CSE 291 - PERSONAL GENOMICS FOR BIOINFORMATICIANS

Problem Set 2 - Ancestry

This homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS2:LASTNAME] by the beginning of class on **Thursday, January 26**. The assignment is worth 10 points total.

Objectives

- Determine ancestry in homogeneous and admixed genomes.
- Predict relationships between related samples.
- Impute missing variants using imputation from a reference panel.

A description of data files for this problem set and several setup steps can be found at PS2 resources.

Part 1: Analyzing ancestry using principal components analysis (4 points)

Overview

A couple of your friends have just received their genomes and would like to learn about their ancestry. Their sample IDs are NA10847, NA18923, and NA19700. With the 1000 Genomes Project as a reference panel, we will use principal components analysis (PCA) to determine the ancestry of these three samples.

The first step of our PCA analysis is to construct a genotype matrix, where each row is a genomic location and each column gives the genotypes for a single sample. To save time, we'll only analyze polymorphic SNPs on chromosome 16 for this exercise. The original VCF file has already been preprocessed for you.

In the ps2 data directory, you'll find:

- The genotype matrix file

  `ps2_pca.genotypes.tab`

- A list of sample IDs, one for each column of the genotype matrix.

  `ps2_pca.samples.txt`

- A file listing the population label for each sample. You'll notice the sample IDs from our friends are missing. To learn what each population label corresponds to, see the 1000 Genomes website

  `ps2_reference_labels.csv`

In the templates directory, you'll find:

- A template python script for performing this analysis.

  `pset2_pca_template.py`

  Copy this to your working directory to edit there at:

/oasis/projects/nsf/csd524/$USER/ps2/code/pset2_pca.py

- An example script for submitting the job to the cluster using SBATCH:

  run_ps2_pca.sh

Exercises

1. **(2 points)** Implement PCA on the genotype matrix. You should write a script that takes as input the genotype matrix, sample names, and sample labels, and outputs a scatter plot of the first two principal components of the data, with each sample colored according to its population label. The template script has already been set up to take these inputs and outputs and make the plot, you just need to edit the function `perform_pca`. However you are free to write your own script if you prefer. You should submit the path to your PCA script and to the resulting figure.
2. **(0.5 points)** You should see three major clusters. What populations to those correspond to? Which groups does the first principal component separate? The second?
3. **(0.5 points)** What population labels should be assigned to NA10847, NA18923, and NA19700?
4. **(1 point)** The run time of PCA grows by some function of the number of samples (and the number of SNPs, which we assume here is constant). Try running your PCA on sample sizes of 100, 500, 1000, 2000, 2500 individuals. Plot the number of samples vs. the run time and describe the relationship. You'll notice the template script has a parameter `--num-samples`. You'll also find the UNIX command `time` useful.
5. **For fun**: Try including your own genome in the PCA analysis. Do you fall where you expect? You can try plotting lower principal components (e.g. PC2 vs. PC3) to get finer resolution about your ancestry.

Part 2: Relative finding (2 points)

Overview

A major reason many people get into personal genomics is to extend knowledge of their family's genealogy. By looking for segments of the genome that are shared between individuals, it is possible to identify close family members (e.g. parent-child, siblings, grandparents) all the way out to distant cousins. 23andMe's database contains more than one million individuals. If you are in a population well represented in 23andMe, the chances are quite high that you will have at least distant cousins in their database.

In this problem, we will explore using identity-by-descent (IBD) to identify relatives in a population. We will focus on 1000 Genomes individuals from the LWK population (the Luhya population from Kenya). These have been preprocessed into the plink file format and can be found in the ps2 data directory at:

ps2_ibd.lwk.ped
ps2_ibd.lwk.map
ps2_ibd.lwk.fam
ps2_ibd.lwk.bed
ps2_ibd.lwk.bim

The first 3 are text files that you can view on the command line, e.g. using:

```
less -S ps2_ibd.lwk.ped
less -S ps2_ibd.lwk.map
less -S ps2_ibd.lwk.fam
```

and the last two are binary formats of those files. To read more about plink file formats, see: http://pngu.mgh.harvard.edu/purcell/plink/data.shtml.

This dataset includes autosomal SNPs for 97 individuals. We would like to perform relative matching in this dataset to identify pairs of related samples.

Exercises

1. **(0.25 points)** What percent of the genome should be shared at IBD=0, 1, or 2 for a parent-child pair? For siblings? For first cousins?
2. **(0.25 points)** Use the plink program to calculate shared IBD for each pair of individuals. An example command is given in the templates directory at:

   ```
   run_ps2_ibd.sh
   ```

   This is not very computationally intensive, and can probably be run without submitting using SBATCH.
3. **(1 point)** Plot percent IBD=0 vs. percent IBD=1 for each pair of samples. Which pairs of samples appear to be a parent and child? Siblings? More distant relatives?
4. **(0.5 points)** You'll notice that all samples have some estimated degree of IBD=2, even those that are unrelated. Speculate why this happens.

Part 3: Imputing missing variants (4 points)

Overview

Most often, a genotype file contains incomplete information about an individual's genome. For instance, 23andMe genotypes 1.5 million SNPs, but there are 3 billion bases in the human genome. As we discussed in class, positions in the genome that are in close proximity are often co-inherited, inducing a correlation between nearby positions. If we can learn from a reference population what those correlations are, we can use a process called *imputation* to infer the missing genotypes.

In this problem, we'll perform imputation on the genome of an admixed individual, evaluate imputation performance under different conditions, and use the imputation results to analyze SNPs not included in the original genotype file.

As in previous problems, the original VCF has been preprocessed for you. In the ps2 data directory, you'll find:

```
ps2_impute.subset.gen.gz
ps2_impute.heldout.vcf.gz
```

Consisting of genotypes for sample NA20340. We will perform imputation on the first file (subset) and evaluate imputation on the true genotypes (heldout) that were removed from the imputation analysis.

Additionally, you'll find IMPUTE2 reference haplotypes in the data directory at:

`imputation/1000GP_Phase3/*`

including reference haplotypes, genetic recombination maps, and additional information about each position. As in pervious problems, we'll focus on chromosome 16 to reduce computation time. You're welcome to perform the analysis on the entire genome if you want.

Exercises

1. **(0.5 point)** Use IMPUTE2 to impute variants on chromosome 16 from position 5e6 to 10e6. An example command is given in the templates directory at:

   `run_ps2_impute.sh`

   Perform imputation using 4 different reference panels: CEU only (European), YRI only (African), CEU+YRI, and all populations except ASW (African American). The example script shows how to access those different reference panels. Which reference panel do you expect to perform the best on an African American genome? Why is it important to include similar population groups to the target genome in the reference panel?

2. **(0.5 point)** How many variants were you able to impute from the original 28,655?

3. **(1 point)** Compare the imputation results to the true results on the held out SNPs for each reference panel. Report the genotype r^2 for each. Do the results match your expectations from the previous problem? The template `run_ps2_impute.sh` contains some commands that may be helpful to combine the various data files needed for this problem.

4. **(2 points)** The "legend" files from the IMPUTE2 reference contain fields reporting the minor allele frequency of each SNP in multiple populations. Determine the genotype r^2 between the held out and imputed genotypes for different minor allele frequency thresholds (e.g. < 0.0001, < 0.001, < 0.01, < 0.05, < 0.1). How does imputation accuracy compare with minor allele frequency? Why would extremely rare variants be particularly difficult to impute?