Ancestry Mapper 2.0

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

September 24, 2016

Abstract

AncestryMapper is an R package that implements the methods described in 'HGDP and HapMap Analysis by Ancestry Mapper Reveals Local and Global Population Relationships.' Magalhães TR, Casey JP, Conroy J, Regan R, Fitzpatrick DJ, et al. PLoS ONE 7(11): e49438. (2012) Ancestry Mapper assigns genetic ancestry to an individual and studies relationships between local and global populations. The method gives each individual an Ancestry Mapper Id (AMid), a genetic identifier comprising genetic coordinates that correspond to its relationship to various reference populations. The AMid metrics have intrinsic biological meaning and provide a tool to measure genetic similarity between world populations.

Contents

1 Package Functions

- **calculateAMidsArith:** calculates and assigns Ancestry Mapper Ids (AMids) to each individual using a new, more precise, arithmetic method

- **calculateAMids:** calculates and assigns Ancestry Mapper Ids (AMids) identical to the older versions of AncestryMapper
- **plotAMids:** produces a heatmap representation of AMids
- **createMedoid:** constructs reference from sample inputs
- **refAdd:** adds user-supplied references or data to a reference file, containing information for order, labelling and color

1.1 calculateAMidsArith

For each individual, ***calculateAMidsArith*** computes the genetic distances amongst that individual and the set of selected references. As input, the function requires a tPED formatted file, a standard file format required by the PLINK software suite.
For details on the format see: http://pngu.mgh.harvard.edu/ purcell/plink/-data.shtml#tped.

It also requires the file 'CorPheno', which contains columns detailing metadata for samples and references such as the order ('Order') of the population as well as the colors to use. An example of which can be found with the package.

Table 1: First 3 Lines of 'CorPheno'

Pheno_Pop	Pheno_Data	Pheno_Region	Colors_Pop	Colors_Region	Colors_Data	Order	UNIQID	Fam
Example_Pop	Fedorova2013	EUR	999999	black	black	140	Example_ID	Example_Fam
Neanderthal.SP	SP	EUR	999999	black	black	140	AltaiNea	AltaiNea
Denisovan.SP	SP	EUR	999999	black	black	140	AltaiDen	DenisovaPinky

A file containing the record of the major and minor alleles as described by dbSNP is needed. An example file for use with the toy data can be found as 'MinMaxFreq.rda' included with the package.

Only SNPs with entries in the 'MinMaxFreq.rda' used will be used in the analysis. A larger MinMaxFreq file containing SNPs used on common Illumina and Affymetrix ChIPs in addition to others is available at: http://bit.ly/1OUstDP

Additionally the user can produce their own 'MinMaxFreq' reference file. This should be a data frame with 2 unnamed columns consisting of a column for each allele and row names corresponding to the rsID. This object must be named MinMaxFreq' and saved as an rda file.

Table 2: First 3 rows of 'MinMaxFreq'

row.names	V1	V2
rs6663840	A	G
rs548726	C	T
rs10803320	C	T

calculateAMidsArith returns a dataframe containing the genetic distance of each individual to the references used. This provides 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 all the references used. 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 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, thus, they provide a global overview on the number of relevant references for each individual.

1.2 calculateAMids

For each individual, ***calculateAMids*** computes the genetic distances amongst that individual and the set of HGDP references (or a set provided by the user). 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/data.shtmlped. It also requires a file containing the ids of the individuals to be used as references, and the population they correspond to.

As output, 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_).

As with calculateAMidsArith, genetic distance is computed as the Euclidean distance normalized by the number of SNPs, between each individual and the references.

The user can include new references in AMids by editing the file 'HGDP_References.txt', inserting the population and the corresponding invidual's name.

A demo file using 500 SNPs is included using data from the HGDP called 'HGDP_500SNPs.ped', this is compatiable with the 'CorPheno' file also included in the package. The small number of SNPs and samples will result is highly noisy plots.

For more information on using ***calculateAMids*** see, ?***calculateAMids***()

1.2.1 Producing an Input File for calculateAMids

The PED file should include individuals that will be taken as the population references, which 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.

To merge a custom ped file with a ped file with the references, users could use PLINK. The commands --bmerge or --merge are used to merge two ped files. In most cases there will be strand inconsistencies, that can be rectified by flipping snps, using the command --flip. SNPs that are CG AT are impossible to determine which strand they are in and as such be removed. Ancestry mapper requires the ped files to be in the 1/2 coding system. 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/2bkWDSQ.

The PED file should include individuals that will be taken as the population references, which 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.

To merge a custom ped file with a ped file with the references, users could use PLINK. The commands --bmerge or --merge are used to merge two ped files. In most cases there will be strand inconsistencies, that can be rectified by flipping snps, using the command --flip. SNPs that are CG AT are impossible to determine which strand they are in and as such be removed. Ancestry mapper requires the ped files to be in the 1/2 coding system. 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/2bkWDSQ.

1.3 plotAMids

plotAMids is used to visualize the relationship amongst individuals and the references. ***plotAMids*** takes as input the dataframe of genetic distances returned by ***calculateAMidsArith***. The user can also provide a file with phenotypes for each individual which will be visible in the plot. Colors used taken from the **BlBrewer** and **RedBl** packages.

1.4 createMedoid

createMedoid constructs an arithmetic reference from a tPED containing only one population with at least 10 individuals. The tPED should contain only

individuals of a specific population and be formatted and prepared as described in the 'Producing a Sample Input File' section.

1.5 refAdd

refAdd adds user-supplied references or data to the CorPheno file, giving population, color and ordering information to cluster the results of samples by population.

2 Producing a Sample Input File

AncestryMapper's references are produced using sequences in the same strand orientation as that used by dbSNP. Thus, all data input files for the functions ***calculateAMidsArith*** and ***createMedoid*** need to be in the same orientation.

If your data is already in the same strand orientation as dbSNP you can skip this section.

The format for any sample data to be analysed by the function ***calculateAMidsArith*** or used to create a reference with ***createMedoid*** is a PLINK tPED.

All files used in this tutorial can be found in the 'extdata' folder with the package. An example BED file, 'All-00', containing the dbSNP alleles for 1000 SNPs is included.

A full version of the 'All-00' BED file covering 53,509,352 SNPs from the dbSNP database can be found at: http://bit.ly/1OUstDP

It is faster to use PLINK BED format for the following steps; at the end the BED file needs to be converted to tPED format.

It is faster to extract only those SNPs being used in your dataset from the All-00 BED and working with the output.

The most efficient way to get data in the same orientation as that used by dbSNP is to to merge the 00-All file with your BED file using the --merge command in PLINK; both files should be in the ACGT format.

```
1 $plink --bfile All-00
2  --bmerge Prep.bed Prep.bim Prep.fam
3  --make-bed --out PrepMerge
```

If there are SNPs in different strand orientations, a list of SNPs affected will be output in a file ending in '-merge.misssnp'.

Flip these using the file ending in '-merge.missnp' in your sample file before merging the two files again.

```
$plink --bfile Prep --flip PrepMerge-merge.missnp --make-bed --out
    PrepFlipped

$plink --bfile All-00
--bmerge PrepFlipped.bed PrepFlipped.bim PrepFlipped.fam
--make-bed --out PrepMerge
```

At the end of this merging, the 'individual' from the All-00 file should be removed using the --remove command in PLINK. A file containing the ID and family ID for the All-00 individual is included as 'IndRm'.

```
$plink --bfile PrepMerge --remove IndRm --make-bed --out PrepMerge
```

SNPs that are CG or AT are invisible to the strand issue. The current AncestryMapper functions exclude them automatically from analysis. Sites with any missingness are also automatically excluded from analysis.

Sites with any missingness are also automatically excluded from analysis. Thus, if you wish to use any SNPs with samples missing you will need to impute replacements or exclude individuals who have missing genotypes.

The final step is to transform the BED file to a tPED file and to output a list of the SNPs used.

```
 $plink --bfile PrepMerge --transpose --recode --write-snplist --out
     PrepFinal
```

It is important to keep a file with the SNPs used in the sample with the ending '.snplist' as ***createMedoid*** will search for all the '.snplist' files corresponding to the path of the tPEDs.

3 Example Data

Example Data in the form of samples and references are provided containing 1000 SNPs with data from numerous populations and datasets containing 147 populations and 591 individuals. Due to the low number of SNPs, the plots will look noisy.

4 Additional Data

Full-sized medoids and a larger dbSNP reference ('00-All') for use with real user samples are currently being hosted at:
http://bit.ly/1OUstDP

5 Tutorial

Below is a short tutorial.

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

```
library(AncestryMapper)

#Path to folder containing population references.
Refs <- system.file("data", package = "AncestryMapper")

#Path to folder containing samples in tPED format.
tpeds <- system.file("extdata", package = "AncestryMapper")

#Path to CorPheno file.
Corpheno <- system.file("extdata", "CorPheno", package = "AncestryMapper")

#Path to dbSNP allele data file.
All00Frq <- system.file("data", "MinMaxFreq.rda", package = "AncestryMapper")
```

Calculate the genetic distance of the samples in your PLINK tPED files to the references.

```
genetic.distance <- calculateAMidsArith(pathTotpeds = tpeds,
                                        NameOut = "Example",
                                        pathToAriMedoids = Refs,
                                        pathAll00 = All00Frq)
```

The next step is to plot the results. In this example we are using the the normalised values for each individual. (*columnPlot* = 'I') For more information on the function arguments see ?plotAMids()

```
plotAMids(AMids = genetic.distance, phenoFile = Corpheno, columnPlot = "I")
```

To plot samples by population or place them in a certain order, they need to be added to the CorPheno file.

If samples are not in the CorPheno file, the samples will be plotted under 'Undefined'. If one or more references aren't in the CorPheno, then they will be plotted at the start of the y axis before the entries with defined orders. An entry with any ID will suffice for references if the user does not wish to add or

use the sample IDs from the data used to create the reference.

The CorPheno can easily be modified at the text level to change things such as population orders, colors or add new entries.

Alternatively the ***refAdd*** function may be more convenient. For more information see ?***refAdd***()

Users can also create their own population references from tPEDs containing individuals from a specific population. This is done using the ***createMedoid*** function.

The required arguments are:

- pathTotpeds: The path to the tPED file(s), one tPED per population
- pathAll00: The path to the file containing the dbSNP alleles

Example:

```
#Path to folder containing samples in tPED format.
tpeds <- system.file("extdata", package = "AncestryMapper")

#Path to dbSNP allele data file.
All00Frq <- system.file("data", "MinMaxFreq.rda", package = "AncestryMapper")

#Give value to chipMan, to give a sense of potential SNP overlap.
manu <- "Illumina"
```

```
createMedoid(pathTotpeds = tpeds, chipMan = manu, pathAll00 = All00Frq)
```

The resulting rda file will be named with the prefix of 'medoidArithmetic_', followed by the name of the tPED used and the number of SNPs. E.g.

medoidArithmetic_Demo_1000_ChipMan.rda

In order for the new reference to be used, it will need to in the path specified.

For more information on the presence of 'ChipMan' at the end of the file name, see ?***createMedoid***()