Processing HapMap III reference data for ancestry estimation

Hannah Meyer

2020-07-07

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 HapMap study [1][2][3] as a reference dataset. Using this reference, population structure down to large-scale continental ancestry can be detected. A step-by-step instruction on how to conduct this analysis is described in this 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 -r $qcdir/plink_log
```

Download and convert Hapmap phase III data

Hapmap phase 3 data (HapMapIII) is available in PLINK text format at ncbi. In addition, a sample file with information about the individuals' ancestry is available and shoud be downloaded as in input for `plinkQC::chec_Ancestry()`. The following code chunk downloads and unzips the data.

```
cd $refdir

ftp=ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2009-01_phaseIII/plink_format/
prefix=hapmap3_r2_b36_fwd.consensus.qc.poly

wget $ftp/$prefix.map.bz2
bunzip2 $prefix.map.bz2

wget $ftp/$prefix.ped.bz2
bunzip2 $prefix.per.bz2

wget $ftp/relationships_w_pops_121708.txt
```

We then convert the PLINK text format to the standardly used PLINK binary format.

```
plink --file $refdir/$prefix \
      --make-bed \
      --out $refdir/HapMapIII_NCBI36
mv $refdir/HapMapIII_NCBI36.log $refdir/log
```

Update annotation

The genome build of HapMap III data is NCBI36. Currently most datasets are updated to CGRCh37 or CGRCh38. In order to update the HapMap III data to the desired build, we use the UCSC liftOver tool. The liftOver tool takes information in a format similar to the PLINK .bim format, the UCSC bed format and a liftover chain, containing the mapping information between the old genome (target) and new genome (query). It returns the updated annotation (newFile) and a file with unmappable variants (unMapped):

```
liftOver oldFile liftover.chain newFile unMapped
```

We first need to download the liftOver tool from https://genome.ucsc.edu/cgi-bin/hgLiftOver and the appropriate liftover chain from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/). We then convert the PLINK .bim format, to the zero-based UCSC bed format.

Hapmap chromosome data is encoded numerically, with chrX represented by chr23, and chrY as chr24. In order to match to data encoded by chrX and chrY, we will have to rename these hapmap chromosomes. Converting to zero-based UCSC format and re-coding chromosome codes can be achieved by:

```
awk '{print "chr" $1, $4 -1, $4, $2 }' $refdir/HapMapIII_NCBI36.bim | \
    sed 's/chr23/chrX/' | sed 's/chr24/chrY/' > \
    $refdir/HapMapIII_NCBI36.tolift
```

[Note: In the official HapMap release, chromosome codes described above, however in the orignal download files (link above), no chr24 detected. I will keep this line in for completeness, but note, when inspecting file that no chr24/chrY are present.]

We use the liftOver tool and the UCSC bed formated annotation file together with the appropriate chain file to do the lift over.

```
liftOver $refdir/HapMapIII_NCBI36.tolift $refdir/hg18ToHg19.over.chain \
    $refdir/HapMapIII_CGRCh37 $refdir/HapMapIII_NCBI36.unMapped
```

After successful liftover, we will be able to extract i) the variants that were mappable from the old to the new genome and ii) their updated positions

```
# ectract mapped variants
awk '{print $4}' $refdir/HapMapIII_CGRCh37 > $refdir/HapMapIII_CGRCh37.snps
# ectract updated positions
awk '{print $4, $3}' $refdir/HapMapIII_CGRCh37 > $refdir/HapMapIII_CGRCh37.pos
```

Update the reference data

We can now use PLINK to extract the mappable variants from the old build and update their position. After these steps, the HapMap III dataset can be used for infering study ancestry as described in the corresponding vignette.

```
plink --bfile $refdir/HapMapIII_NCBI36 \
    --extract $refdir/HapMapIII_CGRCh37.snps \
    --update-map $refdir/HapMapIII_CGRCh37.pos \
    --make-bed \
    --out $refdir/HapMapIII_CGRCh37
mv $refdir/HapMapIII_CGRCh37.log $refdir/log
```

References

1. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437: 1299–320. doi:10.1038/nature04226

2. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851. doi:10.1038/nature06258

3. The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467. doi:10.1038/nature09298