**YangHuang: Deciphering Regional Variations Using SNPs**

Summary of the report:

This project aimed to use PCA on an SNPs data set to obtain information regarding where people come from. PCA, MDS, TNSE and an autoencoder was applied on the data set and visually compared. PCA-reduced data was selected for statistical analysis using machine learning to predict the region where people come from.

Describe the strengths of the report:

The produced figures are informative and aesthetic. The obtained results were extended to a case study which relates the work to the bigger picture.

Describe the weaknesses of the report:

Some minor grammatical errors and technical clarifications needed.

Evaluation on clarity and quality of writing: 4

Introduction: …we applied PCA to investigate how SNPs data “could” inform the region that people come from.

Results- Part 1: …while classical PCA “has/observed a” more intuitive shape

Results- Part 3: …5-fold “cross validation” accuracy to investigate the lower “bound” of…

Discussion: the number of SNPs can be reduced to 512 and still “keep/maintain” the prediction accuracy…

Evaluation on technical quality: 4

The report mentions “all 4 methods can separate regions by SNPs data”, would performing the same statistical analysis with machine learning achieve the same results?

The report mentions “classical PCA got more intuitive shape”, it would be beneficial to elaborate on what is justified as intuitive, especially since clear clusters are observed for all methods.

Unclear what is meant by “random selection”. I’m assuming this is referring to random selection of features. If so, what is the reason for including this?

Overall rating: 4

Confidence on your assessment: 3