Biostat 626 Midterm 2: Exploring Population Structures using Genetic Markers

Due: 11:59 PM, April 26, 2023

Background and Introduction

The primary goal of this project is to apply unsupervised learning algorithms to detect the population structure using the DNA information from samples collected worldwide. Each sample is represented by 2,540 genetic markers (more precisely, single nucleotide polymorphisms or SNPs) from the human genetic diversity panel (HGDP).

A SNP is a common type of genetic mutation that can have two different nucleotides (also known as allels) at a given chromosomal position. A genotype of a person at a given SNP position consists two alleles, one inherited from the mother and the other from the father. For example, a SNP with alleles A and G, can have 3 possible genotypes AA, AG, and GG. A SNP tends to have different allele frequencies (i.e., the frequency of the same reference allele) in different populations. Although each SNP may provide very limited population structure information, the combination of many informative SNPs can potentially provide a strong pattern of population clustering.

Data file format

Download the data file `hgdp.txt` from Canvas. The file is in plain text format. It contains genotype and sampling location information of 927 individuals. Each row represents a single individual with the following simple format

```
Individual_ID location continent geno_SNP1 ... geno_SNP2540
```

The genotypes are coded as the combination of two nucleotides. You may find it convenient to transform the letter coding of the genotype into the dosage coding (i.e., 0,1 and 2) for the required analysis. To do so, taking a SNP with A/G alleles as an example, you may (arbitrarily) specify the A allele as the reference allele and count the number of the reference allele in each genotype. Thus, AA is coded as 2, AG is coded as 1, and GG is coded as 0. If you specify the G allele as the reference allele, the dosage coding will be different, but it will NOT affect the results of the requested analyses.

Questions

Answer following questions by analyzing the HGDP data.

1. Use PCA followed by clustering algorithms to explore the population structure with only the genotype data (i.e., ignore sampling location and continent information).

2. Visualize the cluster structures identified from PCA and clustering analysis, color each sample point using its continental information. (you may wish to plot multiple pairs of PC scores)

3. Comment on the cluster structures identified from the analysis.

4. Find necessary resources to study the emerging technique known as "t-distributed stochastic neighbor embedding", or, t-SNE, apply it to the data set. Compare the t-SNE results to the PCA results.