FSTAT (version 2.9.4), a program (for Windows 95 and above) to estimate and test population genetics parameters.

Jérôme Goudet
Department of Ecology & Evolution,
BB, Lausanne University,
CH-1015 Dorigny
Switzerland.
E-mail: jerome.goudet@ie-zea.unil.ch
`http://www.unil.ch/izea/softwares/fstat.html` [1]

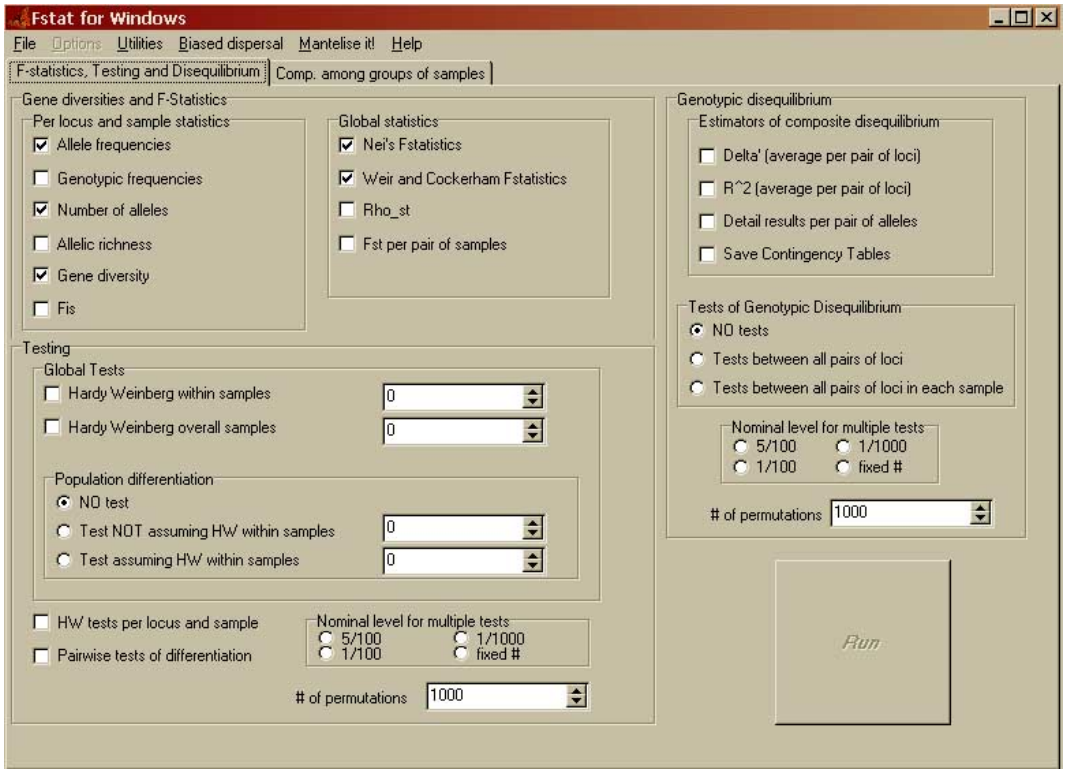
21st November 2003

[1]This software could not have been developed without the financial support of the Swiss National Science Foundation, Grants N° 31-43443.95, 31-55945.98 and 31- .2000. I am indebted to the many users who sent me comments, bugs or cheerful messages that helped improve FSTAT: François Balloux, Lars Berg, Michel Chapuisat, Thierry De Meeüs, Loïc Degen, Greg Douhan, Guillaume Evanno, Arnaud Estoup, Laurent Excoffier, Pierre Fontanillas, Luca Fumagalli, Barbara Giles, Chris Gliddon, Alexandre Hirzel, Michael Krawczak, Martin Lascoux, Lance Barrett-Lennard, Margaret Mackinnon, Patrick Meirmans, Eric Petit, Alan Raybould, Michel Raymond, Max Reuter, François Rousset, Sandrine Trouvé and many others, whose name slipped my mind.

Contents

Chapter 1

Introduction



For evolution to occur, there must be polymorphism on which selection can act. The amount of polymorphism at a locus results from the interaction of the following forces: selection, genetic drift, mutation and migration. By screening genetic markers on their favorite species, population biologists attempt to estimate the strength of each of these forces. The quantities that they are likely to find useful are the allele frequencies and the genetic diversity of the markers, the proportion of each genotype and their expected proportion under the Hardy Weinberg null model. Usually, they will have samples from many locations and might want to know whether these locations differ in allele frequency.

FSTAT is a user-friendly program that estimates and tests all these statistics. Its main strengths compared to other softwares are its ease of use and its exhaustiveness.

I've paid particular attention to the writing of the help file (this document), which will hopefully provide guidelines to a newcomer in the field of population structure analysis.

Do not hesitate to email me if you have suggestions for improvements, discover bugs, or simply think FSTAT is great!!!

Hardware requirements

A PC running Windows 95 or above. It might work with windows 3.1 and wins32, but I have not tried it... I guess a Pentium processor would be a necessity if randomisations are to be carried out. You should also be able to run it on a (recent and powerful) Macintosh with PC emulation software.

Cite as

Goudet, J. 2003. FSTAT (ver. 2.9.4), a program to estimate and test population genetics parameters. Available from `http://www.unil.ch/izea/softwares/fstat.html` Updated from Goudet [1995]

1.1 What the program FSTAT does

From a data set of codominant genetic markers (see input file format), FSTAT calculates the following:

- Number of individuals per sample and loci.
- Allele frequency estimated per sample and overall.
- Observed and expected number of each genotype per sample and locus.
- Unbiased gene diversity per sample and locus.
- Number of alleles sampled per locus and sample, as well as overall.

FSTAT Allelic richness per locus and sample, as well as overall samples.

- F_{IS} per locus and sample, as well as a test of whether it is significantly positive or negative (significant deficit and excess of heterozygotes respectively).
- Nei [1987] estimators of gene diversities and differentiation.
- Weir and Cockerham [1984] F (F_{IT}), θ (F_{ST}) and f (F_{IS}) estimated per allele, per locus and overall. FSTAT also calculates Hamilton [1971] relatedness [relat=$2F_{ST}/(1+F_{IT})$], calculated using an estimator strictly equivalent to Queller and Goodnight [1989]. This measure is the average relatedness of individuals within samples when compared to the whole data set.
- Confidence intervals based on resampling schemes are provided for Weir and Cockerham [1984] statistics:
  - Jackknifing per locus over samples is performed. The resampling unit are the different samples. This procedure is only carried out if there are more than 4 samples.
  - Jackknifing over loci. In this case, the resampling units are the different loci. This procedure is only carried out if there are more than 4 loci.
  - Bootstrapping over loci. Bootstrapping over loci is performed only when there is more than 4 loci in the data set.
- Estimation of R-statistics (Slatkin [1995]), specifically designed for microsatellite undergoing stepwise mutations.

- Estimation of F_{ST} (θ) per pair of samples.
- Overall test whether samples at each locus are in HW equilibrium.
- Test whether the entire data set is in HW equilibrium.
- Test whether samples are differentiated assuming either that there is HW within samples or that there is not (Only one of these 2 tests can be carried out).
- Test whether each sample at each locus is in Hardy-Weinberg (HW) equilibrium.
- Test whether each pair of samples is differentiated. The tests do not assume random mating within samples. A table of significant pairwise differentiation after corrections for multiple testing is produced.

NEW, FSTAT Estimates two multi allelic coefficients of composite disequilibrium, R and Δ'. These coefficients are inspired from Weir [1979, 1996] and Zapata [2000].

- Test whether each pair of loci in each sample and overall is at genotypic equilibrium

FSTAT Test whether groups of samples differ for a large panel of statistics

FSTAT Test whether categories of individuals differ in dispersal rates (four different tests)

- Convert FSTAT format to GENEPOP and vice-versa

FSTAT Performs multiple regression or Partial Mantel tests

1.1.1 What's new in version 2.9.4?

- made the help file a pdf document, printable and searchable.
- added the estimation of 2 coefficients of composite disequilibrium (R and Δ').
- added the possibility of specifying a number of permutations for testing genotypic disequilibrium
- added the possibility of specifying a number of permutations for testing pairwise differentiation and HW in each sample.
- fixed a bug in the testing procedure of the partial mantel test.

todo generalized the bootstrap over loci as a way to obtain confidence intervals for all the statistics.

todo added estimates of two genetic distances, Nei's unbiased distance and Cavalli-Sforza & Edwards Chord distance Nei [1987].

todo Hierarchical F-statistics

todo relatedness between pairs of individuals (Wang's method (Wang [2002])).

todo Convert FSTAT to Nexus format and vice-versa

todo Generates neighbor joining (NJ) trees of the samples, in the newick format (compatible with e.g. TREEVIEW (Page [1996])[`http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`]).

1.2 What the program FSTAT does NOT do

- Write your population genetics papers!
- FSTAT is not suited for dominant data, obtained from e.g. RAPDs or AFLPs. If you have such data, please have a look at HICKORY(Holsinger et al. [2002])[`http://darwin.eeb.uconn.edu/hickory/hickory.html`]
- FSTAT is not suited for sequence data. One of the suitable softwares for this is ARLEQUIN(Schneider et al. [2000])[`http://lgb.unige.ch/arlequin/`].
- FSTAT will not assign individuals to populations. A very good software for this is the STRUCTURE software(Pritchard et al. [2000])[`http://pritch.bsd.uchicago.edu/`].
- check GENEPOP (Raymond and Rousset [1995b]) [`ftp://ftp.cefe.cnrs-mop.fr/genepop/`] for testing for isolation by distance among individuals.
- The following web sites contain links to programs of interest for population genetics analysis:
  - Department of Biological sciences at Louisiana state university `http://www.biology.lsu.edu/general/software.html`
  - Joe Felsenstein list at `http://evolution.genetics.washington.edu/phylip/software.html`

Chapter 2

Running the program

The program is run by double clicking on the icon (a red butterfly) named Fstat294.exe.

The first time you run the program, you will be prompted for 2 numbers, which are the seeds of the Random Number Generator. Any 2 numbers will do. The chosen generator is described in L'Ecuyer [1988]. It is a combination of 2 of the best Multiple Linear Congruential Generator known, and has been thoroughly checked (Goudet [1993]). In future run, the seeds will be loaded from the file FSTAT.INI, created when you exit the program and updated after each run. If for one reason or another, you want to use a set of specific seeds, just edit the file FSTAT.INI or use the `Utility, Reset seeds random number generator` menu.

2.1 Creating an input file

For running FSTAT, it is first necessary to create an input file named `FILENAME.DAT` (where `FILENAME` is anything between 1 and 256 characters) containing the genotypic data, coded numerically, either with a 1, a 2 or a 3-digit number per allele. The file must have the following format:

The first line contains 4 numbers: the number of samples, $np(\leq 200)$, the number of loci, $nl(\leq 600)$, the highest number used to label an allele, $nu(\leq 999)$, and a 1 if the code for alleles is a one digit number (1-9), a 2 if code for alleles is a 2 digit number (01-99) or a 3 if code for alleles is a 3 digit number (001-999). These 4 numbers need to be separated by any number of spaces. There is a further constraint on the total number of individuals in the data set, which needs to be less than 200'000.

The first line is immediately followed by nl lines, each containing the name of a locus, in the order they will appear in the rest of the file.

On line $nl + 2$, a series of numbers as follow:

```
1      0102   0103   0101  0203          0       0303
```

The first number identifies the sample to which the individual belongs, the second is the genotype of the individual at the first locus, coded with a 2 digits number for each allele, the third is the genotype at the second locus, until locus nl is entered (in the example above, $nl = 6$). Missing genotypes are encoded with 0. Note that 0001 or 0100 are not a valid format, that is, both alleles at a locus have to be known, otherwise, the genotype is considered as missing. No empty lines are needed between samples. The number of spaces between genotypes can be anything. Numbering of samples need not be sequential, however, the number of samples np needs to be the same as the largest sample identifier. Samples need not to be ordered. nu needs to be equal to the largest code given to an allele , even if

there are less than nu alleles (e.g. if only alleles 641 and 657 are present, $nu = 657$). An example of 2 valid input files is given in Appendix A.

If you have a file with the GENEPOP format, FSTAT now translate this format into its own. See the `File Conversion menu` under `Utilities`

2.2 LIMITS IN FSTAT

- Maximum number of samples : 200
- Maximum number of loci : 600
- Maximum number of alleles : 999
- Maximum number of individuals : 200'000

It might be difficult memory-wise to run a data set with all these parameters to the maximum, but I guess it would also mean quite some time reading gels!

If you need more capacity, please contact me : jerome.goudet@ie-zea.unil.ch.

Chapter 3

File menu

This is the only accessible menu when you start FSTAT.

3.1 Open submenu

This is where to start! A window will open displaying all the files in the FSTAT directory with the default extension `.DAT`. If the data file is not in this folder, just navigate using the "look in" option of the window until you have located it. Once you have it, double click the file, or click on open.

If many files with the same structure (that is, the same number of samples, loci and individuals) need to be processed, they can all be selected at once[1], and FSTAT will process them with the default options sequentially. This is useful if you want to test a scenario via simulations, using software such as EASYPOP (Balloux [2001]`http://www.unil.ch/izea/softwares/easypop.html`).

Input files with an extension different from `.DAT` will not be opened.

3.2 Save Default Options subMenu

This menu allows you to start FSTAT with your preferred options already selected. You then just need to open a file, and click run!

3.3 Exit subMenu

Obvious, isn't it?!

[1] by clicking on the desired files while keeping the Ctrl key pressed

Chapter 4

Choose options menu

This menu is greyed as long as you have not chosen a data file (menu file, open).

Under this menu, you will select what you want FSTAT to do. Some of the options are selected by default (those with a tick mark). To select an unselected option, or to deselect one, just click on it. Tick marks will then either appears or disappears.

4.1 Label file for pops sub-menu

A text file containing names (labels) for the populations (samples) can be given under this option. The default extension for label files is `.LAB`. Each line should contain the name (label) of one sample, and samples should appear in the same order as in the `.DAT` file. Labels can be any length but will be truncated to six characters in the output files.

4.2 Loci to use sub-menu

It is sometimes useful to focus on only a subset of the collected data. For example, if the study involves different classes of genetic markers such as microsatellites and allozymes, it might be useful to analyse separately each class of markers. Another potential application is if you identify an ill-behaved locus. This sub-menu will let you pick the loci you want to focus on.

After a subset of the loci have been selected, when the `run button` is clicked, you will be prompted for a file name under which to store the reduced data-set.

If you choose the default filename, it will be the original name, followed by -L (for loci), followed by the numbers identifying the selected loci.

The chosen filename will be used for the input data as well as for all the output files, with the appropriate extensions.

4.3 Samples to use sub-menu

It is sometimes useful to focus on only a subset of the collected data. For example, if the study involves samples from different habitats, or regions, one might be interested in focusing on one of these habitats. This sub-menu will let you pick the samples you would like to focus on.

After a subset of the samples have been selected, when the `run button` is clicked, you will be prompted for a file name under which to store the reduced data-set.

If you choose the default filename, it will be the original name, followed by -P (for populations), followed by the numbers identifying the selected samples.

The chosen filename will be used for the input data as well as for all the output files, with the appropriate extensions.

The samples in the new file will appear with the order they had in the original `.dat` file. That is, if you select (in this order) samples 5, 6 2 & 1, the new `.dat` file will contains (in this order) samples 1, 2, 5 & 6. If a label file was associated with the original .dat file, a new label file will be created, with labels corresponding to the sampled populations (and in the right order[1]).

[1] if you would like samples to be identified by a label, you have to first select the file containing the labels, an only then to pick the samples. Proceeding the other way around (unless you have already created a label file with the subset of samples you are interested in) will not associate correctly samples with their respective labels.

Chapter 5

Per locus and sample statistics

5.1 Allele frequencies

For each locus in each sample, will estimate the number of individual typed. The allele frequencies in each sample and overall will then be estimated. For the overall allele frequencies, both the weighted (by sample size) and non-weighted frequencies are reported.

5.2 Genotypic frequencies

If you cheked this box, a file (extension `.X2`, see results files) containing the observed and expected genotypic number for each genotype at each locus in each sample is created. The expected numbers are unbiased estimates based on the hyper geometric distribution:

$$\begin{aligned} A_iA_i &= n\left[p_i\tfrac{(2np_i-1)}{(2n-1)}\right], \\ A_iA_j &= 2n\left[p_i\tfrac{(2np_j)}{(2n-1)}\right] \end{aligned}$$

where A_iA_i represents homozygote genotypes, A_iA_j represents heterozygotes, n the sample size, p_i and p_j the allele frequencies of allele A_i and A_j respectively.

5.3 Number of alleles

Counts the number of different alleles observed at each locus in each sample, and overall samples.

5.4 Allelic Richness

Estimates allelic richness per locus and sample (R_s), and overall samples (R_t). Allelic richness is a measure of the number of alleles independent of sample size, hence allowing to compare this quantity between different sample sizes.

The observed number of alleles in a sample is highly dependant on sample size. To bypass this problem, El Mousadik and Petit [1996] suggested to adapt the rarefaction index of Hurlbert [1971] to population genetics (see also Petit et al. [1998]).

The principle is to estimate the expected number of alleles in a sub-sample of $2n$ genes, given that $2N$ genes have been sampled ($N \leq n$). In FSTAT, n is fixed as the smallest number of individuals typed for a locus in a sample. Allelic Richness is then calculated as:

$$R_s = \sum \left[1 - \frac{\binom{2N-N_i}{2n}}{\binom{2N}{2n}} \right]$$

where N_i is the number of alleles of type i among the $2N$ genes. Note that each term under the sum corresponds to the probability of sampling allele i at least once in a sample of size $2n$. If allele i is so common that we are certain to sample it -when $2n > (2N - N_i)$- the ratio is undefined but the probability of sampling the allele is set to 1.

For R_t, the same sub-sample size n is kept, but N is now the overall samples number of individuals genotyped at the locus under consideration. R_t is reported in the last column.

Differences in allelic richness among groups of populations could be tested, see tab-sheet `Comp. Among groups of samples`.

5.5 Gene diversities

Estimates gene diversity per locus and sample using the unbiased estimator

$$H_S^k = \frac{n_k}{n_k - 1} \left(1 - \sum_i p_{ki}^2 - H_O^k / 2n_k \right)$$

(see [Nei, 1987, eq. 7.39 p. 164]).

5.6 F_{is}

Estimates F_{IS} for each locus and sample as $F_{IS}^k = 1 - \frac{H_O^k}{H_S^k}$. Here, the generic name F_{IS} is given, rather than G_{IS} or f (see below). This is because the statistic is estimated for each sample, in which case the difference between G_{IS} and f (due to the different weightings of varying sample size, see below) vanishes.

Chapter 6

Global statistics

F-statistics are a set of tools devised by Wright [1921, 1969] to partition heterozygote deficiency into a within and an among population component. They are widely used by population biologists to assess levels of structuring in samples of natural populations. F_{IS} measures the heterozygote deficit within populations, F_{ST} among populations (a measure of the Wahlund effect), and F_{IT} the global deficit of heterozygotes.

Estimation of F-statistics has been debated in the literature for the past 30 years, since the early work of Cockerham [1969, 1973] and Nei [1973, 1975]. Advantages and problems of either estimator have been discussed at length (Slatkin and Barton [1989], Chakraborty and Danker-Hopfe [1991], Cockerham and Weir [1993]). A good review of these and other estimators can be found in Excoffier [2001].

FSTAT calculates Nei and Weir & Cockerham estimators of gene diversities and differentiation.

6.1 Nei's estimators of F-statistics

The following statistics, defined in [Nei, 1987, eq. 7.38– 7.43 pp. 164–5] are estimated:

- the observed heterozygosity

$$H_O = 1 - \sum_k \sum_i P_{kii}/np,$$

  where P_{kii} represents the proportion of homozygote i in sample k and np the number of samples.

- the within population gene diversity[1]:

$$H_S = \frac{\tilde{n}}{(\tilde{n}-1)} \left[1 - \sum_i \overline{p_i^2} - \frac{H_0}{2\tilde{n}} \right],$$

  where $\tilde{n} = \frac{np}{\sum_k 1/n_k}$ and $\overline{p_i^2} = \sum_k p_{ki}^2/np$

- the overall gene diversity

$$H_T = 1 - \sum_i \overline{p_i}^2 + \frac{H_S}{(\tilde{n}np)} - \frac{H_0}{(2\tilde{n}np)},$$

  where $\overline{p_i} = \sum_k p_{ki}/np$.

[1] sometimes misleadingly called expected heterozygosity

- the amount of gene diversity among samples $D_{ST} = H_T - H_S$
- $D'_{ST} = \frac{np}{np-1} D_{ST}$
- $H'_T = H_S + D'_{ST}$
- $F_{ST} = \frac{D_{ST}}{H_T}$.[2]
- $F'_{ST} = \frac{D'_{ST}}{H_T}$
- $F_{IS} = 1 - \frac{H_O}{H_S}$

Here, the p_{ki} are unweighted by sample size. These statistics are estimated for each locus and an overall loci estimates is also given, as the unweighted average of the per locus estimates. In this way, monomorphic loci are accounted for (with estimated value of 0) in the overall estimates.

Note that the equations used here all rely on genotypic rather than allelic number and are corrected for heterozygosity (see Nei and Chesser [1983]). A thorough description of many estimators of gene diversity and differentiation is available in the excellent review of Laurent Excoffier (Excoffier [2001])

6.2 Weir & Cockerham estimators of F-statistics

Weir and Cockerham [1984] estimator of F_{IT}, F_{ST} and F_{IS} ($F = \frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_w^2}$, $\theta = \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_w^2}$ and $f = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}$ respectively in FSTAT output) are estimated for each allele, locus and overall. For each locus and overall, FSTAT will also output the different variance components from which F-statistics are estimated, namely the among sample variance component:

$$\sigma_a^2 = \frac{\frac{n_k(p_k-\bar{p})^2}{np-1} - \frac{\sum_k n_k\bar{p}(1-\bar{p}) - \sum_k n_k(p_k-\bar{p})^2 - (1/4)\sum_k n_k H_{Ok}}{\sum_k n_k - np}}{\frac{1}{np-1}\left(\sum_k n_k - \frac{\sum_k n_k^2}{\sum_k n_k}\right)}$$

where ($\bar{p_i} = \frac{\sum_k n_k p_{ki}}{\sum_k n_k}$), the between individual within sample component

$$\sigma_b^2 = \frac{\sum_k n_k\bar{p}(1-\bar{p}) - \sum_k n_k(p_k-\bar{p})^2 - (1/4)\sum_k n_k H_{Ok}}{\sum_k n_k - np} - \frac{\sum_k n_k H_{Ok}}{4\sum_k n_k}$$

and the within individual component

$$\sigma_w^2 = \frac{\sum_k n_k H_{Ok}}{2\sum_k n_k}.$$

For more details about these different quantities, refer to the original paper by Weir and Cockerham [1984] as well as Weir [1996].

If all the samples have the same size, then σ_w^2 is the same as H_O, ($\sigma_w^2+\sigma_b^2$) as H_S, ($\sigma_w^2+\sigma_b^2+\sigma_a^2$) as H'_T, θ as F'_{ST} and f as F_{IS}. This is no longer true if sample sizes are different. Weir & Cockerham (WC) weight allele frequencies according to sample sizes (as would be done in an ANalysis Of VAriance), whereas Nei weights all samples equally, whatever the sample size is. When sample sizes vary a lot, this can lead to large differences between the two families of estimators.

[2] This is not the same as Nei's G_{ST}. Nei's G_{ST} is an estimator of F_{ST} based on allele frequencies only

Nei and WC also treat differently monomorphic loci: Under Nei's family, when a locus is completely monomorphic, F_{IS} and F_{ST} (and F'_{ST}) are all 0, while WC consider that the estimators cannot be defined[3].

A measure of Hamilton [1971] relatedness is also included, calculated using an estimator strictly equivalent to Queller and Goodnight [1989] (see in particular appendix B of this paper), namely: $R = 2\theta/(1+F)$. This measure is the average relatedness of individuals within samples when compared to the whole.

Pamilo [1984, 1985] pointed out that when there is inbreeding, relatedness is biased and proposed an "inbreeding corrected relatedness", labelled relatc in FSTAT, and estimated as: $R_c = [R - 2F/(1+F)]/[1 - 2F/(1+F)]$ One should be wary however, that this estimator is not bounded between 0 and 1. In particular when working with species undergoing partial selfing, relatc seems to produce invalid results. On the other hand, when the population is structured, relatc adequately removes the increase in relatedness due to this structuring (see Chapuisat et al. [1997] for an example)

Differences among groups for H_O (σ^2_w), H_S (σ^2_b+σ^2_w), f, θ, relat and relatc could all be tested, see tab-sheet Comp. Among groups of samples.

6.2.1 Jackknife variance and Bootstrap confidence interval

Jackknifing over samples and loci and bootstrapping over loci are automatically performed for the statistics F, θ, f and relat if the number of samples and the number of loci are sufficient (more than 4). See Raymond and Rousset [1995a] for a discussion of some problems encountered with bootstrapping (their argument is buttressed on the data set `diploid.dat` given in the appendix) when there is not a sufficient number of loci, and why randomisations are better suited.

6.3 R_{ST}

R_{ST} is an estimator of gene differentiation accounting for variance in allele size and defined for genetic markers undergoing a stepwise mutation model (Slatkin [1995]). It is not worth bothering about unless you are pretty sure that mutation follows a stepwise mutation model quite strictly, and that mutation can not be neglected compared to others forces such as migration (Balloux et al. [2000], Balloux and Goudet [2002]). FSTAT estimates R_{ST} for each locus following Rousset [1996]. This estimator does not depend on the number of samples. It also outputs the different components of variance of allele size: V_a for among samples, V_b for among individuals within samples and V_w for within individuals. $V_t = V_a + V_b + V_w$ is thus an unbiased estimate of the total overall variance in allele size.

Three estimators of overall loci R_{ST} are provided:

1. one according to Rousset [1996], where each locus is weighted by its amount of allelic variance.

2. Because variances in allele size commonly vary by orders of magnitude among loci, FSTAT also gives the estimator of Goodman [1997]. Calling M_o and V_o the mean and variance in allele size at a locus, Goodman [1997] suggested to use the centred normalised allele size $(x - M_o)/\sqrt{V_o}$ instead of allele size x in the estimation of R_{ST}. Individual loci R_{ST} will not be affected by this centering-rescaling, but the overall loci estimator will, because each individual

[3] it is sometimes found in the literature that F_{ST} or F'_{ST} cannot be negative. While this is true for the statistic G_{ST} because this statistic does not include a correction for heterozygosity (Nei [1973]), it is wrong for subsequent versions accounting for heterozygosity. For details on this issue and more about the comparison between Nei and WC estimators, Cockerham and Weir [1993]

locus variance component will be divided by its locus allelic size variance (since $s^2(aX+b) = a^2 s^2(X)$). Goodman's R_{ST} will thus be expressed as:

$$UR_{ST} = \frac{\sum_{i=1}^{l} V_a^i / V_o^i}{\sum_{i=1}^{l} V_t^i / V_o^i}$$

where l represents the number of Loci. V_t is an unbiased estimator of V_o, so $V_t \approx V_o$, and therefore $\sum_{i=1}^{l} V_t^i / V_o^i \approx l$ (In fact, one could argue that V_t should be used for the total variance instead of V_o). This leaves us with :

$$UR_{ST} \approx 1/l \sum_{i=1}^{l} \frac{V_a^i}{V_t^i} = R_{STU},$$

the average of the individual Loci R_{ST}.

3. This last un-weighted estimate (R_{STU}) is also produced by FSTAT for completeness.

For a review of the properties of these 3 estimators compared to F_{ST}, and some advice on when to use them, see Balloux and Goudet [2002].

6.4 F_{ST} per pair

Calculates the multilocus Weir and Cockerham [1984] estimator of F_{ST} (θ) between all pairs of samples. Two output files are produced. One with the extension `.FST`, containing the table of pairwise θ. The other with the extension `.MAT`, containing 2 half tables. The first one gives, for each pair of samples, the sum of the three variance components (σ_a^2, σ_b^2 and σ_w^2). The second half table gives σ_a^2.

Since F_{ST} is closely related to a genetic distance (Reynolds et al. [1983]), the first table can be used in Mantel tests (e.g., Manly [1985]). Rousset [1997] has elucidated the appropriate transforms to be used on F_{ST} and geographic distances if one wants to test for a linear association between these two quantities[4].

[4] in the old days (v 1.2), FSTAT produced a file containing the so-called "Nm" values. These values were simply a function of pairwise F_{ST} , namely 1/(4 F_{ST}) - 1/4. In many cases however, the assumptions necessary to transform F_{ST} into a number of migrants are not fulfilled (Whitlock and McCauley [1999]). To avoid misuse, FSTAT does not produce these values anymore

Chapter 7

Testing

7.1 Introduction

All the tests in FSTAT are randomisation based. The principle behind randomisation based tests is the following: Data sets fitting the null hypothesis to be tested are generated by randomising the appropriate units (alleles, genotypes... see below). A statistic is calculated on these randomised data sets and its value compared to the statistic obtained from the observed data set. The proportion of statistics from randomised data sets that give a value as large or larger than the observed provides an unbiased estimation of the probability that the null hypothesis is true (P-value of the test). These tests are carried out for each locus individually and then over all loci. Bonferroni procedures should be applied when using individual population tests or individual locus tests (Rice [1989]).

Units of randomisation are NOT the same for the different tests:

For HW within samples

Alleles are permuted among individuals, within samples. Then a statistic has to be chosen to compare the randomised data sets to the observed. In FSTAT, this statistic is F_{IS} (f).

For HW overall

Alleles are permuted among samples. The statistic used to compare the randomised data sets to the observed is F_{IT} (F)

For population differentiation

Two types of tests can be carried out: If there is strong evidence that there is HW within samples, then it is valid to permute alleles among samples to test for population differentiation, since alleles can be considered as independent. On the other hand, if HW is rejected within samples, alleles within an individual are not independent anymore. In this case, permuting genotypes among samples is the only valid permutation scheme[1].

In either case, contingency tables of alleles by samples are generated, and the statistic that FSTAT uses to classify them is the log-likelihood statistic G (Goudet et al. [1996]):

[1]Only complete multilocus genotypes are randomised, to make sure that the combined test over loci is valid. If you'd like to use all possible single locus genotype to estimate the probability of differentiation for individual loci, you'll need to create separate file for each locus (using the options, loci to use sub-menu)

$$G = -2\sum_{k=1}^{np}\sum_{i=1}^{nu} n_{ik}\log\left(\frac{n_{ik}}{n_k\overline{p_i}}\right),$$

where n_{ik} is the number of allele i in sample k, n_k is the number of alleles (twice the number of individuals) in sample k, and $\overline{p_i}$ is the frequency of allele i in the whole dataset for the overall test, and in the two samples considered for the pairwise test (with $np = 2$ in this case). To test for population differentiation among all loci, the following statistic is used to classify contingency tables:

$$G = -2\sum_{l=1}^{nl}\sum_{k=1}^{np}\sum_{i=1}^{nu} n_{ikl}\log\left(\frac{n_{ikl}}{n_{kl}\overline{p_{il}}}\right),$$

where l indexes loci.

Two values are given for the P-values in the above options. These values correspond to the proportion of randomised data that gave: (i) a value larger or equal than the observed, (ii) a value strictly larger than the observed. These two values are in general very close one to the other. But sometimes they are quite different, and there are cases where they can cover the whole interval 0:1. These cases correspond to loci with very little polymorphism, therefore bringing little information about the population structure or the mating system. In any case, the P-value that should be reported is the first one.

Test overall loci

The tests presented above are all designed for one locus. But how can one combine the results of these tests to obtain a P-value overall loci? One option is to use Fisher's procedure to combine probabilities (Fisher [1954]) or more appropriately a test based on the symmetry of the distribution of the P-values obtained from the different loci (see Goudet [1999]). But both these tests give the same weights to the different loci. One could wonder whether it is justified to give the same weight to a bi-allelic, poorly polymorphic locus and to one highly polymorphic with 10 alleles. An alternative is to construct a statistic accounting for all the loci, such as the sum of the statistics obtained from individual loci, or a function of this sum, like the mean. This is what FSTAT does. The tests are carried out either on the sum or on a weighted average of the statistic obtained from each locus (the weighting being the same as the one used to estimate the overall loci statistic). In this way, loci with a higher polymorphism receive more weights. So for the tests based on f, the statistic used is the overall loci f, namely:

$$f = \frac{\sum^{nl}\sum^{nu}\sigma_b^2}{\sum^{nl}\sum^{nu}(\sigma_b^2+\sigma_w^2)},$$

where the first sum is over loci and the second over alleles. For the test based on F, it is:

$$f = \frac{\sum^{nl}\sum^{nu}\left(\sigma_a^2+\sigma_b^2\right)}{\sum^{nl}\sum^{nu}(\sigma_a^2+\sigma_b^2+\sigma_w^2)},$$

and for the tests of differentiation based on the likelihood ratio statistic, FSTAT uses as an overall test statistic the sum of individual loci G-values

$$G = -2\sum_{l=1}^{nl}\sum_{k=1}^{np}\sum_{i=1}^{nu} n_{ikl}\log\left(\frac{n_{ikl}}{n_{kl}\overline{p_{il}}}\right),$$

A brief description of the performance of this last test is given in Petit et al. [2001].

7.2 Global tests

7.2.1 Hardy Weinberg Within samples

Alleles are randomised among individuals, within samples. f is used to classify contingency tables of alleles against alleles in each sample.

7.2.2 Hardy Weinberg Overall samples

Alleles are randomised over the whole dataset. F is used to classify contingency tables of alleles against alleles in the whole dataset.

7.2.3 Population Differentiation NOT assuming Hardy Weinberg within

Rather than alleles, genotypes are randomised among samples. Contingency tables of alleles against samples are obtained from the randomised data sets, and the log-likelihood statistics G is used to classify them.

My advice is NOT to assume HW within samples for testing population differentiation, unless you are certain that HW is realised within these samples.

7.2.4 Population Differentiation assuming Hardy Weinberg within

Alleles are randomised over the whole data set. Contingency tables of alleles against samples are obtained from the randomised data sets, and the log-likelihood statistics G is used to classify them.

7.3 Tests per sample or pair of samples

7.3.1 HW tests per locus and sample

Alleles are randomised among individuals, within samples. F_{IS} is used to classify tables. This option will report 2 tables of P-values. The first table corresponds to tests for deficit in heterozygotes, the second for excesses of heterozygotes.

The nominal level for multiple tests box allows you to define table wide levels of significance at 5%, 1% or 0.1%, or to specify a given number of permutations.

7.3.2 Pairwise tests of differentiation

For each pair of samples, multi-loci genotypes are randomised between the two samples. The overall loci G-statistic is used to classify tables[2]. The nominal level for multiple tests box allows you to define table wide levels of significance at 5%, 1% or 0.1%, or to specify a given number of permutations.

For pairwise tests of differentiation, The results are outputted to a file with the extension `-PP.PVL` (for Per Pair P-VaLues). This file contains two tables. The first

[2]This option does not produce individual loci pairwise P-values

gives the non-adjusted P-value for each pair. The second gives the pairwise significance after standard Bonferroni corrections[3]. "***" corresponds to significance at the 0.1% nominal level, "**" significance at the 1% nominal level and "*" significance at the 5% nominal level. "NS" stands for non-significant and "NA" for not available.

[3]the reported significance levels are after strict (not sequential) Bonferroni corrections based on the indicative adjusted P-value

Chapter 8

Composite disequilibrium

When two or more loci are present, it is of interest to verify whether alleles at the different loci assort independently. The classical measure of gametic disequilibrium is $D = p_{AB} - p_A p_B$ where p_{AB} represents the frequency of gamete carrying allele A at the first locus and B at the second, and p_A and p_B represents the frequency of alleles A and allele B respectively (Lewontin and Kojima [1960]). With genotypic data, estimation of gametic disequilibrium is impossible unless one is willing to assume Hardy-Weinberg at each locus (Weir [1979]) because gametic type in the double heterozygotes cannot be inferred. This is where the composite disequilibrium is useful.

8.1 Estimating composite disequilibrium

Δ_{AB} is the classical estimator of composite disequilibrium described by Weir [1979, 1996]. Given 2 polymorphic loci, alleles A and B at the first and second locus respectively, and calling $\bar{A}$ and $\bar{B}$ the non A and non B alleles, we have the following contingency table of counts for genotypes, with a total of n two locus genotypes:

	AA	$A\bar{A}$	$\bar{A}\bar{A}$
BB	n_1	n_2	n_3
$B\bar{B}$	n_4	n_5	n_6
$\bar{B}\bar{B}$	n_7	n_8	n_9

The double heterozygotes (n_5) can be produced by 2 types of pairs of gametes: $(AB)(\bar{A}\bar{B})$ and $(A\bar{B})(\bar{A}B)$. It is therefore not possible to estimate the proportion of gametes AB, but it is straightforward to obtain the count of (AB) gametes plus the count of "non gametic" (i.e one allele residing in each parental gamete) (A/B). This sum is simply $n_{AB} = 2n_1 + n_2 + n_4 + n_5/2$. If the 2 loci assort independently, the expected proportion of gametic and non-gametic $(AB), (A/B)$ is $2p_A p_B$. This leads to the following estimator of composite linkage disequilibrium:

$$\Delta_{AB} = \frac{n_{AB}}{n} - 2p_A p_B$$

An unbiased estimator of Δ_{AB} is $\widehat{\Delta_{AB}} = n/(n-1)\Delta_{AB}$ (Weir [1996]).

Δ_{AB}, like its gametic equivalent D_{AB} suffers from a major drawback: it is bounded between $[-2 \times \min(p_A p_B, p_{\bar{A}} p_{\bar{B}}), 2 \times \min(p_A p_{\bar{B}}, p_{\bar{A}} p_B)]$. In order to make the range of the composite genotypic disequilibria less dependent on allelic frequencies, the following statistic, similar to the D' measure of gametic disequilibrium (e.g. Hedrick [2000]), can be defined:

$$\begin{aligned}\Delta'_{AB} &= \frac{\Delta_{AB}}{2 \times \min(p_A p_B, p_{\bar{A}} p_{\bar{B}})}, \Delta_{AB} < 0 \\ \Delta'_{AB} &= \frac{\Delta_{AB}}{2 \times \min(p_A p_{\bar{B}}, p_{\bar{A}} p_B)}, \Delta_{AB} \geq 0\end{aligned}$$

Weir [1979, 1996] suggested another standardization, leading to a coefficient akin to a correlation coefficient, hence its name R_{AB}:

$$R_{AB} = \frac{\Delta_{AB}}{\sqrt{(p_A p_{\bar{A}} + D_A)(p_B p_{\bar{B}} + D_B)}}$$

where $D_A = \frac{n_{AA}}{n} - p_A^2$ and $D_B = \frac{n_{BB}}{n} - p_B^2$ represent excess homozygosity at locus A and B respectively. $(p_A p_{\bar{A}} + D_A)$ is the expression for the variance of the frequency of allele A given its frequency p_A and departure from HW D_A. Hence, the square root of he product of this quantity for alleles A and B is the product of the standard deviation, and since Δ_{AB} is the covariance of the frequency of allele A and B, R_{AB} is indeed akin to a correlation coefficient.

When the loci are multi allelic, the number of disequilibrium values per pair of loci is $n * m$, with n alleles at the fist locus and m at the second. It seems worthwhile to combine these into a single measure for the locus. A natural multi-allelic estimator consists in taking the weighted average of the $|\Delta'_{AB}|$, where the weight is the expected frequency of the gamete, $p_A p_B$ (really, it should be $2p_A p_B$, dividing the total by 2):

$$\Delta' = \sum_{i=1}^{n} \sum_{j=1}^{m} p_i p_j |\Delta'_{ij}|$$

The statistical properties of multiallelic Δ' have not been investigated, but those of the equivalent estimator of gametic disequilibrium D' have been thoroughly investigated by Zapata [2000], Zapata et al. [2001]. In particular, Zapata et al. [2001] showed that under many conditions, the distribution of multi-allelic D' do not deviate from normality.

An overall samples estimator of Δ' is also provided[1] as:

$$\Delta' = \frac{\sum_{k=1}^{np} n_k \Delta'_k}{\sum_{k=1}^{np} n_k}$$

I am unaware of multi-allelic estimators of R^2. Following the logic used to obtain one for Δ', I suggest the following:

$$R^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} p_i p_j R^2_{ij}$$

An overall samples estimator of R^2 is also provided:

$$R^2 = \frac{\sum_{k=1}^{np} n_k R^2_k}{\sum_{k=1}^{np} n_k}$$

If check-box `Delta'` is checked, values of Δ' estimated for each pair of locus in each sample are saved in a file with the extension `-dp.cd`.

Similarly if check-box `R^2` is checked, values of R^2 estimated for each pair of locus in each sample are saved in a file with the extension `-r2.cd`.

[1] perhaps a better weight would be $n_k H^k_{Sl1} H^k_{Sl2}$, where H^k_{Sl1} represents gene diversity at the first locus in the kth sample

If check-box `detail results` is checked, a file with the extension `-details.cd` is created. It contains, for each sample and and pair of alleles for each pair of locus the following information:

- N, the number of 2 locus genotypes in the sample for the pair of locus considered
- p_A, the frequency of the allele considered at the first locus
- p_B, the frequency of the allele considered at the second locus
- $c_A = p_A(1-p_A) + P_{AA} - p_A^2$, where P_{AA} is the frequency of homozygote AA.
- $c_B = p_B(1-p_B) + P_{BB} - p_B^2$
- $n_{AB} = 2n_1 + n_2 + n_4 + n_5/2$
- $\widehat{\Delta_{AB}} = \frac{n}{n-1}(n_{AB}/n - p_A p_B)$
- Δ'_{AB}
- R_{AB}

Last, it might be useful to have a list of all the contingency tables of genotypes by genotypes. If you check the box `Save contingency Tables`, a file with the extension `-tables.ld` will be created. It can be quite large, as it will contain for each pair of loci in each sample the cross-table of 2 locus genotypes.

8.2 Testing for genotypic disequilibrium

This option allows testing the significance of association between *genotypes* at pairs of loci in each sample[2]. The statistic used to test the tables is the log-likelihood ratio G-statistic[3]. Clearly, only individuals typed at both loci enter the table.

The P-value of the test is obtained as follows. Genotypes at the 2 loci are associated at random a number of times and the statistic is recalculated on the randomised data set. The P-value is estimated as the proportion of statistics from randomised data sets that are larger or equal to the observed.

An overall sample statistic is obtained by summing (over samples) the G-statistics obtained in each sample. The overall test is obtained by comparing this overall statistic with that obtained from randomised tables (randomisation occurring of course only within samples).

If the option `Tests between all pairs of loci in each sample` is chosen, then the P-value for each pair of loci in each sample as well as overall is produced. The number of randomisations is fixed by the nominal level for multiple test radiogroup box. It will be $20 \times np \times nl \times (nl-1)/2$, $100 \times np \times nl \times (nl-1)/2$ or $1000 \times np \times nl \times (nl-1)/2$ randomisations for a 5%, 1% or 0.1% nominal level, respectively (by nominal, I mean at the desired level after Bonferroni corrections). With 10 samples and 10 loci, these numbers are 9'000, 45'000 and 450'000 respectively. For polymorphic markers such as microsatellites, these permutations can take an exceedingly long time.

[2]This test correspond to hypothethis P_4 in Zaykin et al. [1995]: Two-locus genotype frequency is product of two one-locus genotypes frequencies. It is therefore valid when the two loci are not in HWE

[3]more accurately, the only part of this statistic that changes when randomising tables: $\sum_{i\leq j}^{n}\sum_{k\leq l}^{m} x_{ijkl}\log(x_{ijkl})$, where x_{ijkl} represents the number of individuals in the sample with genotype ij at the first locus and genotype kl at the second locus

In most case you are only interested in overall samples level of significance. You should then choose the option `Tests between all pairs of loci`. The number of randomisations for this option is fixed by the nominal level for multiple test radio-group box. It will be $20 \times nl \times (nl-1)/2$, $100 \times nl \times (nl-1)/2$ or $1000 \times nl \times (nl-1)/2$ randomisations for a 5%, 1% or 0.1% nominal level, respectively. With 10 loci, these numbers are 900, 4'500 and 45'000 respectively. It can still take some time for loci with a large number of alleles, as for each randomisation, the test needs to reconstruct the cross-table. With 20 alleles at the 2 loci for instance, the dimension of the table to scan is 210×210!.

Because the number of randomisations can turn out to be very large, I added the possibility to fix this number, using the `fixed number` option combined with the `number of permutations` box (1000 by default).

Chapter 9

Run button

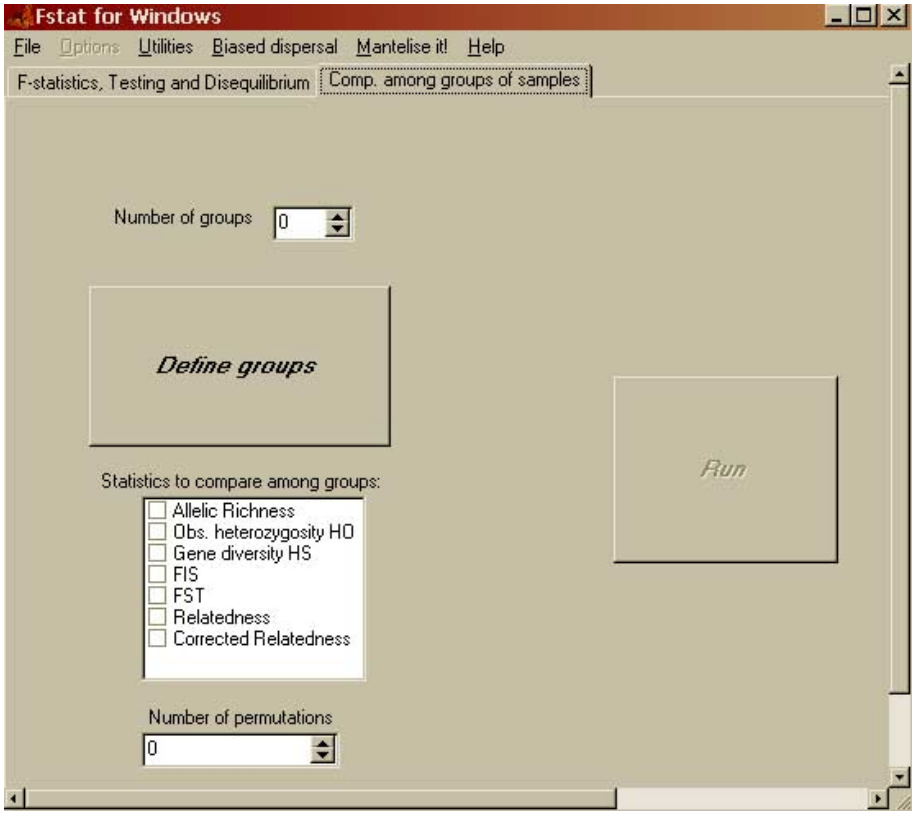
This button is greyed as long as you have not chosen a data file

Once you have selected all the options you want, just click here to run the analysis. The results are written (appended if the file already exists) to the `.OUT` result files.

If you have selected a subset of samples and/or loci, when you click on this menu, FSTAT will prompt you for an alternative filename for the subset of data. The default is the original filename, followed by "-L" and numbers corresponding to the selected loci, followed by "-P" and numbers corresponding to the selected samples. It is a good idea to change this default filename, that could turn out to be very long, and not very meaningful after a few weeks without using the data set!.

Chapter 10

Comp. Among groups of samples



Under this tab-sheet, tests for difference among groups of populations for a number of statistics (allelic richness, Observed heterozygosity, Gene diversity, F_{IS}, F_{ST}, relatedness and corrected relatedness) can be carried out. The number of groups can be anything between 2 and 12. Each group needs to be made of at least 2 samples. If only two groups are compared, the tests could be one or two sided, otherwise, they are two sided.

10.1 Principle of the tests

FSTAT first calculates for each group the average (over samples and loci) of the chosen statistic x (where x could be allelic richness R_s, Observed heterozygosity H_O, Gene diversity H_S, F_{IS}, F_{ST}, relatedness or inbreeding corrected relatedness). If only 2 groups are chosen and a one sided probability is requested, then the difference in x between group 1 and group 2 ($x_1 - x_2$) is taken. Otherwise, the following quantity is calculated:

$$OS_x = \sum_{i=1}^{ng-1} \sum_{j=i+1}^{ng} (x_i - x_j)^2,$$

where ng represents the number of groups. To assess the significance of the statistic OS_x, a permutation scheme is used. Whole samples are allocated at random to the different groups (keeping the number of samples in each group constant), and S_x is calculated from the randomised data set. It is important to remember that for these tests, *the unit of observation is the sample, not the individual*. The P-value of the test is then taken as the proportion of randomised data sets giving a larger S_x than the observed OS_x.

10.1.1 Number of groups

This is the first thing you need to define. This number is constrained to be between 2 and 12.

10.1.2 Define groups

On clicking this button, a window containing 2 lists appears. The list on the left contains all the samples. If you choose to give a label to your different samples (menu option, label file for pops) then their name appears. Select the samples (at least 2, otherwise the test will be meaningless) belonging to the first group, click on the ">" sign, and the samples selected will move from the list on the left to the list on the right. Click OK. The same window reappears to define the second, and subsequent groups. Be careful not to select the same sample for two different groups! If you made a mistake, click the "Cancel" button.

10.1.3 Type of tests for two groups

If only 2 groups were chosen, then a radio group box appears next to "number of groups" after you've define the 2 groups. You could choose here whether you want to test the null hypothesis of no differences against the alternative hypothesis $x_1 > x_2$, or against a two sided alternative.

10.1.4 Statistics to compare among groups

Allelic richness R_S is calculated as described in the per locus and sample statistics section. R_S^k for group k is taken as the (non weighted) average over loci and samples of group k.

Observed heterozygosity, gene diversity, F_{IS}, F_{ST}, relatedness and corrected relatedness are calculated as described in the Global statistics section. These statistics are all weighted by sample size.

10.1.5 Number of permutations

The minimum is 200, the maximum 15'000. Try 1'000 first (it can take very long, particularly if allelic richness is estimated for a large sub-sample). If the P-value you've obtained is close to the nominal level, increase the number of permutations.

10.1.6 Run

Run the tests. Results will be stored in a file with the name `filename_test.out`. Each time you run a new test, results are appended to this file.

Hint: the units of randomisation for these tests are the samples. To increase Degrees of freedom (and therefore the tests' power), it is better to compare many samples with perhaps fewer individuals rather than the opposite. I guess if you are using FSTAT now, it might be already too late to change the sampling design.

Chapter 11

Biased dispersal

11.1 Introduction

This option tests for biases in dispersal among two apriori defined groups of individuals using information from codominant genetic markers. See Goudet et al. [2002] for details of the principle and methods, and a power analysis of the different tests. This option can also be used to test for other things than dispersal such as different levels of inbreeding (using the observed heterozygosity H_O, see below), colour, size, parasitic state etc. . . .

11.2 Design of the tests

The tests assume an animal species with non-overlapping generations where dispersal occurs at the juvenile stage, before reproduction. They further assume that individuals are sampled post dispersal. Under these conditions, several statistical descriptors of an individual's genotype can be used to indicate biases in dispersal. In what follows, I use the example of sex-biased dispersal, but any other grouping would work as well.

11.2.1 Assignment index AI

This statistic was first introduced in Favre et al. [1997]. For each individual j in locality k, the program calculates the probability P_{kij} that its genotype at locus i should appear in the kth sample as the squared frequency of the allele if the individual is homozygous, or twice the product of the frequencies of its two alleles if it is heterozygous. If the loci segregate independently, then the probability of occurrence of a multi-locus genotype is the product of the probabilities of the individual loci. Because populations can contain very different levels of gene diversity, the multi-locus probabilities of individuals in different populations are not directly comparable. To remove this problem, the average probability of the sample is subtracted from the individual multi-locus probability, after log-transformation to avoid rounding errors with very small numbers. This gives the following formula for AI_c, the corrected assignment index of individual j in sample k:

$$AI_{ckj} = \log\left[\prod_{i=1}^{l} P_{kij}\right] - 1/n \sum_{j=1}^{n} \log\left[\prod_{i=1}^{l} P_{kij}\right]$$

for l loci and n individuals in the kth sample. The distribution of AI_c will therefore be centered around 0. A positive value indicates a genotype more likely

than average to occur in its sample (likely a resident individual), while a negative value indicates a genotype less likely than average (potentially a disperser).

mean of AI_c (mAI_c) Because immigrants tend to have lower AI_c values than residents, under sex biased dispersal, the average index for the sex that disperses most is expected to be lower than that for the more philopatric sex. A t-statistic is used for the test: $t = \frac{mAI_c^p - mAI_c^d}{\sqrt{s^2_{AI_c^p}/n_p + s^2_{AI_c^d}/n_d}}$,where n_p and n_d are the number of individuals in the more philopatric and the more dispersing group respectively.

variance of AI_c (vAI_c) Because members of the dispersing sex will include both residents (with common genotypes) and immigrants (with rare genotypes), vAI_c for the sex dispersing most should be largest.

11.2.2 F_{ST}

F_{ST} is a statistic expressing the proportion of the total genetic variance that resides among populations (Hartl and Clarck [1997]). Allelic frequencies for individuals of the sex dispersing most should be more homogeneous than those for individuals of the more philopatric sex. We therefore expect F_{ST} for the more philopatric sex to be higher than that of the more dispersing sex. Among the available estimators of F_{ST}, we choose Weir and Cockerham [1984], because it is the most commonly used and it is also unbiased.

11.2.3 F_{IS}

F_{IS} is a statistic describing how well the genotype frequencies within populations F_{IT} with Hardy Weinberg expectation (Hartl and Clarck [1997]). If only males disperse, the males sampled from a single patch will be a mixture of two populations, residents and immigrants; due to the Wahlund effect, the sample should show a heterozygote deficit and a positive F_{IS}. In general, members of the dispersing sex should therefore display a higher F_{IS} than the more philopatric sex. Among the several estimators of F_{IS}, we also choose Weir and Cockerham [1984].

Relatedness

relatedness is related to F_{ST} as relat=2 F_{ST} /(1+ F_{IT}). A test based on relatedness has essentially the same properties as one based on F_{ST}, and is provided here for convenience.

11.2.4 H_O

The observed heterozygosity, H_O, is NOT expected to change with the dispersal status (nor is the total gene diversity H_T, providing it is calculated with a weighting proportional to the size of each group). But H_O should differ among inbred and outbred individuals. It might be of interest to test whether an individual's status is linked to inbreeding. For instance, one could test whether parasitized individuals tend to be more inbred than healthy ones. If you have such categories, then a test based on H_O is of interest. See Trouvé et al. [2003] For an application of this test.

11.2.5 H_S

The within group gene diversity H_S should be largest for the group dispersing most.

11.2.6 Testing

To test whether these statistics differ significantly between the two sexes, a randomisation approach is used. Under the null hypothesis that males and females disperse equally, the four statistics do not depend on the variable 'sex'. Letting X_d and X_p be the statistic of interest for the dispersing and the philopatric sex respectively, the program proceeds as follows for one sided tests.

1. It first calculates the statistic for each sex over all populations and either take the difference for F_{IS}, H_S and H_O (for H_O the group supposed to be most outbred is assimilated to the group dispersing most); the t-statistic defined above for mAI_c; the difference for F_{ST} and relatedness; or the ratio for vAI_c.
2. It randomly assigns a sex to each of the multi-locus genotypes (keeping the genotypes in their original sample, and the sex ratio in each sample constant).
3. It recalculates the appropriate difference or ratio for the randomised data set.
4. steps 2/ to 3/ are repeated numbperm times.

The probability that dispersal is unbiased by sex is then estimated as the proportion of times where the relevant statistic is larger or equal to the observed one. The two-tailed test are constructed under the same principle using either the absolute value of the differences, or the ratio of the largest to smallest variance.

11.3 Running the program

11.3.1 Input file format

It is a slight modification of the GENEPOP format (Raymond and Rousset [1995b])[1]. Here is a brief exemple:

```
5 8 64
loc-3ks, loc-17, loc-3T7, loc-57, loc-72, loc-23, loc-9, loc-53
Pop
F,0, 4041 2627 3234 2828 2830 2122 2223 2022
M,0, 3940 2929 3232 3333 2430 2022 2329 0
M,0, 4040 2727 3232 2833 2731 0 2222 2020
F,0, 4059 2429 3434 3233 2430 1920 2325 2323
F,0, 4059 2429 3232 3233 2430 1920 2329 2023
F,0, 4040 2429 3232 3233 3030 0 0 2222
F,0, 4041 2829 3234 2931 2428 1919 2223 2023
M,0, 4059 2930 3434 3233 2425 1920 2531 2020
M,0, 4040 2429 3234 3233 0 0 2323 2023
Pop
F,0, 4041 2629 3334 3031 2428 2021 2326 2021
F,0, 4040 2735 3434 2833 2528 2023 2326 2020
M,0, 4040 2929 3333 2931 2424 2022 2223 2020
M,0, 4040 2629 3334 3030 2425 2020 2326 2121
F,0, 3940 2929 3234 2932 2425 1919 2326 2023
M,0, 4040 2729 3333 2931 2430 0 2323 2023
M,0, 4040 2626 3434 2931 2830 1921 2326 2021
F,0, 4040 2629 3334 3031 2425 2020 2223 2121
F,0, 4041 2727 3234 3232 2428 2021 2326 2123
M,0, 4040 2729 3233 3031 2430 2022 2223 1920
Pop
F,0, 4040 2729 3234 2933 2425 1919 2223 2023
M,0, 4040 2727 3434 2929 2425 1919 2223 1821
M,0, 4041 2729 3434 0 2528 1922 2123 2323
M,0, 4040 2727 3234 2933 2527 1919 2326 1821
F,0, 4040 2429 3333 2929 2324 2223 2323 1818
M,0, 4040 2929 3232 2933 2427 1919 0 2121
M,0, 4040 2929 3434 2933 2525 1919 2326 2121
M,0, 4041 2727 3434 2932 2528 1919 2326 2121
M,0, 4041 2929 3434 3333 2528 1919 2226 2121
```

The format has to be the following:

- The file name has the extension `.GEN`. No other extension will be recognised by the program.
- The first line contains 3 numbers:

[1] See menu `utilities, file conversion` for creating GENEPOP format using FSTAT

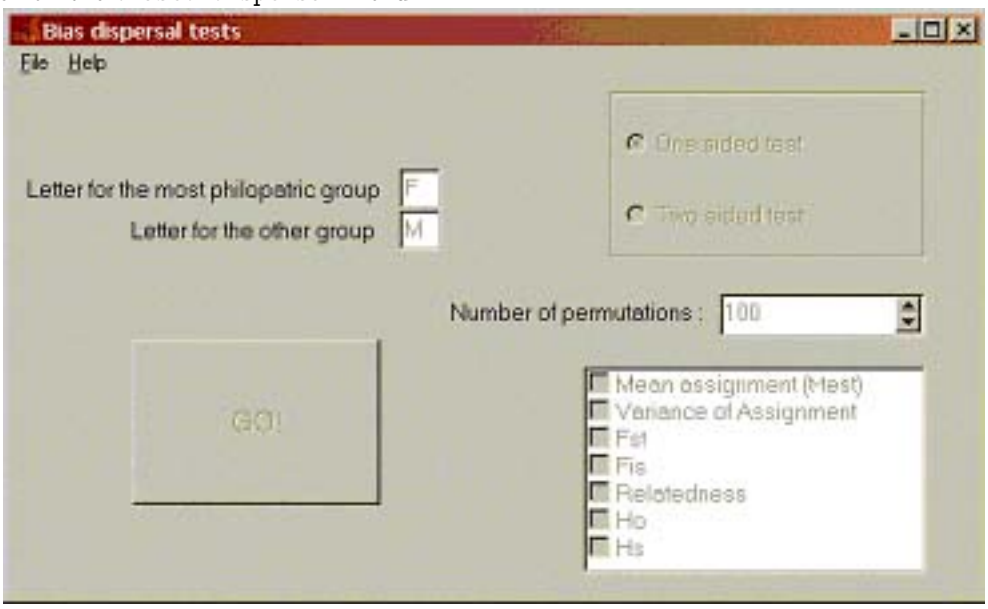
  - the number of samples,
  - the number of loci, and
  - the highest index used for an allele.

  Alleles are to be encoded as 2 number digits (from 01 to 99).

- The second line contains the name of the loci, separated by comas.
- The third line contains the keyword 'Pop' indicating the start of a new sample.
- The following lines contains the multilocus genotypes for each individual, preceded by a capital letter defining the group to which it belongs, a comma, another descriptor of the individual (could be any string of characters excluding a comma) and a second comma. The capital letters defining the group could be anything but a 'P', the program will generate an error message if you use a 'P'.
- Each new sample is separated from the previous by the keyword 'Pop'. See the example of input data set.

11.3.2 Running the program

Click on the `biased dispersal` menu



Here is a screen shot of the program windows. To run the program, do the following:

1. choose the file containing the data under the `File, open` menu.
2. type in each box (`Letter for the most philopatric group` and `Letter for the other group`) the letter used in the data file to define members of the two groups
3. choose whether you want a one or a two-sided test
4. choose which test(s) statistic you want to use
5. enter the number of randomizations for the tests (between 100 and 10'000)
6. click the `GO` button.

11.3.3 Results

The results are stored in a file with the extension `.res` and the same name as the input file. An example of such a file is given in Appendix B. The program also produces 4 other files:

- One is with the extension `.per` and contains the results of all the permutations. The first line of the file contains the observed values for the differences between groups of the different statistics (in the order Mean ass, Var ass, F_{ST} and F_{IS}, relatedness, H_O and H_S), preceded by the number 0. The remaining lines contain the same information for each of the (numbperm-1) randomisations. With this file, you can generate the distribution of the statistics under the null hypothesis, as is done in the article describing the tests (Goudet et al. [2002]).
- Three are files with the extension `.dat` and with the data reformatted so that they can be analysed with FSTAT . One of these files contains the whole data set, the two others contain the data for each group, with the letter corresponding to the group identifier appended to the file name.

Chapter 12

Mantelize it!

Under this menu, one can carry out multiple regression or partial Mantel test. This latter option is available only providing the data originated from distance matrices. While Raufaste and Rousset [2001] have criticized partial mantel, these tests are still useful (rewrite)

12.1 Matrices to Columns

In order to perform a multiple regression or a (partial) mantel test with FSTAT (see menu multiple regression and partial mantel), the data must be in a columnar mode. But often the data from distance matrices to perform mantel tests are in matrix form, this is for instance the case of pairwise F_{ST} values produced by FSTAT. The menu 'Matrices to Columns' allows you to combine the data from different files in matrix form (whole matrices, with 0 on the diagonal) into a single file with the data in columnar form (the half matrix below the diagonal is read, implying that datum [2,1] is first read, followed by data [3,1] & [3,2] etc. until datum [n,n-1] where n is the dimension of the matrices). When prompted for "File names for data in matrix form", select all the files containing data you want to see appearing in the multiple regression or partial mantel analysis. Clearly, these files need to have the same dimension (same number of lines and columns). Multiple selections can be achieved by holding the Ctrl key while clicking on the different files. The program then prompts the user for the name of the columnar mode file. No specific extension is required for this file. An example of such a file is given in the distribution. See Example of input file for multiple regression and partial mantel

12.2 Multiple regression and partial mantel

This is an adaptation of Manly [1997] code for multiple regression, described in the chapter "regression analysis" of the book. The modification I made only concerns analyses with data taken from symmetrical matrices (pairwise F_{ST} values for instance). If you have such data, the program detects it (by counting the number of observations: if the data are from a distance matrix, then they should be $n(n-1)/2$ observations in the file). This will then modify the type of randomisations to be carried out when testing significance of the multiple regression.

Once you have chosen a file with data in the appropriate format (Example of input file), you need to first pick the dependant variable, usually a genetic distance. Once it is chosen, you can pick up to 10 explanatory variables. If the number of observations is compatible with the data being taken from a distance matrix, a check-box appears allowing you to specify the appropriate permutation scheme.

You also have the possibility to save residuals in a file. The default number of permutations for the test is set at 2000. Once everything is selected, click on run. The results appear in the window at the lower right hand of the panel, while the two upper right panels show scatterplots of the estimates and residuals, and fitted and observed values. By left-clicking the graph, you will see the residuals (right panel) and the estimates (left panel) as a function of the estimates (right panel) or and the observed values (left panel).

The result file is structured as follows:

A brief description of the input file is given (name and comments). Then follows the number of randomisations used in the test.

The total sum of square of the dependant variable is then given, followed by the result of the regression: the (partial) correlation of each explanatory variable with the dependant variables, the coefficient associated with each explanatory variable and the sum of squares explained by the explanatory variables.

The overall percentage of the variance explained by the model (R-squared is then given).

Last, the P-value for the coefficient associated with each variable is given, together with the P-value associated with the proportion of variance explained by each explanatory variable. These P-value are given as percentage, that is, a 5 would mean 5% of the randomisation gave values as large or larger than the observed.

Each time a new test is run, it is appended at the end of the result file.

Chapter 13

Utilities menus

13.1 Reset seeds Random number generator

A dialog box appears, allowing you to reset the seeds of the random number generator. This is equivalent to editing the contents of the file `FSTAT.INI`

13.2 File conversion

This menu allows you to convert FSTAT format to GENEPOP and vice-versa. GENEPOP data files need to have the extension `.GEN` in order to be recognised by FSTAT.

The conversion from GENEPOP to FSTAT will not work if the label of an individual (the text before the comma on each line of the GENEPOP format) starts with the characters "POP", "pop" or any case combination of these three letters.

If the label starts with the letter "P", make sure that it is longer than 3 characters (spaces are valid characters, so the label "PXX", where X stands for a space or any other character [0..9; a..z; A..Z], is valid).

Chapter 14

Files of results

Results will be stored in a file named `FILENAME.OUT`, located in the same folder as the input file.

This file first contains a line saying when the analysis has been run. On the following lines, one find the identifier of the sample, then, for each locus, the size of each sample, followed by the frequency per sample and overall frequency of each allele present at the locus. The rest of the output file should be self explanatory.

If you requested that pairwise F_{ST} be estimated under the menu choose options, F-statistics, F_{ST} per pair 2 files are created:

- A file `FILENAME.FST`. This file contains the estimated pairwise F_{ST}
- A File `FILENAME.MAT`. It contains two half matrices. The first one corresponds to the pairwise sum of variance components (σ_a^2, σ_b^2 and σ_w^2) representing the denominator of Cockerham estimator of F_{ST}. The second corresponds to the pairwise variance component σ_a^2, the numerator of Cockerham estimator of F_{ST}.

If you requested that genotypic frequencies be calculated under the menu choose options, gene diversities, genotypic frequencies, a File FILENAME.X2 will be created. It contains for each genotype at each locus the observed and expected genotypic count, where the expected genotypic count is calculated using unbiased estimation:

$$\begin{aligned} A_iA_i &= n\left[p_i\tfrac{(2np_i-1)}{(2n-1)}\right], \\ A_iA_j &= 2n\left[p_i\tfrac{(2np_j)}{(2n-1)}\right] \end{aligned}$$

where A_iA_i represents homozygote genotypes, A_iA_j represents heterozygotes, n the sample size, pi and pj the allele frequencies of allele A_i and A_j respectively.

If you have selected the sub-option, Testing, Pairwise Pop. diff. NOT assuming HW within submenu, a file `FILENAME-PP.PVL` containing two tables is generated. The first gives the non-adjusted P-value for each pair. The second gives the pairwise significance after standard Bonferroni corrections. "***" corresponds to significance at the 0.1% nominal level, "**" significance at the 1% nominal level and "*" significance at the 5% nominal level. "NS" stands for non-significant and "NA" for not available.

If you have requested to estimate the composite disequilibrium Δ', a file with extension `-dp.cd` will be created. Similarly, if you have requested the composite disequilibrium R^2, a file with extension `-r2.cd` will be generated. If you requested a

detailed output, a file `-details.cd` will be generated. Last, if you requested FSTAT to save contingency tables, a file with extension `-tables.ld` will be generated.

All the separators in the output files are tabs, which allows direct importation of the results into commercial packages such as the Microsoft spreadsheet Excel, therefore facilitating printing and graphical representation of the results.

Chapter 15

Special use of Fstat

15.1 Haploid data

While FSTAT has been written for diploid data sets, it could also be used for haploid species. To do so, it is necessary to encode haploid genotypes as diploid genotypes, by writing twice the identifier of the allele. A brief example follows:

```
   3  4  4  1
loc-1
loc-2
loc-3
loc-4
1 11 22 11 11
1 22 22 11 22
1 11 33 11 22
1 22 22 33 11
2 11 11 11 11
2 11 22 11 22
2 11 11 11 11
3 22 11 11 11
3 33 11 11 11
3 22 11 33 11
```

Note that all genotypes are homozygous. When running this file, F_{IS} (f) and F_{IT} (F) will be meaningless. F_{ST} (θ) will however be an appropriate measure of the Wahlund effect. The file `HAPLO.DAT`, given in Weir [1996] 'Genetic Data Analysis', is distributed with the example files.

CAUTION : one should not mix haploid and diploid data in the same data file. If you have autosomal markers and mitochondrial or Y chromosome markers, do not put them in the same file but rather create two files, one for the autosomal loci, the other for the haploid loci. The same goes for haplo-diploid species. This is to insure that your overall loci statistics have some meaning!

15.2 Only one sample

In this case, FSTAT will output the statistics it can estimate, namely allele frequencies, gene diversities, number of alleles sampled, F_{IS} if requested, test of F_{IS} if requested, as well as genotypic disequilibrium statistics and tests.

15.3 Pooling samples

One may want to calculate F-statistics for different levels of grouping of samples, and follow the changes of, e.g. F_{IS} (f) with different levels of pooling. An example of such a procedure is given in Goudet et al. [1994], Goudet [1993]. To allow for this, one needs to modify the input file by renumbering samples according to pooling, and modifying the number of samples np on the first line.

15.4 Missing data

The situation may arise where some loci could not be screened for all the samples. FSTAT handles this situation properly. But I can only urge users to avoid data sets with many missing genotypes. In particular when some samples have no individuals typed at one or more loci, you might run into all sorts of problems. You've certainly heard from statistical courses that balance sampling is the best. Well, It really is!

Appendix A

Examples of input files

One digit encoding for data file Two digits encoding for data file Label file

A.1 One digit encoding for input data file

A file encoded with 1 digit number (This is the example file DIPLOID.DAT, given in Weir's (1990) 'Genetic Data Analysis'.

```
6  5  4  1
loc-1
loc-2
loc-3
loc-4
loc-5
    1   44 43 43 33 44
    1   44 44 43 33 44
    1   44 44 43 43 44
    1   44 44  0 33 44
    1   44 44 24 34 44
    1   44 44  0 43 44
    1   44 44 43 43 44
    1   44 44  0 43 44
    2   44 44 33 32 44
    2   44 33 44 43 44
    2   44 43 44 43 44
    2   44 44 33 33 44
    2   44 43 44 44 44
    2   44 44 44 22 44
    2   44 44 43 43 44
    2   44 44 44 44 44
    3   44 44 44 43 44
    3   44 44 44 44 44
    3   44 44 43 21 44
    3   44 44 33 43 44
    3   44 44 43 21 44
    4   44 44 43 44 44
    4   44 44 43 43 44
    4   44 44 43 43 44
    4   44 44 43 44 44
    4   44 44 43 44 44
    4   44 44 44 33 44
    4   44 44 44 44 44
    5   44 44 44 21 44
    5   44 44 44 33 44
    5   44 44 43 43 44
    5   44 44 43 43 44
    5   44 44 44 44 44
    5   44 44 44 43 44
    5   44 44 43 43 44
    5   44 44 44  0 44
    5   44 43 44 43 44
    6   44 44 44 43 44
    6   44 44 43 33 44
    6   44 44 44 32 44
    6   44 44 43 41 44
    6   44 44 44 44 44
    6   44 44 44 42 44
    6   44 44 44 43 44
```

A.2 Two digits encoding for input data file

The same file with alleles encoded with a 2 digit number:

```
6  5  4  2
loc-1
loc-2
loc-3
loc-4
loc-5
1       0404 0403 0403 0303 0404
1       0404 0404 0403 0303 0404
1       0404 0404 0403 0403 0404
1       0404 0404    0 0303 0404
1       0404 0404 0204 0304 0404
1       0404 0404    0 0403 0404
1       0404 0404 0403 0403 0404
1       0404 0404    0 0403 0404
2       0404 0404 0303 0302 0404
2       0404 0303 0404 0403 0404
2       0404 0403 0404 0403 0404
2       0404 0404 0303 0303 0404
2       0404 0403 0404 0404 0404
2       0404 0404 0404 0202 0404
2       0404 0404 0403 0403 0404
2       0404 0404 0404 0404 0404
3       0404 0404 0404 0403 0404
3       0404 0404 0404 0404 0404
3       0404 0404 0403 0201 0404
3       0404 0404 0303 0403 0404
3       0404 0404 0403 0201 0404
4       0404 0404 0403 0404 0404
4       0404 0404 0403 0403 0404
4       0404 0404 0403 0403 0404
4       0404 0404 0403 0404 0404
4       0404 0404 0403 0404 0404
4       0404 0404 0404 0303 0404
4       0404 0404 0404 0404 0404
5       0404 0404 0404 0201 0404
5       0404 0404 0404 0303 0404
5       0404 0404 0403 0403 0404
5       0404 0404 0403 0403 0404
5       0404 0404 0404 0404 0404
5       0404 0404 0404 0403 0404
5       0404 0404 0403 0403 0404
5       0404 0404 0404    0 0404
5       0404 0403 0404 0403 0404
6       0404 0404 0404 0403 0404
6       0404 0404 0403 0303 0404
6       0404 0404 0404 0302 0404
6       0404 0404 0403 0401 0404
6       0404 0404 0404 0404 0404
6       0404 0404 0404 0402 0404
6       0404 0404 0404 0403 0404
```

A.3 Example of label file

A file of labels for the samples in diploid.dat. Each line contains the name (label) of a sample:

```
Stade de France
Twickenham
Arms Park
Millenium
Lansdowne road
Murrayfield
```

Only the first six characters (those in bold) will be used in the output files

One digit encoding for data file Two digits encoding for data file

A.4 Example of file format for testing biased dispersal

File exampsb.gen of the distribution.

```
File exampsb.gen of the distribution.

5 8 64
loc-3ks, loc-17, loc-3T7, loc-57, loc-72, loc-23, loc-9, loc-53
Pop
F,0, 4041 2627 3234 2828 2830 2122 2223 2022
M,0, 3940 2929 3232 3333 2430 2022 2329 0
M,0, 4040 2727 3232 2833 2731 0 2222 2020
F,0, 4059 2429 3434 3233 2430 1920 2325 2323
F,0, 4059 2429 3232 3233 2430 1920 2329 2023
F,0, 4040 2429 3232 3233 3030 0 0 2222
F,0, 4041 2829 3234 2931 2428 1919 2223 2023
M,0, 4059 2930 3434 3233 2425 1920 2531 2020
M,0, 4040 2429 3234 3233 0 0 2323 2023
Pop
F,0, 4041 2629 3334 3031 2428 2021 2326 2021
F,0, 4040 2735 3434 2833 2528 2023 2326 2020
M,0, 4040 2929 3333 2931 2424 2022 2223 2020
M,0, 4040 2629 3334 3030 2425 2020 2326 2121
F,0, 3940 2929 3234 2932 2425 1919 2326 2023
M,0, 4040 2729 3333 2931 2430 0 2323 2023
M,0, 4040 2626 3434 2931 2830 1921 2326 2021
F,0, 4040 2629 3334 3031 2425 2020 2223 2121
F,0, 4041 2727 3234 3232 2428 2021 2326 2123
M,0, 4040 2729 3233 3031 2430 2022 2223 1920
Pop
F,0, 4040 2729 3234 2933 2425 1919 2223 2023
M,0, 4040 2727 3434 2929 2425 1919 2223 1821
M,0, 4041 2729 3434 0 2528 1922 2123 2323
M,0, 4040 2727 3234 2933 2527 1919 2326 1821
F,0, 4040 2429 3333 2929 2324 2223 2323 1818
M,0, 4040 2929 3232 2933 2427 1919 0 2121
M,0, 4040 2929 3434 2933 2525 1919 2326 2121
M,0, 4041 2727 3434 2932 2528 1919 2326 2121
M,0, 4041 2929 3434 3333 2528 1919 2226 2121
Pop
F,0, 4060 3031 1734 3134 2830 2224 2428 2324
F,0, 4157 2830 1734 2834 1332 2124 2330 1919
M,0, 4040 2335 3333 3032 2328 2224 2527 1922
M,0, 4041 2331 3234 3034 2532 2324 2527 1922
M,0, 4041 2832 3334 3034 2832 2224 2730 1919
F,0, 4158 2635 1732 3131 1323 2225 2426 1920
F,0, 4040 2729 3034 3131 2328 2225 2527 1919
F,0, 4040 2735 3334 3032 2323 2222 2525 1922
F,0, 4054 2628 1732 2929 2931 2023 2526 1920
M,0, 4041 2428 3434 2734 2532 2425 2325 1922
Pop
F,0, 4040 2629 3430 2830 2931 2323 2324 2222
M,0, 4041 2832 3434 2530 2932 2024 2529 1920
M,0, 4040 2328 3432 3031 2929 2020 2426 1920
M,0, 4041 2628 3034 2930 2929 2023 2426 2022
M,0, 4052 2731 3428 2530 3030 2024 2323 1920
M,0, 4141 2832 3434 2530 2932 2024 2323 1920
M,0, 4052 2832 3428 3134 3030 2022 2323 1920
F,0, 4360 2627 1729 2930 1213 1924 2426 1921
F,0, 4153 3332 3428 2534 2932 2224 2127 1920
F,0, 4040 2832 3433 2829 2929 2425 2126 1920
M,0, 4040 2826 3034 3930 2929 2023 2424 1922
```

A.5 Example of input file for multiple regression and partial mantel

```
Beginning of file YANOM.ALL of the distribution

Put here any comments you want
       171       3
var0 (YANOM.GEN)
var1 (YANOM.ANT)
var2 (YANOM.GEO)
35.00000    96.00000     9.00000
44.00000    147.00000   28.00000
38.00000    88.00000    20.00000
47.00000    295.00000   152.00000
56.00000    318.00000   161.00000
59.00000    309.00000   178.00000
52.00000    284.00000   149.00000
65.00000    339.00000   158.00000
67.00000    348.00000   175.00000
30.00000    158.00000    7.00000
57.00000    253.00000   169.00000
76.00000    258.00000   175.00000
80.00000    260.00000   184.00000
50.00000    235.00000   102.00000
43.00000    253.00000   95.00000
65.00000    289.00000   172.00000
69.00000    301.00000   178.00000
69.00000    321.00000   187.00000
52.00000    277.00000   98.00000
60.00000    270.00000   91.00000
68.00000    120.00000    7.00000
69.00000    507.00000   253.00000
74.00000    571.00000   259.00000
63.00000    583.00000   276.00000
61.00000    519.00000   232.00000
65.00000    455.00000   238.00000
88.00000    472.00000   330.00000
79.00000    487.00000   327.00000
34.00000    488.00000   244.00000
46.00000    560.00000   250.00000
46.00000    570.00000   267.00000
```

Appendix B

Files of results from the input file in Appendix 1.

B.1 DIPLOID.OUT

```
*****************************************************************************************************************************
*   The following results were generated the 09.08.2001 at 13:44:36 with \textsc{Fstat} for windows, V2.9.3 from file diploid2.dat. *
*****************************************************************************************************************************
           Stade   Twicke  Arms P  Millen  Lansdo  Murray   All_W  All_UW
       Locus: loc-1
        N       8       8       5       7       9       7
   p:   4   1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000


       Locus: loc-2
        N       8       8       5       7       9       7
   p:   3   0.063   0.250   0.000   0.000   0.056   0.000   0.068   0.061
   p:   4   0.938   0.750   1.000   1.000   0.944   1.000   0.932   0.939


       Locus: loc-3
        N       5       8       5       7       9       7
   p:   2   0.100   0.000   0.000   0.000   0.000   0.000   0.012   0.017
   p:   3   0.400   0.313   0.400   0.357   0.167   0.143   0.280   0.297
   p:   4   0.500   0.688   0.600   0.643   0.833   0.857   0.707   0.687


       Locus: loc-4
        N       8       8       5       7       8       7
   p:   1   0.000   0.000   0.200   0.000   0.063   0.071   0.047   0.056
   p:   2   0.000   0.188   0.200   0.000   0.063   0.143   0.093   0.099
   p:   3   0.688   0.375   0.200   0.286   0.438   0.357   0.407   0.390
   p:   4   0.313   0.438   0.400   0.714   0.438   0.429   0.453   0.455


       Locus: loc-5
        N       8       8       5       7       9       7
   p:   4   1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000


************************************************

   Gene diversity per locus and population :
   loc-1    0.000   0.000   0.000   0.000   0.000   0.000
   loc-2    0.125   0.411   0.000   0.000   0.111   0.000
   loc-3    0.600   0.482   0.550   0.476   0.292   0.262
   loc-4    0.446   0.688   0.800   0.452   0.643   0.714
   loc-5    0.000   0.000   0.000   0.000   0.000   0.000

************************************************

   number of alleles sampled :
   loc-1        1       1       1       1       1       1       1
   loc-2        2       2       1       1       2       1       2
   loc-3        3       2       2       2       2       2       3
   loc-4        2       3       4       2       4       4       4
   loc-5        1       1       1       1       1       1       1

************************************************

   Allelic Richness per locus and population
   based on min. sample size of:    5 diploid individuals.
   loc-1    1.000   1.000   1.000   1.000   1.000   1.000   1.000
   loc-2    1.625   1.992   1.000   1.000   1.556   1.000   1.526
   loc-3    3.000   1.999   2.000   2.000   1.931   1.934   2.093
   loc-4    1.999   2.964   4.000   1.999   3.250   3.648   3.035
   loc-5    1.000   1.000   1.000   1.000   1.000   1.000   1.000

************************************************

   Fis Per population :
   loc-1       NA      NA      NA      NA      NA      NA
   loc-2    0.000   0.391      NA      NA   0.000      NA
```

```
    loc-3   -0.667   0.741   0.273  -0.500  -0.143  -0.091
    loc-4   -0.400   0.273   0.000   0.368  -0.167   0.000
    loc-5       NA      NA      NA      NA      NA      NA

    All     -0.494   0.446   0.111  -0.077  -0.142  -0.024

**************************************************

    P-value for Fis within samples.
    based on :  3000 randomisations.
    Indicative adjusted nominal level (5\%) for one table is :    0.00167

    Proportion of randomisations that gave a LARGER Fis than the observed:
    loc-1       NA      NA      NA      NA      NA      NA
    loc-2   1.0000  0.3950      NA      NA  1.0000      NA
    loc-3   1.0000  0.0807  0.6300  1.0000  1.0000  1.0000
    loc-4   1.0000  0.2223  0.7030  0.4500  0.8533  0.6703
    loc-5       NA      NA      NA      NA      NA      NA

    All     1.0000  0.0137  0.4757  0.8217  0.8710  0.6957

    Proportion of randomisations that gave a SMALLER Fis than the observed:
    loc-1       NA      NA      NA      NA      NA      NA
    loc-2   1.0000  0.9880      NA      NA  1.0000      NA
    loc-3   0.1233  1.0000  0.9547  0.3193  0.8310  0.9230
    loc-4   0.3977  0.9233  0.7657  0.9817  0.4297  0.6733
    loc-5       NA      NA      NA      NA      NA      NA

    All     0.0503  0.9953  0.8163  0.6247  0.3630  0.6287

**************************************************
    Nei's estimation of heterozygosity

    LocName     Ho      Hs      Ht     Dst    Dst'     Ht'     Gst    Gst'     $G_{IS}$
    loc-1    0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
    loc-2    0.081   0.109   0.117   0.008   0.009   0.118   0.065   0.077   0.258
    loc-3    0.476   0.443   0.445   0.002   0.002   0.446   0.004   0.005  -0.074
    loc-4    0.613   0.623   0.635   0.013   0.015   0.638   0.020   0.024   0.016
    loc-5    0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000

    Overall  0.234   0.235   0.239   0.004   0.005   0.240   0.018   0.022   0.005

**************************************************
    Weir & Cockerham (1984) estimation of Fit (CapF), Fst (theta) and Fis (smallF).
    relat is Relatedness estimated following Queller & Goodnight (1989)
    relatc is relatedness inbreeding corrected following Pamilo (1984, 1985)
    sig_a, sig_b and sig_w are the component of variance
    among samples, among individuals within samples and within individuals respectively.


     For locus : loc-2
    Allele    Capf   Theta  Smallf   Relat  Relatc   Sig_a   Sig_b   Sig_w
        3    0.303   0.069   0.251   0.107
        4    0.303   0.069   0.251   0.107
      All    0.303   0.069   0.251   0.107  -0.668   0.009   0.030   0.091

     For locus : loc-3
    Allele    Capf   Theta  Smallf   Relat  Relatc   Sig_a   Sig_b   Sig_w
        2    0.006   0.037  -0.032   0.073
        3   -0.017  -0.010  -0.006  -0.021
        4   -0.045   0.017  -0.063   0.036
      All   -0.030   0.005  -0.035   0.009   0.067   0.002  -0.015   0.439

     For locus : loc-4
    Allele    Capf   Theta  Smallf   Relat  Relatc   Sig_a   Sig_b   Sig_w
        1   -0.030   0.046  -0.080   0.096
        2    0.186   0.011   0.177   0.019
        3    0.007   0.048  -0.043   0.095
        4    0.027   0.002   0.025   0.004
      All    0.037   0.024   0.013   0.047  -0.026   0.015   0.008   0.605

**************************************************

         Over all loci
          Capf   Theta  Smallf   Relat  Relatc   Sig_a   Sig_b   Sig_w
         0.042   0.022   0.020   0.043  -9.990   0.026   0.023   1.135

**************************************************

    Rst Over all samples estimated following Rousset (1996) and Goodman (1997)

           Rst    siga    sigb    sigw   amean   astdev
    loc-2    0.069     0.0     0.0     0.0    3.93   0.2521
    loc-3    0.042     0.0    -0.0     0.3    3.70   0.4861
    loc-4    0.016     0.0     0.2     0.4    3.27   0.8130

      Rst over loci   Weighted     Goodman  Unweighted
               Rst:     0.0260      0.0426      0.0425

**************************************************

**************************************************
    Jackknifing over populations.

     For locus : loc-2
          Capf   Theta  Smallf   Relat
     Total   0.461   0.117   0.357   0.207   Means
             0.303   0.100   0.226   0.162   Std. Err.

     For locus : loc-3
          Capf   Theta  Smallf   Relat
     Total  -0.032   0.003  -0.025  -0.012   Means
             0.222   0.046   0.264   0.108   Std. Err.

     For locus : loc-4
          Capf   Theta  Smallf   Relat
     Total   0.035   0.023   0.014   0.046   Means
```

```
           0.098   0.058   0.106   0.111   Std. Err.

************************************************

     Jackknifing over loci.

         Capf   Theta  Smallf   Relat
     total   0.031   0.021   0.010   0.041   Means
           0.046   0.010   0.038   0.018   Std. Err.

************************************************
           Bootstrapping over Loci.

           95\% Confidence Interval.

         CapF   theta  Smallf   Relat
       -0.030   0.005  -0.035   0.009
        0.303   0.069   0.251   0.107

           99\% Confidence Interval.

         CapF   theta  Smallf   Relat
       -0.030   0.005  -0.035   0.009
        0.303   0.069   0.251   0.107


************************************************

    Randomising alleles within samples.
    Testing for Hardy-Weinberg within samples.
    Statistic used to classified tables is smallf (Fis).
    based on :  5000 randomisations.

loc-2   Prop rand. larger than observed in [       0.39600     0.01320]
loc-3   Prop rand. larger than observed in [       0.64380     0.47100]
loc-4   Prop rand. larger than observed in [       0.50840     0.37720]
All Loci    Prop rand. larger than observed in [       0.41180     0.39980]

************************************************
    Randomising alleles overall samples

    Test based on :  5000 randomisations.

    Testing for Hardy-Weinberg overall using the statistic CapF (Fit)

loc-2   Prop rand. larger than observed in [       0.02300     0.02260]
loc-3   Prop rand. larger than observed in [       0.55260     0.55200]
loc-4   Prop rand. larger than observed in [       0.32860     0.32860]
All Loci    Prop rand. larger than observed in [       0.27340     0.27340]

************************************************
    Randomising genotypes among samples.

    test based on :  5000 randomisations.

    testing for population differentiation,
    NOT ASSUMING RANDOM MATING within samples.
    Statistic used is the log-likelihood G (Goudet et al, 1996)


loc-2   Prop rand. larger than observed in [       0.15000     0.14620]
loc-3   Prop rand. larger than observed in [       0.31900     0.31860]
loc-4   Prop rand. larger than observed in [       0.22500     0.22500]
All Loci    Prop rand. larger than observed in [       0.13380     0.13380]

********************************************************
    P-value for genotypic disequilibrium
    based on   6000 permutations.
    Adjusted P-value for 5\% nominal level is :   0.000833
    Adjusted P-value for 1\% nominal level is :   0.000167

        Stade   Twicke  Arms P  Millen  Lansdo  Murray     All
loc-1 X loc-2             NA          NA          NA          NA          NA          NA          NA
loc-1 X loc-3             NA          NA          NA          NA          NA          NA          NA
loc-1 X loc-4             NA          NA          NA          NA          NA          NA          NA
loc-1 X loc-5             NA          NA          NA          NA          NA          NA          NA
loc-2 X loc-3        1.00000     0.77000          NA          NA     1.00000          NA     0.81650
loc-2 X loc-4        0.36633     1.00000          NA          NA     1.00000          NA     0.97733
loc-2 X loc-5             NA          NA          NA          NA          NA          NA          NA
loc-3 X loc-4        1.00000     0.42883     0.34117     0.33250     1.00000     0.51867     0.21267
loc-3 X loc-5             NA          NA          NA          NA          NA          NA          NA
loc-4 X loc-5             NA          NA          NA          NA          NA          NA          NA
```

B.2 Files obtained from the "F_{ST} per pair" option

```
DIPLOID.FST

     0.0000  0.0207  0.0665  0.1101  0.0704  0.1068
     0.0207  0.0000 -0.0390  0.0119 -0.0185 -0.0131
     0.0665 -0.0390  0.0000 -0.0006  0.0147 -0.0104
     0.1101  0.0119 -0.0006  0.0000  0.0342  0.0378
     0.0704 -0.0185  0.0147  0.0342  0.0000 -0.0525
     0.1068 -0.0131 -0.0104  0.0378 -0.0525  0.0000


DIPLOID.MAT

     1.3980
     1.3177  1.4438
     1.1724  1.2953  1.0982
     1.1452  1.2730  1.1690  1.0243
```

```
    1.1648  1.2850  1.1157  0.9898  0.9666

    0.0290
    0.0876 -0.0563
    0.1291  0.0155 -0.0006
    0.0806 -0.0235  0.0171  0.0351
    0.1245 -0.0168 -0.0116  0.0374 -0.0508
```

B.3 DIPLOID.X2

```
Observed and expected genotype frequencies

    Observed for locus : loc-1
    0404      8   8   5   7   9   7

    Expected for locus : loc-1
    0404      8.00    8.00    5.00    7.00    9.00    7.00

    Observed for locus : loc-2
    0303      0   1   0   0   0   0
    0304      1   2   0   0   1   0
    0404      7   5   5   7   8   7

    Expected for locus : loc-2
    0303      0.00    0.40    0.00    0.00    0.00    0.00
    0304      1.00    3.20    0.00    0.00    1.00    0.00
    0404      7.00    4.40    5.00    7.00    8.00    7.00

    Observed for locus : loc-3
    0202      0   0   0   0   0   0
    0203      0   0   0   0   0   0
    0204      1   0   0   0   0   0
    0303      0   2   1   0   0   0
    0304      4   1   2   5   3   2
    0404      0   5   2   2   6   5

    Expected for locus : loc-3
    0202      0.00    0.00    0.00    0.00    0.00    0.00
    0203      0.44    0.00    0.00    0.00    0.00    0.00
    0204      0.56    0.00    0.00    0.00    0.00    0.00
    0303      0.67    0.67    0.67    0.77    0.18    0.08
    0304      2.22    3.67    2.67    3.46    2.65    1.85
    0404      1.11    3.67    1.67    2.77    6.18    5.08

    Observed for locus : loc-4
    0101      0   0   0   0   0   0
    0102      0   0   2   0   1   0
    0103      0   0   0   0   0   0
    0104      0   0   0   0   0   1
    0202      0   1   0   0   0   0
    0203      0   1   0   0   0   1
    0204      0   0   0   0   0   1
    0303      3   1   0   1   1   1
    0304      5   3   2   2   5   2
    0404      0   2   1   4   1   1

    Expected for locus : loc-4
    0101      0.00    0.00    0.11    0.00    0.00    0.00
    0102      0.00    0.00    0.44    0.00    0.07    0.15
    0103      0.00    0.00    0.44    0.00    0.47    0.38
    0104      0.00    0.00    0.89    0.00    0.47    0.46
    0202      0.00    0.20    0.11    0.00    0.00    0.08
    0203      0.00    1.20    0.44    0.00    0.47    0.77
    0204      0.00    1.40    0.89    0.00    0.47    0.92
    0303      3.67    1.00    0.11    0.46    1.40    0.77
    0304      3.67    2.80    0.89    3.08    3.27    2.31
    0404      0.67    1.40    0.67    3.46    1.40    1.15

    Observed for locus : loc-5
    0404      8   8   5   7   9   7

    Expected for locus : loc-5
    0404      8.00    8.00    5.00    7.00    9.00    7.00
```

B.4 Files obtained from the "testing pairwise pop differentiation" option

```
             Twicke       Arms P       Millen       Lansdo       Murray
Stade       0.14467      0.02733      0.03467      0.21000      0.03733
Twicke                   0.31867      0.18667      0.40400      0.05733
Arms P                                0.27800      0.61067      0.53467
Millen                                             0.43733      0.29000
Lansdo                                                          0.95400

P-values obtained after :     1500 permutations
Indicative adjusted nominal level (5\%) for multiple comparisons is :   0.003333

             Twicke       Arms P       Millen       Lansdo       Murray
Stade            NS           NS           NS           NS           NS
Twicke                        NS           NS           NS           NS
Arms P                                     NS           NS           NS
Millen                                                  NS           NS
Lansdo                                                               NS
```

B.5 Example of result file from the comp. Among group of samples tab sheet file diploid-test.out of the distribution

```
file diploid-test.out of the distribution

*********************************************************************************************************************
*   The following results were generated the 03.08.2001 at 16:26:37 with \textsc{Fstat} for windows, V2.9.3 from file diploid2.dat. *
*********************************************************************************************************************
Samples of group   1 (G1) : pop1, pop2, pop3,
 Allelic richness:    1.772 Ho:    0.241 Hs:    0.273 Fis:    0.117 Fst:    0.013 Rel:    0.024 Relc:   -0.264
Samples of group   2 (G2) : pop4, pop5, pop6,
 Allelic richness:    1.621 Ho:    0.214 Hs:    0.197 Fis:   -0.083 Fst:    0.005 Rel:    0.011 Relc:    0.153

The two-sided P-values obtained after   5000 permutations are :
 Allelic Richness:  0.19440
               Ho:  0.80540
               Hs:  0.10380
              Fis:  0.70200
              Fst:  0.96200
              Rel:  0.96200
             Relc:  0.70200
```

B.6 Example of result file from the biased dispersal menu

```
File exampsb.res of the distribution

results are for data stored in filename: examp
The tests are two sided and based on 1000 randomisations.
Fis for M:  -0.0280; Fst for M:   0.2109
Fis for F:  -0.0045; Fst for F:   0.0769
Fis overall:  -0.0069; Fst overall:   0.1377
NB M:     27; mean assignment:    1.51030; var assignment:    4.11501
NB F:     22; mean assignment:   -1.85355; var assignment:   16.25338
p_value for assignment Ttest:   0.0020
p_value for variance assignment test:   0.0060
p_value for Fst test:   0.0020
p_value for Fis test:   0.7240
```

B.7 Example of result file from the multiple regression and partial mantel menu

```
File YANOM.RES of the distribution

Put here any comments you want data from the file :
E:\fstatdev\Fstat2.9.3\data\multiple regression\YANOM.ALL

P-values are given after   2000 randomizations.


The total sum of square for variable var0 (YANOM.GEN) is :30312.4219

  Correlation (Partial if # expl. variables > 1), Coefficient (Beta) and Sum of squares (SS) for the observed data:
  Variable     (Partial) Corr.           Beta             SS
-----------------------------------------------------------------
var1 (YANOM.ANT)            0.299551       0.029754      2719.9512
-----------------------------------------------------------------
Error sum of squares: 27592.4707

Percent of the variance explained by the model (R_squared):       8.97

Percent. point for absolute regression coefficients (2-sided) P(Beta)
 and percent. point for extra sums of squares P(SS)

  Variable   P(Beta)     P(SS)
-------------------------------
var1 (YANOM.ANT)      0.05      0.05
-------------------------------
Percent point for the error sum of square      0.05
```

Bibliography

F. Balloux. Easypop (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, 2001. URL `http://www.unil.ch/izea/softwares/easypop.html`.

F. Balloux, H. Brünner, N. Lugon-Moulin, J. Hausser, and J. Goudet. Microsatellites can be misleading: an empirical and simulation study. *Evolution*, 54:1414–1422, 2000. URL `http://www.unil.ch/popgen/research/reprints/ballouxetal_evolution_2000.pdf`.

F. Balloux and J. Goudet. Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology*, 11:771–783, 2002. URL `http://www.unil.ch/popgen/research/reprints/balloux&goudet_mec_2002.pdf`.

R. Chakraborty and H. Danker-Hopfe. *Statistical Methods in Biological and Medical Sciences*, chapter Analysis of population structure: A comparative study of different estimators of Wright's fixation indices. Elsevier Science Publishers, 1991.

M. Chapuisat, J. Goudet, and L. Keller. Microsatellites reveal high population viscosity and limited dispersal in the ant F*ormica paralugubris*. *Evolution*, 51:475–482, 1997. URL `http://www.unil.ch/popgen/research/reprints/chapuisatetal_evolution_1997.pdf`.

C.C. Cockerham. Variance of gene frequencies. *Evolution*, 23:72–84, 1969.

C.C. Cockerham. Analysis of gene frequencies. *Genetics*, 74:679–700, 1973.

C.C. Cockerham and B.S. Weir. Estimation of gene-flow from F-statistics. *Evolution*, 47:855–863, 1993.

A. El Mousadik and R.J. Petit. High level of genetic differentiation for allelic richness among populations of the argan tree [A*rgania spinosa* (l.) skeels] endemic to morocco. *Theoretical and Applied Genetics*, 92:832–839, 1996.

L. Excoffier. *Handbook of statistical genetics*, chapter Analysis of population subdivision. Wiley and Sons, Ltd, 2001.

L. Favre, F. Balloux, J. Goudet, and N. Perrin. Female-biased dispersal in the monogamous mammal C*rocidura russula*: evidence from field data and microsatellite patterns. *Proceeding of the Royal Society of London B*, 264:127–132, 1997. URL `http://www.unil.ch/popgen/research/reprints/favreetal_prsb_1997.pdf`.

R. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 12 edition, 1954.

S.J. Goodman. Rst calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and a determining their significance. *Molecular Ecology*, 6:881–885, 1997. URL `http://helios.bto.ed.ac.uk/evolgen/rst/rst.html`.

J. Goudet. *The genetics of geographically structured populations*. PhD thesis, University of Wales at bangor, 1993. URL `http://www.unil.ch/popgen/research/reprints/`.

J. Goudet. Fstat (vers.1.2): a computer program to calculate F-statistics. *Journal of Heredity*, 86:485–486, 1995. URL `http://www.unil.ch/izea/softwares/fstat.html`.

J. Goudet. An improved procedure for testing key innovations. *American Naturalist*, 53:549–555, 1999. URL `http://www.unil.ch/popgen/research/reprints/goudet_amnat_1999.pdf`.

J. Goudet, T. De Meeüs, A.J. Day, and C.J. Gliddon. *Genetics and Evolution of Aquatic Organisms*, chapter The different levels of population structuring of dogwhelks, *Nucella lapillus*, along the south Devon coast. Chapman and Hall, London, 1994. URL `http://www.unil.ch/popgen/research/reprints/`.

J. Goudet, N. Perrin, and P. Waser. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Molecular Ecology*, 11:1103–1114, 2002. URL `http://www.unil.ch/popgen/research/reprints/goudetetal_mec_2002.pdf`.

J. Goudet, M. Raymond, T. Demeeus, and F. Rousset. Testing differentiation in diploid populations. *Genetics*, 144:1933–1940, 1996. URL `http://www.unil.ch/popgen/research/reprints/goudetetal_genetics_1996.pdf`.

W.D. Hamilton. *Man and Beast: Comparative Social Behavior*, chapter Selection of selfish and altruistic behaviour in some extreme models, pages 57–91. Eisenberg and Dillon, washington, DC, 1971.

D.L. Hartl and A.G. Clarck. *Principles of Population Genetics*. Sinauer Associates, third edition, 1997.

P.W. Hedrick. *Genetics of Population*. Jones and Bartlett, second edition, 2000.

K. E. Holsinger, P. O. Lewis, and D. K. Dey. A bayesian method for analysis of genetic population structure with dominant marker data. *Molecular Ecology*, 11:1157–1164, 2002. URL `http://darwin.eeb.uconn.edu/hickory/hickory.html`.

S.H. Hurlbert. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52:577–586, 1971.

P. L'Ecuyer. Efficient and portable random number generators. *Communications of the ACM*, 31:147–157, 1988.

R.C. Lewontin and K. Kojima. The evolutionary dynamics of complex polymorphism. *Evolution*, 14:458–472, 1960.

B.J.F. Manly. *The Statistics of Natural Selection*. Chapman and Hall, 1985.

B.J.F. Manly. *Randomization and Monte Carlo methods in biology*. Chapman et Hall., second edition, 1997.

M. Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA*, 70:3321–3323, 1973.

M. Nei. *Molecular Population Genetics and Evolution*. Elsevier, 1975. URL `http://www.bio.psu.edu/People/Faculty/Nei/Lab/BOOK.pdf`.

M. Nei. *Molecular Evolutionary Genetics*. Columbia University Press, 1987.

M. Nei and R.K. Chesser. Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, 47:253–259, 1983.

R. D. M. Page. Treeview: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12:357–358, 1996. URL `http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`.

P. Pamilo. Genotypic correlation and regression in social groups: multiple alleles, multiple loci and subdivided populations. *Genetics*, 107:307–320, 1984.

P. Pamilo. Effect of inbreeding on genetic relatedness. *Hereditas*, 103:195–200, 1985.

E. Petit, F. Balloux, and J. Goudet. Sex biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. *Evolution*, 55:635–640, 2001. URL `http://www.unil.ch/popgen/research/reprints/petitetal_evolution_2001.pdf`.

R.J. Petit, A. El Mousadik, and O. Pons. Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, 12:844–855, 1998.

J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000. URL `http://pritch.bsd.uchicago.edu/`.

D.C. Queller and K.F. Goodnight. Estimating relatedness using genetic markers. *Evolution*, 42:258–275, 1989.

N. Raufaste and F. Rousset. Are partial mantel tests adequate? *Evolution*, 55: 1703–1705, 2001.

M. Raymond and F. Rousset. An exact test for population differentiation. *Evolution*, 49:1280–1283, 1995a.

M. Raymond and F. Rousset. Genepop (version 1.2): population genetics software for exact tests and oecumenicism. *Journal of Heredity*, 86:248–249, 1995b. URL `ftp://ftp.cefe.cnrs-mop.fr/genepop/`.

J. Reynolds, B.S. Weir, and C.C. Cockerham. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, 105:767–779, 1983.

W.R. Rice. Analysing tables of statistical tests. *Evolution*, 43:223–225, 1989.

F. Rousset. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142:1357–1362, 1996.

F. Rousset. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145:1219–1228, 1997.

S. Schneider, D. Roessli, and L. Excoffier. *Arlequin: A software for population genetics data analysis*. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, 2.000 edition, 2000. URL `http://lgb.unige.ch/arlequin/`.

M. Slatkin. A measure of population subdivision based on microsatellite allele frequency. *Genetics*, 139:457–462, 1995.

M. Slatkin and N.H. Barton. A comparison of three methods for estimating average levels of gene flow. *Evolution*, 43:1349–1368, 1989.

S. Trouvé, L. Degen, F. Renaud, and J. Goudet. Population structure of the freshwater snail L*ymnaea truncatula*: the importance and consequences of selfing. *Evolution*, 57:In Press, 2003. URL `http://www.unil.ch/popgen/research/reprints/trouveetal_evolution_2003_pp.pdf`.

J. Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160:1203–1215, 2002.

B.S. Weir. Inferences about linkage disequilibrium. *Biometrics*, 35:235–254, 1979.

B.S. Weir. *Genetic data analysis II*. Sinauer, second edition, 1996.

B.S. Weir and C.C. Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.

M.C. Whitlock and D. McCauley. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, 82:117–125, 1999.

S. Wright. Systems of mating. *Genetics*, 6:111–178, 1921.

S. Wright. *Evolution and the genetics of populations. II. The theory of gene frequencies*, volume 2. University of Chicago Press, 1969.

C. Zapata. The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. *Evolution*, 54:1809–1812, 2000.

C. Zapata, C. Carollo, and Rodriguez S. Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Annals of Human Genetics*, 65:395–406, 2001.

D. Zaykin, L. Zhivotovsky, and B.S. Weir. Exact tests for association between alleles at arbitrary numbers of loci. *Genetica*, 96:169–178, 1995.