ASAFE (Ancestry Specific Allele Frequency Estimation)

Qian Zhang

2016-10-17

Contents

1 Introduction: What ASAFE does

The ASAFE (Ancestry Specific Allele Frequency Estimation) package contains a collection of functions that can be used to carry out an EM algorithm to estimate ancestry-specific allele frequencies for a bi-allelic genetic marker (e.g. a SNP) from genotypes and ancestry pairs, when each diploid individual's genotype phase relative to ancestry pair is unknown. If there are three ancestries, a = 0, 1, or 2 (e.g. African, European, or Native American), ASAFE functions can be used to estimate three probabilities, P(Allele 1|Ancestry a), a $\in \{0, 1, 2\}$, at each marker. ASAFE function algorithm_1snp_wrapper() can be applied to a matrix of ancestries and a matrix of genotypes for 3-way admixed diploid individuals at bi-allelic markers. Ancestries at different markers need not be phased with respect to each other, and genotypes at different markers need not be phased with respect to each other. Denoting each marker's alleles 0 and 1, algorithm_1snp_wrapper() outputs estimates of three ancestry-specific allele 1 frequencies for each marker.

1.1 ASAFE in the Context of a Larger Genetic Analysis Workflow

The following file in the ASAFE R package gives a diagram illustrating a genetic analysis workflow involving ASAFE: inst/ASAFE_Visual.pdf. An example script that performs two steps in the workflow, involving phasing of admixed genotypes with BEAGLE and then obtainment of local ancestry estimates and re-phased genotypes via RFMIX, is here: inst/scripts/bgl_then_rfmix.sh.

2 Input Files

2.1 Ancestry File

Your ancestry file should give admixed individuals' phased ancestries as a rectangular matrix with the following rows, columns, and entries:

2.1.1 Rows

- 1st row: Header line with column names
- Subsequent rows: 1 row corresponds to 1 marker

2.1.2 Columns

- 1st column: Marker ID
- 1 column per chromosome, with two consecutive columns per individual, corresponding to the individual's pair of homologous chromosomes. For example, if there are 3 admixed individuals, then the first row that gives column names might be

Marker ADM1 ADM1 ADM2 ADM2 ADM3 ADM3

- Columns should be separated by whitespace (i.e. spaces or tabs)

2.1.3 Entries

- For an entry that is not in the Marker ID column, an entry can take value 0, 1, or 2, which are arbitrary labels for three ancestries

To read your file into R, use a command like this:

```
ancestries <- read.table(file = "your_ancestry_file.txt", header = TRUE)
```

We have provided a subset of the full data set of simulated ancestries that was used in the ASAFE paper. [1] This data set is stored as a matrix adm_ancestries_test.

2.2 Genotype File

Your genotype file should give unphased admixed individuals' genotypes as a rectangular matrix with the following rows, columns, and entries:

2.2.1 Rows

- 1st row: Header line with column names
- Subsequent rows: 1 row corresponds to 1 marker

2.2.2 Columns

- 1st column: Marker ID
- 1 column per person. For example, if there are 3 admixed individuals, then the first row that gives column names might be

ID ADM1 ADM2 ADM3

- Columns should be separated by whitespace (i.e. spaces or tabs)
- Individuals must be listed in the same order in the genotype file as in the ancestry file

2.2.3 Entries

- For an entry that is not in the Marker ID column, an entry can take value 0/0, 0/1, 1/0, or 1/1, where 0 and 1 are arbitrary labels for a bi-allelic SNP's two alleles
- A slash "/" indicates an unphased genotype, so 0/1 and 1/0 are the same unphased genotype

To read your file into R, use a command like this:

```
genotypes <- read.table(file = "your_genotypes_file.txt", header = TRUE)
```

We have provided a subset of the full data set of simulated genotypes that was used in the ASAFE paper. [1] This data set is stored as a matrix adm_genotypes_test.

For details about data matrices that come with the ASAFE package, load the ASAFE package into R with the command "library(ASAFE)" and type "?" followed immediately by the name of the matrix, for example "?adm_ancestries_test".

3 Functions

These are the functions you might want to use:

- algorithm_1snp()
- algorithm_1snp_wrapper()

For information about these functions (e.g. their inputs, outputs, and examples for usage), see the function's man page by doing the following: Load the ASAFE package into R with the command "library(ASAFE)" and type "?" followed immediately by the name of the function, for example "?algorithm_1snp".

If you are interested in other functions that are not listed above, for instance functions that the above functions call, see the .R file that implements the function you're interested in for comments describing the function.

4 Reproducibility

Because Bioconductor requires that an R package pass "R CMD check" and "R CMD build" in less than 5 minutes each, and use at most 2 Gb of RAM to run code in this vignette, this Reproducibility section has been cut. Please see the version of the R package on GitHub (http://biostatqian.github.io/ASAFE/) for a vignette with a complete Reproducibility section.

5 Try ASAFE Out on a Small Data Set

We demonstrate how ASAFE package functions can be used in an analysis, by generating ancestry-specific allele 1 frequency estimates for a small data set. Ancestry and genotype data for the SNPs are respectively contained in matrices adm_ancestries_test and adm_genotypes_test.

```
# Clear workspace and load ASAFE
rm(list=ls())
library(ASAFE)

# adm_ancestries_test is a matrix with
# Rows: Markers
# Columns: Marker ID, individuals' chromosomes' ancestries
# (e.g. ADM1, ADM1, ADM2, ADM2, and etc.)

# adm_genotypes_test is a matrix with
# Rows: Markers
# Columns: Marker ID, individuals' genotypes (a1/a2)
# (e.g. ADM1, ADM2, ADM3, and etc.)

# Making the rsID column row names

row.names(adm_ancestries_test) <- adm_ancestries_test[,1]
row.names(adm_genotypes_test) <- adm_genotypes_test[,1]

adm_ancestries_test <- adm_ancestries_test[,-1]
adm_genotypes_test <- adm_genotypes_test[,-1]

# alleles_list is a list of lists.
# Outer list elements correspond to SNPs.
# Inner list elements correspond to 250 people's alleles
# with no delimiter separating alleles.
alleles_list <- apply(X = adm_genotypes_test, MARGIN = 1,
                      FUN = strsplit, split = "/")

# Creates a matrix:
# Alleles for chromosomes (ADM1, ADM1, ..., ADM250, ADM250) x (SNPs)
alleles_unlisted <- sapply(alleles_list, unlist)

# Change elements of the matrix to numeric
alleles <- apply(X = alleles_unlisted, MARGIN = 2, as.numeric)

# Apply the EM algorithm to each SNP to obtain
# ancestry-specific allele frequency estimates for all SNPs in
# matrices alleles and adm_ancestries_test.
#
# Columns correspond to markers.
# Rows correspond to ancestries 0, 1, and then 2.
# Entries in rows 2 through 4
# give P(Allele 1 | Ancestry a), a = 0, 1, or 2 for a marker.

adm_estimates_test <- sapply(X = 1:ncol(alleles), FUN = algorithm_1snp_wrapper,
                         alleles = alleles, ancestries = adm_ancestries_test)
```

```
adm_estimates_test
```

```
##      [,1]                    [,2]                    [,3]
## [1,] "rs0"                   "rs1"                   "rs4"
## [2,] "2.17620769780578e-08" "0.105263153761591"    "1.18131694231919e-09"
## [3,] "0.168278615987883"    "3.83154290109415e-09" "0.057142851651153"
## [4,] "0.334748301823412"    "2.35375994316977e-10" "4.92885078071704e-09"
##      [,4]                [,5]                    [,6]
## [1,] "rs7"               "rs14"                  "rs19"
## [2,] "0.815381149051769" "2.92397652269074e-08" "5.79951848126169e-22"
## [3,] "0.728970436943625" "8.78469964260191e-08" "0.017142818073981"
## [4,] "0.999999980824761" "0.00649340666737419"  "4.43964502125075e-08"
```

6 Citation

ASAFE: Ancestry-Specific Allele Frequency Estimation Qian S. Zhang; Brian L. Browning; Sharon R. Browning Bioinformatics 2016; doi: 10.1093/bioinformatics/btw220