AFLPsim Manual (ver. 0.4-1)

Francisco Balao & Juan Luis García-Castaño
University of Seville

December 23, 2014

Contents

1 Theorical background **1**
1.1 Hybrid simulation 1
1.2 Genome scan methods 2
1.3 Demographic evolution in hybrid zones 4

2 Exporting results **4**

3 Examples **4**
3.1 Example 1. Simulating hybridization with selection and genomic scan for BxA individuals 5
3.2 Example 2. Simulating F_1, F_2 and backcross hybrids from two divergent species under neutral selection 6
3.3 Example 3. Simulating demographical evolution under hybridization 8

4 Obtaining and citing AFLPsim **9**

1 Theorical background

This document clarifies the use of the package AFLPsim for simulating dominant marker profiles in hybridizing populations. In addition, a new genome scan approach for hybrids is explained.

1.1 Hybrid simulation

The AFLPsim package includes a one-step simulation algorithm based on phenotypic frequencies for each fragment. Firstly, the algorithm simulates allelic frequencies for i markers in each parental population out of independent theoretical beta distributions (with $\alpha = \beta = 0.5$, following [Wright, 1931]; Figure 1). Later, the algorithm calculated the expected offspring phenotypic frequencies for each fragment under neutrality (i.e., Hardy-Weinberg equilibrium and lack of linkage disequilibrium). For the F_1 hybrids, the expected phenotypic frequency of each band is

$$E(f_{F_1}) = p_A + p_B - p_A p_B$$

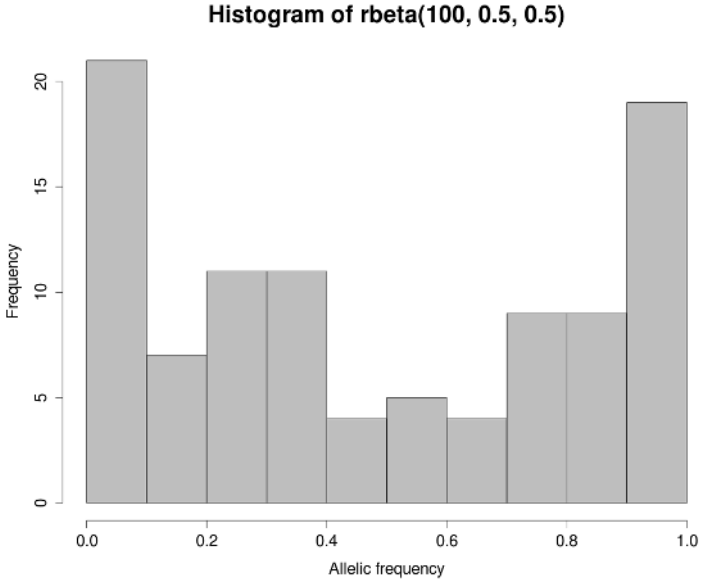

Figure 1: Simulated allelic frequencies from a theoretical beta distribution with $\alpha = \beta = 0.5$.

where p_A is the frequency of the presence-allele in the parental population A ($p_A = 1 - \sqrt{(}1 - frequency\,of\,the\,band)$), and p_B is the frequency of the presence-allele in the parental population B ($p_B = 1 - \sqrt{(}1 - frequency\,of\,the\,band)$). Expected frequencies are used to create hybrid progeny profiles ($N = 1000$) from i binomial distributions (one for each locus). For other hybrid classes we follow the same development. For example, for backcrosses with parental A, the expected frequency, based again on parental allele frequencies, is

$$E(f_{BxA}) = \frac{(3p_A + p_B - p_A^2 - p_A p_B)}{2}$$

A phenotypic selection coefficient (s) can be applied, which alters the expected neutral frequency obtaining, after selection, a new expected one. This coefficient was 0 when there was no selection and it varied following negative and positive directional selection (from -1 to $+\infty$); finally, the resulting progenies were rescaled to the total resulting offspring. The following formula summarizes both steps

$$E(f_{F_1}) = \frac{w(p_A + p_B - p_A p_B)}{(w(p_A + p_B - p_A p_B) + (1 - p_A)(1 - p_B))}$$

where w is 'fitness', which relates to the selection coefficient (s) through the equation $w = 1 + s$.

1.2 Genome scan methods

We developed a new method (bal&gar-ca) based on the departure of the theoretically expected frequencies for each band in each hybrid category under neutral

Figure 2: Animation showing the steps of the bal&gar-car method.

introgression. In contrast to the approach proposed by [Gagnaire et al., 2009], our method takes into account the sampling error in the parental estimated values, which could seriously bias the test. [Gagnaire et al., 2009] tested the frequencies from the sampled hybrids against these expected values with a binomial test [Sokal and Rohlf, 1981]. However, errors when sampling parental frequencies should be taken into account, and our method does so. Instead of calculating only one expected value we calculated four expected frequencies, which delimited a small portion of the neutral expectation surface. To define this area, confidence intervals at $\sqrt{(}1-\alpha)$ were calculated for each parent using the very conservative Clopper-Pearson exact procedure ([Clopper and Pearson, 1934]; [Brown et al., 2001]), based both on the number of counts for each fragment and the number of individuals sampled. Every combination of one interval end from one parental and one interval end from the other parental led to an expected value within the neutral expectation surface, and the four values delimited a $(1-\alpha)$ probability portion of it. To test if a specific locus behaves as an outlier, the average real offspring value is confronted against these four estimated values, considering the two following possibilities: (i) if it was within the two most extreme values, we concluded the fragment was not under selection; (ii) if not, we chose the closest of the four frequencies to be the expected value of the binomial test. Finally, a False Discovery Rate (FDR) correction [Benjamini and Hochberg, 1995] was applied to avoid false positive detection due to type I errors (q-values $\alpha = 0.05$ were considered significant). This genome scan algorithm is summarized in Figure 2.

1.3 Demographic evolution in hybrid zones

A heuristic model of demographic evolution in hybrid zones was developed by [Epifanio and Philipp, 2000]. This model simulates the proportion of parentals, F_1, F_x and backcrosses (with both parentals) individuals for each generation. The composition of each taxon following admixture and hybridization depends on three independent variables: (1) the initial proportion of parental taxa; (2) the fitness gradient among parental and hybrid taxa; and, (3) the strength of assortative mating among these taxa. Composition at any time (t, in generations) is calculated by multiplying initial abundance by relative fitness and probability of mating using the general equation

$$S_G = \frac{(\phi_t * \omega_G)}{(\sum \phi_t * \omega_G)}$$

where S_G is the proportion of a taxon G, surviving to reproduction, ϕ_t is the frequency of the taxon at the beginning of the generation t, before selection, and ω_G is the fitness of the taxon. The expected contribution of a taxon to the subsequent generation is determined by the general equation

$$\phi_{t+1} = S_G * M$$

where M is a matrix of mating preference to account for assortative mating.

2 Exporting results

AFLPsim functions do not require external input files out of the R environment. However, for simulation of hybrids from user-specified parental data, this should be loaded to R as a `matrix` or a `data.frame`. Simulation results can be readily used by multivariate and phylogenetic methods of other R packages. The function `sim2genind` exports `hybridsim` object to `genind` object for packages `ade4` [Dray and Dufour, 2007] and `adegenet` [Jombart, 2008]. Our package is also able to export data formatted for ARLEQUIN (`sim2arlequin` [Excoffier et al., 2005]) and POPGENE (`sim2popgene` [Yeh and Boyle, 1997]) to estimate summary statistics from the data set. Simulation results can also be exported to assignment software such as STRUCTURE (`sim2structure`, [Falush et al., 2007]) and NEWHYBRIDS (`sim2newhybrids`, [Anderson, 2008]).
In addition, AFLPsim contains functions that produce graphics for visualizing the expected frequencies under neutrality for loci under selection across the different hybrid classes. Our package finally includes a function that plots the results of the demographic evolution model in a hybrid zone.

3 Examples

To demonstrate the capacities of AFLPsim, we have created three illustrative examples, which can be easily reproduced.

3.1 Example 1. Simulating hybridization with selection and genomic scan for BxA individuals

In this example, we carry out the simulation of two parental populations of 100 individuals and 100 BxA hybrids for a total of 300 markers using the `hybridsim` function. Negative selection is simulated with s = -0.999 for 15 out of the 300 markers (we can simulate negative selection using a $-1 < s < 0$).Firstly, we need to load AFLPsim and set the random seed number (arbitrarily to 1234567 for reproducibility).

```
> require(AFLPsim)
> set.seed(1234567)
> BxAhybrid<-hybridsim(Nmarker=300, Na=100, Nb=100, Nbxa=100,
+ type='selection', hybrid='BxA', S=-0.999, Nsel=15)
```

Then we performe a genomic scan with the 'bal&gar-ca' method setting the type parameter to the correct hybrid class (i.e. BxA hybrids).

```
> outlier<-gscan(BxAhybrid, type='BxA', method='bal&gar-ca')
```

```
[1] 7.91544e-44
[1] 0.8263916
[1] 1.049247e-71
[1] 0.8415584
[1] 8.089165e-92
[1] 1
[1] 1
[1] 1
[1] 5.0459e-90
[1] 0.4493354
[1] 1
[1] 0.9194567
[1] 2.065094e-33
[1] 0.6903591
[1] 2.488801e-91
[1] 0.6326109
[1] 1.104209e-15
[1] 1
[1] 5.20453e-105
[1] 0.3877579
[1] 4.335256e-74
[1] 1
[1] 0.7611835
[1] 4.05538e-55
[1] 2.65595e-53
[1] 0.9200181
[1] 0.5486732
[1] 1
[1] 0.6876739
[1] 0.5496625
[1] 1.306521e-34
[1] 0.5151991
```

```
[1] 0.01293191
[1] 1
[1] 1
[1] 1
[1] 0.6720631
[1] 1
[1] 0.5925196
[1] 0.68449
[1] 1
[1] 1
[1] 3.102551e-33
[1] 0.4310621
[1] 1
[1] 4.304945e-09
[1] 0.6754158
[1] 0.8839897
```

The results of the genomic scan analysis are saved in a data object (outlier), which is used as the input for the `plot.sim` function to generate a plot of the outlier markers and the expected frequencies under neutrality (Figure 4).

```
> plot.hybridsim(BxAhybrid,hybrid='BxA', markers=outlier$Outliers)
```

3.2 Example 2. Simulating F_1, F_2 and backcross hybrids from two divergent species under neutral selection

In this second example, we carry out the simulation of 50 hybrid individuals of each hybrid class (F_1, Backcross to Parental A and Backcross to Parental B) from two parental profiles from two provided files ('SpeciesA.txt' and 'SpeciesB.txt', included in the package) with 1000 AFLP markers.

```
> SpeciesA<-read.table(system.file('/files/SpeciesA.txt',
+ package = 'AFLPsim'), header=TRUE, row.names=1)
> SpeciesB<-read.table(system.file('/files/SpeciesB.txt',
+ package = 'AFLPsim'), header=TRUE, row.names=1)
```

The hybridization is simulated under neutrality with the `hybridize` function.

```
> set.seed(1234567)
> hybridswarm<-hybridize(SpeciesA, SpeciesB, Nf1=50,
+ Nbxa=50, Nbxb=50, type='neutral', hybrid=c('F1', 'BxA', 'BxB'))

########Neutral hybridization########
```

The result of the hybridization is saved to an object (hybridswarm) and exported to the NEWHYBRIDS format.

```
> sim2newhybrids(hybridswarm, filename= 'testnewhybrids.txt')
```

In addition, we calculate the hybrid index of the simulated hybrids using the wrapper for the 'est.h' function of the introgress package. A histogram of the estimates is plotted to visualize the three different hybrid classes (Figure 5).

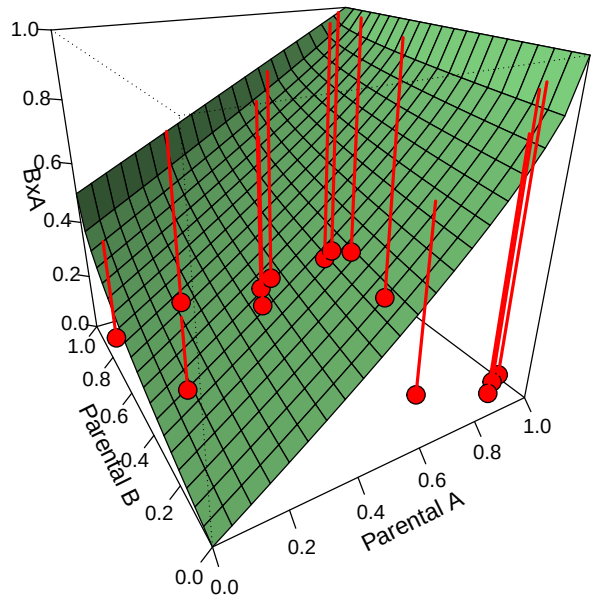

Figure 3: Three-dimensional scatter plot showing significant outlier loci detected by the `gscan` function for the backcrosses to parental A (BxA) simulated in Example 1. The green-coloured surface shows the theoretical probability of observing a dominant marker as a function of the band presence frequency in each parental species. The difference between the observed and the theoretical band frequency is represented with a vertical line joining both values.

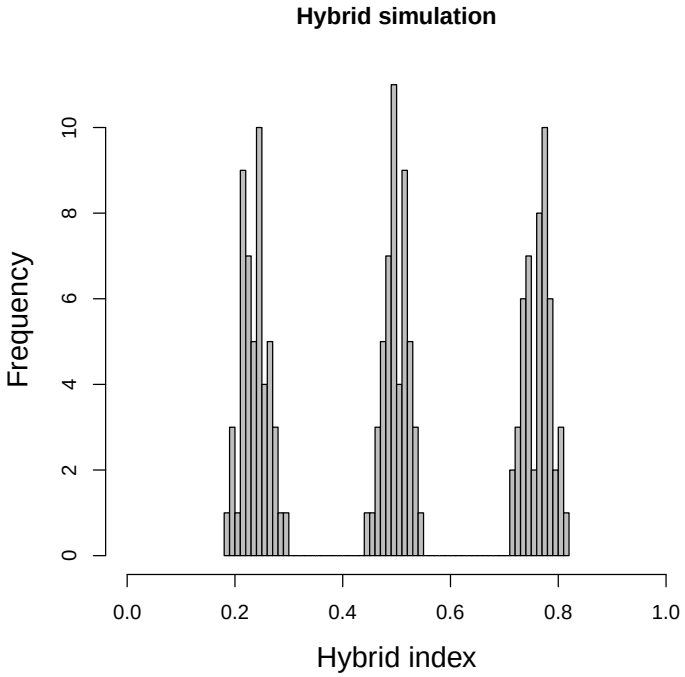

Figure 4: Histogram of the maximum likelihood hybrid index estimates of the F_1 and simulated backcross individuals from Example 2.

```
> hind<-hybridindex(hybridswarm)

prepare.data is working; this may take a moment
Processing data for 150 individuals and 1000 loci.
est.h is working; this may take a few minutes

> hist(hind$h, col='grey', xlim=c(0,1), breaks=50,
+ xlab='Hybrid index', main ='Hybrid simulation', cex.lab=1.4)
```

3.3 Example 3. Simulating demographical evolution under hybridization

Finally, we are going to simulate hybridization on one population. For the initial frequencies we need to create a vector with the frequencies of Parental A, Parental B, F_1, Backcross to Parental A, Backcross to Parental B and F_x. In our example, we fix Parental A and Parental B initial frequencies to 0.5.

```
> freqinit<-c(0.5,0.5,0,0,0,0)
```

Then, we create a matrix of assortative mating using the matrix function, and allow crosses among all taxa with the same propability.

```
> matingmat<-matrix(1,ncol=6,nrow=6)
```

In this example, parentals have similar fitness but F_1 individuals' is lower than parentals'. Here, we want asymmetrical introgression and breakdown to occur after F_1 hybrids, with posterior hybrid generations (F_x) and Backcross to Parental B (BxB) being sterile. Moreover, Backcross to Parental A (BxA) would have a similar fitness to F_1 individuals. Hence, fitness would be modified as following:

```
> fitness<-c(1,1,0.5,0.5,0,0)
```

We obtain a matrix with the frequency of each taxon in eight generations. The results was that Parental A dominates the hybrid zone after eight generations, displacing the other parental and the hybrids.

```
> set.seed(1234567)
> results<-demosimhybrid(freqinit, matingmat, fitness)
```

We use the `plot.demosim` function to visualize the demographic evolution (Figure 6):

```
> plot.demosimhybrid(results)
```

4 Obtaining and citing AFLPsim

AFLPsim is a package of the statistical programming environment R, which is available for all computing platforms from the Comprehensive R Archive Network (CRAN, http://cran.r-project.org). A stable version of AFLPsim is also available on CRAN and can be downloaded from within R while connected to the Internet by entering the following commands at the prompt:

```
> install.packages('AFLPsim')
> library(AFLPsim)
```

AFLPsim package depends on introgress R library to run `hindex` function. BAYESCAN software [Foll and Gaggiotti, 2008] must be installed in the system (see the help provided with the package for more information). In addition to the stable version on CRAN, a development version is available on Github repository
(https://github.com/fbalao/AFLPsim); this version can be installed from within R by using the `devtools` package [Wickham and Chang, 2013]:

```
> install.packages('devtools')
> library(devtools)
> install_github('AFLPsim', username='fbalao')
```

Scientists using AFLPsim in a published paper should cite this article. Citation information can be obtained by typing at the command prompt:

```
> citation('AFLPsim', auto=T)
```

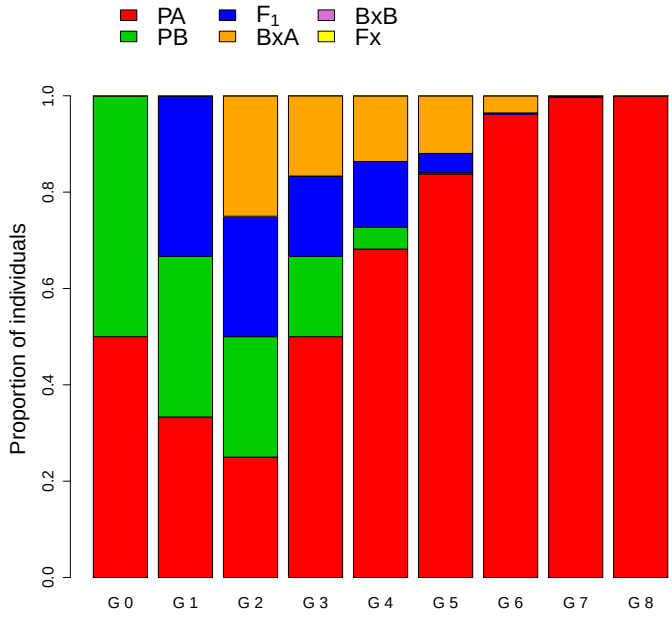

Figure 5: Simulated demographic evolution of a hybrid zone under similar initial proportions of the parentals, using the model of [Epifanio and Philipp, 2000] (Example 3). Each bar represents the relative proportion of each parental and hybrid category (see legend) in the area over 8 generations, until Parental B (PB) dominates the population.

```
To cite package 'AFLPsim' in publications use:

  Francisco Balao and Juan Luis García-Castaño
  (2014). AFLPsim: Hybrid simulation and genome scan
  for dominant markers. R package version 0.3-4.
  http://www.r-project.org,
  http://personal.us.es/fbalao/software.html

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {AFLPsim: Hybrid simulation and genome scan for dominant markers},
    author = {Francisco Balao and Juan Luis García-Castaño},
    year = {2014},
    note = {R package version 0.3-4},
    url = {http://www.r-project.org,
http://personal.us.es/fbalao/software.html},
  }
```

References

[Anderson, 2008] Anderson, E. C. (2008). Bayesian inference of species hybrids using multilocus dominant genetic markers. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1505):2841–2850.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

[Brown et al., 2001] Brown, L. D., Cai, T. T., and Anirban, D. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117.

[Clopper and Pearson, 1934] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

[Dray and Dufour, 2007] Dray, S. and Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22:1–20.

[Epifanio and Philipp, 2000] Epifanio, J. and Philipp, D. (2000). Simulating the extinction of parental lineages from introgressive hybridization: the effects of fitness, initial proportions of parental taxa, and mate choice. *Reviews in Fish Biology and Fisheries*, 10(3):339–354.

[Excoffier et al., 2005] Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1:47–50.

[Falush et al., 2007] Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7:574–578.

[Foll and Gaggiotti, 2008] Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, 180(2):977–93.

[Gagnaire et al., 2009] Gagnaire, P. A., Albert, V., Jónsson, B., and Bernatchez, L. (2009). Natural selection influences AFLP intraspecific genetic variability and introgression patterns in Atlantic eels. *Molecular Ecology*, 18(8):1678–1691.

[Jombart, 2008] Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24:1403–1405.

[Sokal and Rohlf, 1981] Sokal, R. R. and Rohlf, F. J. (1981). *Biometry : the principles and practice of statistics in biological research*. W.H. Freeman, New York.

[Wickham and Chang, 2013] Wickham, H. and Chang, W. (2013). *devtools: tools to make developing R code easier*. R package version 1.4.1.

[Wright, 1931] Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

[Yeh and Boyle, 1997] Yeh, F. and Boyle, T. (1997). Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany*, 129:157.