FABSIM

ANNA RAMÍREZ-SORIANO AND FRANCESC CALAFELL

JULY 2008

F_{ST} is widely used to find genes under local selection by comparing the F_{ST} value of a single locus against genome-wide, empirical values. However, empirical distributions suffer from ascertainment bias caused by the protocol used to select SNPs, which affects the shape of the distribution. An alternative is working with simulated distributions, but this procedure also produces unreliable distributions as F_{ST} is highly dependant on the demographic history of the samples, and simulations do not take into account ascertainment bias. Provided that there is an increasing amount of information on the demographic history of populations, we have developed a software that applies ascertainment bias on simulated sequences and calculates F_{ST} on them. Moreover, we also used our program to generate several simulated F_{ST} distributions with different ascertainment biases and have compared them against the F_{ST} values found in an empirical database.

The program has been developed using Java version 6.0 and compiled under Windows, but it can also be used in Linux with graphic environment. To run FABSIM, just double-click the .jar file or, if working from the command line, write java –jar FABSIM.jar.

Reference: Ramirez-Soriano, A., and F. Calafell, FABSIM: A software for generating fst distributions with various ascertainment biases. Submitted.

INFILE INFORMATION

GENERAL INFORMATION

Input files are entered to the program selecting them from their location using the Browse button. All infile names must have an extension separated from the name by a dot. However, FABSIM is not strict in the extension name.

FABSIM accepts three different infile formats, the ones produced by ms (http://home.uchicago.edu/~rhudson1/; Hudson, 2002), cosi (http://www.broad.mit.edu/~sfs/cosi/; Schaffner, *et al.* 2005), and SelSim (http://www.stats.ox.ac.uk/~spencer/SelSim/Controlfile.html; Spencer, *et al.* 2004). Depending on the input format, FABSIM requires additional information, as detailed below.

INFILE FORMATS

ms format

The ms format is characterised by having a header and, below, the samples separated by a space and a double bar (//).

The header has the name of the program, the number of chromosomes per sample, the number of runs, the parameter for running ms and the seed used:

```
ms <chromosomes> <runs> -s/-t <various parameters>
<seed>

ms 5 20 -s 10
111
```

Of the header, the only information that will be actually used by FABSIM is the number of chromosomes per sample and the number of runs.

The samples start with the double bar. After that, two lines indicate the number of segregating sites and their positions. Finally the chromosomes are listed, one per line:

```
//
segsites: 10
positions:  0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
```

As ms provides the relative position of each segregating site in a scale from 0 to 1, FABSIM requires the simulated sequence length from the user. Absolute positions are obtained multiplying the relative position by the length of the fragment and rounding to the nearest integer.

cosi format

cosi provides two files for each simulated population, named out.hap and out.pos, which contain the haplotypes and the information for each segregating site respectively. If run as provided, cosi only simulates one sample per file. However, FABSIM is able to process files containing multiple samples separated by a blank line (this must be implemented in both files).

A multiple-sample files can be obtained using the script run_cosi, which can be dowloaded from http://www.snpator.com/public/downloads/aRamirez/FABSIM. run_cosi is a command-line working script which runs under php (http://www.php.net/downloads.php). It requires to be placed in the same folder than coalescent.exe (the cosi's executable file) and the input file, infile.1. It requires as parameters the number of samples, the infile without .1 and a label, as follows:

```
php run_cosi.php 5 infile TEST
```

The cosi output haplotypes file only contains the sample or samples as follows:

```
0       1       2 2 1 2 2 2 2 1 2 2
1       1       1 2 2 1 2 2 2 2 2 1
2       1       2 2 2 2 2 2 2 2 2 1
3       1       2 1 1 2 1 2 2 1 2 2
4       1       2 2 2 2 2 1 2 1 2 2
5       1       2 2 2 2 2 2 2 2 2 1
```

The first column states the chromosome number, the second the population label and afterwards come the segregating sites, each position separated by a blank space.

The file containing the information per each segregating sites contains the SNP number, the population label, the position of the site, and the frequency of each allele:

```
SNP   CHROM   CHROM_POS    ALLELE1 FREQ1   ALLELE2 FREQ2
1       1       127.3788     1       0.1154  2       0.8846
2       1       215.9448     1       0.0000  2       1.0000
3       1       229.8352     1       0.0000  2       1.0000
4       1       623.0247     1       0.4231  2       0.5769
5       1       463.2629     1       0.1538  2       0.8462
```

When using infiles in a cosi output format, the user must introduce in the program the two files provided for each population included in the analysis. The name of the haplotype files must contain a label followed by a dash, a number indicating the population code and a point:

```
TC-1.testCosi.1
TC-2.testCosi.1
```

Information files must have the same label followed by a dot, the Word "pos", a dash and the number indicating the population code:

```
TC.pos-1.testCosi.1
TC.pos-2.testCosi.1
```

FABSIM requires the number of samples (iterations) in the file from the user. The positions are rounded for analysis to the nearest integer.

SelSim format

Output files from SelSim start with a blank line followed by a header that contains the name of the control file, the seed, and the type of output used. FABSIM only accepts "sequences" files:

```
SelSimCON.txt   -1147959592  Sequences
```

The header is followed by the samples, separated by a space. Samples start with a line with a double bar (//) followed by the number of chromosomes, the number of segregating sites and the sequence length. Next the positions of each segregating site are specified and after a blank line the samples, with each locus separated by blank spaces. After another blank line the time of the marginal genealogy underlying each position is provided, and finally the total time in all marginal trees which has not mutated, as follows:

```
        //5   11   2000
        1  35  132  285  299  330  463  525  781  1528  1703

        1  0  0  1  0  0  0  0  1  0  0
        1  1  0  0  0  0  0  0  0  0  1
        1  0  0  0  0  0  0  0  0  0  1
        1  0  0  1  0  0  0  0  1  0  0
        1  0  0  0  0  0  1  0  1  0  0

        3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093
3.33093   3.33093   3.33093

        6621.9
```

FABSIM requires the number of samples (iterations) in the file from the user.

CALCULATIONS

ASCERTAINMENT BIAS

Seven different ascertainment biases (or no bias) can be applied to data. Each one of them requires information to be added by the user. The appropriate fields required turn white rather than grey after the bias to be applied is selected by the user.

More than one bias can be applied to data at the same time, except if "none" is selected. Other incompatibilities are listed when applicable. If more than one bias is selected, they are applied to the sample in the order FABSIM displays them (which is also the same in which they are explained here).

Except if the contrary is said, in case that more than one population is introduced for analysis the SNPs are selected over only one population (determined by the user), but the bias is applied over all populations. That is, the SNPs deleted from the chosen population are deleted from all populations.

Discovery sample per gene

Discovery sample per gene assumes that only some chromosomes (a subsample of *d* size) of a sample of size *n* have been resequenced, and the segregating sites found on them have been genotyped on the whole sample *n*.

When this bias is activated by the user, *d* sequences are randomly selected over the total number of sequences in the sample, and only those SNPs that are polymorphic in these *d* sequences are kept.

The information required to apply this bias is the population where the *d* sample is to be selected from and the *d* sample size. The latter must be an integer between 0 and *n*.

This bias cannot be applied together with *Discovery sample per SNP*.

Discovery sample per SNP

Discovery sample per SNP works similarly to *discovery sample per gene*, but a different *d* sample is chosen for each locus.

As above, the information required to apply this bias is the population where the *d* sample is to be selected from and the *d* sample size. The latter must be an integer between 0 and *n*.

This bias is incompatible with *Discovery sample per gene*.

SNPs polymorphic in a given population

SNPs polymorphic in a given population keeps only those SNPs that are polymorphic in the selected populations.

FABSIM requires from the user the population in which to select the polymorphic loci.

This bias cannot be applied together with *SNPs polymorphic in all populations*.

SNPs polymorphic in all populations

SNPs polymorphic in all populations keeps only those SNPs which are polymorphic in all the populations entered.

No parameters are needed from the user in this bias.

This bias cannot be applied together with *SNPs polymorphic in a given population*.

MAF > threshold

The *MAF > thresold* bias discards all the SNPs that have a minor allele frequency (MAF) under the threshold provided by the user.

The information required to apply this bias is the population to ascertain and the threshold. The last must be a positive number smaller than 0.5.

Fixed SNP spacing

The *Fixed SNP spacing* bias selects one SNP every given number of bases. To do so, FABSIM selects randomly one segregating site among the *x* first base pairs. From this first selected SNP, it counts the position that is found *x* base pairs further. If in this new position there is a segregating site FABSIM selects it; otherwise, it selects the nearest one, either upstream or downstream of the new position. FABSIM proceed as explained until the new position is found outside the simulated fragment.

The information required to apply this bias is the population from which the SNPs should be ascertained and *x*, the spacing in basepairs. The latter must be a positive integer.

This bias cannot be applied together with *Variable SNP spacing*.

Variable snp spacing

In this bias, the user can specify different segments in the simulated sequence in which different SNP spacings will be applied, as described above for *Fixed SNP spacing*.

The information required to apply this bias is the population from which the SNPs should be ascertained and a file containing the different SNP densities. This file must have two columns: the first indicates the position in bp and the second the spacing for SNPs in this fragment:

```
200	50
300	10
500	70
```

In this example we consider a simulated 500-bp region. For its first 200 bp, a SNP is chosen every 50 bp. From the SNPs located between position 200 and 300, one SNP is chosen every 10 bp. From position 300 to the end, the distance between SNPs is 70 bp.

This bias cannot be applied together with *Fixed SNP spacing*.

STATISTICS

FABSIM can calculate F_{st}, minor (MAF) and derived (DAF) allele frequencies, and a number of neutrality tests on simulation data. Several statistics can be computed together in the same execution of the program.

F_{ST}

F_{st} can be calculated according to several parameters. On one hand, it can be corrected or not by the different sample sizes between populations. On the other hand, FABSIM can output the F_{st} for each SNP in the sample, per gene, or both.

The number of populations to compare is not limited. However, all of them need to have the same number of segregating sites, located in the same positions.

MAF and DAF

FABSIM calculates minor (MAF) and derived (DAF) allele frequencies for all the SNPs in all the samples of the simulations.

To calculate DAF, FABSIM assumes as ancestral the locus coded as ‘0’ of ms and SelSim and the locus coded as ‘2’ for cosi, as stated in the documentation of the programs.

Neutrality statistics

The neutrality statistics included in FABSIM are the number of segregating sites, the number of pairwise differences, the number of singletons, Tajima’s *D* (Tajima, 1989); Fu and Li’s *D*, *F*, *D** and *F** (Fu, *et al.* 1993); Fay and Wu’s *H* (Fay, *et al.* 2000), R_2 (Ramos-Onsins, *et al.* 2002), Fu’s F_s (Fu, 1997), *Dh* (Nei, 1987)(equation 8.4 replacing 2n by n), Wall’s *B* and *Q* (Wall, 1999), Kelly’s Z_{nS} (Kelly, 1997), Rozas’ Z_A and *ZZ* (Rozas, *et al.* 2001) and extended haplotype homozygosity *EHH* (Sabeti, *et al.* 2002, Ramirez-Soriano, *et al.* 2008,).

OUTPUT FILES

OUTFILE FORMATS

FABSIM output file do not have an uniformly formatted content, given the diversity of the results FABSIM can produce. However, two general predefined formats are provided: information per sample and tabulated data.

Information per sample

Information per sample shows the information for all samples linearly, separating them by a blank line. Each sample starts with a line stating the sample number (e.g. SAMPLE 1), and is followed by a list with the desired statistic values. Examples of this format for each statistic are shown below.

F_{ST}

The F_{st} output file in the information per sample format has three different appearances depending on if the user wants the information per snp, per sample or both. In any case the first two lines, which are shared between this and the tabulated format, show the populations that are being compared (first) and the legend. Next come the F_{st} value for each SNP or the average, maximum and minimum F_{st} of the gene, depending on what it has been required. If both per gene and per locus F_{st} are requested, per locus F_{st} is given first, as shown in the example:

```
Fst comparison between: TC-1.testCosi.1 TC-2.testCosi.1
np = not polymorfic, fixed position

SAMPLE 1
Position 1        Fst value: 0.044446
Position 2        Fst value: 0.000000
Position 3        Fst value: 0.030770
Position 4        Fst value: 0.249997
Position 5        Fst value: 0.117649
Position 6        Fst value: 0.000000
Position 7        Fst value: 0.400003
Position 8        Fst value: np
Position 9        Fst value: 0.142855
Position 10       Fst value: 0.025975
Average Fst: 0.112411    Max Fst: 0.400003       Min Fst: 0.000000
```

MAF and DAF

If MAF and DAF are computed, for each sample information on every locus is displayed sequentially in three lines corresponding to the SNP number, the MAF, and the DAF:

```
SAMPLE 1
Snp 1
        Maf: 0.1111111111
        Daf: 0.1111111111
Snp 2
        Maf: 0.2222222222
        Daf: 0.2222222222
Snp 3
        Maf: 0.1111111111
        Daf: 0.1111111111
Snp 4
```

```
        Maf: 0E-10
        Daf: 0E-10
Snp 5
        Maf: 0.4444444444
        Daf: 0.5555555556
```

Neutrality statistics

In the case of neutrality statistics, for each sample the outfile starts with a line containing some statistics describing the variability of the sequences. Next the neutrality statistics appear, classified according to whether they belong to Class I (based on the mutation spectrum of frequencies) or to Class II (based on haplotypes):

```
SAMPLE 1
Sequences: 9    Seg. sites: 4    Pi: 1.388889    Singletons: 2
Class I Statistics
Tajima's D: -0.228839
Fu and Li D*: -0.264179    Fu and Li F*: -0.284088
Fu and Li D: -0.467128     Fu and Li F: -0.483865
R2: 0.157288
Fay and Wu H: 0.777778
Class II Statistics
Fu's Fs: -1.686055
EHH average: 8.000000    EHH maximum: 8.000000
Dh: 0.805556
Wall's B: 0E-8    Wall's Q: 0E-8
ZnS: 0.116805     Za: 0.075890     ZZ: -0.040914
```

Tabulated data

The tabulated data format shows as many columns as satistics plus one first colum with the sample number, separated by tabulators. The first line is a header stating what each column is. Examples of this format for each statistic are shown below.

F_{ST}

The tabulated output file for F_{st}, as in the previous format, has three different appearances depending on the calculation chosen and contains the two lines showing the populations that are being compared (first) and the legend. If both SNP and gene information are displayed, the outfile will look as follows, with the average, maximum and minimum F_{st} added in three columns next to the last SNP in the sample:

```
Fst comparison between: TC-1.testCosi.1 TC-2.testCosi.1
np = not polymorfic, fixed position

sample  snp     fst         average_fst     max_fst min_fst
1       1       0.044446
1       2       0.000000
1       3       0.030770
1       4       0.249997
1       5       0.117649
1       6       0.000000
1       7       0.400003
1       8       np
1       9       0.142855
1       10      0.025975        0.112411        0.400003        0.000000
```

If information on SNPs is required exclusively only the first three columns are shown. Instead, if the information asked is F_{st} per gene, the "snp" and "fst" columns are not displayed.

MAF and DAF

MAF and DAF output tabulated format has four columns which, from left to right, correspond to the sample and locus number, MAF and DAF.

```
Sample  SNP     maf           daf
1       1       0.3333333333  0.3333333333
1       2       0.3333333333  0.3333333333
1       3       0.4444444444  0.4444444444
1       4       0.3333333333  0.3333333333
1       5       0.3333333333  0.6666666667
```

Neutrality statistics

The neutrality statistics outfile has 21 columns, the first for the sample and the next for the descriptors and the statistics, in the same order as in the information per sample format:

```
Sample  seq       segsites pi        singl     Ts_D      FL_D2     FL_F2     FL_D      FL_F      R2        FW_H
        Fs        EHH_a     EHH_m    Dh        W_B       W_Q       ZnS       Za        ZZ
1       9         4         1. 388889           2         -0. 228839          -0.264179           -0.284088
        -0.467128           -0. 483865          0. 157288           0.777778            -1.686055
        8.000000            8.000000            0. 805556           0E-8      0E-8      0.116805
        0.075890            -0.040914
```

OUTFILE NAMES

With the exception of F_{st}, the name of the output is formed based on the name of the infile. The outfile name, then, is the infile without extension followed by a dash and an abbreviation for the calculation done. Its extension depends on the predefined outfile format selected.

In the case of F_{st}, the outfile name without extension must be provided by the user. FABSIM will use this name to code it as explained.

The codes for the calculations and the formats are:

CALCULATIONS	
_fst	F_{st}
_maf	MAF and DAF
_stats	neutrality statistics
FORMATS	
.smp	information per sample
.tab	tabulated data

Examples

The output file obtained from calculating neutrality statistics on a file named simulations.inp, specifying the information per sample format, would be named simulations_stats.smp.

If the user wants to calculate F_{st} and MAF and DAF on a set of simulations from two populations which are in the files population1.out and population2.out, and obtain the results in a tabulated format, it first must provide a name for the F_{st} output file. Let's say the given name is *populations*. FABSIM will then output three outfiles:

populations_fst.tab
population1_maf.tab
population2_maf.tab

REFERENCES

Fay, J. C. and Wu, C. I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405-13.

Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915-25.

Fu, Y. X. and Li, W. H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693-709.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337-8.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. *Genetics* **146:** 1197-206.

Nei, M. 1987. *Molecular evolutionary genetics.* Columbia University Press, New York.

Ramirez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F., and Navarro, A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179:** 555-567.

Ramos-Onsins, S. E. and Rozas, J. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19:** 2092-100.

Rozas, J., Gullaud, M., Blandin, G., and Aguade, M. 2001. DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics* **158:** 1147-55.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832-7.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15:** 1576-1583.

Spencer, C. C. A. and Coop, G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20:** 3673-3675.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-95.

Wall, J. D. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74:** 65-79.