DOGA ULUPINAR

ANCESTRY MAPPING

BIOLOGICAL BACKGROUND AND RELEVANCE

- Genetic make up, specifically SNPs within populations are more closely shared than across populations
- Control for population stratification in genetic association studies
- Understand how ethnic differences affects disease susceptibility
- Insight on which genes are more favorable in different populations

COMPUTATIONAL FORMULATION OF PROBLEM

- Input: Genotype data (n individuals by m SNPs)
- Output: Assign global ancestry to each individual
  - Ancestry = {African,Asian,European,American}
- Benchmark: Accuracy (F1 score) and Runtime

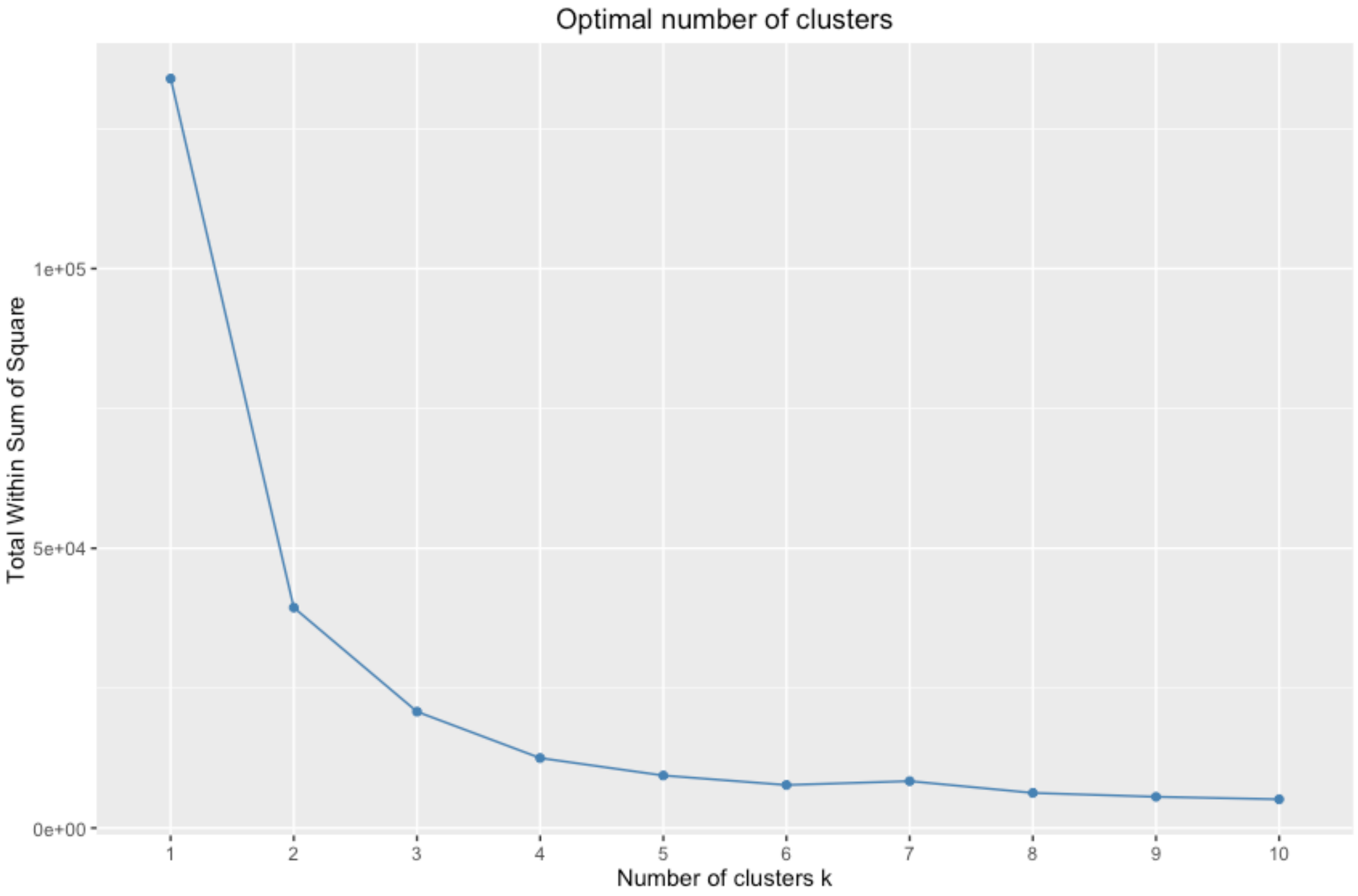
$$F1score = 2 * \frac{precision * recall}{precision + recall}$$

- Difficulties: High dimensionality of data (large amount of SNPs) and size of data

THE DATASET AND ASSUMPTIONS MADE

- Dataset: 1000 Genomes Project
  - 1092 Individuals
  - Chromosome 20, 21 and 22 ~ 1.5 million SNPs
- Assumptions
  - Number of Populations is known
  - Population of each individual is not known

CHECKING ASSUMPTIONS OF KNOWN 4 POPULATIONS



BASELINE METHOD – KMEANS ACROSS WHOLE CHROMOSOME

- Objective: Minimize distance between each point and the center of the cluster that this individual is assigned to

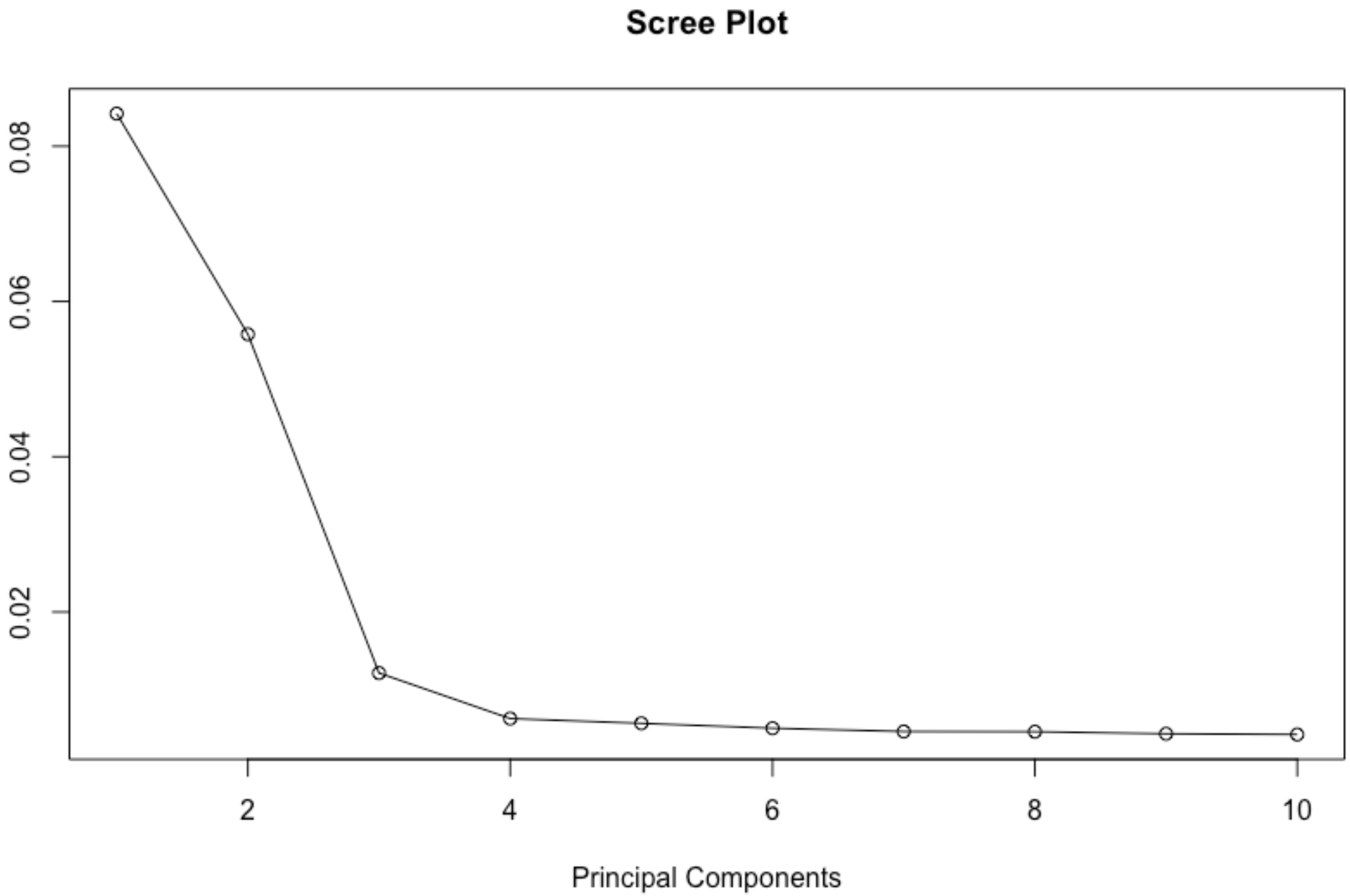
$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

- Starting points matter, multiple restarts with different start seeds
- Time Complexity:
  O(#iterations * #clusters * #individuals * #SNPs)
- Space Complexity: O((#individuals + #clusters) * # SNPs)

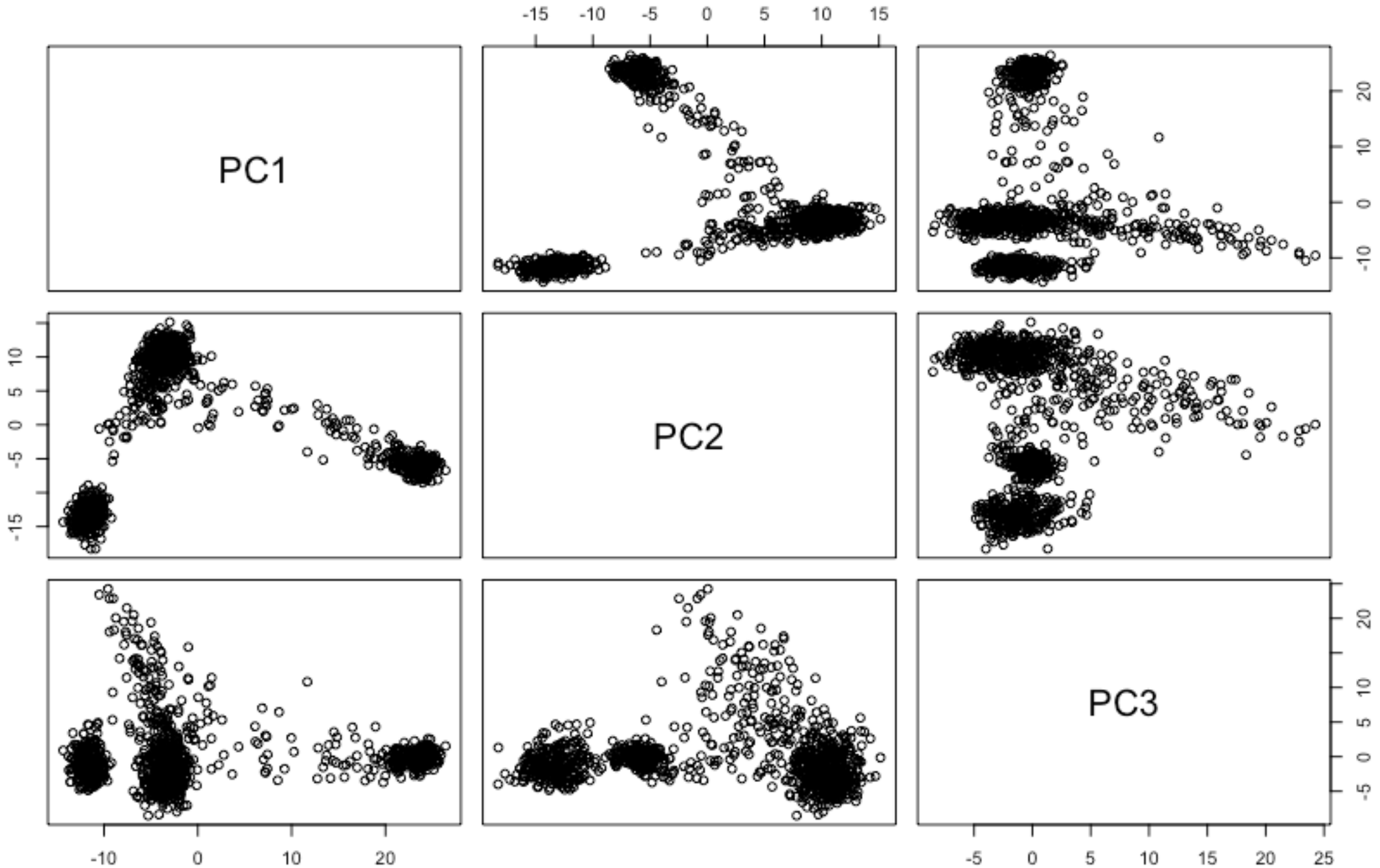
BETTER ANCESTRY MAPPING (B.A.M.)

- Reduce Dimensionality using Principal Component Analysis (PCA)
  - Returns orthogonal “dimensions” of highest variance, principal components through eigen decomposition of covariance matrix
  - Time Complexity: $O(\min(\#snps^3, \#individuals^3))$
- Benefits of PCA
  - Easier to visualize and interpret data
  - Reduce dimensionality of data to decrease runtime and memory usage

PCA APPLIED ON CHROMOSOME 22



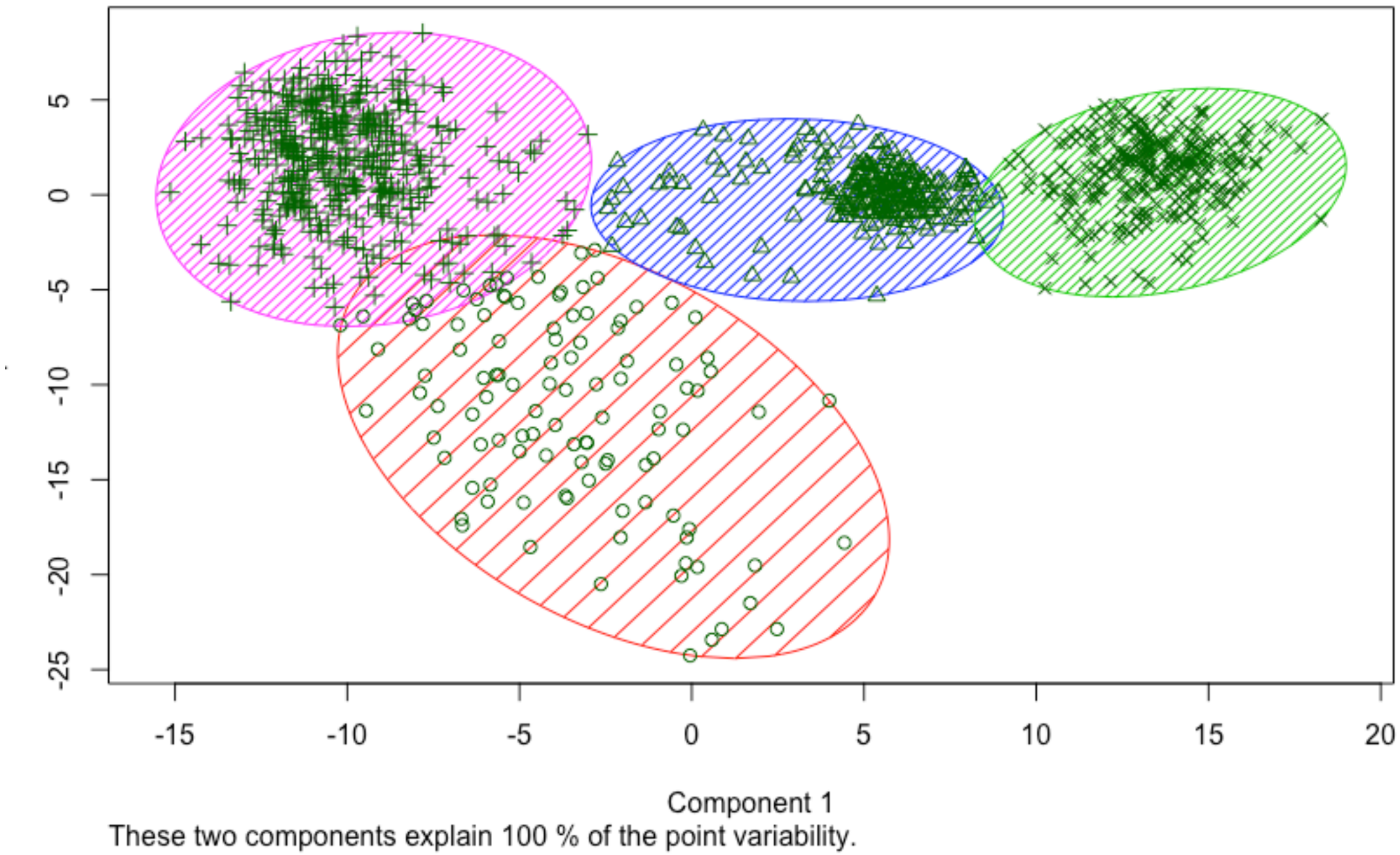
PCA APPLIED ON CHROMOSOME 22



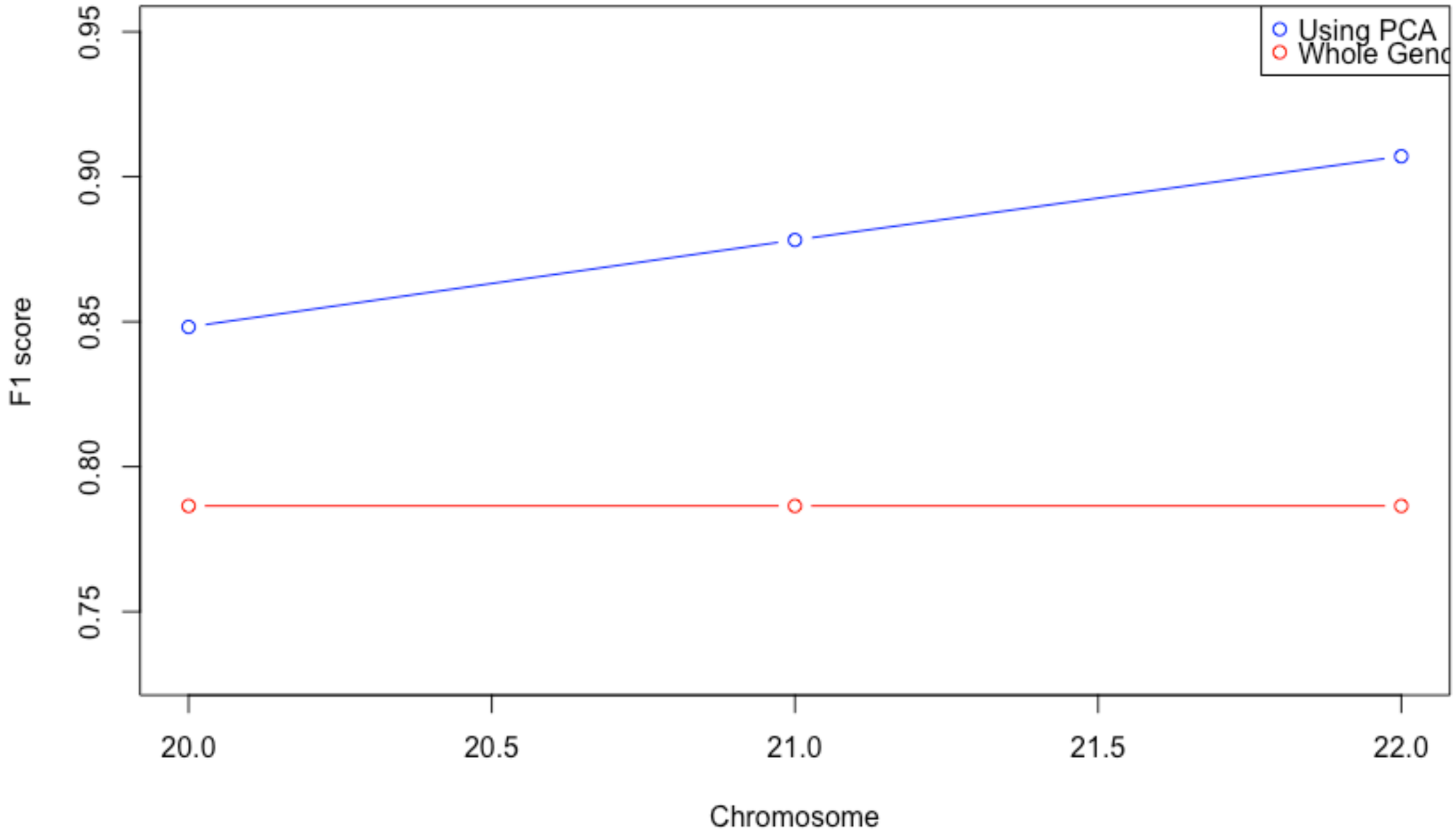
BIG PICTURE

- Compute PCA for each chromosome (can be distributed)
- Determine optimum number of principal component using scree plot, and plot of pairs
- Run Kmeans on aggregate principal components from each chromosome
- Train and predict using kfold cross validation

TWO DIMENSIONAL VISUALIZATION OF KMEANS

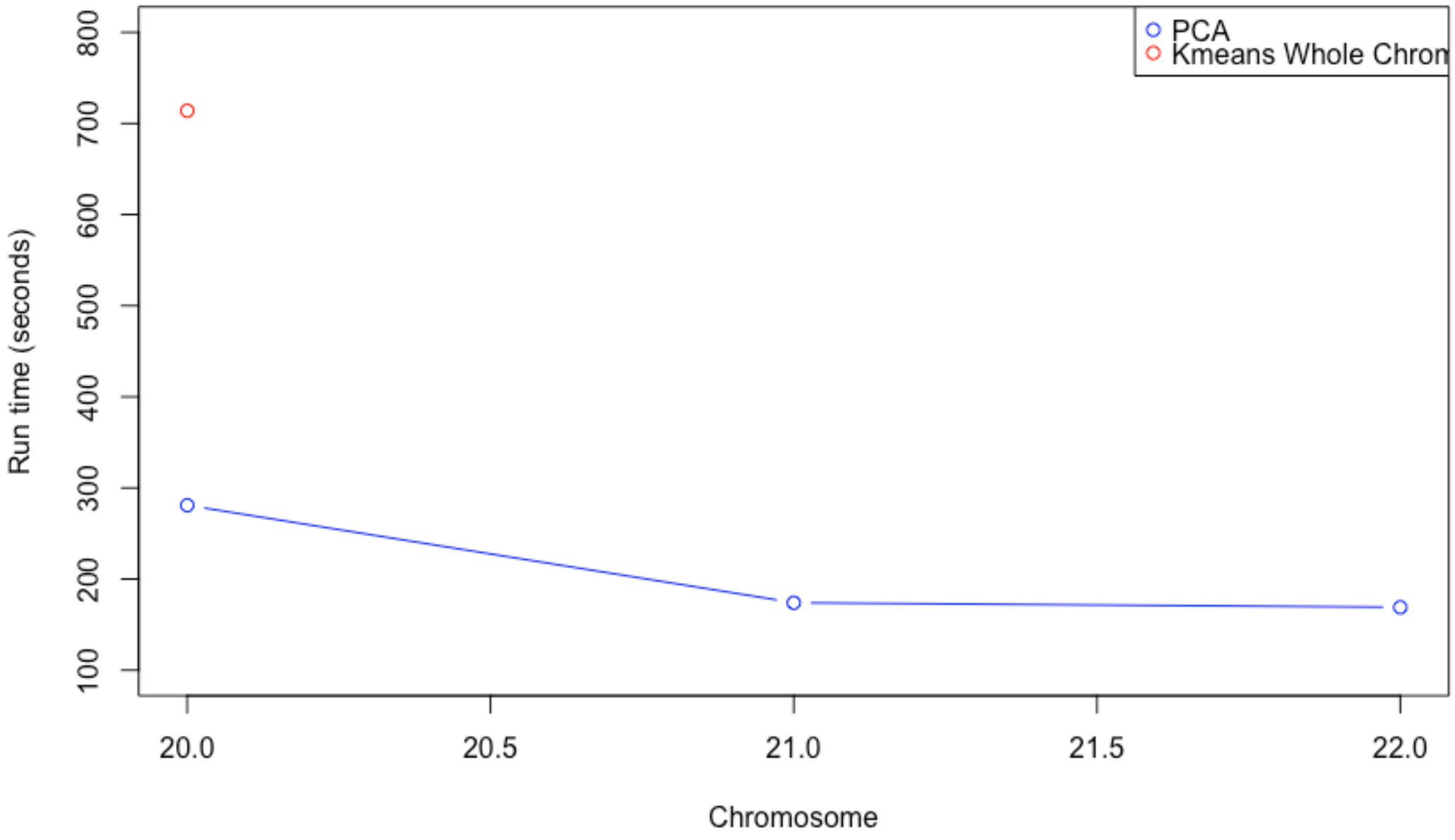


BENCHMARK – F1 SCORE



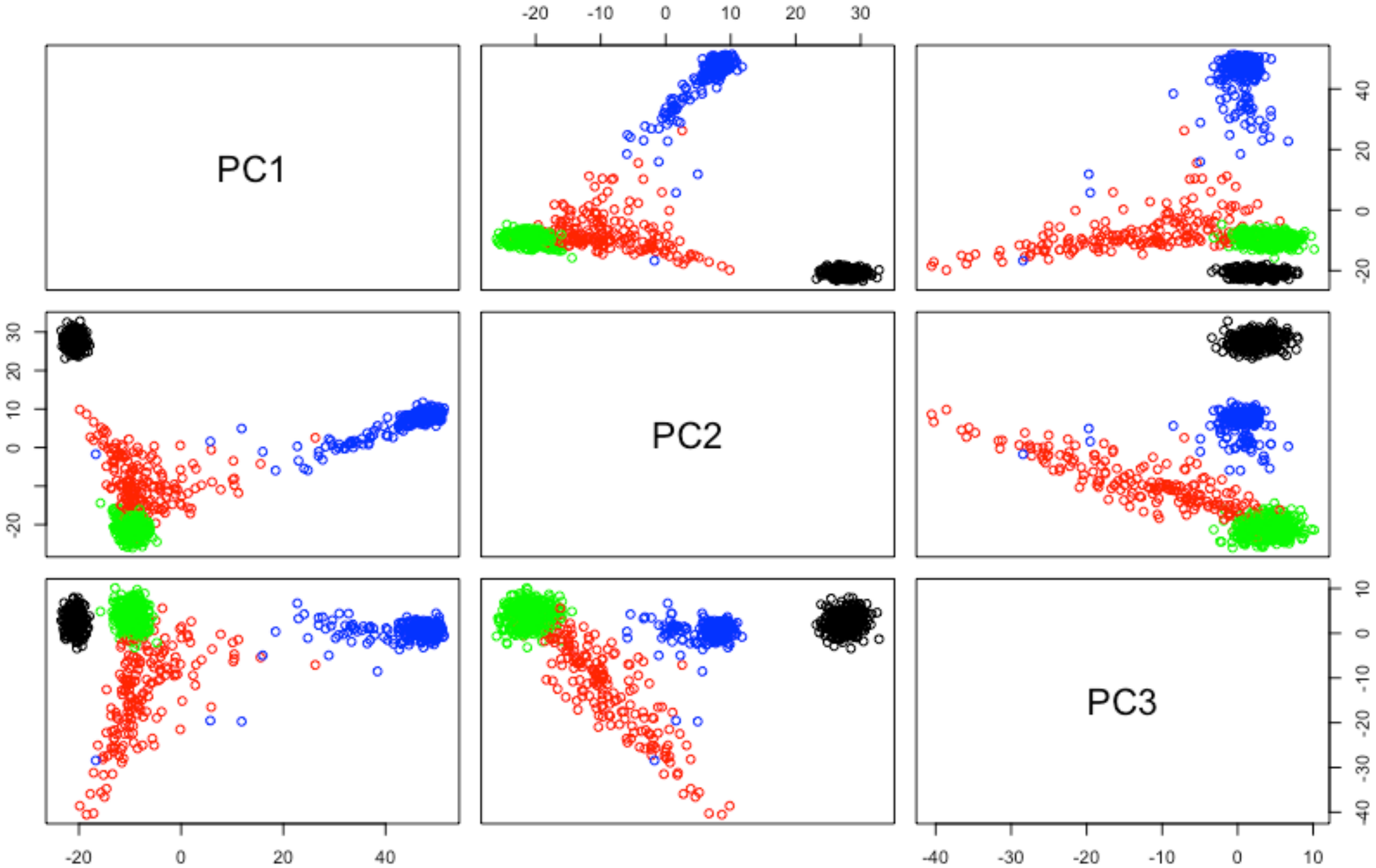
NOTE: Kmeans was only able to run on 1 chromosome

BENCHMARK – RUN TIME



NOTE: Kmeans was only able to run on 1 chromosome

PCAS PERFORMANCE CLASSIFYING DATA



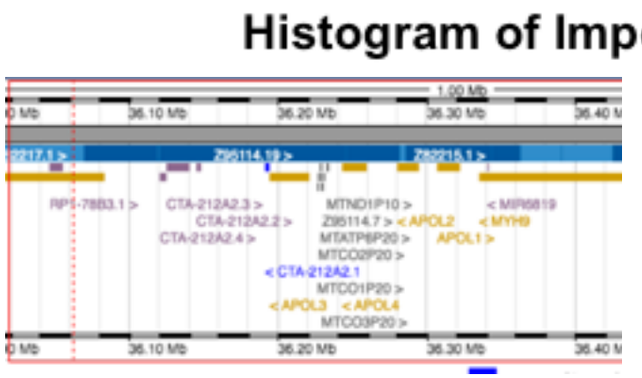
Blue: African

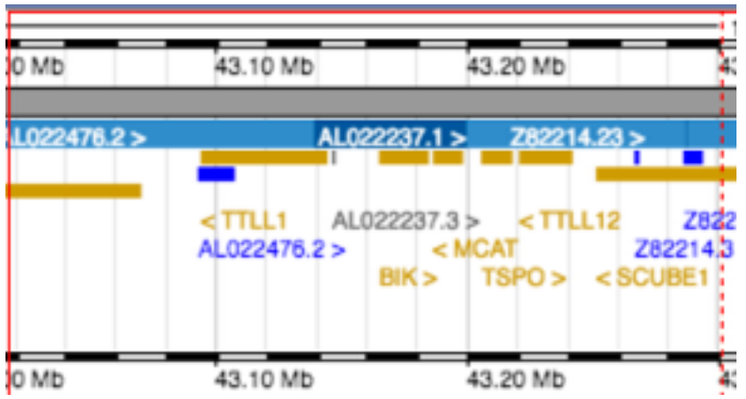
Green: European

Black: Asian

Red: American

SPARSE PCA


1.00 Mb
36.10 Mb
36.20 Mb
36.30 Mb
36.40 Mb
RP5-788J3.1 >
CTA-212A2.3 >
CTA-212A2.2 >
CTA-212A2.4 >
MTND1P10 >
Z95114.7 >
< APOL2
< MYH9
< MIR6819
MTATP6P20 >
APOL1 >
MTCO2P20 >
< CTA-212A2.1
MTCO1P20 >
< APOL3
< APOL4
MTCO3P20 >


43.10 Mb
43.20 Mb
AL022476.2 >
AL022237.1 >
Z82214.23 >
< TTLL1
AL022237.3 >
< TTLL12
AL022476.2 >
< MCAT
BIK >
TSPO >
< SCUBE1

Histogram of Important SNPs
Frequency
0
10
20
30
40
2.0e+07
2.5e+07
3.0e+07
3.5e+07
4.0e+07
4.5e+07
5.0e+07
Genetic Position

DISCUSSION AND FUTURE WORK

- Kmeans clustering on the entire chromosome seems to suffer from the high dimensionality of the data
- Reducing each chromosome to its principal components and aggregating based increases F1 score and can be done concurrently
- Europeans and Americans populations are genetically very similar, a more sophisticated probabilistic approach like admixture can provide for better results