Preprocessing of Human Origins (plus Pacific) data

Alejandro Ochoa and John D. Storey

2019-04-30

Introduction

This document has two goals:

- To explain how the full Human Origins and Pacific datasets (public and non-public) were processed and merged.
- To construct a version of this dataset consisting of public data only, for the community to independently validate our conclusions. The following files are the result, and are available as part of this repository (https://github.com/StoreyLab/human_differentiation_analysis):

```
data/human_origins_and_pacific_public.bed
data/human_origins_and_pacific_public.bim
data/human_origins_and_pacific_public.fam
```

Software requirements

These instructions assume a standard unix-like terminal, such as those of most modern Linux and MacOS.

In addition, you will need the following binaries (click for links):

- `plink2`, version 2.00 alpha.
- `plink`, version 1.9. This is for **merging** two BED files, a step currently unsupported in `plink2`.
- `eigensoft`, version 7.2.1. Needs compilation. Only `convertf` is used.

These binaries will be assumed to be in your terminal's `PATH`.

There are several R scripts provided in the `scripts/` directory of this repository. These script require the packages `readr`, `tibble` (both available on CRAN) and `genio`. The latter is available on GitHub and can be installed by running these commands in an R session:

```
install.packages("devtools") # if needed
library(devtools)
install_github("OchoaLab/genio", build_opts = c())
```

Download public data

First let's switch to a `data` directory where all of these file are going. On a terminal, type:

```
# create directory if needed
mkdir data
cd data
```

The public data is available on the Reich Lab Datasets website. Here are terminal commands to download the two specific files of interest:

```
# URL base (shared by both links)
BASE='https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files'
# Genotypes of ancient individuals and fully public Affymetrix Human Origins
# present-day individuals analyzed in ...

# ... Lazaridis et al. Nature 2016
wget $BASE/NearEastPublic.tar.gz

# ... Skoglund et al. Nature 2016
wget $BASE/SkoglundEtAl2016_Pacific_FullyPublic\ \(3\).tar.gz
```

Extract the files from these archives:

```
tar -xzf NearEastPublic.tar.gz
tar -xzf SkoglundEtAl2016_Pacific_FullyPublic\ \(3\).tar.gz
```

Obtaining full (non-public) data

The non-public data is not available online. If you are interested you must draft a letter to David Reich at Harvard agreeing to certain usage restrictions. Instruction can found in the `README` file that is part of the `NearEastPublic.tar.gz` archive we downloaded above. Be sure to request the public data for both the Lazaridis et al. **and** Skoglund et al. publications, as they are provided separately.

The following instructions, applied only to the public data in this document, can be easily applied to the full datasets just by replacing the right file names.

Convert data to plink BED format

The data we just downloaded is in Eigensoft's own format. However, our pre-processing and analysis requires the plink BED format. The program `convertf` that is part of Eigensoft is used here to convert files. Since the command is rather verbose, we wrote a wrapper function that we can more easily apply to several datasets. Let's load the function we need on the terminal:

```
. ../scripts/geno_to_bed.bash
```

The function assumes that `convertf` is on the `PATH`. Let's apply it to our two datasets:

```
geno_to_bed HumanOriginsPublic2068
geno_to_bed SkoglundEtAl2016_Pacific_FullyPublic
```

Although the original individual annotation tables (`*.ind` files) contain subpopulation labels, these are not preserved in the `plink` tables (`*.fam` files). This `R` script adds the subpopulation labels to the `*.fam` files using the first column (technically "family IDs"), which did not have any useful information before.

```
Rscript ../scripts/ind_to_fam.R HumanOriginsPublic2068
Rscript ../scripts/ind_to_fam.R SkoglundEtAl2016_Pacific_FullyPublic
```

The public Pacific data now has subpopulation lables in the first column, but unfortunately these are also present in the second (Sample ID) column, where they are concatenated with the actual sample IDs. The following script cleans up the second column of this file in this specific format:

```
Rscript ../scripts/fam_clean_pacific.R SkoglundEtAl2016_Pacific_FullyPublic
```

We note that the full *Pacific* data that we received was already in plink BED format, and the individual annotations table (FAM file) was already in the correct format, so no modifications were needed.

Merging main Human Origins and Pacific datasets

This command merges our two datasets into a single set of plink BED files. This is the only step that requires `plink` version 1.9:

```
plink \
    --keep-allele-order \
    --indiv-sort none \
    --bfile HumanOriginsPublic2068 \
    --bmerge SkoglundEtAl2016_Pacific_FullyPublic \
    --out human_origins_and_pacific_public_pre

# cleanup
rm human_origins_and_pacific_public_pre.log
```

This is not our final file yet (hence the `_pre` suffix). We shall filter loci and individuals in the following steps.

List of loci in intersection of datasets

These datasets have non-overlapping individuals that were genotyped using the same microarray platform. However, the Pacific dataset has more stringent quality controls, so fewer loci appear in that data. We will keep only the intersection of loci.

First let's look at the data dimensions, which we obtain on the terminal just by counting lines on the annotation tables. The number of lines in the BIM file is the number of loci, while the number of lines in the FAM file is the number of individuals:

```
# public data
wc -l HumanOriginsPublic2068.{bim,fam}
# 621799 HumanOriginsPublic2068.bim
#   2068 HumanOriginsPublic2068.fam
wc -l SkoglundEtAl2016_Pacific_FullyPublic.{bim,fam}
# 597573 SkoglundEtAl2016_Pacific_FullyPublic.bim
#     74 SkoglundEtAl2016_Pacific_FullyPublic.fam
wc -l human_origins_and_pacific_public_pre.{bim,fam}
# 621799 human_origins_and_pacific_public_pre.bim
#   2142 human_origins_and_pacific_public_pre.fam
```

What `plink` did is keep all loci, inserting missing values for all Pacific genotypes at their missing loci.

This `plink2` command returns the list of loci present in the Pacific data:

```
plink2 \
    --bfile SkoglundEtAl2016_Pacific_FullyPublic \
    --write-snplist \
    --out loci_pacific

# cleanup
rm loci_pacific.log

# this confirms that all IDs are unique in this data!
```

```
wc -l loci_pacific.snplist
# 597573
uniq loci_pacific.snplist | wc -l
# 597573
```

Filters for individuals

We remove a handful of individuals to simplify our figures. The main filter is to remove individuals from singleton subpopulations (those with only one individual). This script identifies those samples and stores them in the file `rm_fam.txt`:

```
Rscript ../scripts/singleton_fams.R human_origins_and_pacific_public_pre rm_fam.txt
```

Opening `rm_fam.txt` shows these contents so far:

```
Ignore_Adygei(relative_of_HGDP01382)
Lapita_Tonga
Saami_WGA
```

The data so far also has some ancient individuals from the subpopulation `Lapita_Vanuatu`, which we remove by appending that code to the list of families to remove. On the terminal:

```
echo Lapita_Vanuatu >> rm-fam.txt
```

Lastly, we also remove the `AA` subpopulation (African Americans), as they are not a native subpopulation.

```
echo AA >> rm-fam.txt
```

Filter into final dataset

This command removes individuals from the above subpopulations, keeps only loci present in the Pacific dataset, and additionally removes fixed loci (`--mac 1`) and loci outside of autosomes:

```
plink2 \
    --bfile human_origins_and_pacific_public_pre \
    --extract loci_pacific.snplist \
    --remove-fam rm_fam.txt \
    --autosome \
    --mac 1 \
    --make-bed \
    --out human_origins_and_pacific_public

# cleanup, including intermediate data
rm human_origins_and_pacific_public.log
rm rm_fam.txt
rm loci_pacific.snplist
rm human_origins_and_pacific_public_pre.{bed,bim,fam}
```

This command simplifies subpopulation labels in the FAM file. It merges all Gujarati samples (by removing A,B,C,D suffixes):

```
perl -p -i -e 's/Gujarati[A-D]/Gujarati/' human_origins_and_pacific_public.fam
```

The final dataset has these dimensions:

```
wc -l human_origins_and_pacific_public.{bim,fam}
# 582835 human_origins_and_pacific_public.bim
#   2139 human_origins_and_pacific_public.fam
```