Genetic Data Clean Up

Simple Project

pyGenClean

January 11th, 2016

Contents

List of Figures

List of Tables

1 Background

The aim of this project is to perform data QC prior to genetic analysis.

2 Methods

The automated script *pyGenClean* version 1.8.0 [1] was used to launch the analysis. *Plink* version v1.07 [2] was used for the data cleanup procedure. The following files were used as input.

- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.bed`
- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.bim`
- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.fam`

Briefly, the clean up procedure was as follow:

1. Checks sample's ethnicity using reference populations and IBS [`check_ethnicity`]. The script uses pairwise IBS matrix as a distance metric to identify cryptic relatedness among samples and sample outliers by multi-dimensional scaling (MDS).
2. Check sample's gender using Plink [`sex_check`]. The script identifies any individual with discrepancies between phenotype and genotype data for sex. Individuals with sex error are to be investigated.

3 Results

3.1 Ethnicity

Using 48,813 markers and a multiplier of 1, there was a total of 63 outliers of the CEU population. Figure 1 shows the first two principal components of the MDS analysis, where outliers of the CEU population are shown in grey.

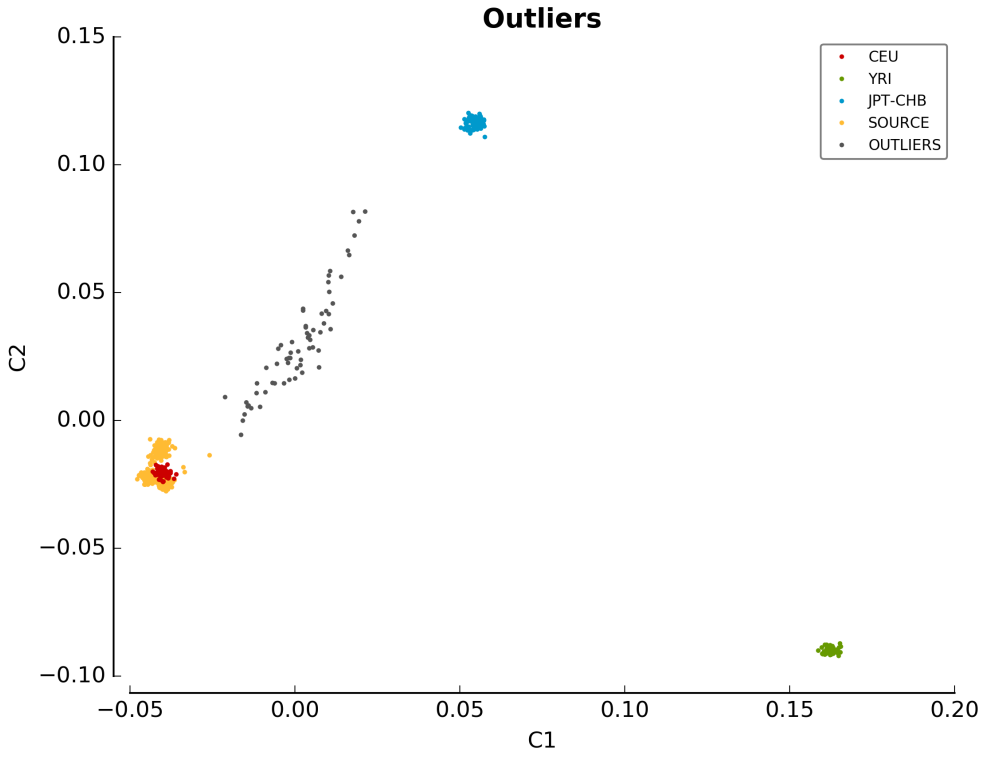

Figure 1: MDS plots showing the first two principal components of the source dataset with the reference panels. The outliers of the CEU population are shown in grey, while samples of the source dataset that resemble the CEU population are shown in orange. A multiplier of 1 was used to find the 63 outliers.

3.2 Gender check

Using F thresholds of 0.7 and 0.3 for males and females respectively, 4 samples had gender problem according to Plink. Table I summarizes the gender problems encountered during the analysis.

Table I: Summarization of the gender problems encountered during Plink's analysis. HET is the heterozygosity rate on the X chromosome. %NOCALL is the percentage of no calls on the Y chromosome.

FID	IID	PEDSEX	SNPSEX	STATUS	F	HET	%NOCALL
HG00361	HG00361	2	0	PROBLEM	0.5195	0.1492	N/A
NA20506	NA20506	2	1	PROBLEM	0.8644	0.0421	N/A
NA20530	NA20530	2	1	PROBLEM	0.8174	0.0567	N/A
NA20533	NA20533	2	0	PROBLEM	0.3422	0.2042	N/A

3.3 Summary results

Table II: Summary information of the data cleanup procedure.

	Markers	Samples
Initial number of markers	585,895	
Initial number of samples	360	
Number of markers used for MDS analysis	48,813	
Number of CEU outliers	63	
Number of samples with gender problem		
- no genetic gender	2	
- discordant gender	2	
Final number of markers	585,895	
Final number of samples	360	

4 Conclusions

After the genetic data clean up procedure, a total of 360 samples and 585,895 markers remained. The following files are available for downstream analysis.

- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.bed`
- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.bim`
- `pyGenClean_test_data/1000G_EUR-MXL_Human610-Quad-v1_H.fam`

For a list of markers and samples that were excluded during the data clean up procedure, refer to the files `excluded_markers.txt` and `excluded_samples.txt`, respectively.

References

[1] Lemieux Perreault LP, Provost S, Legault MA, Barhdadi A, Dubé MP: **pyGenClean: efficient tool for genetic data clean up before association testing.** *Bioinformatics* 2013, **29**(13): 1704–1705

[2] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *American Journal of Human Genetics* 2007, **81**(3): 559–575

[3] Goo J, Matthew F, Kurt NH, Jane MR, Kimberly FD, Gonçalo RA, Michael B, Hyun Min K: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.** *The American Journal of Human Genetics* 2012, **91**(5): 839–848