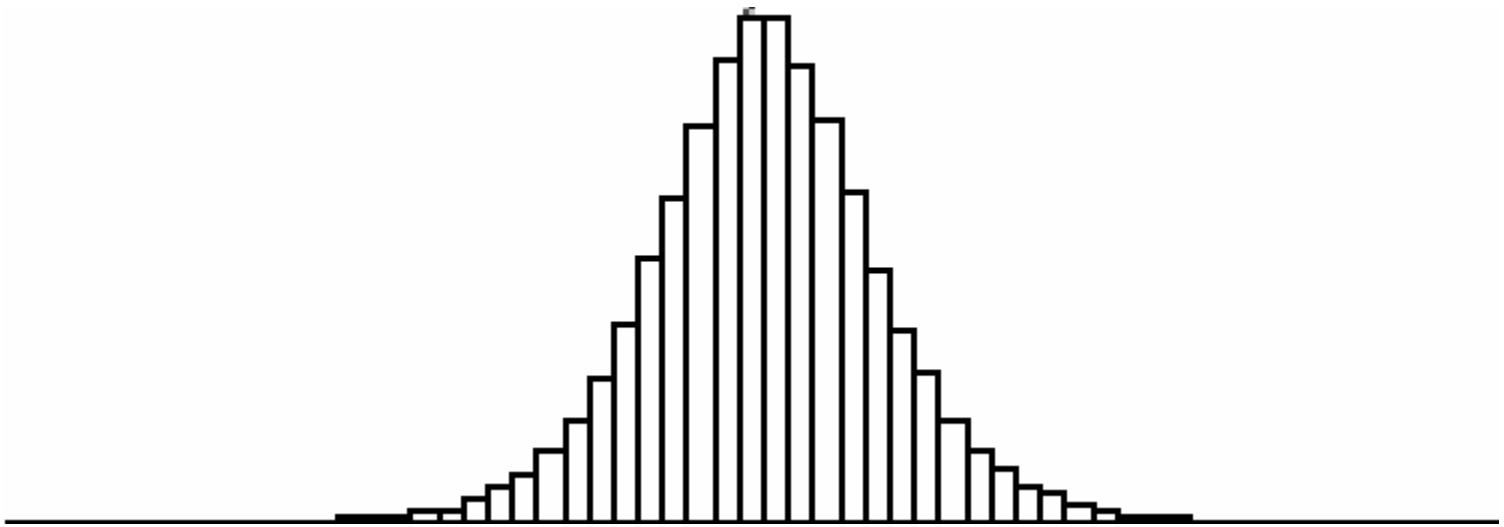
EIGENSTRAT

PC individual "scores"



EIGENSTRAT

SNP “loadings”, PC1



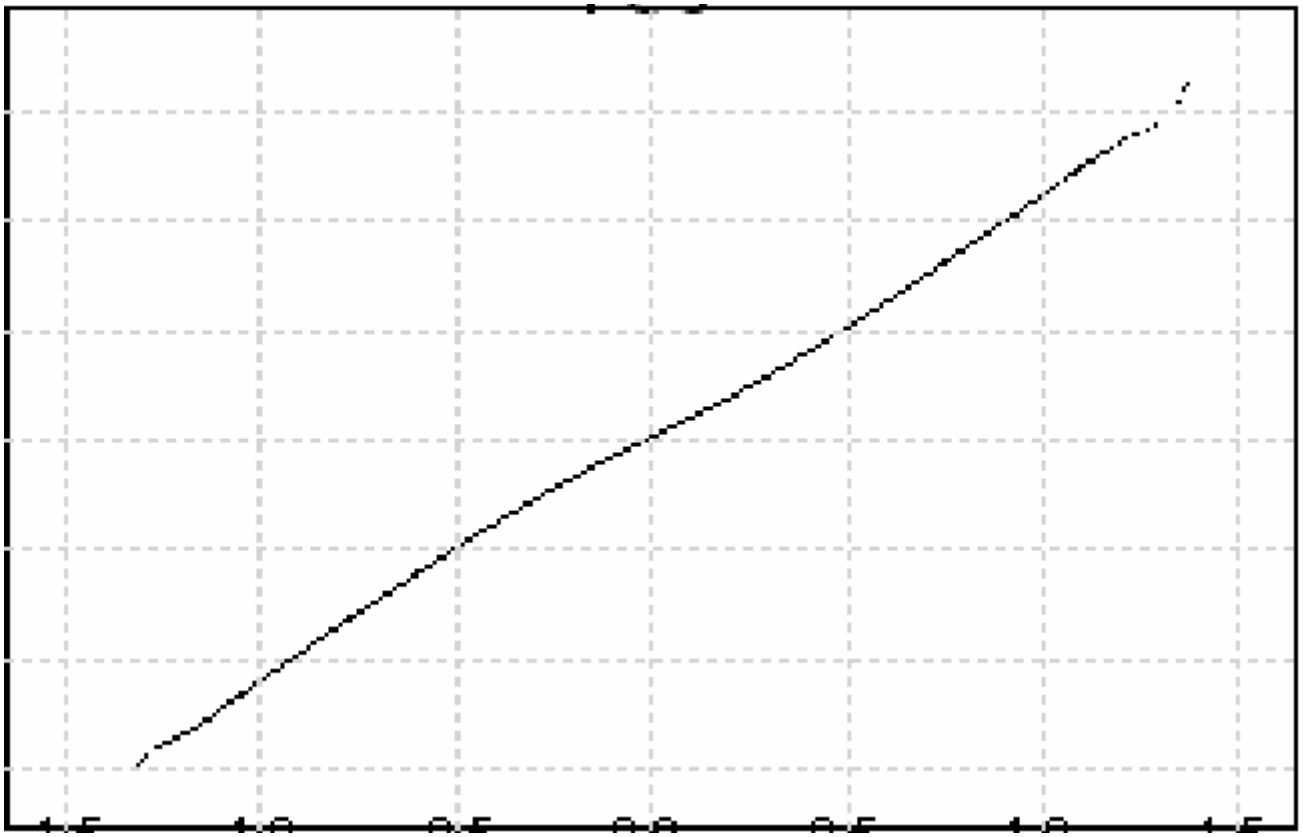
PC1 SNP loading distribution

Whole genome contributes

EIGENSTRAT

SNP "loadings", PC1

Whole genome contributes

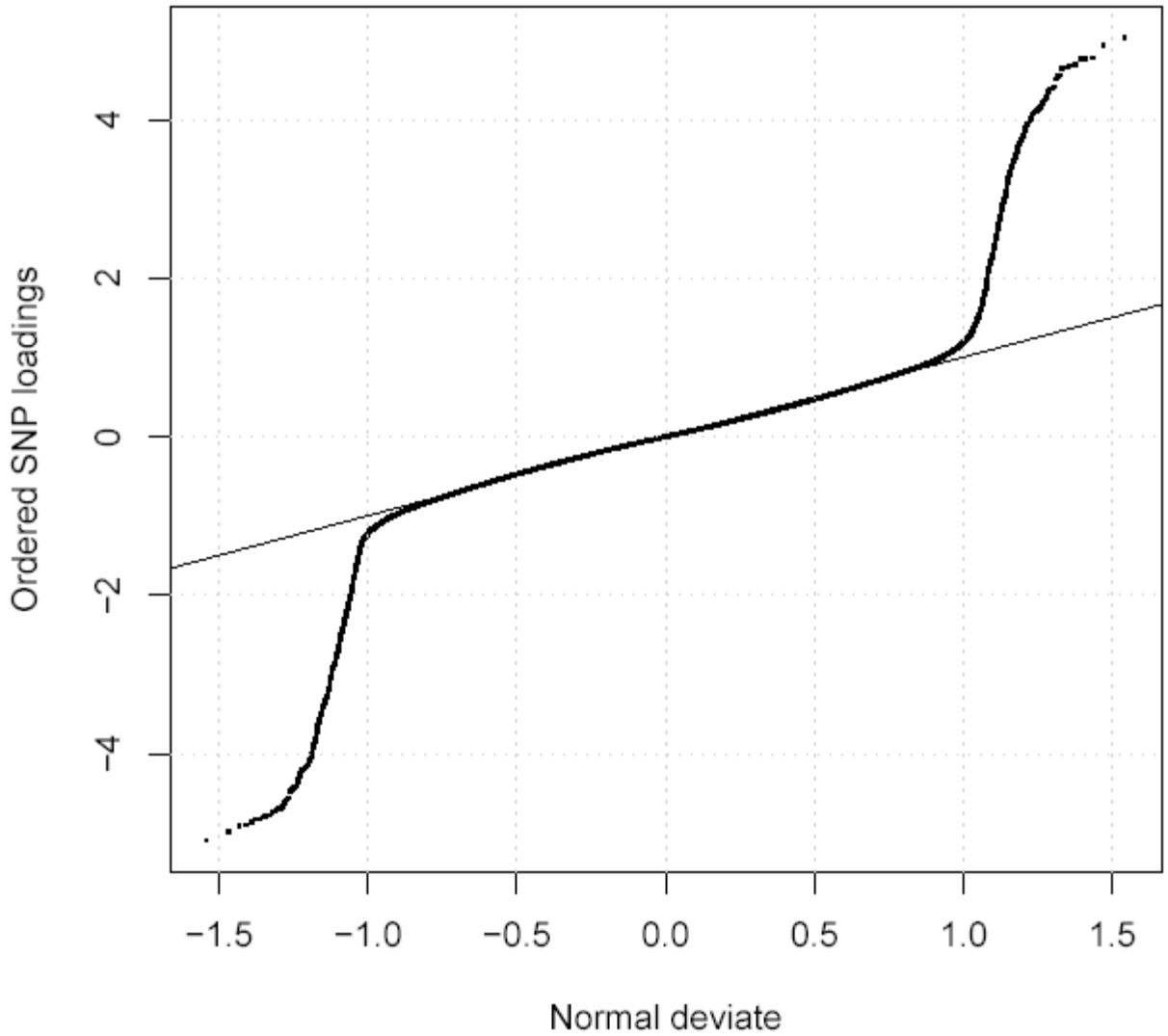


PC1 SNP loading Q-Q plot

EIGENSTRAT

SNP "loadings", PC2

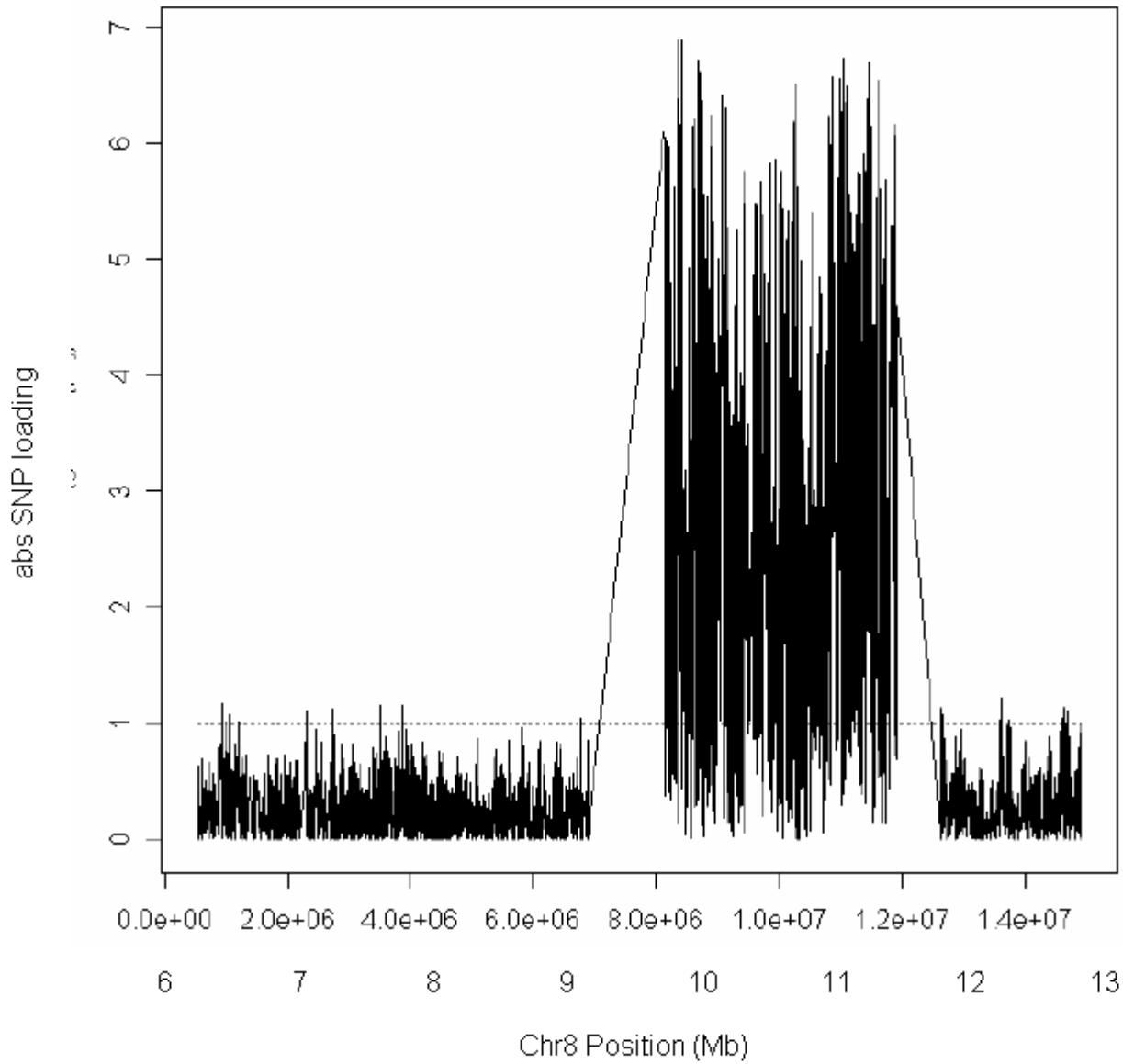
Only *part* of the genome contributes



EIGENSTRAT

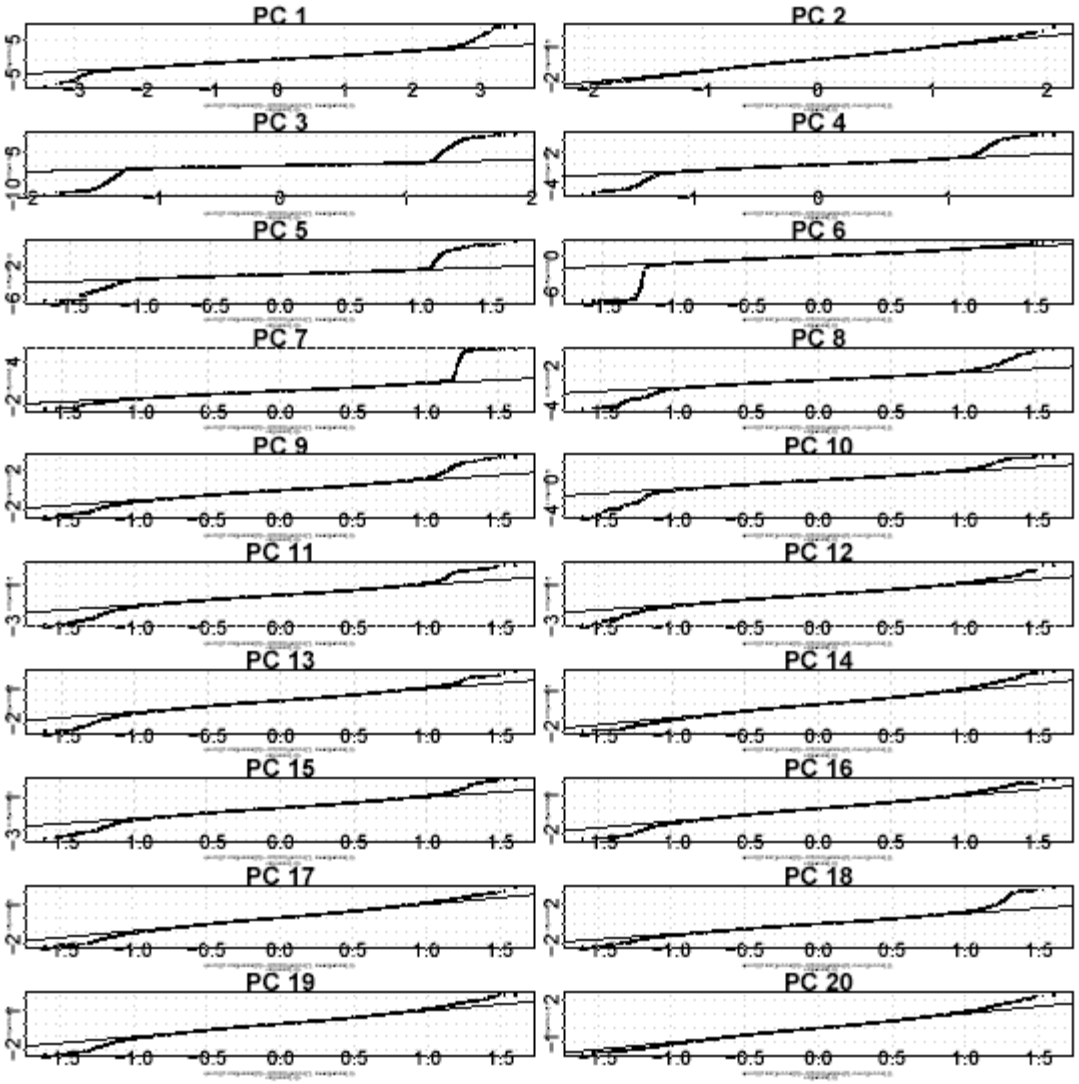
PC2 driven by known ~4Mb inversion poly on Chr8

Characteristic LD pattern revealed by SNP loadings



EIGENSTRAT

PC axis types revealed by SNP loading Q-Q plots in Illumina iControl dataset



EIGENSTRAT

Extended EIGENSTRAT procedure corrects for local LD

1) Known high-LD regions excluded
2) SNPs thinned using LD criterion
   - $r^2<0.2$
   - Window size = 1500 contiguous SNPs
   - Step size = 150
3) Each SNP regressed on the previous 5 SNPs, and the residual entered into the PCA analysis
4) Iterative removal of outlier SNPs and/or outlier individuals
5) Nomination of axes to use as covariates based on Tracy Widom statistics
6) Enter significant PC axes as covariates in a logistic or linear regression:

   Phenotype = g(const. + $\boldsymbol{\beta}$*covariates + γ*SNP j genotype) + ε

EIGENSTRAT

Guidance on use of EIGENSTRAT

Phase-change in ability to detect structure: $F_{st} = 1/\sqrt{nm}$

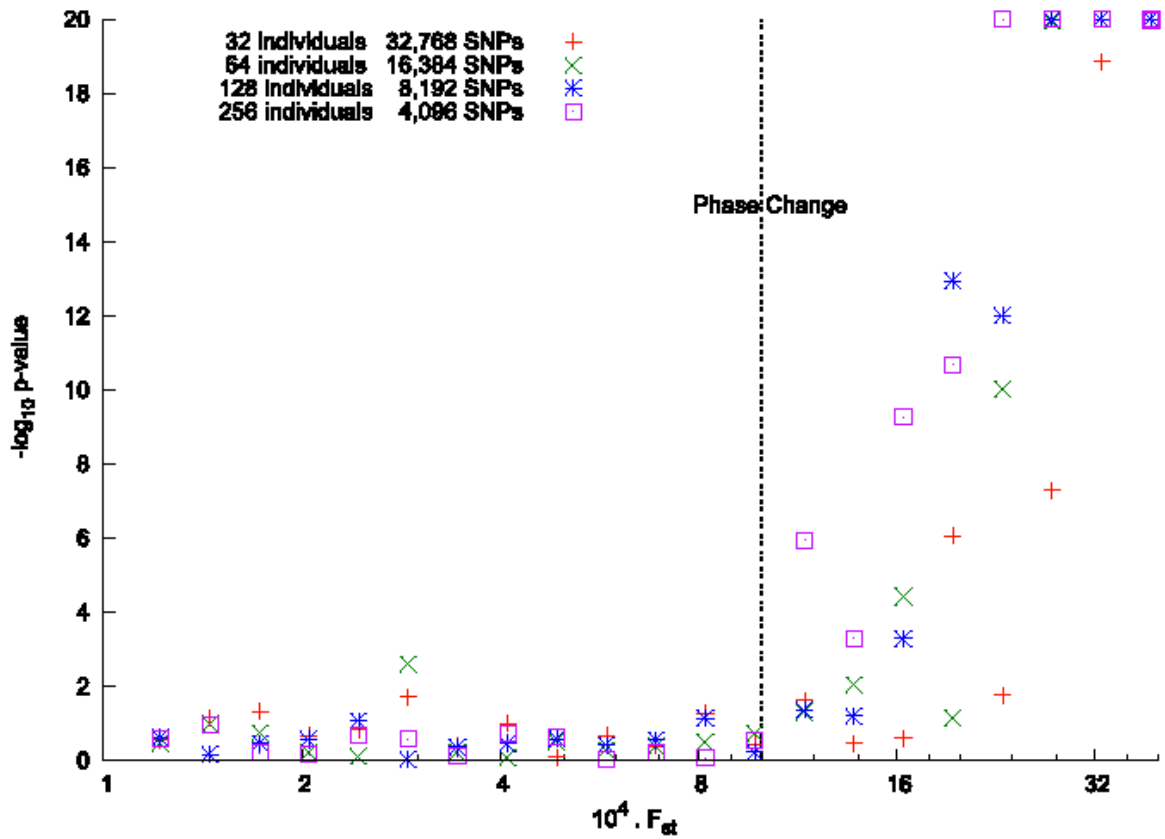


Figure 6. The BBP Phase Change

We ran a series of simulations, varying the sample size m and number of markers n but keeping the product at $mn = 2^{20}$. Thus the predicted phase change threshold is $F_{ST} = 2^{-10}$. We vary F_S and plot the log p-value of the Tracy–Widom statistic. (We clipped $-\log_{10} p$ at 20.) Note that below the threshold there is no statistical significance, while above threshold, we tend to get enormous significance.

doi:10.1371/journal.pgen.0020190.g006

Patterson et al. (2006, PLoS Genet 2: e190)

EIGENSTRAT

Number of SNPs needed for EIGENSTRAT to work

Supplementary Table 2: Simulations using *M* SNPs

N=1000, F_{ST}=0.01, α=0.0001, 'lactase-type' SNPs

M	False positive rate	Correlation of top axis
100	0.0826	68.4%
200	0.0079	80.9%
500	0.0016	90.8%
1,000	0.0007	94.8%
2,000	0.0002	97.4%
5,000	0.0001	99.0%
10,000	0.0001	99.5%
20,000	0.0001	99.7%
50,000	0.0001	99.9%
100,000	0.0001	99.9%

Price et al. (2006, Nat Genet 38: 904)

Take-home messages

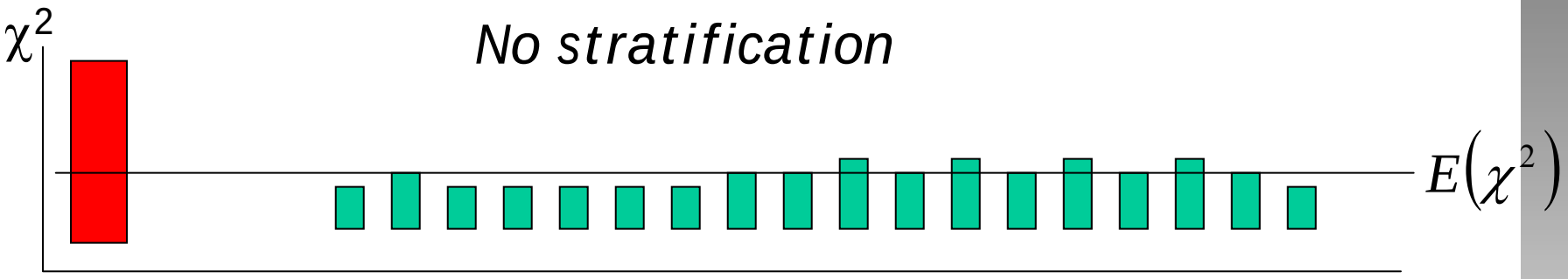
- EIGENSTRAT work very well with >2000 SNPs
  - Clinal/admixture model seems to work well in practice
  - Other more computationally demanding methods don't achieve huge power increases

- Genomic Control works well with <200 SNPs
  - Still has a place in smaller studies (GWAS replication, candidate gene)
  - Also copes with mismatched Case/Control designs (e.g. centralized control resources)

PLINK Practical

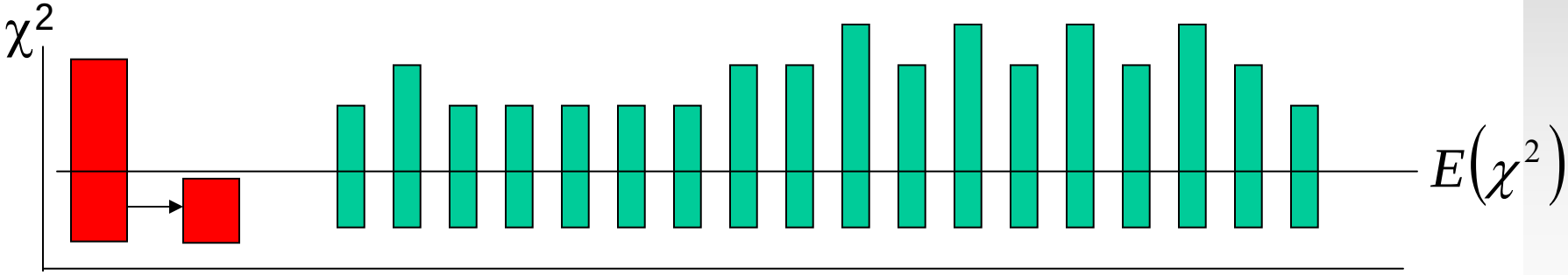
Genomic control

No stratification



Test locus

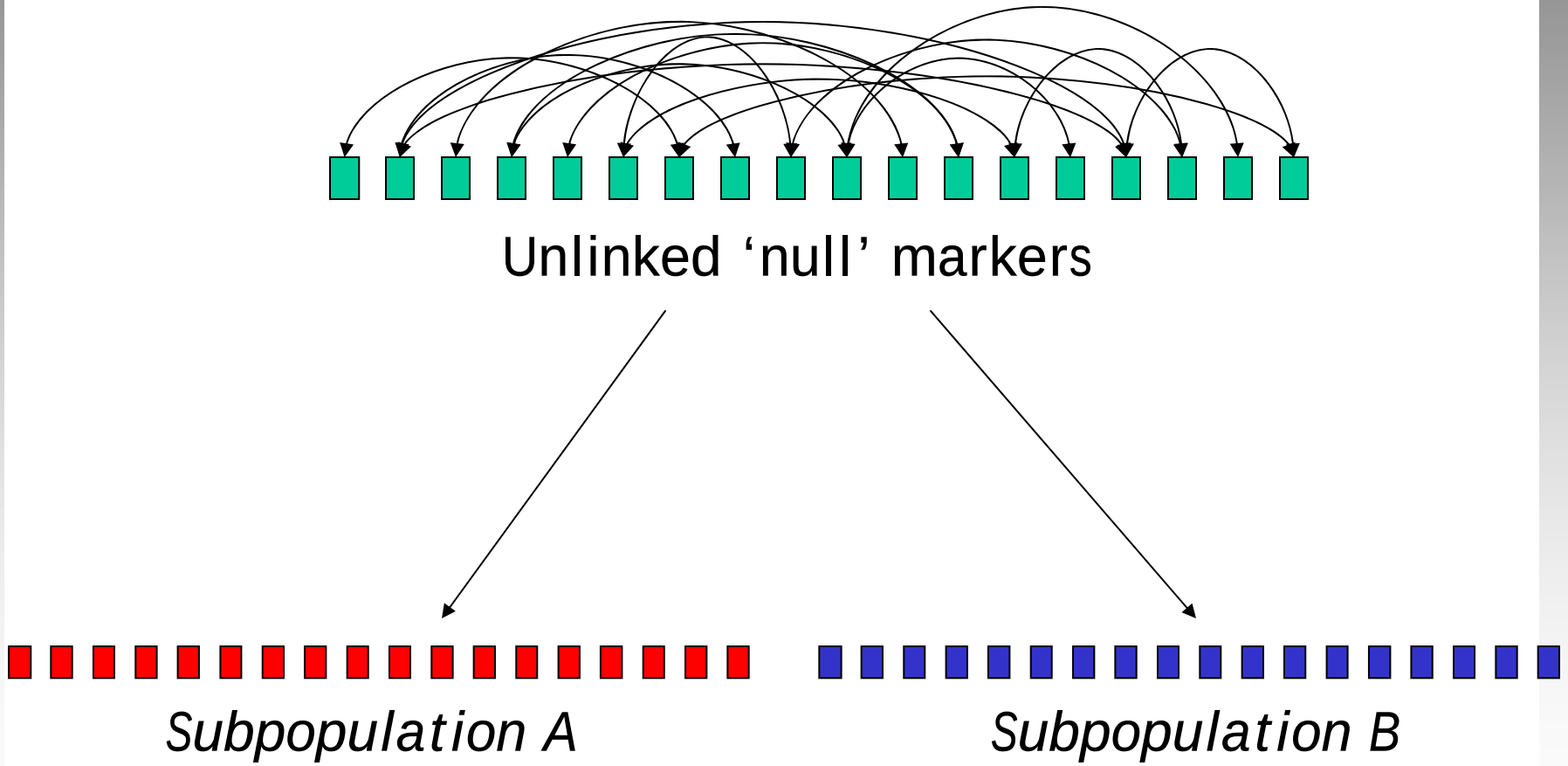
Unlinked 'null' markers



Stratification $\rightarrow$ adjust test statistic

Structured association

LD observed under stratification



Identity-by-state (IBS) sharing

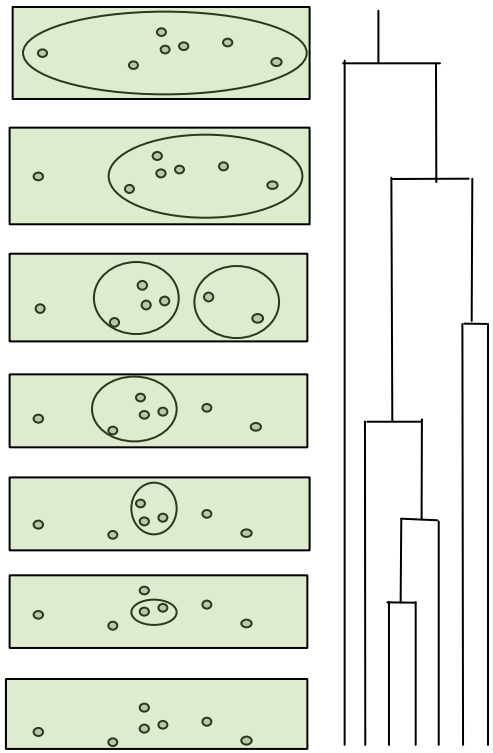
Pair from same population

Individual 1	A/**C**	G/**T**	**A/G**	A/A	**G/G**
	|	|	| |		| |
Individual 2	C/**C**	T/**T**	**A/G**	C/C	**G/G**
IBS	1	1	2	0	2

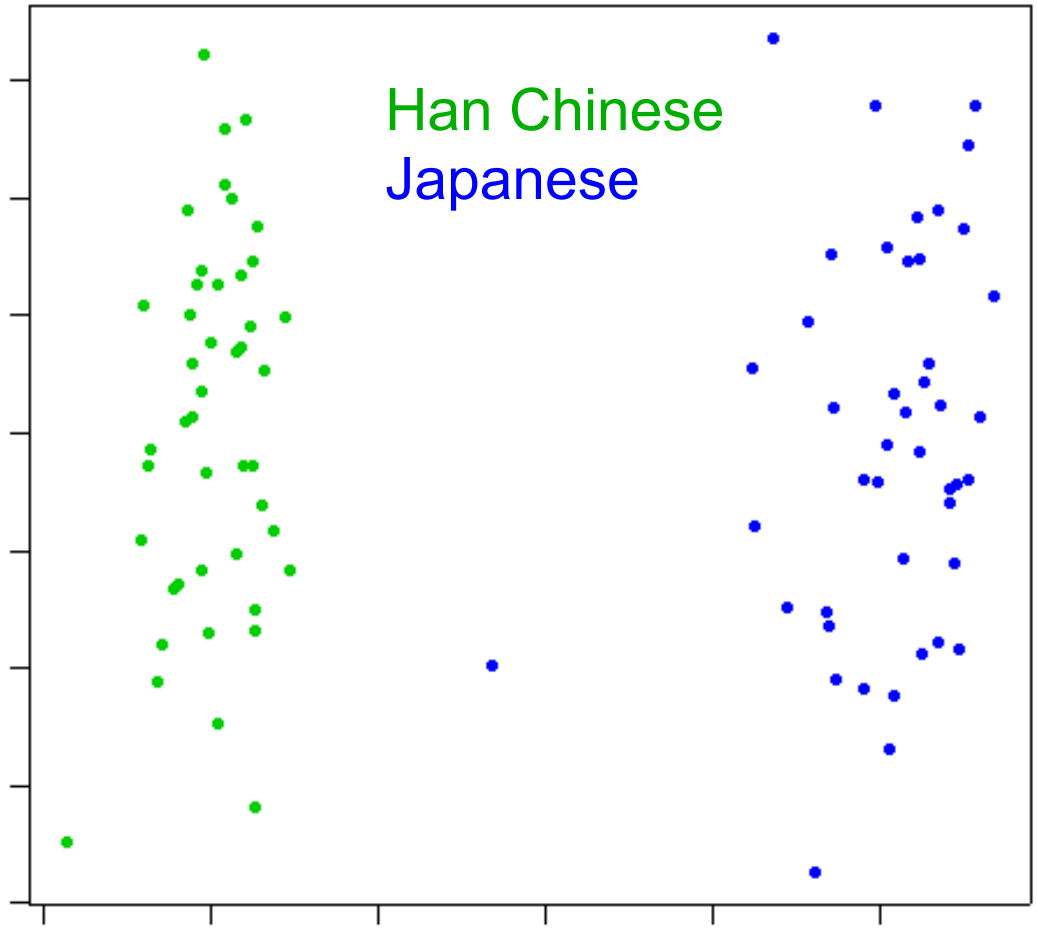
Pair from different population

Individual 3	A/**C**	G/G	A/A	A/A	G/**G**
	|				|
Individual 4	C/**C**	T/T	G/G	C/C	A/**G**
IBS	1	0	0	0	1

Empirical assessment of ancestry



Complete linkage IBS-based hierarchical clustering



Multidimensional scaling plot: ~10K random SNPs

Population stratification: LD pruning

Perform LD-based pruning

```
plink --bfile example --indep 50 5 2
```

Window size in SNPs
Number of SNPs to shift the window
VIF threshold

Spawns two files: plink.prune.in (SNPs to be kept) and plink.prune.out (SNPs to be removed)

Population stratification: Genome-file

Generates plink.genome

```
plink --bfile example –-genome --extract plink.prune.in
```

Extracts only the LD-pruned SNPs from the previous command

The genome file that is created is the basis for all subsequent population based comparisons

Population stratification: IBS clustering

Perform IBS-based cluster analysis for 2 clusters

```
plink --bfile example –-cluster --K 2 --extract
plink.prune.in --read-genome plink.genome
```

In this case, we are reading the genome file we generated

Clustering can be constrained in a number of other ways

cluster size, phenotype, external matching criteria, patterns of missing data, test of absolute similarity between individuals

Population stratification: MDS plotting

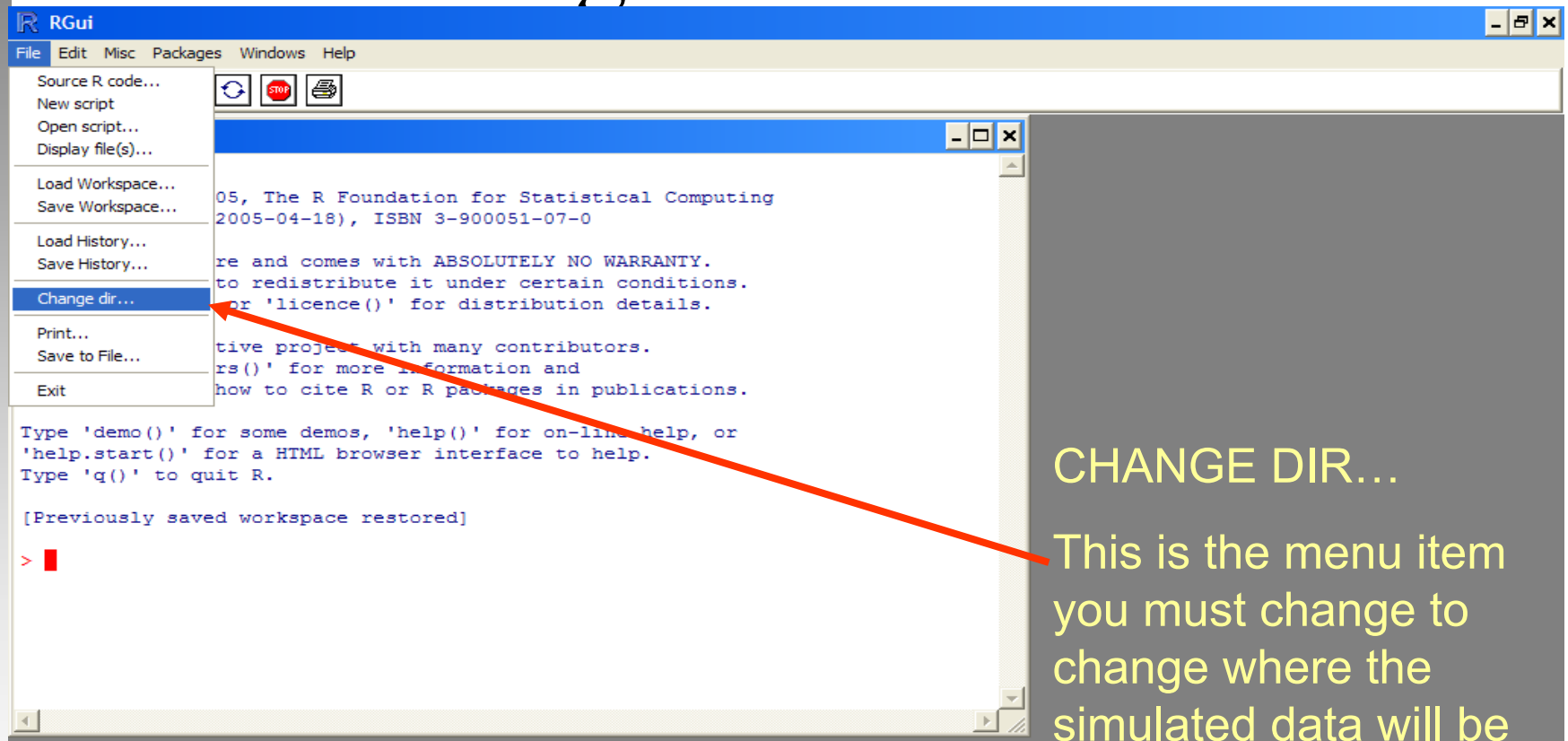
Telling plink to run cluster analysis

```
plink --bfile example --cluster --mds-plot 4 --K 2 --extract plink.prune.in --read-genome plink.genome
```

Calculating 4 mds axes of variation, similar to PCA

We will now use R to visualize the MDS plots. Including the --K 2 command supplies the clustering solution in the mds plot file

Plotting the results in R

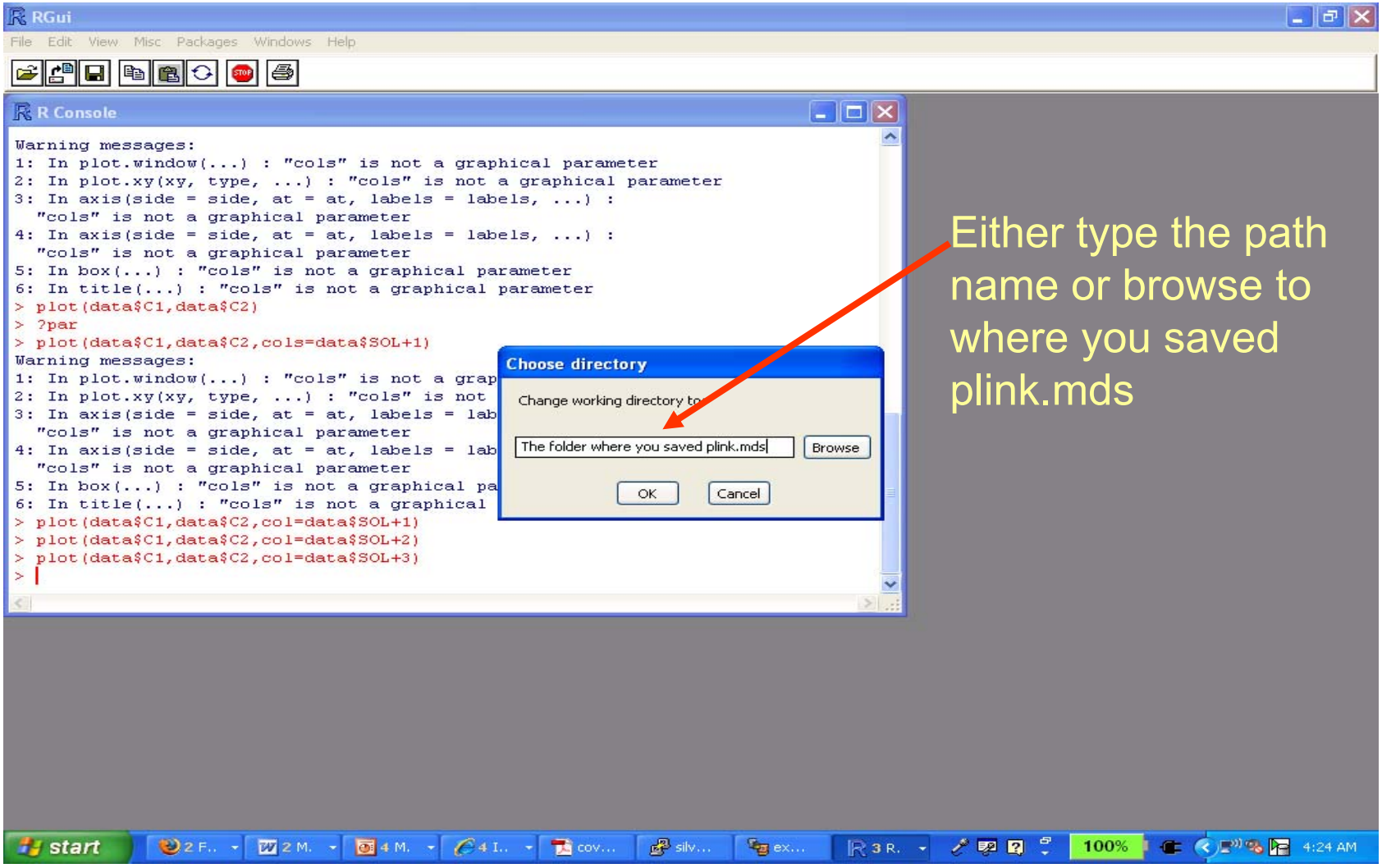


CHANGE DIR…

This is the menu item you must change to change where the simulated data will be placed

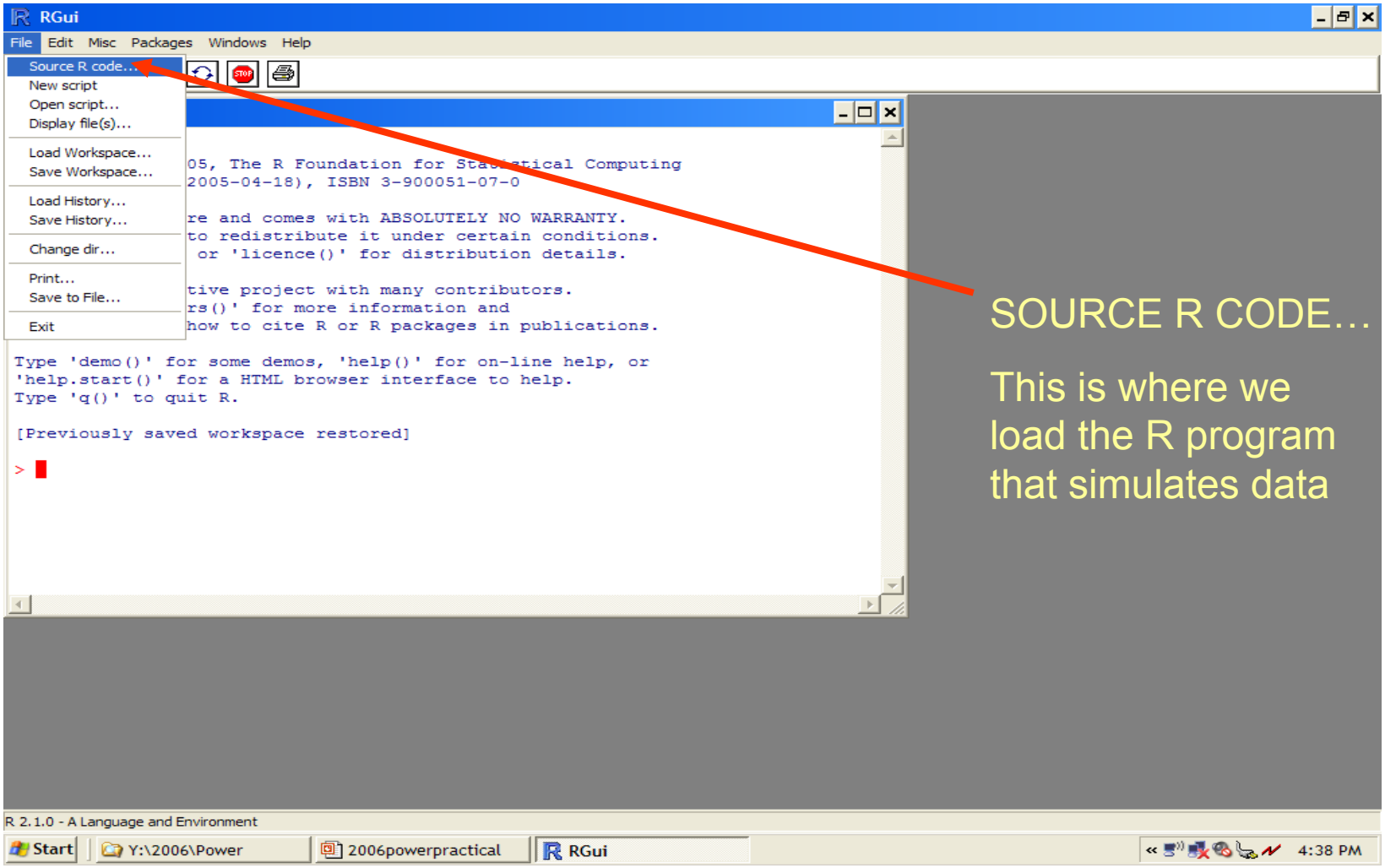
Note you must have the R console highlighted

Picture of the dialog box



Either type the path name or browse to where you saved plink.mds

Running the R script



SOURCE R CODE…

This is where we load the R program that simulates data

Screenshot of source code selection