Processing 1000 Genomes reference data for ancestry estimation

Hannah Meyer

2024-08-22

Contents

Introduction

Genotype quality control for genetic association studies often includes the need for selecting samples of the same ethnic background. To identify individuals of divergent ancestry based on genotypes, the genotypes of the study population can be combined with genotypes of a reference dataset consisting of individuals from known ethnicities. Principal component analysis (PCA) on this combined genotype panel can then be used to detect population structure down to the level of the reference dataset.

The following vignette shows the processing steps required to use samples of the 1000 Genomes study [1],[2] as a reference dataset. Using the 1000 Genomes reference, population structure down to large-scale continental ancestry can be detected. A step-by-step instruction on how to conduct this ancestry analysis is described in this Ancestry estimation vignette.

Workflow

Set-up

We will first set up some bash variables and create directories needed; storing the names and directories of the reference will make it easy to use updated versions of the reference in the future. Is is also useful to keep the PLINK log-files for future reference. In order to keep the data directory tidy, we'll create a directory for the log files and move them to the log directory here after each analysis step.

```
refdir=~/reference
mkdir -p $refdir/plink_log
```

PLINK software

In addition to PLINK v1.9, which is a requirment for the `plinkQC` package, we will also need PLINK v2 for processing the downloaded the dataset. In the following, when `plink` is invoked, this corresponds to v1.9, whereas `plink2` corresponds to v2.

Download and decompress 1000 Genomes phase 3 data

1000 Genomes phase III (1000GenomesIII) is available in PLINK 2 binary format at https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3. In addition, a sample file with information about the individuals' ancestry is available and should be downloaded as in input for `plinkQC::check_ancestry()`. The following code chunk downloads and decompresses the data. The genome build of these files is the same as the original release of the 1000GenomesIII, namely CGRCh37.

NB: CGRCh38 positions in vcf format can be found here. The remainder of this vignette will however look at the data processing required for the 1000GenomesIII available in PLINK 2 binary format.

NB: the links to the files below are the three boldfaced links on this page: https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3. The dropbox links have been updated in the past, which means the links below where outdated. Please refer to the original site and open an issue on github if you notice a change. Thank you!

```
cd $refdir

pgen=https://www.dropbox.com/s/j72j6uciq5zuzii/all_hg38.pgen.zst?dl=1
pvar=https://www.dropbox.com/s/vx09262b4k1kszy/all_hg38.pvar.zst?dl=1
sample=https://www.dropbox.com/s/2e87z6nc4qexjjm/all_hg38.psam?dl=1

wget $pgen
mv 'all_hg38.pgen.zst?dl=1' all_hg38.pgen.zst
plink2 --zst-decompress all_hg38.pgen.zst > all_hg38.pgen

wget $pvar
mv 'all_hg38.pvar.zst?dl=1' all_hg38.pvar.zst

wget $sample
mv 'all_hg38.psam?dl=1' all_hg38.psam
```

Convert 1000 Genomes phase 3 data to plink 1 binary format

We then convert the PLINK 2 binary format to the (at the moment) more standardly used PLINK 1 binary format.

```
plink2 --pfile $refdir/all_hg38 vzs\
      --max-alleles 2 \
      --make-bed \
      --allow-extra-chr 0 \
      --out $refdir/all_hg38
mv $refdir/all_hg38.log $refdir/plink_log/all_hg38_convert.log
```

After these steps, the 1000 Genomes dataset can be used for inferring study ancestry as described in the corresponding Ancestry estimation vignette.

References

1. 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81. doi:10.1038/nature15394
2. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526: 75–81. doi:10.1038/nature15393