AncestryMapper

Tiago R Magalhães, Darren J. Fitzpatrick, Eoghan O'Halloran

February 3, 2015

Abstract

AncestryMapper is an R package that implements the methods described in Magalhães TR, Casey JP, Conroy J, Regan R, Fitzpatrick DJ, et al. (2012) HGDP and HapMap Analysis by Ancestry Mapper Reveals Local and Global Population Relationships. PLoS ONE 7(11): e49438.

Contents

1 Introduction

Knowledge of human origins, migrations and expansions is greatly enhanced by the availability of large datasets of genetic information from different populations and by the development of bioinformatic tools used to analyze the data.

Ancestry Mapper assigns genetic ancestry to an individual and studies relationships between local and global populations. The principle function of the method gives each individual an Ancestry Mapper Id (AMid), a genetic identifier comprising 51 genetic coordinates that correspond to its relationship to the Human Genome Diversity Project (HGDP). The AMid metrics have intrinsic biological meaning and provide a tool to measure genetic similarity between world populations.

2 Package Functions

The package consists of two functions:

- `calculateAMids`: calculates and assigns Ancestry Mapper Ids (AMids) to each individual
- `plotAMids`: produces a heatmap representation of AMids

2.1 calculateAMids

For each individual, calculateAMids computes the genetic distances amongst that individual and the set of HGDP references. As input, the function requires a PED formatted file. PED formatting is the standard file format required by the PLINK software suite. For details on the format see `http://pngu.mgh.harvard.edu/~purcell/plink/`. It also requires a file containing the references with columns in the order of population, reference and order.

As output, `calculateAMids` returns a dataframe containing the genetic distance of each individual to the all HGDP references. We provide the raw distance measures (starting with the prefix C_) and indices (normalized Values, starting with the prefix I_).

The genetic distance is computed as the Euclidean distance normalized by the number of SNPs, between each individual and the 51-HGDP-based references. AMids for a single individual from any dataset can be computed provided there is a reasonable overlap between the set of SNPs for that individual and the HGDP references. The AMids can take values from 0 to 2. In our experience, the values are in the range 0.4 to 1.1.

The normalized values of the distances are such that the highest reference is scored as 100, the lowest as 0 and all others adjusted accordingly. These indices place the individual in the genomic map, forcing it to be committed to one reference, even if the absolute similarities, as indicated by the euclidean distances, are not very big. Thus, they provide a global overview on the number of relevant references for each individual.

The user can include new references in AMids by editing the file “HGDP_References.txt”, inserting the invidual’s name, population it corresponds to and the order it should appear in the AMid.

2.2 plotAMids

The function `plotAMids` is used to visualize the relationship amongst individuals and the 51 HGDP references. `plotAMids` takes as input the dataframe of genetic distances returned by `calculateAMids`. The user can also provide a file with phenotypes for each individual which will be visible in the plot. The colors for the plot are from the `BlBrewer` and `RedBl` packages but are hard coded so there are no dependencies.

3 Producing a PED file

The PED file should include individuals that will be taken as population references that will be used to calculate the ancestry mapper indexes (AMIds) for the user dataset. In our original work we used as references the 51 populations included in the Human Genome Diversity Project. The HGDP dataset can be obtained at `http://hagsc.org/hgdp/files.html`.

In order to merge the HGDP references with a PED formatted file use the `--merge` command in PLINK; both files should be in the ACGT format. In most cases there will be strand inconsistencies. To rectify this use the `--flip` command. SNPs that are CG AT are invisible to the strand issue and should be removed. Identify them in the output of the `--freq` command, and remove them with `--exclude` command. Ancestry mapper requires the 1/2 coding system. To convert from ACGT coding, use the `--recode12` command.

The individual Ids are taken as the second column of the ped file; these ids should be unique.

We have produced a bed file with the references for the 51 HGDP populations, with 630,597 snps; the file is named HGDP_51RefAM_AutosomalSnps_630597_ACGT and can be obtained at http://bit.ly/1vnZzCT.

The python 3 script for Linux 'py_merge_HGDPAncestryMapperRefs_AncMap.py', also present in the folder, merges the HGDP reference bed file with any user bed PLINK file. The user should edit the script providing the path to both files, the working folder, the name of the merged file, and whether sex chromosomes should be removed.

4 Future Releases

In future releases, it is anticipated two additional functions will be added to the package, 1) a clustering function in order to group individuals and reference samples, and 2) a function to add custom references to the HGDP reference panel. We are also currently working to expand the population references to close to 200 world-wide populations.

5 Tutorial

```
library(AncestryMapper)
```

The first step is to call the example data files distributed with the package. These files are:

- `HGDP.References`: the 51 HGDP population references.
- `HGDP.500SNPs`: 500 randomly selected SNPs from the HGDP data set.
- `HGDP.Phenotypes`: A file detailing phenotypes of the individuals.

```
 HGDP.References <- system.file('extdata','HGDP.References.txt',package='AncestryMapper')

HGDP.500SNPs <- system.file('extdata', 'HGDP.500SNPs.ped', package='AncestryMapper')

HGDP.Phenotypes <- system.file('extdata', 'HGDP.Phenotypes.txt', package='AncestryMapper')
```

The second step calculates the genetic distances using `calculateAMids`. This returns a dataframe detailing the genetic distance of each individual to the 51 HGDP references.

```
  genetic.distance <- calculateAMids(pedtxtFile=HGDP.500SNPs,
fileReferences=HGDP.References)
```

The plot function produces a genomic map detailing the genetic distance of individuals and the HGDP references. The plot function can incorporate phenotypes as an option. their are multiple options for the plot function (see `?plotAMids`)

```
 plotAMids(AMids=genetic.distance, phenoFile=HGDP.Phenotypes)

 plotAMids(AMids=genetic.distance, phenoFile='')
```

The plot is directed to the R plotting device but can be saved in a variety of formats, e.g., pdf, png, and tiff.