Manual for EBhybrids v0.991

Ida Moltke and Matthew Stephens

September 2015

Contents

1 Introduction

This is a manual for EBhybrids version 0.991. EBhybrids is a collection of R functions for analyzing microsatellite data from samples with ancestry from either or both of two (sub)species. It provides functionality to easily estimate the posterior probability that a given sample is a hybrid between the two (sub)species and posterior probabilities of it being different hybrid types, such as F1 and F2.
For example, the R functions were developed for a study about African elephants [Mondol et al. 2015] in which a large number of elephant samples were genotyped at 16 autosomal loci. There are two subspecies of the African elephant; forest elephant and savanna elephant and most of the genotyped samples were either pure forest elephants or pure savanna elephants. However, a fraction of the samples, were potential hybrids between the two subspecies. For each of these samples EBhybrids was used to estimate the posterior probabilities of the sample being one of 6 hybrid types: (pure) forest, (pure) savanna, F1 (the offspring of a forest elephant and a savanna elephant), F2 (the offspring of two F1s), backcross forest (the offspring of a forest elephant and an F1) and backcross savanna (the offspring of a savanna elephant and an F1)[1]. Also, it was used to estimate the posterior probability of each sample being a hybrid (estimated as HP=1-Pr(hybrid type= forest|data)-Pr(hybrid type=savanna|data)).
Even though the code was developed for a study about elephants genotyped at 16 specific loci, EBhybrids can also be used for other diploid species or elephants genotyped at other markers. Note however that this will require that some parameters have to be re-estimated (allelic drop out rates), so the underlying model reflects the data being analyzed.

2 Getting started

2.1 Technical requirements to your computer

EBhybrids is a set of R functions. It works on all platforms (windows, linux and Mac), but R has to be installed. If you do not have R installed then first download and install R using the instructions here: `http://cran.r-project.org/doc/manuals/R-admin.html`

2.2 How to make the code ready for use

All you need to do is to unpack the file EBhybrids_v0.991.tar.gz with some unpack program. For instance, in unix, you can run the following command in you terminal window:

Installation commands

```
tar xfvz EBhybrids_v0.991.tar.gz
```

The unpacked folder will be called EBhybrids and is structured as follows: The "src" subfolder contains the R code. The "data" subfolder contains the example dataset which is used in this manual and the parameter values estimated for elephants in Mondol et al. (2015). Finally, the "man" subfolder contains this document.

[1] In principle the underlying model supports other hybrid types as well (e.g. third generation hybrids), however the current implementation does not.

3 File formats

3.1 Input files

To use EBhybrids, three input files are required: a genotype file, an ancestry proportion file, and a file with marker and (sub)species specific allelic dropout rate estimates. For elephants typed at the same 16 loci as the samples in Mondol et al. (2015) the allelic dropout rates from that study are available in the data folder. In all other cases you have to provide these estimates yourself.

3.1.1 The genotype file

The genotype file should contain unphased genotypes for all samples. It should be in Structure format, i.e. for each sample it should contain two lines; one for each chromosome copy. Each of these lines should as minimum contain the sample ID in the first column and the genotypes of the M genotyped markers in the last M columns.
For example a genotype file for three samples can look like this:

```
E1  71  93  70 -999  152  157  145   98  177  220  162  226 -999  143 171 145  146
E1  71  95  70 -999  160  157  155   98  177  220  174  226 -999  149 185 145  154
E2  76  95  84  187  160  153  151   94  163  223  166  226  245  149 181 139  146
E2  76  95  84  191  160  153  153  108  165  223  182  228  249  151 185 157  146
E3  76  89  82 -999  156  151  151  104  171 -999 -999 -999  245 -999 195 145 -999
E3  76  99  82 -999  156  153  155  108  271 -999 -999 -999  245 -999 199 149 -999
```

Here, for each of the three samples there are two lines and each of these lines have the sample's ID in the first column, the sample's location ID in the second column and the sample's genotype data from 16 microsatellite markers in the last 16 columns.
Note that missing genotypes can be indicated as either -9 or -999, but not both.

3.1.2 The ancestry proportion file

The ancestry proportion file is a file with estimated ancestry from each of the two potential ancestral (sub)species for all the samples. The file should consist of a single line for each sample, containing 2 numbers: the proportion of ancestry that is estimated to be from one of the (sub)species (for elephants this first subspecies should be 'forest' if you are using the allelic drop out rate estimates provided in the data folder) followed by the proportion of ancestry that is estimated to be from the other (sub)species. The order of the lines in this file should match the order in the genotype data file, so the first line contains ancestry proportion for the first sample in the genotype file etc.

An example of the content of an ancestry proportion file for three samples is

```
0.02 0.98
0.99 0.01
0.92 0.08
```

The ancestry proportions are only used to select the 'reference samples' from which the (sub)species specific allele frequencies are estimated. The exact ancestry proportions are therefore not important; what is important is whether one of ancestry proportions is above a certain threshold indicating that it has ancestry (almost) only from one of the (sub)species and thus can be used for estimating allele frequencies for that (sub)species (a threshold of 0.95 was used in Mondol et al. (2015), but the threshold is chosen by the user).
The proportions can be estimated using e.g. Structure [Pritchard et al. 2000]. Alternatively, if you have prior knowledge about which of the sample of "pure", then you can produce the file manually by adding a line with "1 0" for the pure samples from one (sub)species (this subspecies should be 'forest' if you are using the allelic drop out rate estimates provided in the data folder), the line "0 1" for the pure samples from the other (sub)species, and the line "0.5 0.5" for the rest (basically indicting they are not certain to be pure samples). However, unless you are very sure of your assignments, we recommend that the proportions are estimated using e.g. Structure.

3.1.3 The allelic dropout rate file

The allelic dropout rate file contains an estimate of the allelic drop out rate for each marker for each (sub)species. It should have one row for each marker with 3 numbers in it. First an estimate for each of the two (sub)species and then an estimate for loci with mixed ancestry (in Mondol et al. (2015), we used the mean of the estimates for the two subspecies as the last estimate). An example can be seen in the data folder where the file Mondoletal_allelicdropoutrates.txt contains the the allelic drop out rates that were estimated and used in Mondol et al. (2015).

3.1.4 Important comments about the input data

It is important to note that the estimates of posterior probabilities are highly dependent on the allele frequencies estimated. The input dataset should therefore contain numerous reference samples from each (sub)species, and samples that you are not certain a pure should not be included as such. Also, we recommend that you use reference samples from locations close to the sampling locations of the potential hybrids, since this will give rise to allele frequency estimates that better reflect the source populations for the potential hybrids. Finally, the posterior probabilities are also dependent on the dataset being representative in terms of the fraction of different hybrid types (including the pure ones). Hence it is a good idea to include reference samples in fractions that are representative of the area the potential hybrids are from (so the dataset is a close to a random sample from the area as possible).

3.2 Output files

The output is per default given as R matrices. But EBhybrids also provide a function for writing out these matrices to files. For examples of both matrices and files, see next section.

4 How to run an analysis

EBhybrids is a set of R functions, hence it is run by first opening R, then sourcing (reading into R) the functions and finally using these functions. Below is an example of a full analysis.

4.1 Example of use

First we open R with the src folder as work directory and 'source' the code, so all functions are available:

Make functions available

```
source("inferencefunctions.R")
```

We then read in the three input files:

Read in input data

```
## Read in genotype data
genodat <- read.table("../data/smallexample_geno.txt",as.is=T)

## Read in ancestry proportions
ancestryprops <- read.table("../data/smallexample_ancestryprops.txt",as.is=T)

## Read in allelic drop out rates
allelicdropoutrates = t(read.table("../data/Mondoletal_allelicdropoutrates.txt"))
```

Subsequently, we set the necessary additional parameters; the number of markers, the threshold for when to consider a sample "pure" and the genotyping error rate:

Set additional parameters

```
nmarkers = ncol(allelicdropoutrates)
ancestrythreshold=0.95
errorprob=0
```

Now we are ready to perform the inference. We first calculate likelihoods, P(data|hybrid type), for 6 different hybrid types: pure (sub)species 1, pure (sub)species 2, F1, F2, backcross to (sub)species 1, and backcross to (sub)species 2. Then, based on these likelihoods, we estimate posterior probabilities, P(data|hybrid type), for the same 6 hybrid types and HPs; posterior probabilities that the samples are hybrids (of any type):

Make inference

```
## - Calc likelihoods
lls = getloglikes(genodat,nmarkers,ancestryprops,ancestrythreshold,errorprob,allelicdropoutrates)

## - Estimate posteriors
hybridtypeposteriorinfo = getallposteriors(lls)
HPs = gethybridposteriors(hybridtypeposteriorinfo)
```

Now 'lls' contains a line for each sample with the log of the likelihoods for the 6 hybrid types described above. Similarly, 'hybridtypeposteriorinfo$posteriors' contains a line for each sample with the posteriors

for the 6 hybrid types. Finally, HPs contains the probability for each sample to be a hybrid. For example:

View results

```
> hybridtypeposteriorinfo$posteriors
                Pure1        Pure2 F1 F2 Bx1 Bx2
ForestE1  1.000000e+00 1.876741e-12  0  0   0   0
ForestE2  1.000000e+00 6.483287e-11  0  0   0   0
ForestE3  1.000000e+00 3.174554e-10  0  0   0   0
ForestE4  1.000000e+00 8.395608e-13  0  0   0   0
ForestE5  1.000000e+00 3.169808e-09  0  0   0   0
UnknownE1 1.041153e-04 9.998959e-01  0  0   0   0
UnknownE2 7.166475e-04 9.992834e-01  0  0   0   0
SavannaE1 4.538531e-11 1.000000e+00  0  0   0   0
SavannaE2 1.240423e-09 1.000000e+00  0  0   0   0
SavannaE3 3.613286e-11 1.000000e+00  0  0   0   0
SavannaE4 1.246257e-12 1.000000e+00  0  0   0   0
SavannaE5 9.084518e-14 1.000000e+00  0  0   0   0
```

We can write these results to file if we want. One way to do this is to use the function 'writePosteriors'. This function takes the name of the file and the posteriors you want to write to file and writes the posteriors to both a txt file and a csv file. For example:

Write results to files

```
## - Write out hybrid type posteriors to both a txt file and csv file
writeResults2files("smallexample_HybridTypePosteriors",hybridtypeposteriorinfo$posteriors)

## - Write out HPs to both a txt file and csv file
writeResults2files("smallexample_HPs",HPs)
```

produces the files smallexample_HybridTypePosteriors.txt, smallexample_HybridTypePosteriors.csv, smallexample_HPs.txt and smallexample_HPs.csv. The first one looks like this:

Content of smallexample_HybridTypePosteriors.txt

```
SampleID Pure1 Pure2 F1 F2 Bx1 Bx2
ForestE1 0.999999999998123 1.87674149267333e-12 0 0 0 0
ForestE2 0.999999999935167 6.48328683760275e-11 0 0 0 0
ForestE3 0.999999999682545 3.17455417262481e-10 0 0 0 0
ForestE4 0.99999999999916 8.39560776398524e-13 0 0 0 0
ForestE5 0.999999996830192 3.16980787002802e-09 0 0 0 0
UnknownE1 0.000104115273318623 0.999895884726681 0 0 0 0
UnknownE2 0.000716647450876046 0.999283352549124 0 0 0 0
SavannaE1 4.53853092543472e-11 0.999999999954615 0 0 0 0
SavannaE2 1.24042293926445e-09 0.999999998759577 0 0 0 0
SavannaE3 3.61328615743512e-11 0.999999999963867 0 0 0 0
SavannaE4 1.24625698529427e-12 0.999999999998754 0 0 0 0
SavannaE5 9.08451796357836e-14 0.999999999999909 0 0 0 0
```

If you want the numbers rounded e.g. so you get at most 6 digits after the comma you can add that

as a last argument to the function:

Write rounded results to files

```
writeResults2files("smallexample_HybridTypePosteriors_rounded",hybridtypeposteriorinfo$posteriors,6)
```

The resulting txt file will then look like this:

Content of smallexample_HybridTypePosteriors_rounded.txt

```
SampleID Pure1 Pure2 F1 F2 Bx1 Bx2
ForestE1 1 0 0 0 0 0
ForestE2 1 0 0 0 0 0
ForestE3 1 0 0 0 0 0
ForestE4 1 0 0 0 0 0
ForestE5 1 0 0 0 0 0
UnknownE1 0.000104 0.999896 0 0 0 0
UnknownE2 0.000717 0.999283 0 0 0 0
SavannaE1 0 1 0 0 0 0
SavannaE2 0 1 0 0 0 0
SavannaE3 0 1 0 0 0 0
SavannaE4 0 1 0 0 0 0
SavannaE5 0 1 0 0 0 0
```

5 Miscellaneous

5.1 More details about the underlying model and method

For more details see Mondol et al. (2015).

5.2 Warranty

This code comes under no warranty.

5.3 Bugs and questions

If you find a bug in the code, please let us know. You can contact us per email: ida@binf.ku.dk.

5.4 Citation

If you use this software for an analysis please cite Mondol et al. (2015).

References

[Mondol et al. 2015] Mondol S, Moltke I, Hart J, Keigwin M, Brown L, Stephens M, Wasser SK. *New evidence for hybrid zones of forest and savanna elephants in Central and West Africa.* 2015 (in review).

[Pritchard et al. 2000] Pritchard JK, Stephens M and Donnelly P. *Inference of population structure using multilocus genotype data.* Genetics 155(2):945–959, 2000.