Practical: Population structure

Alicia Martin
Postdoctoral Research Fellow

December 10, 2016

Objectives

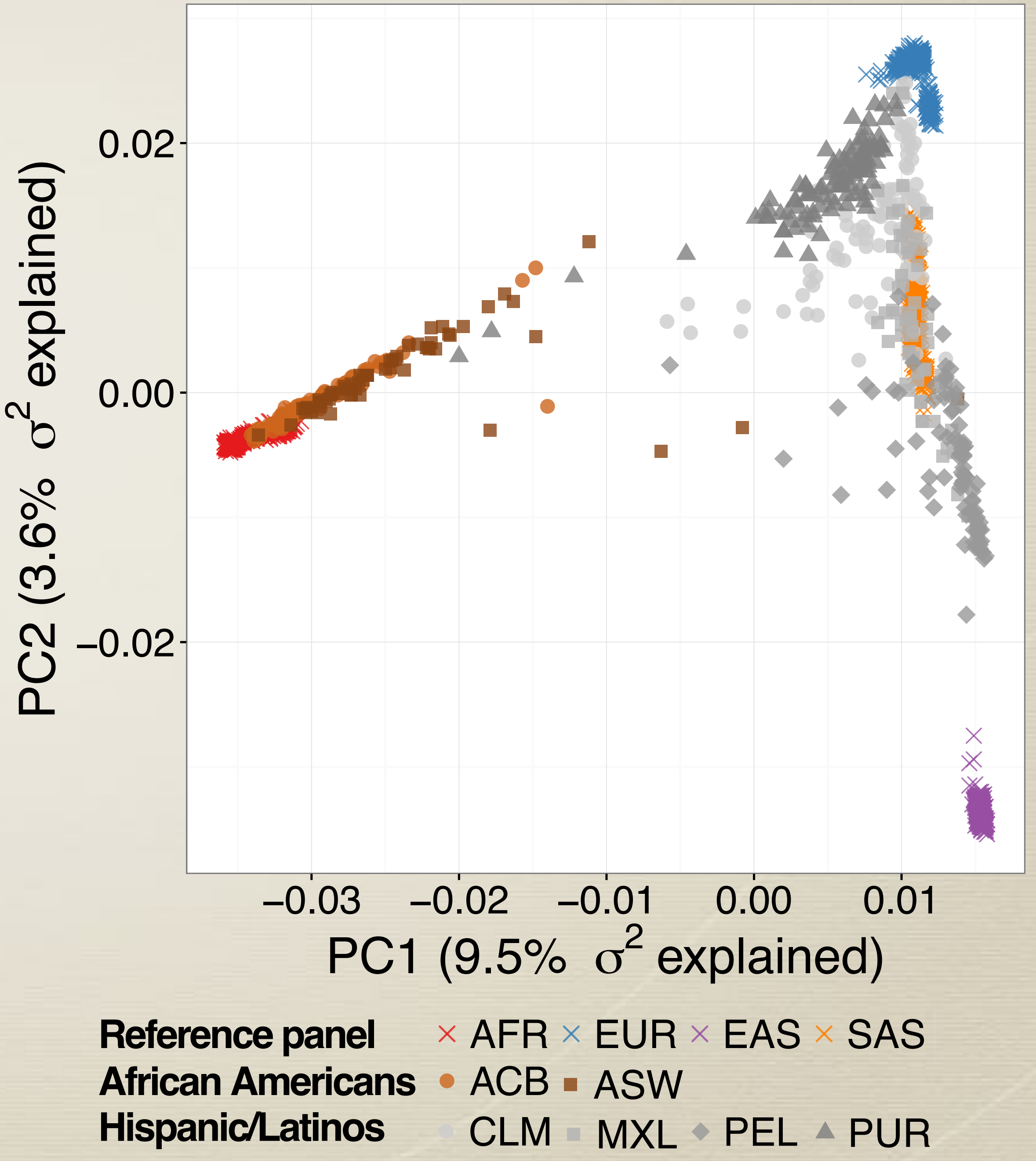
- PCA
- AS-PCA
- ADMIXTURE

Useful summaries

- `dim, length, ncol, nrow` - get size of your data
- `head, tail` - get top/bottom 6 rows
- `summary` - summarize you dataset
- `typeof` - determine type (e.g. data.frame = list, matrix, vector=numeric, character, etc)
- `?typeof` - gives help for using typeof, can be used for any function
- `example(typeof)` - provides example of function

PCA

- I ran PCA on QC'd 1000 Genomes genotype data with smartpca in EIGENSTRAT
- You could do the same using prcomp in R
- `?prcomp`



Principal components analysis (PCA)

$$m = \text{markers}$$

$$n = \text{individuals}$$

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix}$$ (centered, scaled rows)

$$X = USV^T$$ Singular Value Decomposition

U is $m \times m$, where $U^T U = I$

S is $m \times n$ singular value matrix

V is $n \times n$, where $V^T V = I$

Principal components analysis (PCA)

$$\begin{aligned}
X^T X =& V S^T U^T U S V^T \\
=& V S^T S V^T \\
X^T X V =& V S^T S V^T V \\
=& V S^T S \\
X^T X v_k =& \lambda_k v_k
\end{aligned}$$

λ_k are eigenvalues of $X^T X$

v_k are eigenvectors of $X^T X$

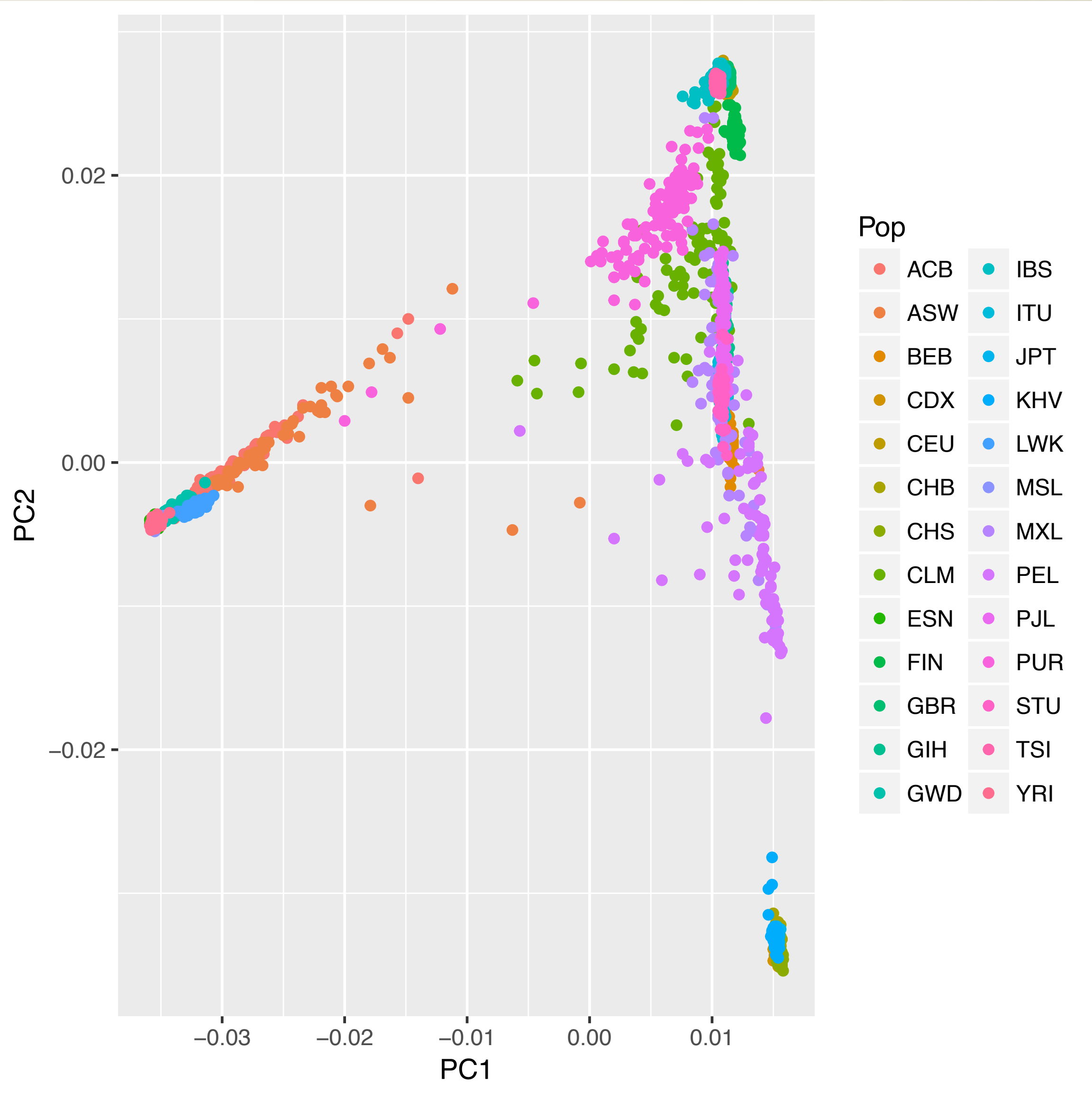
- Multidimensional scaling (MDS) is mathematically equivalent to PCA/SVD
- These are insufficient to correct for rare variants!

Load/process PCA data

- Run through “load” part of the script
- Explore data types/structures using summary functions outlined previously (e.g. head, dim, etc)
- Understand evec/eval output
- Reformat and rename PCA columns, make simple PCA plot

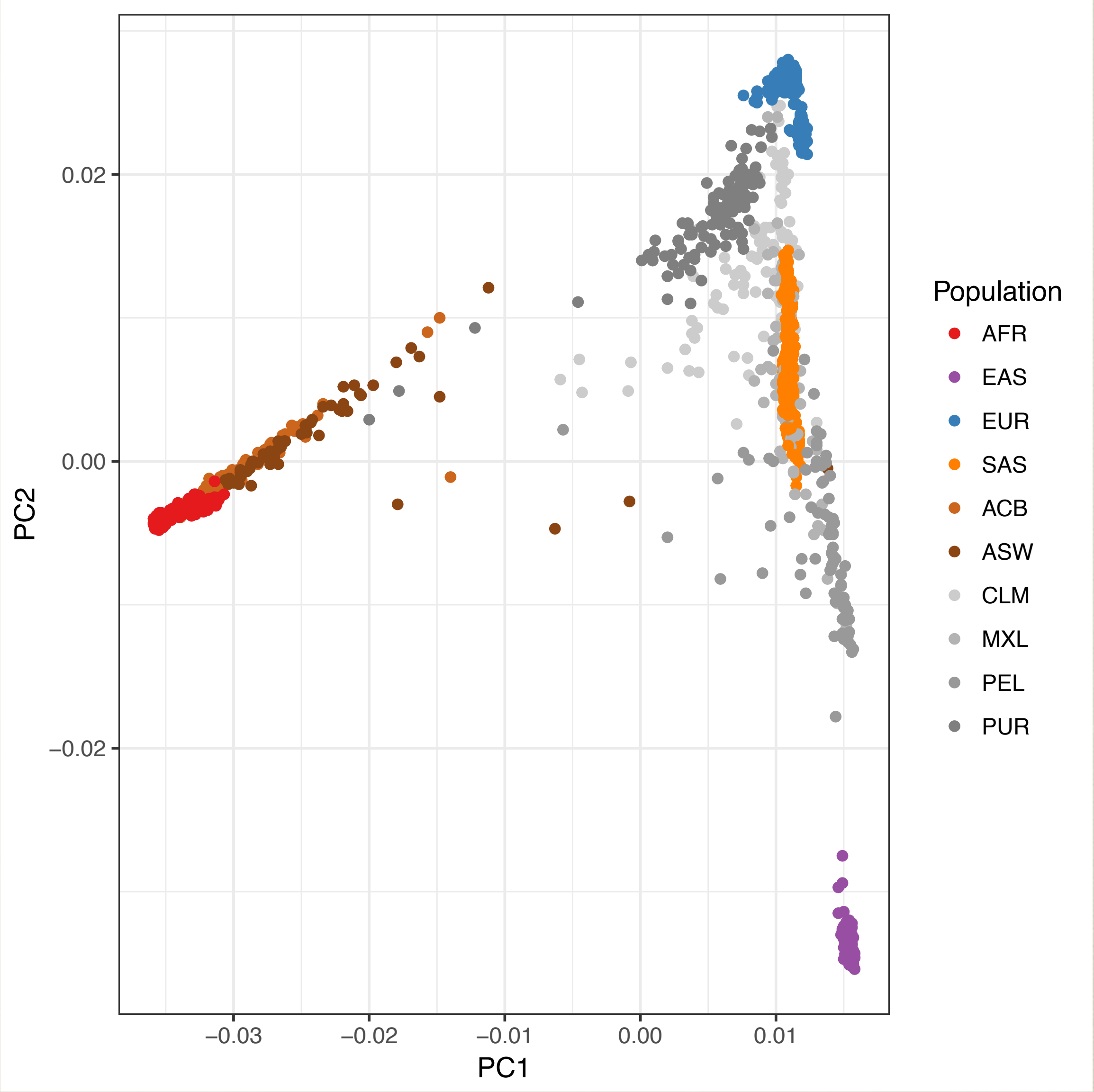
Make simple PCA plot

- Reformat columns
- Rename columns
- Make simple plot
- Next: Fix number of population labels, legend label, background color, etc

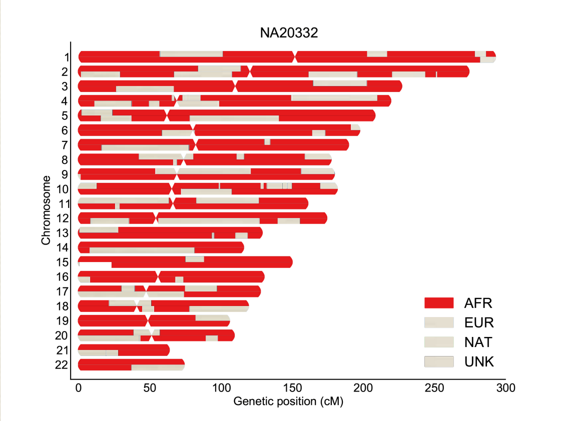


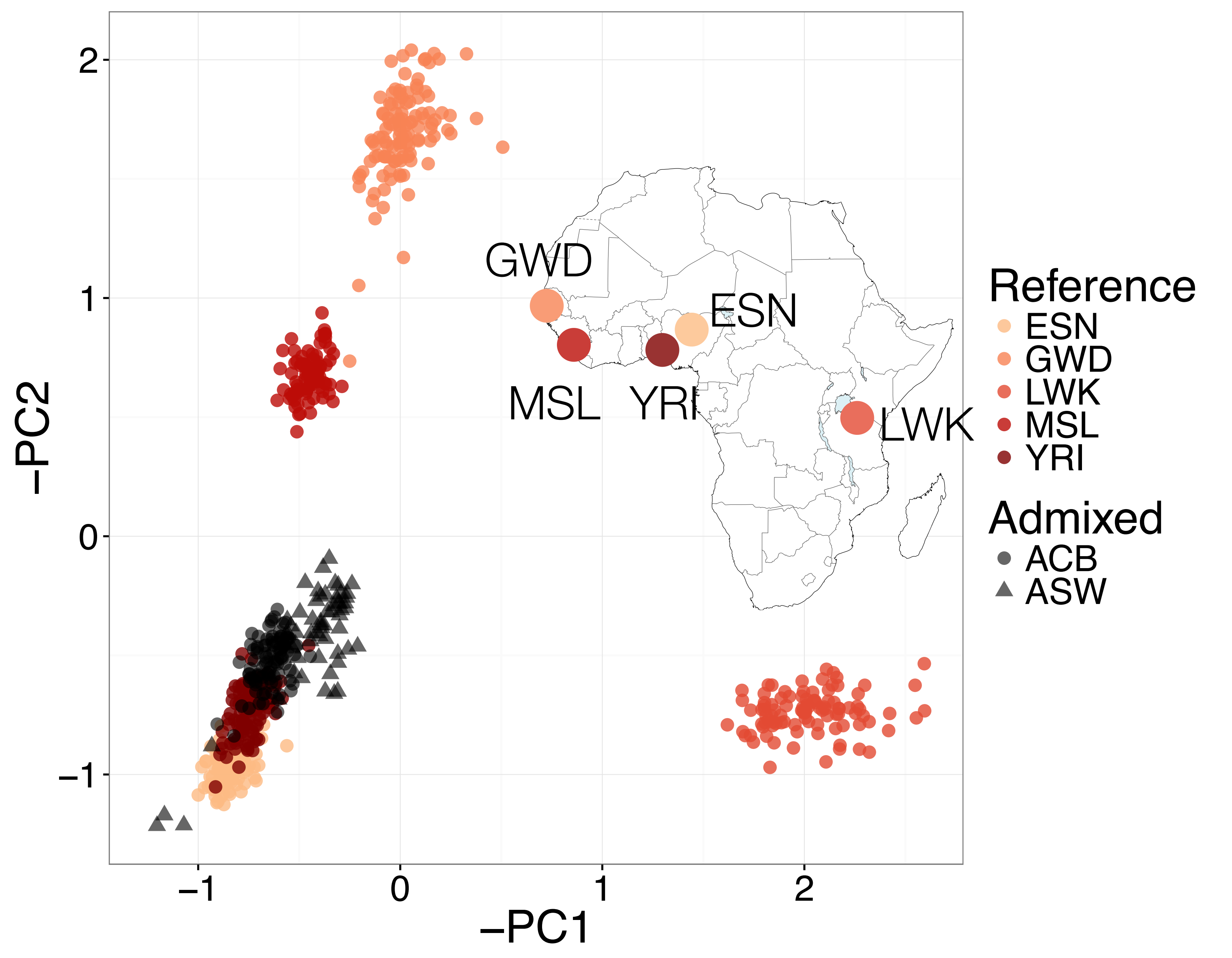
Make final PCA plot

- Merge PCA loadings and colors
- Make a label column, including continent or admixed population label
- Make a color vector by population
- Plot PCA



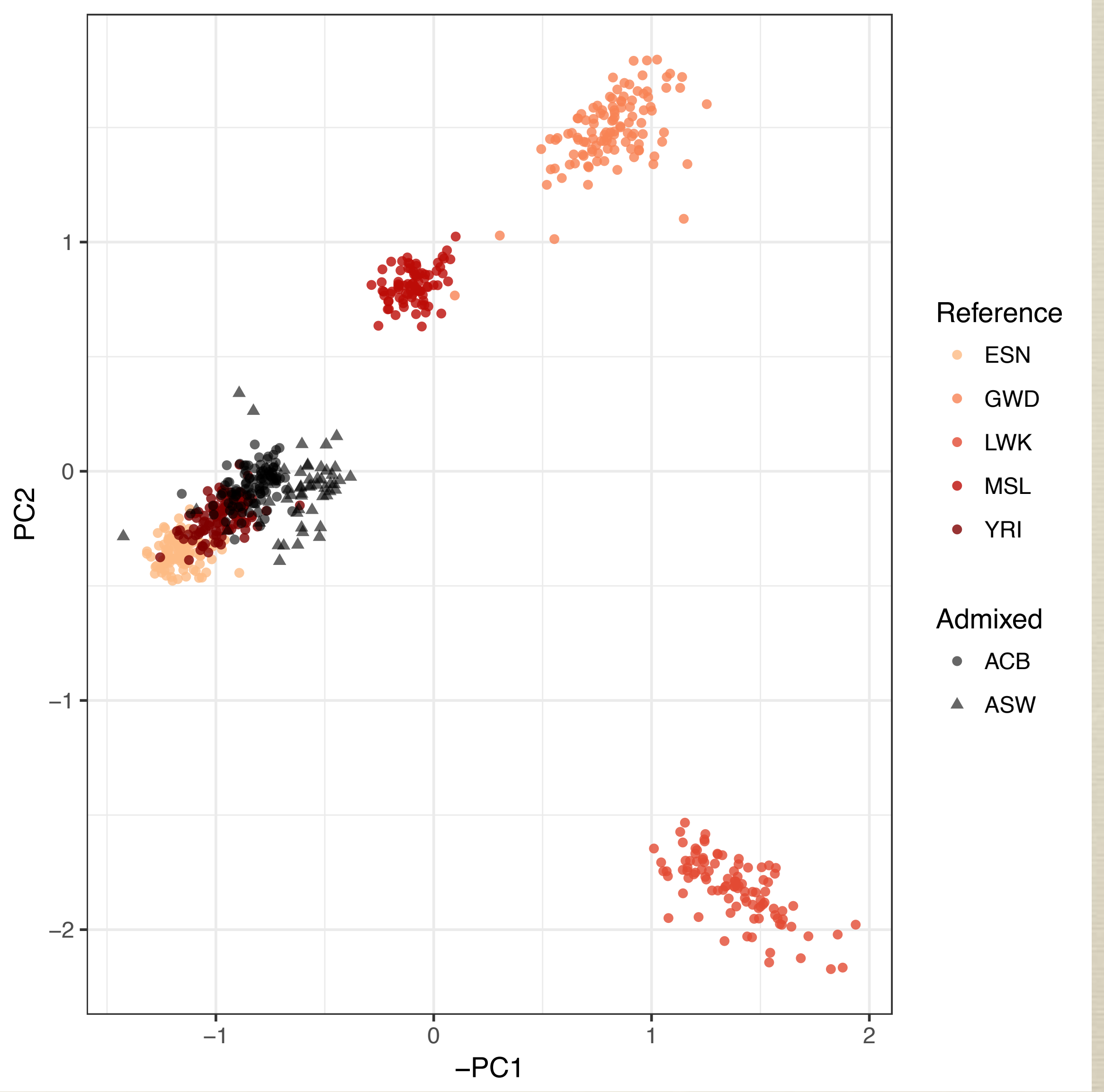
Plot AS-PCA


NA20332
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
AFR
EUR
NAT
UNK
0
50
100
150
200
250
300
Genetic position (cM)


2
1
0
−1
−PC2
−1
0
1
2
−PC1
GWD
ESN
MSL
YRI
LWK
Reference
● ESN
● GWD
● LWK
● MSL
● YRI
Admixed
● ACB
▲ ASW

Steps to plot ASPCA

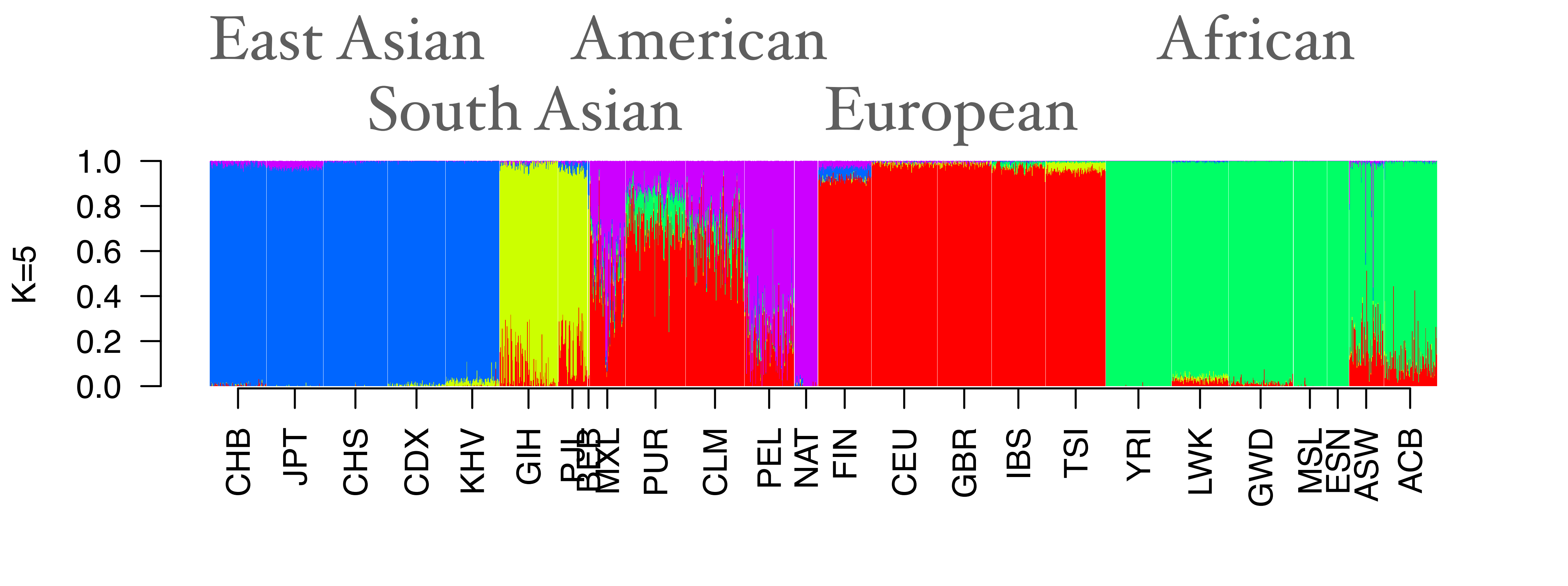
- Average PCs across haplotypes
- Add information about population, continents, etc
- Add plot color info



ADMIXTURE plots

- I ran ADMIXTURE for several predefined number of clusters (k) via: admixture [plink file] [k]

- Plot this for multiple values of k. What do you see?


East Asian
American
African
South Asian
European
K=5
CHB
JPT
CHS
CDX
KHV
GIH
PJL
BEB
MXL
PUR
CLM
PEL
NAT
FIN
CEU
GBR
IBS
TSI
YRI
LWK
GWD
MSL
ESN
ASW
ACB