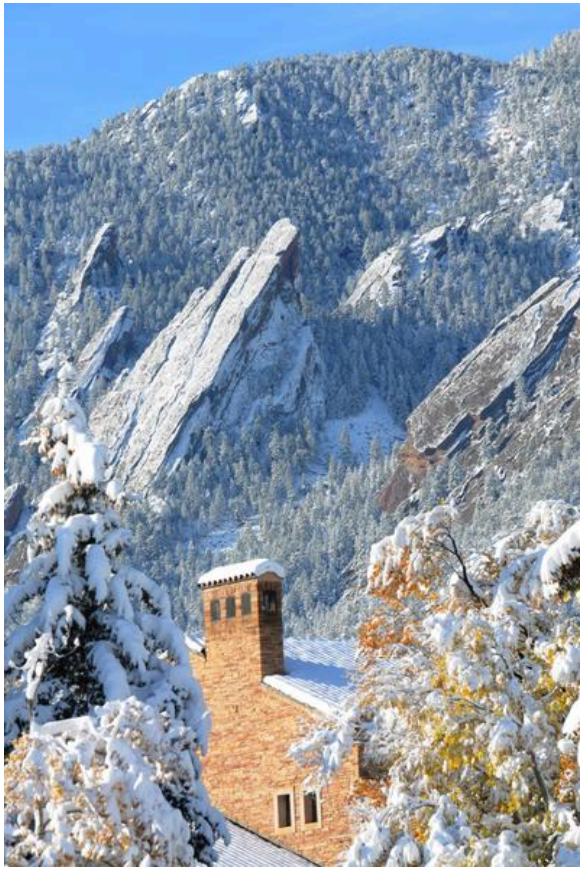
Lecture 2: Quality Control

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado

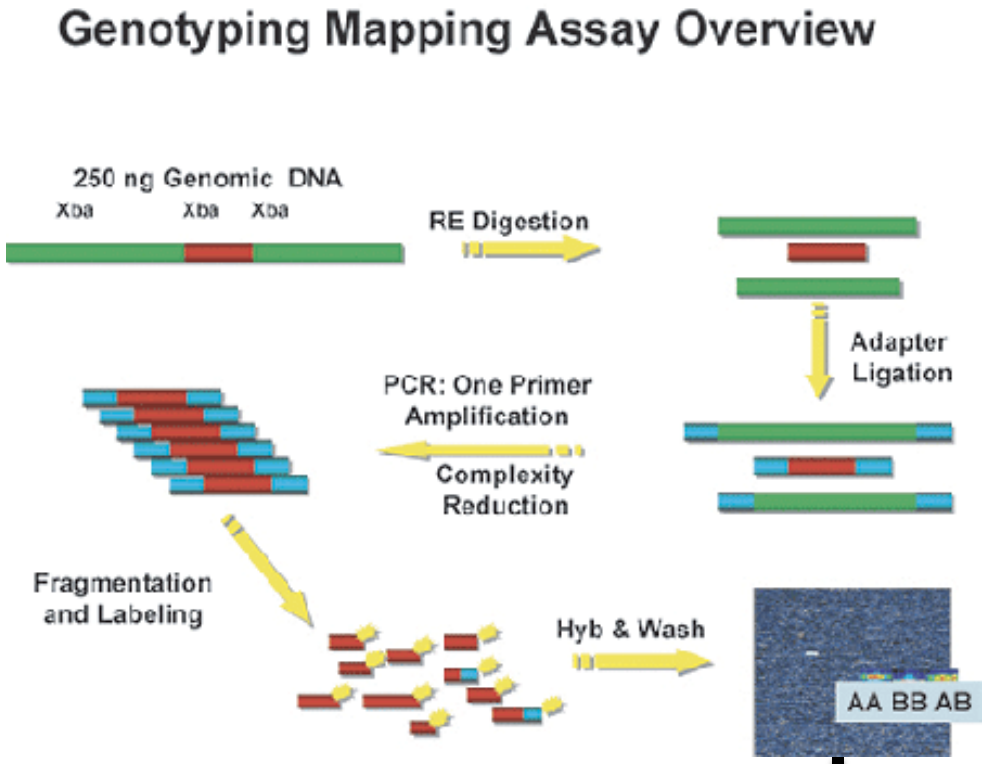


QC Overview

- Genotyping
- Quality Control for Samples
  - Genotyping Call Rate and Heterozygosity
  - Sex Check
  - Relatedness Measures
- Quality Control for Genetic Markers
  - Missing Data Rate
  - Hardy-Weinberg Equilibrium
- Genetic Ancestry

Genotyping

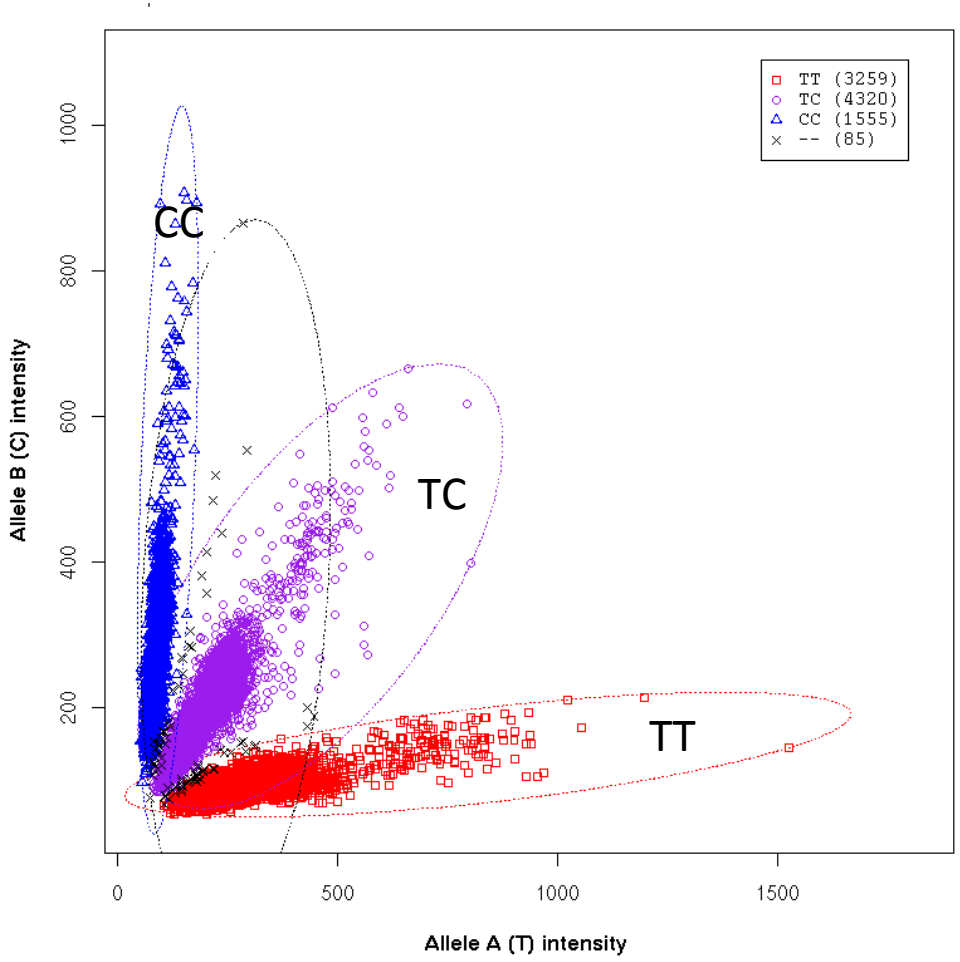
Genotyping



Affymetrix System

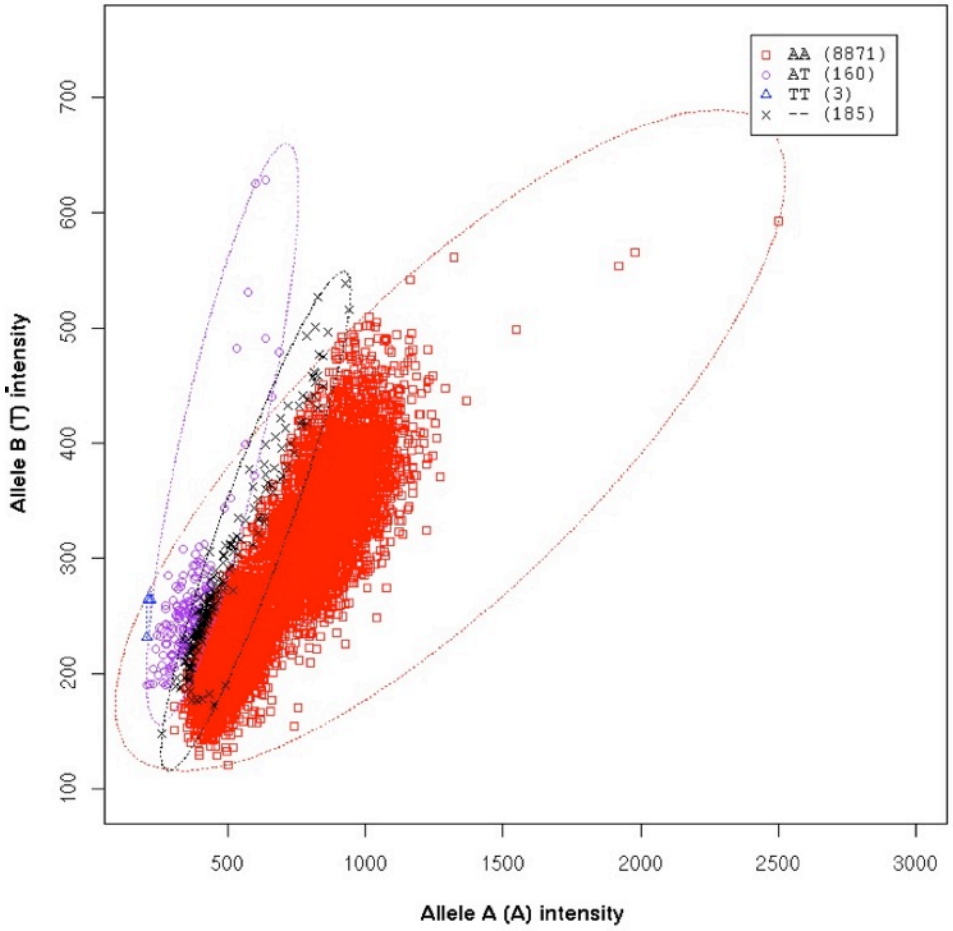
Intensity Files

Intensity to Genotypes



When things go well…

Intensity to Genotypes



When things **don't** go well…

Intensity to Genotypes

- Can’t possibly visibly inspect ALL plots
- Scrutinize any significant or “interesting” associations by going back to the plots
- Just something to be aware of…

Strand Orientation

- Strand unambiguous
  - A/G alleles -> T/C alleles
- Strand ambiguous
  - A/T and C/G

Strand Reporting

- Probe/Target
  - Generically as A/B
  - Illumina/Affymetrix
- Plus (+) / Minus (-)
  - 5’ end at the tip of the short arm = plus
  - 1000 genomes, HapMap
- FWD / REV
  - Based upon submitted flanking DNA sequence
  - dbSNP (NCBI)
- TOP/BOT
  - Based upon flanking sequences
  - Illumina

Strand orientation

http://browser.1000genomes.org

Example:

- rs932402

Raw files (Illumina GenomeStudio)

```
[Header]
GSGT Version    1.6.3
Processing Date 6/28/2011 12:25 PM
Content         HumanOmni1-Quad_v1-0_B.bpm
Num SNPs        1140419
Total SNPs      1140419
Num Samples     672
Total Samples   672
File    1 of 17
[Data]
SNP Name        Sample ID       Allele1 - Top   Allele2 - Top       GC Score
200006  1481-01A_30011425_A01   G                   G               0.8273
200052  1481-01A_30011425_A01   T                   T               0.9487
200053  1481-01A_30011425_A01   A                   A               0.6645
200070  1481-01A_30011425_A01   G                   G               0.9347
200078  1481-01A_30011425_A01   C                   G               0.6885
200087  1481-01A_30011425_A01   A                   A               0.8085
```

GC Score

- GenCall Score
  - Confidence “measure” on the quality of genotypes
  - GC < 0.15 are typically set to “missing”
    - By individual for each SNP

Quality Control for Samples

Genotyping Call Rate

- Per sample (individual) rate
- Number of non-missing genotypes divided by the total number of successfully genotyped markers
- Typical thresholds vary
  - 95-97% call rate

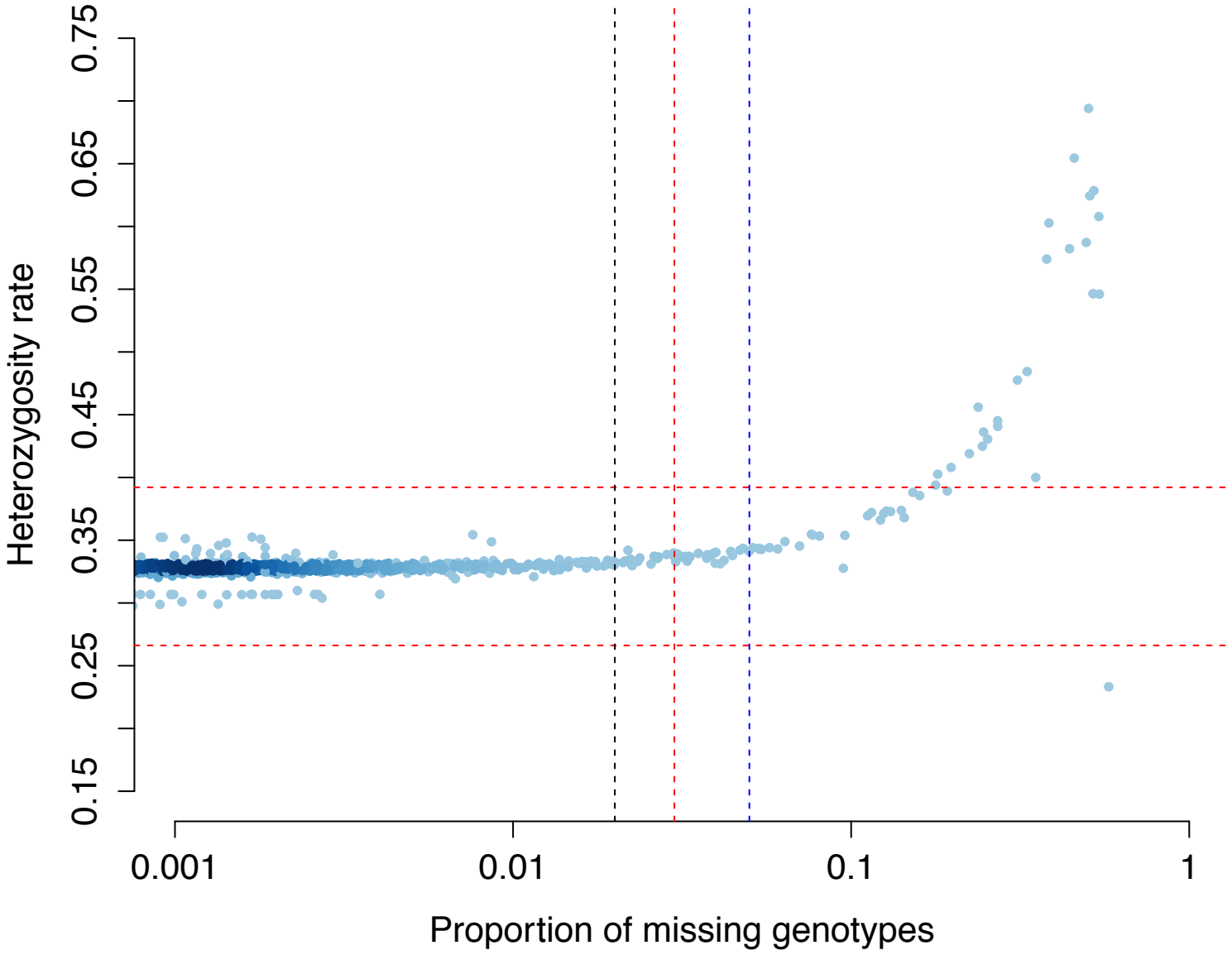
Heterozygosity Rate

- Per sample (individual) rate
- Number of (non-missing – homozygous) genotypes divided by number of homozygous genotypes
- Excess heterozygosity
  - Possible sample contamination
- Less than expected heterozygosity
  - Possibly inbreeding

Graphical inspection

- Often, genotyping call rate and heterozygosity rate are plotted to visualize data

Genotyping Call Rate by Heterozygosity

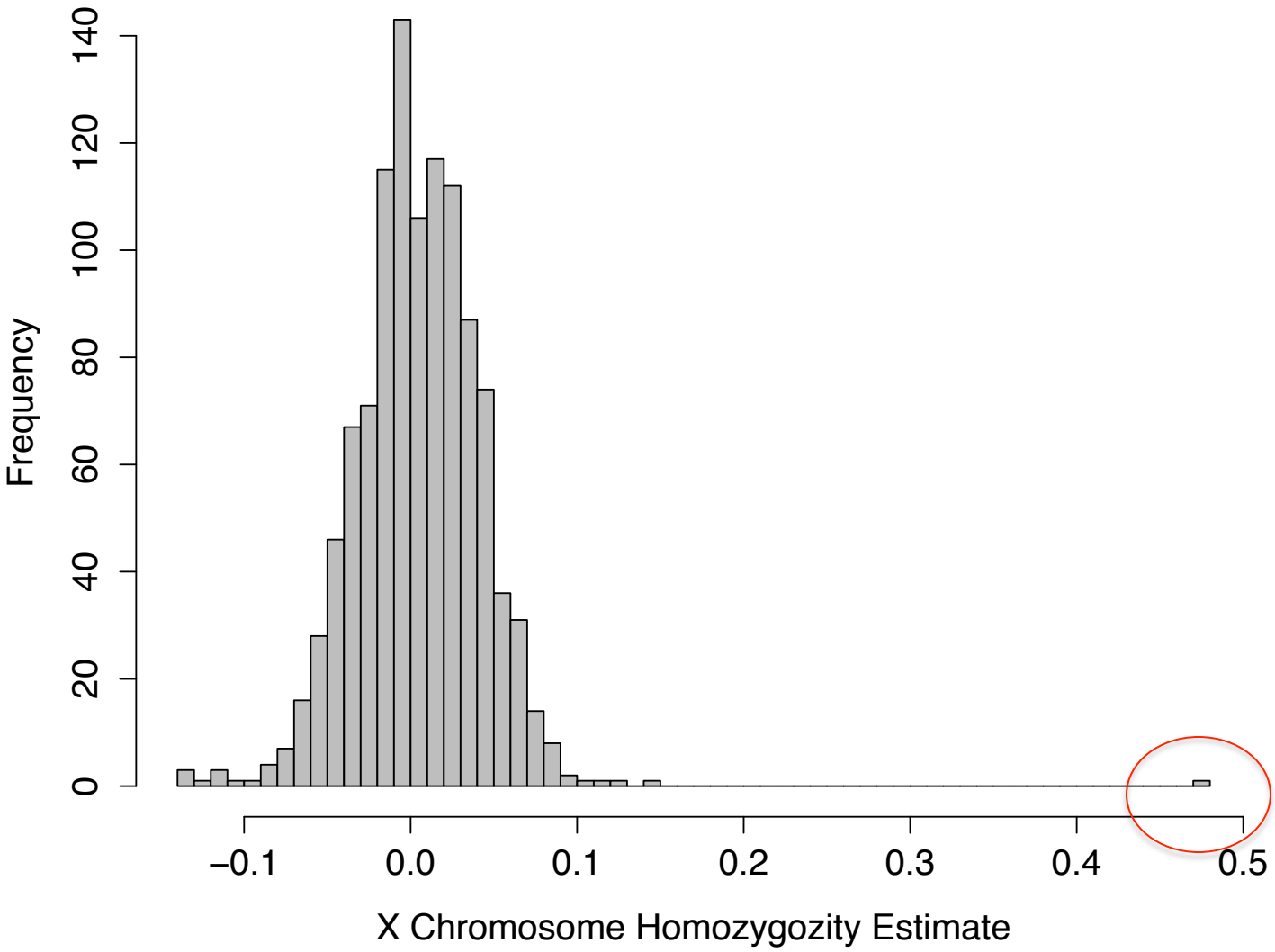


Sex Checks

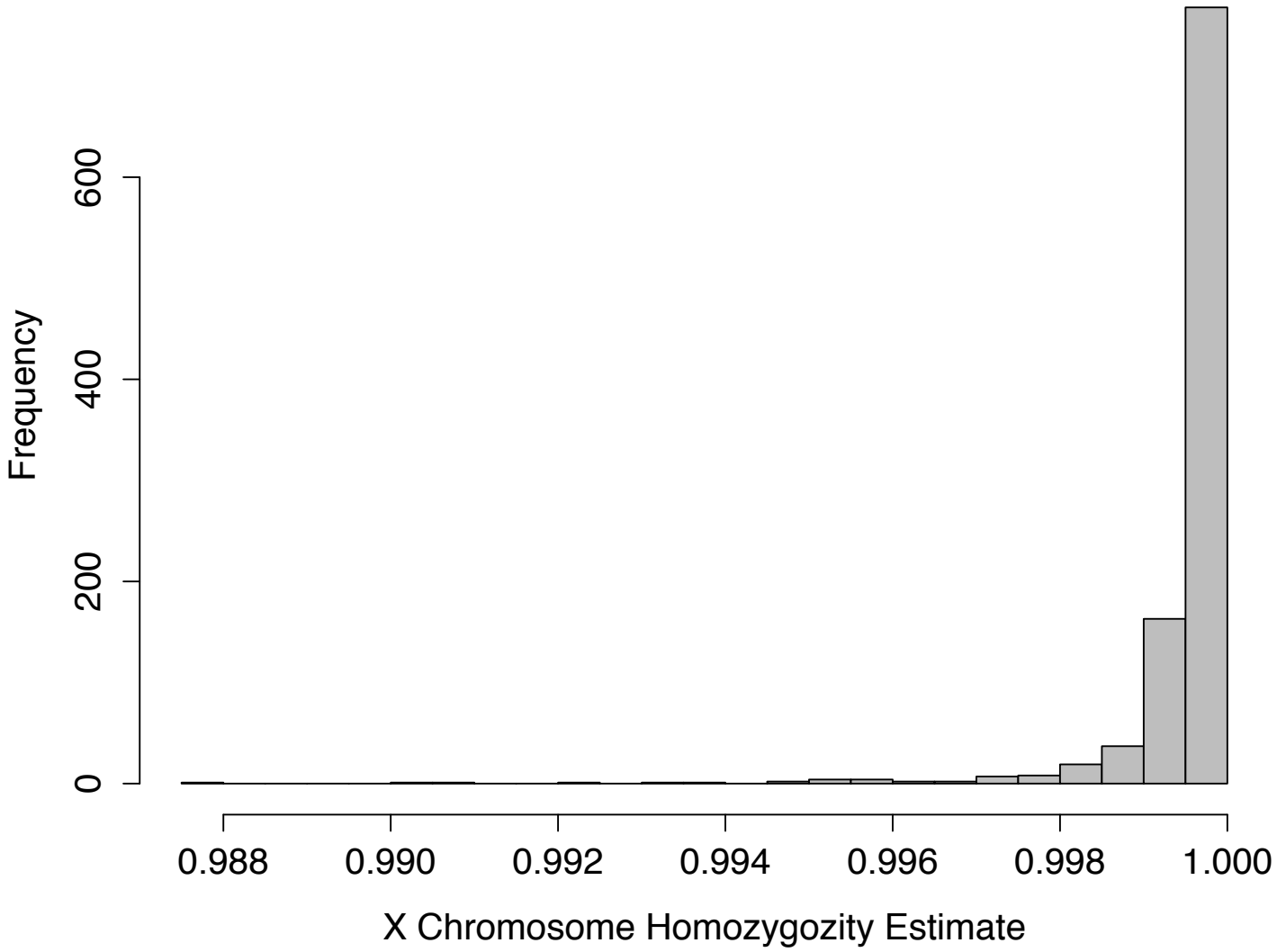
X-Chromosome Homozygosity

- Inbreeding estimate (F)
  - Theoretical range between -1 and 1
- E(Female) ~ 0
  - Two copies of X
- E(Male) ~ 1
  - One copy of X
- Can be used to identify miscoded sex or sample mix-ups, etc.

Females


All Female Samples
Frequency
0
20
40
60
80
100
120
140
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
X Chromosome Homozygozity Estimate

Males


All Male Samples
Frequency
0
200
400
600
0.988
0.990
0.992
0.994
0.996
0.998
1.000
X Chromosome Homozygozity Estimate

Genetic Relatedness

Identity by State (IBS)

ac bd

How many alleles are in common?

IBS = 0

Identity by State (IBS)

ac ad

How many alleles are in common?

IBS = 1

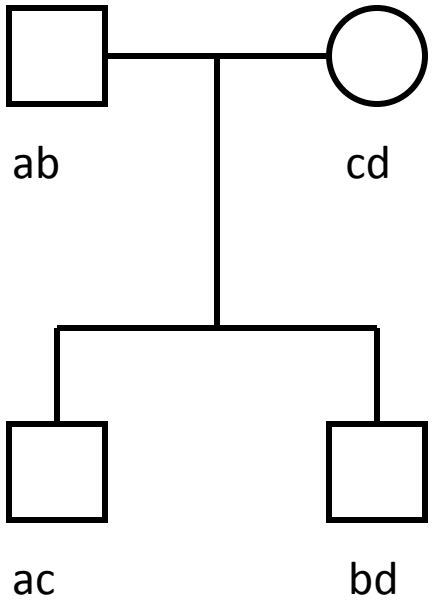
Identity by State (IBS)

ac ac

How many alleles are in common?

IBS = 2

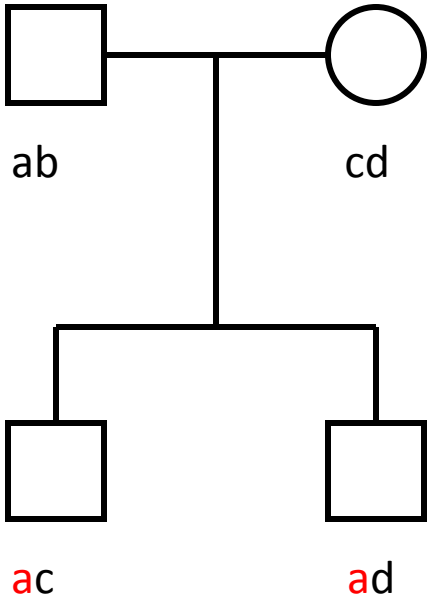
Identity by Descent (IBD)



How many alleles are common by descent?

IBD = 0

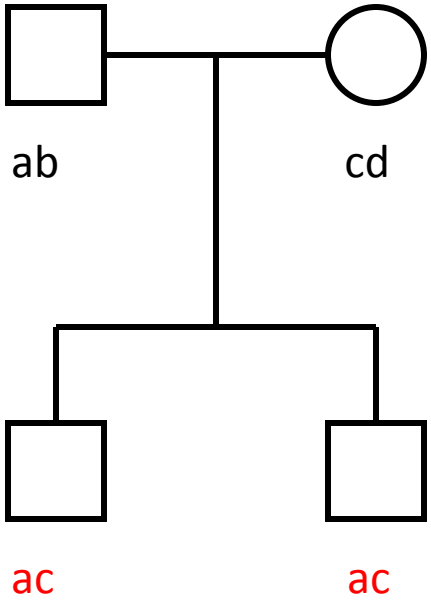
Identity by Descent (IBD)



How many alleles are common by descent?

IBD = 1

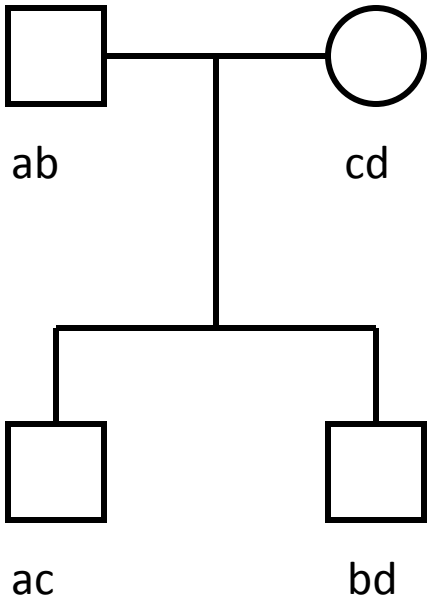
Identity by Descent (IBD)



How many alleles are common by descent?

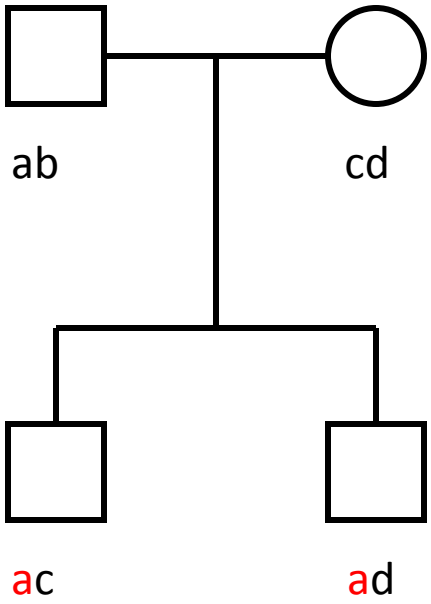
IBD = 2

IBS and IBD



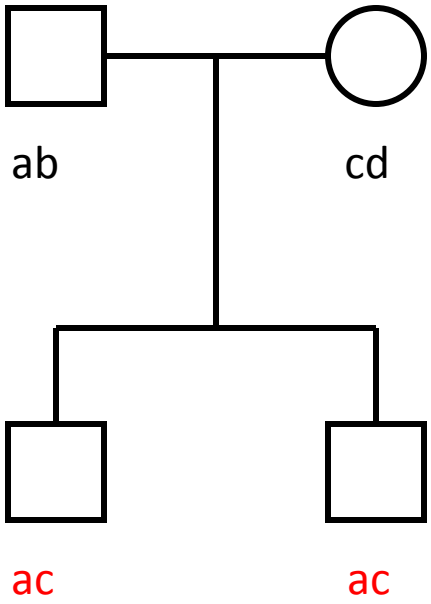
IBS = 0
IBD = 0

IBS and IBD



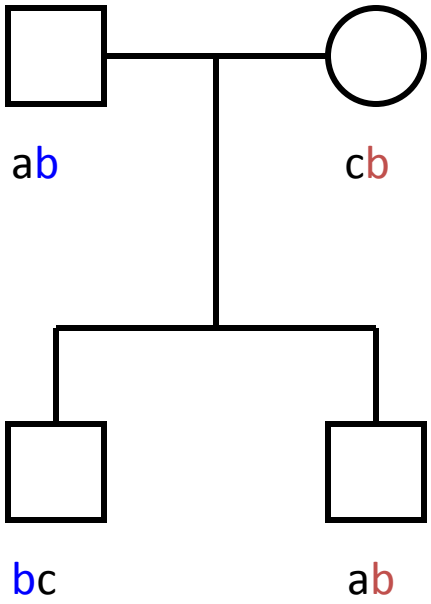
IBS = 1
IBD = 1

IBS and IBD



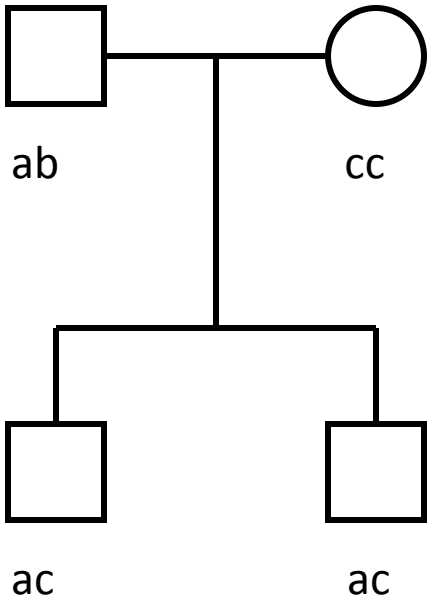
IBS = 2
IBD = 2

Ambiguous IBD



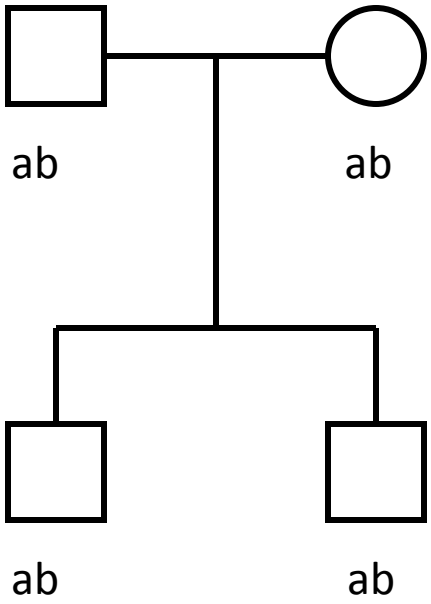
IBS = 1
IBD = 0

Ambiguous IBD



IBS = 2
IBD = ?

Ambiguous IBD



IBS = 2
IBD = ?

IBD Probabilities

	Probability of Sharing IBD Alleles		
Relative Pair	π_0	π_1	π_2
MZ Twins	*0*	*0*	*1*
Full Sibs	*0.25*	*0.50*	*0.25*
Parent-Offspring	*0*	*1*	*0*
First Cousin	*0.75*	*0.25*	*0*
Grandparent-Grandchild	*0.50*	*0.50*	*0*
Half-Sibs	*0.50*	*0.50*	*0*
Avuncular	*0.50*	*0.50*	*0*

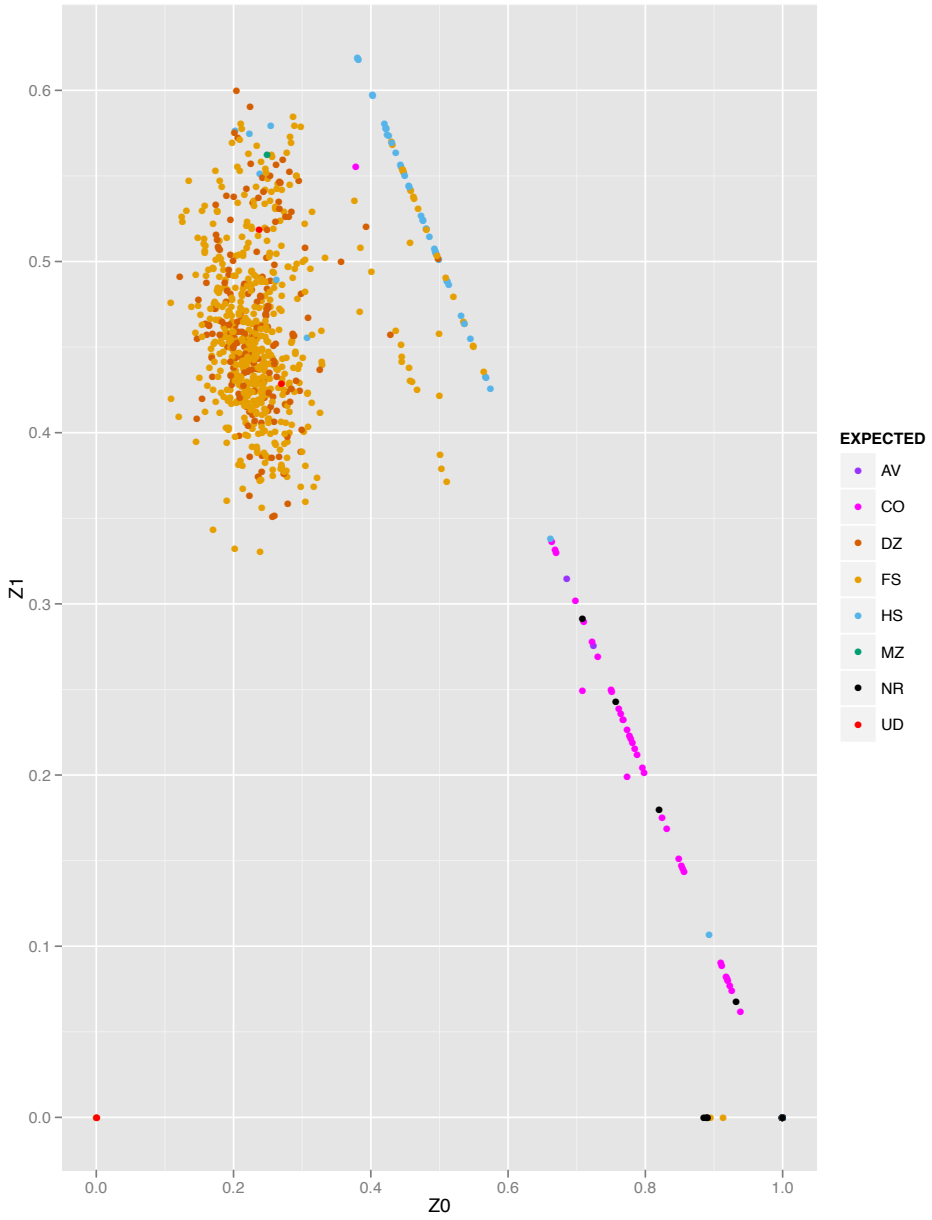
Checking Relationships

- Mean IBD = [(Z1/2) + Z2]
- "PI_HAT" in PLINK
  - PI_HAT = 1 for duplicates/MZ twin pairs
  - PI_HAT = 0.5 for sibling pairs
  - Etc.

Be warned!

- Using PLINK for relationships (PI_HAT) can be biased with a diverse sample
  - Two people may appear related (PI_HAT > 0.10) but are merely of the same ethnic group
- Alternatives
  - Kinship-based Inference for GWAS
    - KING
  - Relatedness Estimation in Admixed Populations
    - REAP

Checking Relationships



Using Relatedness Measures

- Identify related subjects
  - When they should be unrelated
- Identify unrelated subjects
  - When they should be related
- Duplicates
  - IBS/PI_HAT =1

Quality Control for Genetic Markers

Sample Removal

- Low genotyping call rate: < 97%
- High/low heterozygozity rate
  - (visual inspection)
- Unresolved sex-check
- Duplicate samples (relationship measures)

Missing Data Rate (Markers)

- Per marker
- Number of samples missing the marker genotype divided by the total number of successfully genotyped samples at that marker
- Typically, markers with missing data > 5% are removed
  - This is flexible however

Missing Data Rate Patterns

- Chi-square test
  - Cases vs Controls
  - Samples from population X vs population Y
  - Blood vs Saliva
  - Etc.

Hardy Weinberg Assumptions

- Diploid organisms
- Infinite population size
- Non-overlapping generations
- Random mating
- No selection, mutation or migration

Testing for HWE

1. Calculate the allele frequency (p)
   - Using observed genotype counts

2. Calculate the expected genotype counts
   - Using the allele frequency (p)

3. Compare the observed to the expected counts
   - χ^2 test

HWE Example

- Observed Genotypes

Genotype	*GG*	*Gg*	*gg*
Frequency	*12*	*2*	*8*

Step I

- Observed Genotypes

Genotype	*GG*	*Gg*	*gg*
Frequency	*12*	*2*	*8*

1. *Calculate the allele frequency (p):*

$$p = \frac{2(12)+2}{2(22)} = 0.59$$

Step II

- Observed Genotypes

Genotype	*GG*	*Gg*	*gg*
Frequency	*12*	*2*	*8*

2. Calculate the expected genotype counts:

$$E(GG) = np^2 = 22(0.59^2) = 7.66$$

$$E(Gg) = n2pq = 22(0.59)(1-0.59) = 10.64$$

$$E(gg) = nq^2 = 22((1-0.59)^2 = 3.68$$

Step III

- Observed Genotypes

Genotype	*GG*	*Gg*	*gg*
Frequency	*12*	*2*	*8*

3. Compare the observed and expected counts:

$$\chi_1^2 = \frac{(12-7.66)^2}{7.66} + \frac{(2-10.64)^2}{10.64} + \frac{(8-3.69)^2}{3.69} = 14.50$$

REJECT THE NULL!

Reasons for HW Deviations

- Genotyping Error
- Subdivided Population
  - Excess homozygotes = “Wahlund Effect”
- Any violations of the HW assumptions

HWE on the Genome-Wide Scale

- Threshold for significance
  - 10^{-3} to 10^{-6}

Marker Removal

- Missing Data > 5%
- Missing data rate associated with case status
- HWE test p-value < 0.001
- No reliable map location
- Monomorphic (no variation)
- Low frequency
  - For example, < 1% or < 5%

Genetic Ancestry

Genetic Ancestry

- Falls under QC to an extent, but also used in the analysis
- Most ancestry measures are based upon "relatedness" estimation
  - IBS, IBD
- Use pair-wise relationship measures to capture population level information

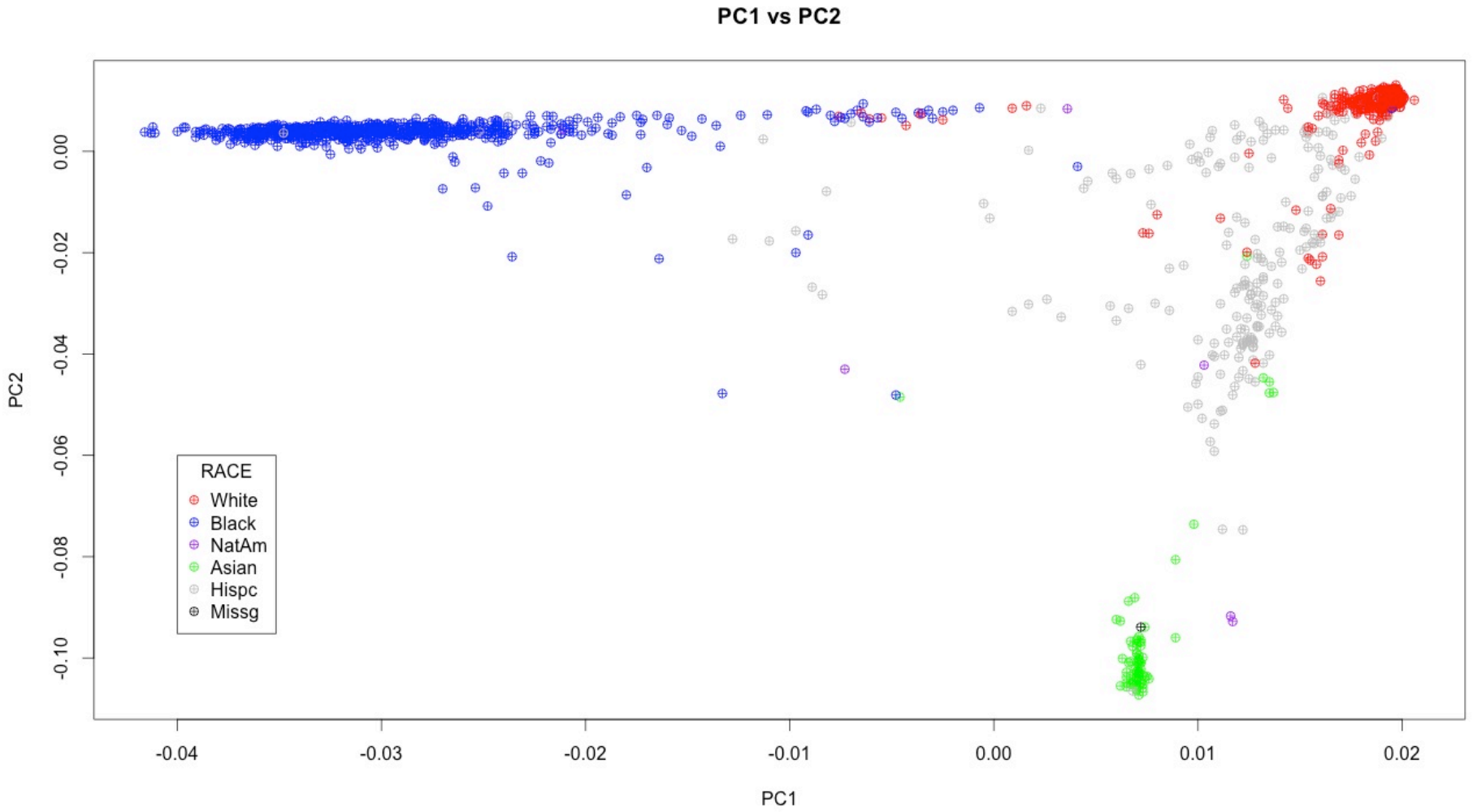
Genetic Ancestry Tools

- PLINK
  - Multidimensional Scaling (MDS)
- Eigenstrat
  - Principal Components
- STRUCTURE
  - Cluster-based algorithm
- ADMIXTURE
  - ML approach (similar to STRUCTURE)

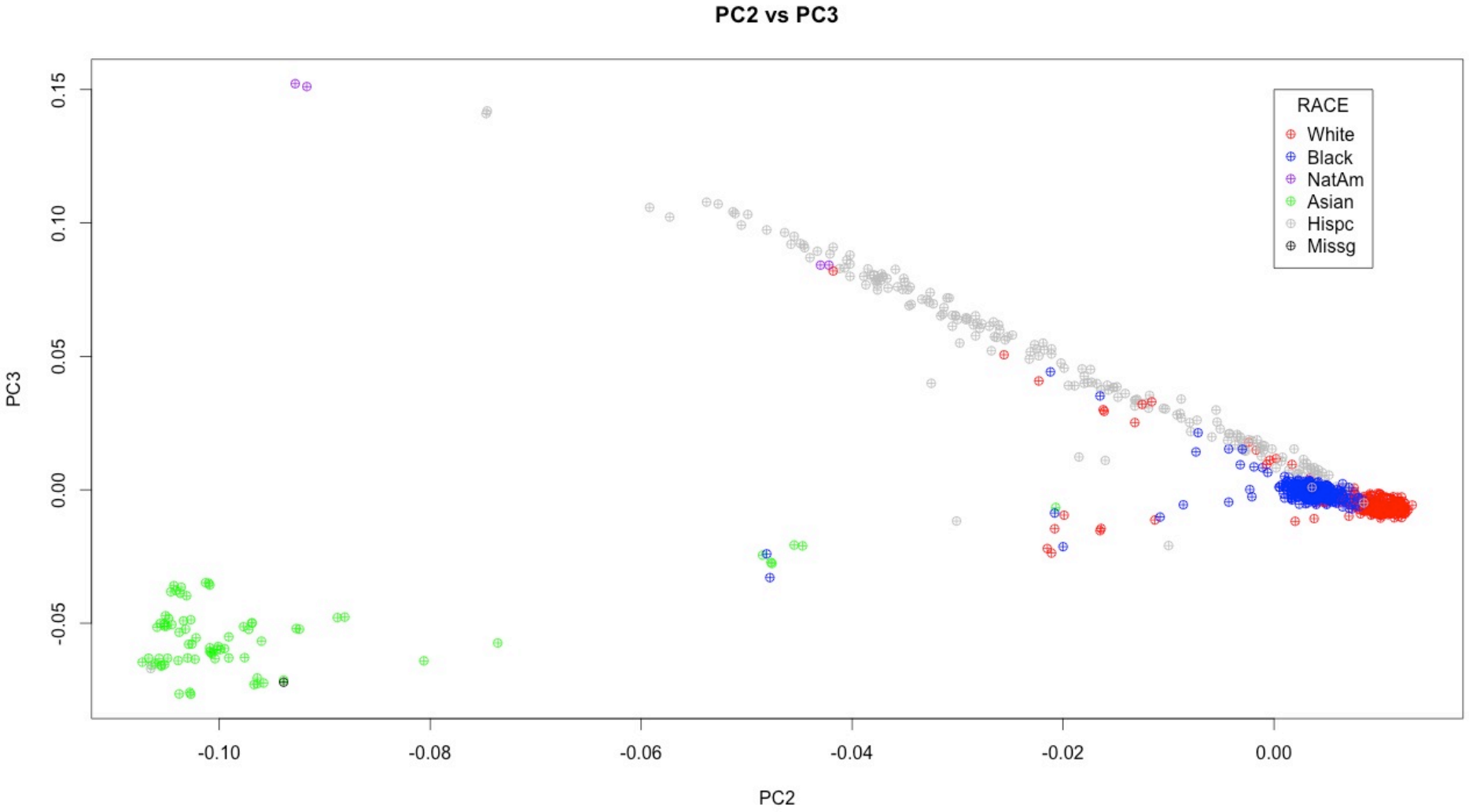
Using PLINK for Ancestry

- MDS Approach
- Using IBS/IBD information
- Generate any number of principal coordinates
  - Typically, at least 10 (some up to 20)

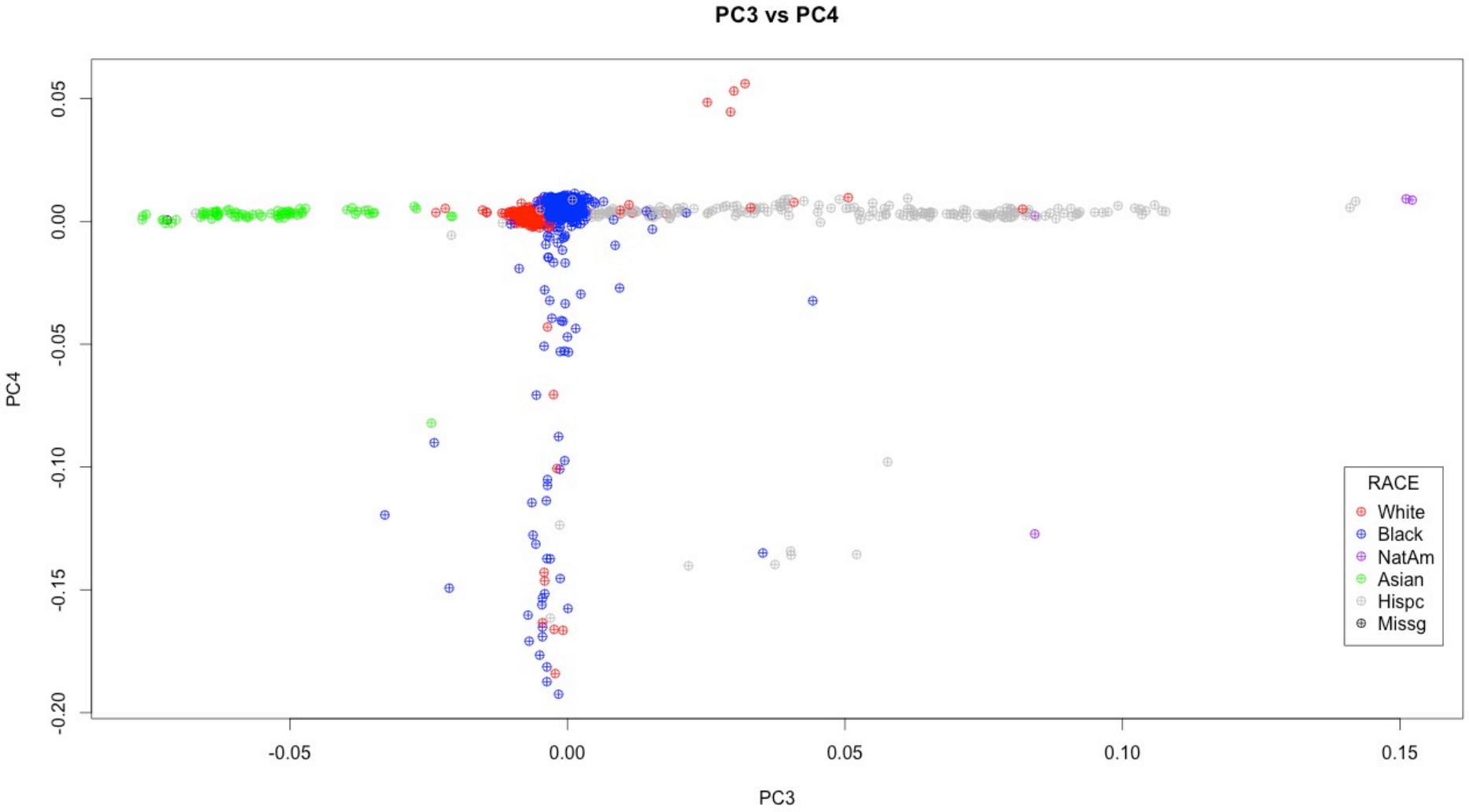
MDS Plot


PC1 vs PC2
PC2
PC1
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02
RACE
White
Black
NatAm
Asian
Hispc
Missg

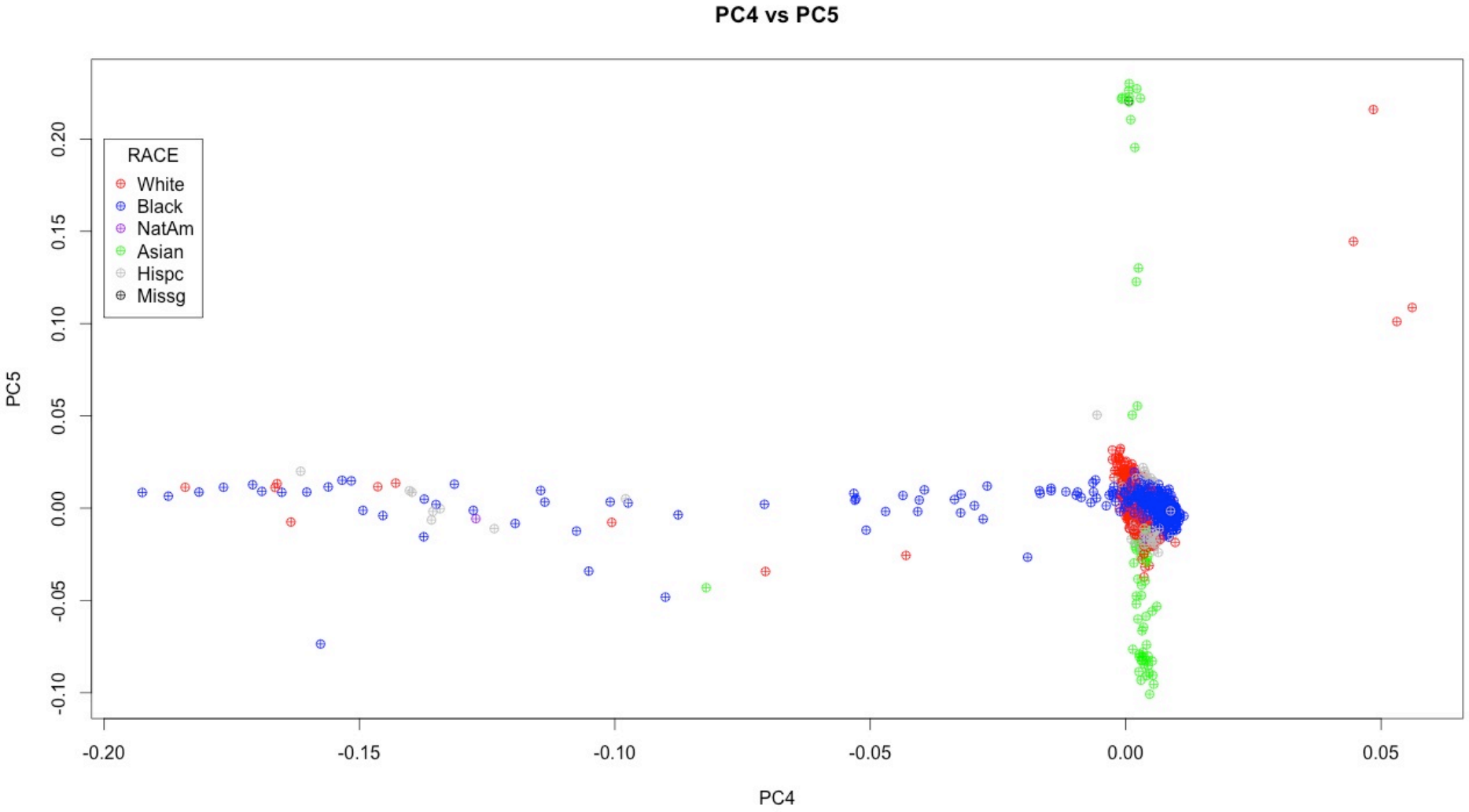
MDS Plot


PC2 vs PC3
RACE
White
Black
NatAm
Asian
Hispc
Missg
PC3
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
PC2

MDS Plot


PC3 vs PC4
PC4
PC3
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.05
0.00
0.05
0.10
0.15
RACE
White
Black
NatAm
Asian
Hispc
Missg

MDS Plot

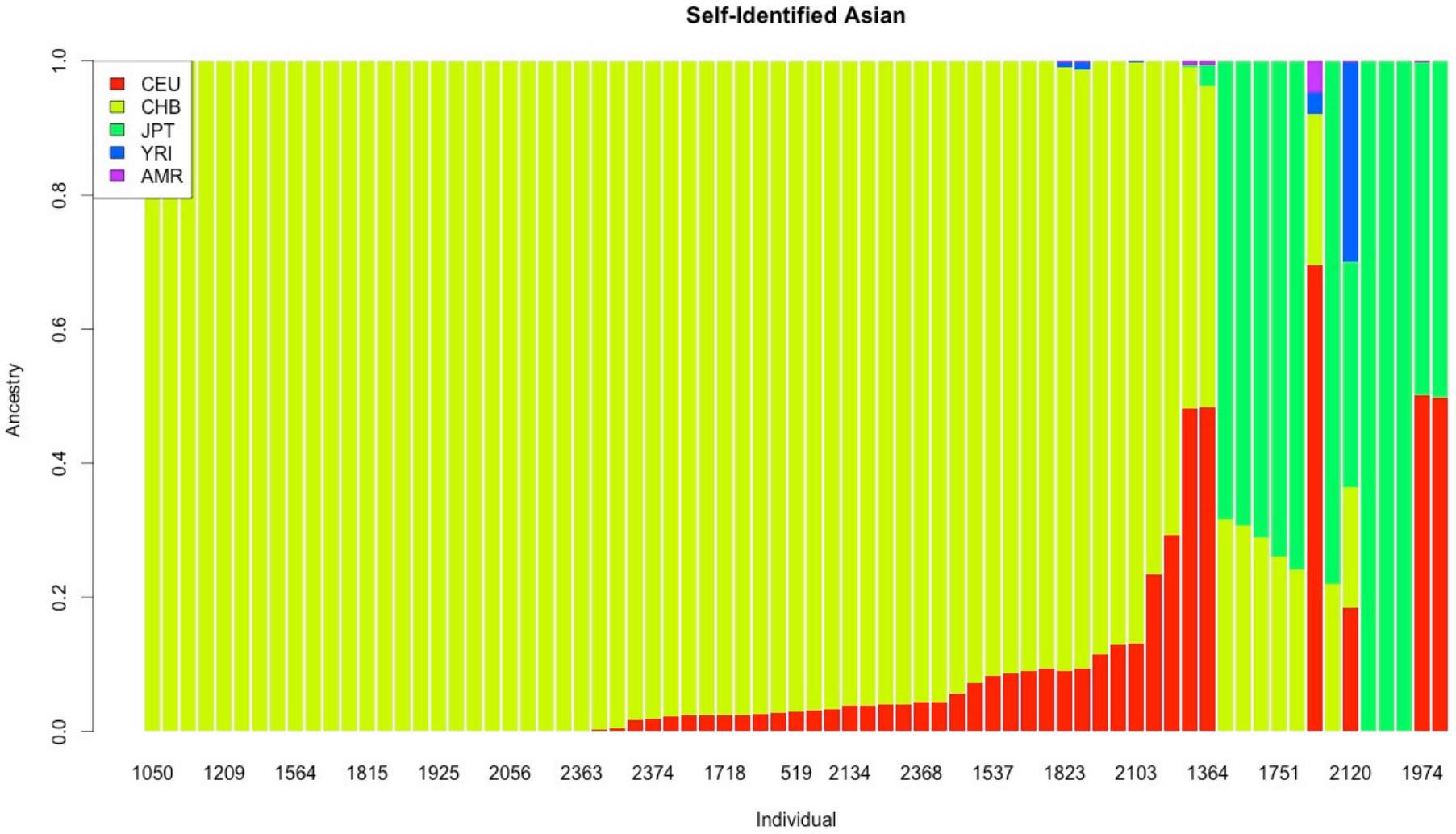

PC4 vs PC5
RACE
White
Black
NatAm
Asian
Hispc
Missg
PC5
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.20
-0.15
-0.10
-0.05
0.00
0.05
PC4

ADMIXTURE

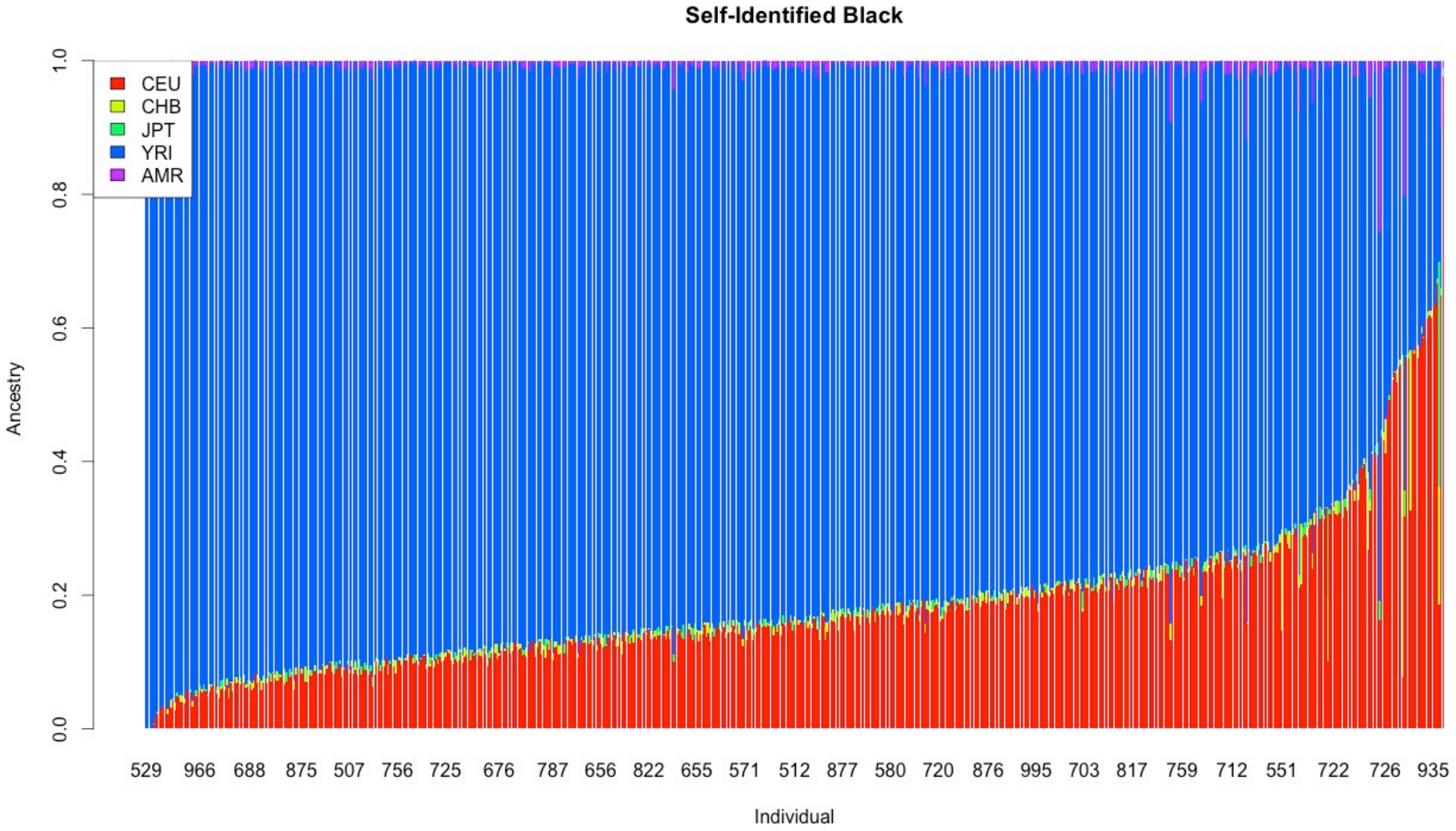
- Similar to STRUCTURE (but faster)
- ML estimation of individual ancestries

https://www.genetics.ucla.edu/software/admixture/

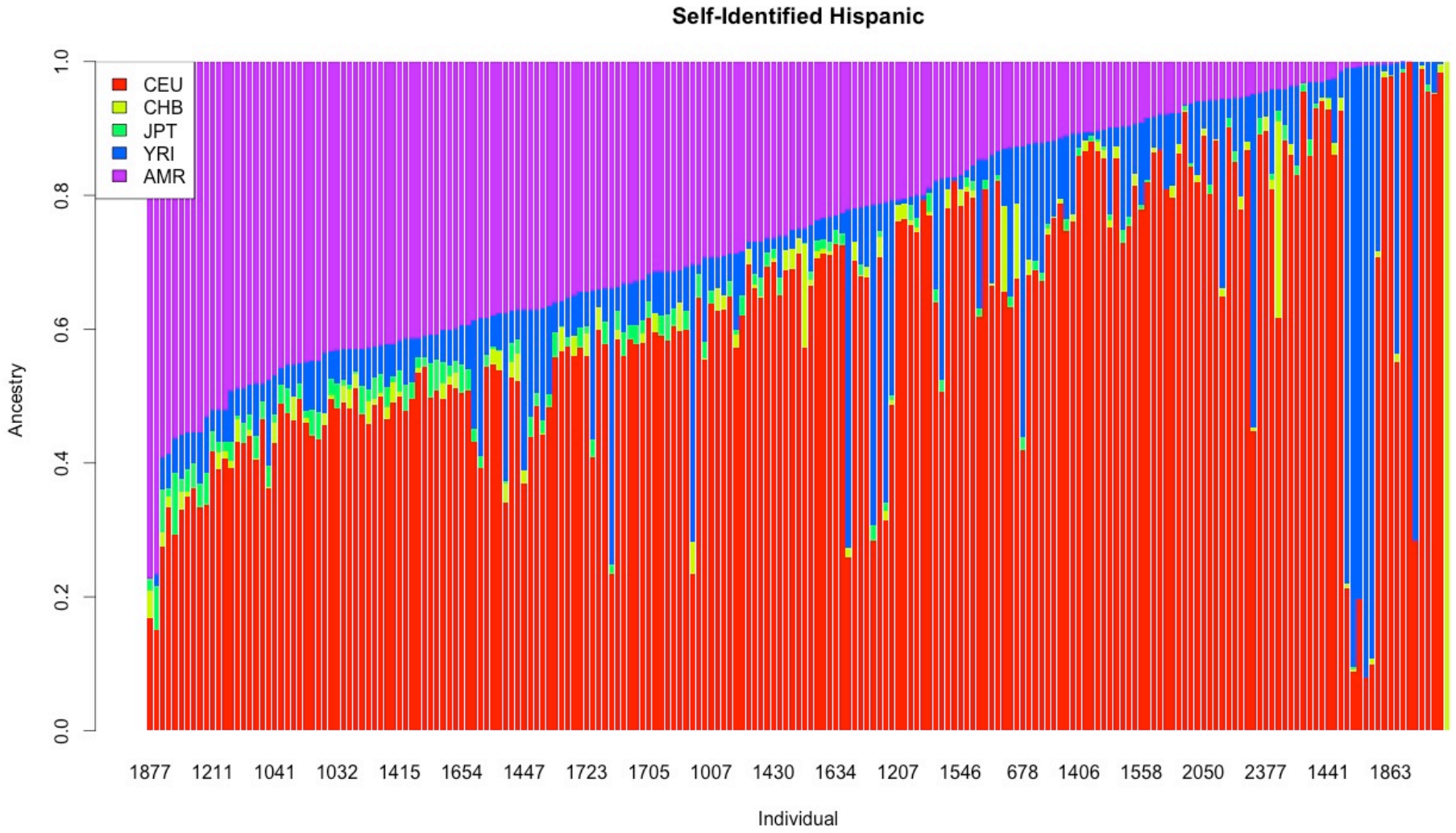
ADMIXTURE - Asian



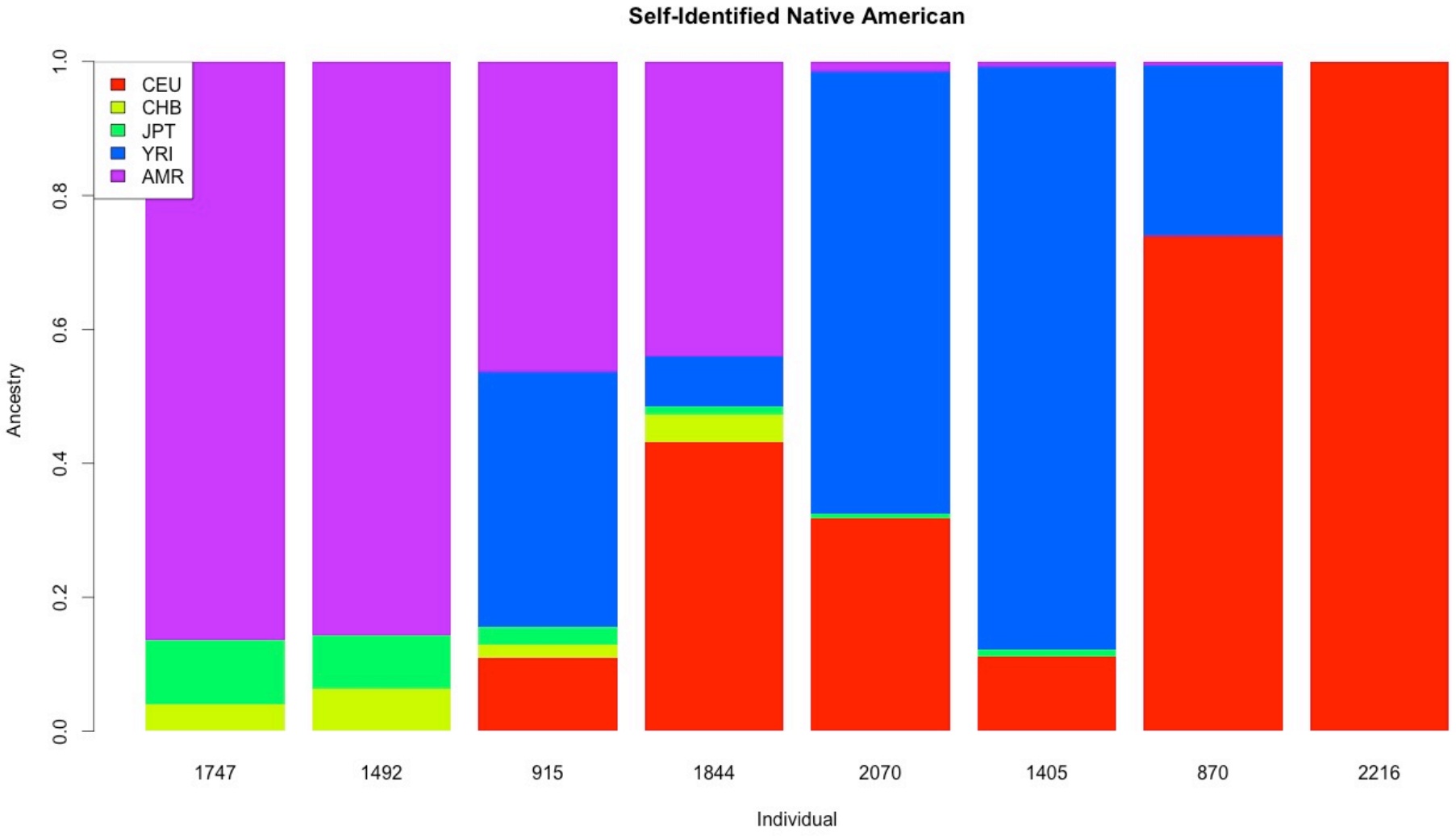
ADMIXTURE – African American



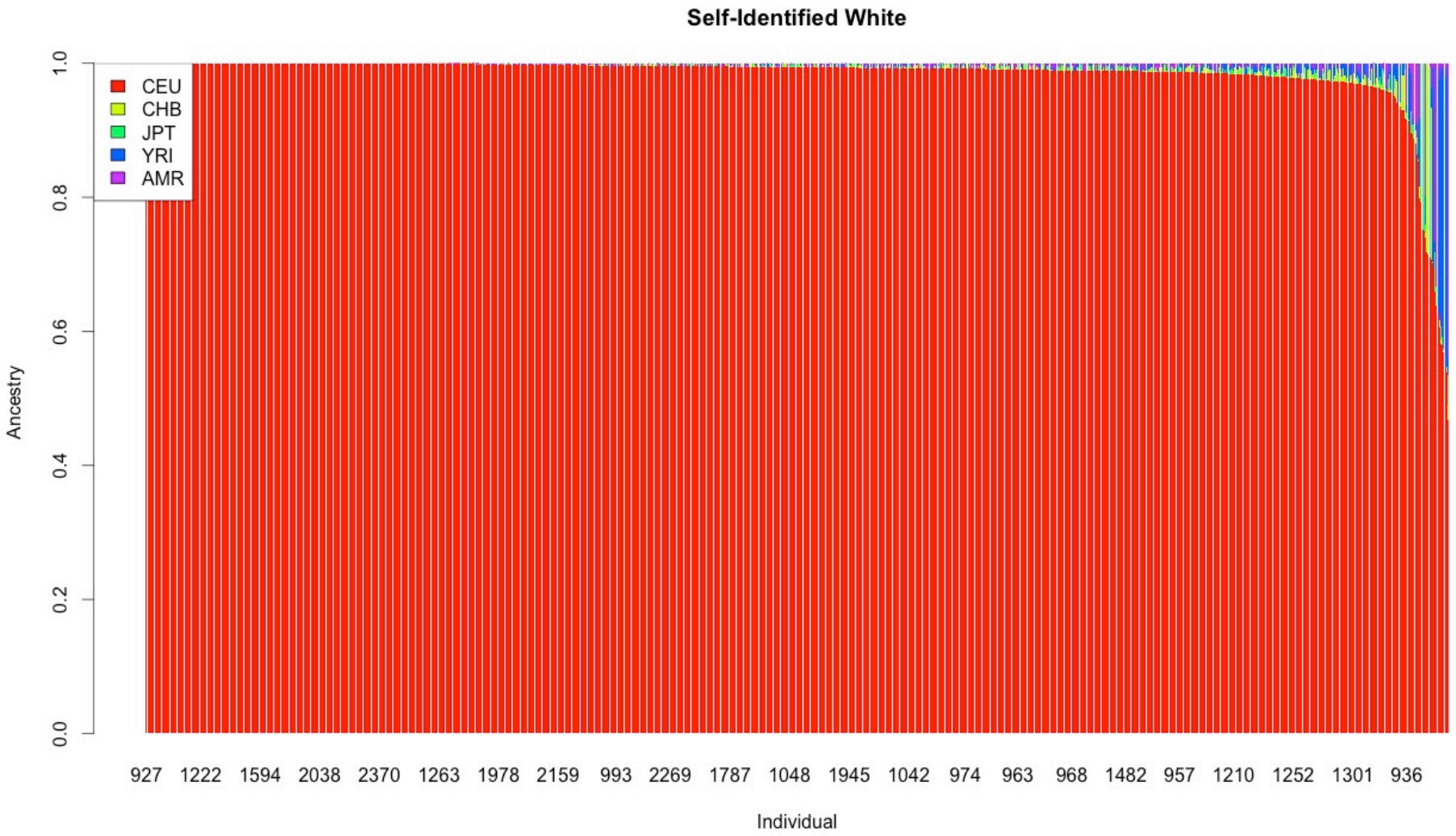
ADMIXTURE - Hispanic



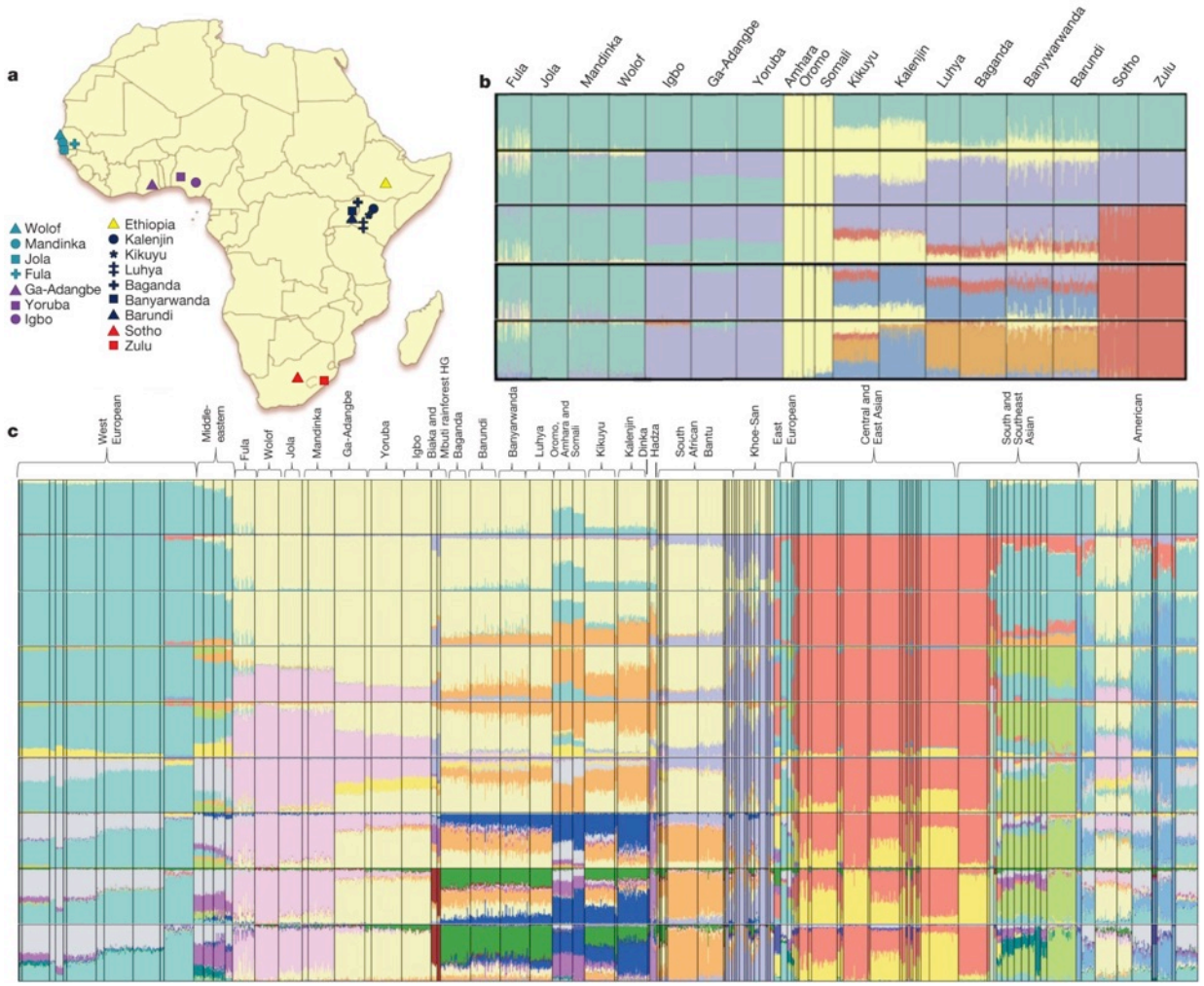
ADMIXTURE – Native American



ADMIXTURE – White



Populations studied in the AGVP.



D Gurdasani *et al. Nature* **000**, 1-6 (2014) doi:10.1038/nature13997

nature

Applications

- Control for population stratification
  - More on this later
- Construct a homogenous sample
  - Beyond self-report ethnicity
- "Control" for the genetic influences
- Identify mislabeled samples

Next up…

- Tutorial 2