

User guide

EASYPOP_{version 2.0.1}

A software for population genetics simulations

Author: François Balloux

Available at: <http://www.unil.ch/dee>

Department of Genetics
University of Cambridge
Downing Site
Downing Street
Cambridge CB2 3EH, UK
E-mail: fb255@gen.cam.ac.uk

July 2006

1. Introduction.....	3
1.1 Purpose of EASYPOP	3
1.1 Applications	3
1.3 What can EASYPOP 2.0.1 do	4
1.4 How does EASYPOP work?.....	5
2. Getting started.....	7
3. Input.....	8
3.1 Available memory	8
3.2 Simulation parameters	9
3.2.1 Haploids	9
3.2.2 Diploids (one sex).....	9
3.2.4 Haplodiploids	10
3.2.5 Number of populations.....	11
3.2.6 Number of individuals	11
3.2.7 Migration.....	11
3.2.8 Same migration scheme over all simulation?.....	13
3.2.9 Linkage.....	13
3.2.10 Mutation models	14
3.2.11 Starting variability	14
3.2.12 Partial simulation results	15
3.2.13 Retrieving the complete genealogy.....	15
3.2.13 Replicates	16
4. Output files.....	17
5. General problems	19
5.1 How to obtain EASYPOP	19
5.2 Debugging and troubleshooting	19
5.3 How to cite EASYPOP	19
6. References.....	20
7. Acknowledgements	20

1. Introduction

1.1 Purpose of EASYPOP

If there are many programs for population genetic data analysis, less effort have been devoted to simulation software. One reason is that until recently computers were not powerful enough to simulate multilocus datasets of a reasonable size. There are many situations where simulated data are useful. EASYPOP has been written to fill this gap. Some applications are listed below.

- Exploration of problems in population genetics too complex to be tracked analytically
- Construction of null hypothesis against which real data can be statistically tested; In population genetics, the null hypothesis is generally random mating. Sometimes it can be useful to generate alternative null hypotheses. Datasets simulated under various conditions allow testing data from real populations against simulated null hypothesis.
- Test of new population genetics tools; simulations allow working on controlled simulated datasets with numerous replicates
- Test population genetics software under non-trivial conditions
- Education for undergraduate and graduate students: various simulated datasets can be provided to exercise the treatment of population genetic data analysis

1.1 Applications

As examples of possible applications of EASYPOP, I have listed below a few references of papers that used the program.

Balloux F, W Amos and T Coulson 2004 Does heterozygosity estimate inbreeding in real populations. *MOL ECOL* **13**: 3021-3031

Balloux F, Lehmann L, de Meeus T. 2003. The population genetics of clonal and partially clonal diploids. *GENETICS* **164**:1635-1644

Hardy OJ, Charbonnel N, Freville H, et al. 2003. Microsatellite allele sizes: A simple test to assess their significance on genetic differentiation. *GENETICS* **163**: 1467-1482

Koskinen MT, Haugen TO, Primmer CR 2003 Contemporary fisherian life-history evolution in small salmonid populations
NATURE **419**: 826-830

Fisher MC, Rannala B, Chaturvedi V, et al. 2002 Disease surveillance in recombining pathogens: Multilocus genotypes identify sources of human *Coccidioides* infections P
NATL ACAD SCI USA **99**: 9067-9071

Goudet J, Perrin N, Waser P 2002 Tests for sex-biased dispersal using bi-parentally inherited genetic markers *MOL ECOL* **11**: 1103-1114

1.3 What can EASYPOP 2.0.1 do?

EASYPOP is an individual based model intended to simulate datasets under a very broad range of conditions. It can simulate haploid, diploid or haplodiploid data. For diploids there is the choice between hermaphrodites or sexuals. For hermaphrodites, the proportion of clonal reproduction and selfing can be chosen, whereas for sexuals, complex breeding structures can be simulated (e.g. monogamy with a given proportion of extra-pair matings). The number of individuals can be selected for each population and dispersal is sex-specific. There are various migration models such as two-dimensional stepping stone or hierarchical island model. In addition there is an isolation-by-distance option which works with the coordinates of the populations on any number of dimensions. There are also several mutation models implemented, which are particularly oriented on the simulation of microsatellite loci. Genotypes are real multilocus, (i.e. there are not independent replicates for each locus). All mutation parameters can be set individually for each locus.

EASYPOP is able to handle very large simulations on standard personal computers and is limited only by the memory of the machine. The computer code has been optimized for maximum speed. This allows running very large simulations on personal computers in a reasonable amount of time. In order to fit to analytical expectations in particular for variances, the functions implemented in EASYPOP are probabilistic and not deterministic. In other words, the simulations rely on the generation of random numbers.

This approach allows generating independent replicates, which are different realisation of a same expectation.

The inputs of the program are limited to the parameters chosen by the user for any simulation. The outputs are genotypes after a number of generations. These files can be directly analyzed either on FSTAT (Goudet 1996), GENEPOP (Raymond & Rousset 1996) or ARLEQUIN (Schneider *et al.* 1997). In addition there is a file which gives heterozygosities and F -statistics each ten generation. This file can be used to follow these quantities over time. From version 1.8.2 onwards, it is also possible to retrieve a file for the complete genealogy of the simulation.

1.4 How does EASYPOP work?

The basic procedure is built as a Markov chain. There are always two matrices, one for time t and for $t+1$. When a new generation starts, matrix t is full and matrix $t+1$ is empty. EASYPOP then randomly draws an individual from matrix t (a female if there are two sexes) irrespective of the population. For instance in an example with two populations of size 1000 (p_1) and 100 (p_2) respectively, the program draws a random number between 1 and 1100. If the random number drawn was 1003, this will correspond to an individual (female if two sexes) of the second population (p_2). In a simulation with sexual recombination, EASYPOP will also draw a mate (random number between 1 and 100, size of p_2). Then the routine creates a new individual, which has a chance of $1-m$ (m being the migration rate) of filling position 1 of p_2 in the matrix $t+1$ and a chance of m to get in p_1 . This procedure goes on until both populations are full. Sometimes an individual is allocated to a population, which is already full, and will then be discarded. When the matrix $t+1$ is complete it replaces matrix t . Generating individuals in that way allows avoiding the problems of replacement of immigrants by resident genotypes (or vice versa), since there is no order for filling the populations or creating the immigrant and resident genotypes, thus fitting to analytical expectations while keeping the number of individuals constant. The migration procedure is based on a matrix with the number of populations both in rows and columns, which contains the probabilities for migrating from population i to j . For instance the diagonal gives the probabilities for staying at home.

EASYPOP uses a lot of random numbers. In the present program they are drawn from the generator proposed by L'Ecuyer (1988). It combines two of the best multiplicative linear congruential Generators known and has passed all the tests for random number generators (Goudet 1995).

2. Getting started

There are two versions available of EASYPOP (v. 2.0.1). The first is for PC running under Windows the second and the second for the OSX exploitation system.



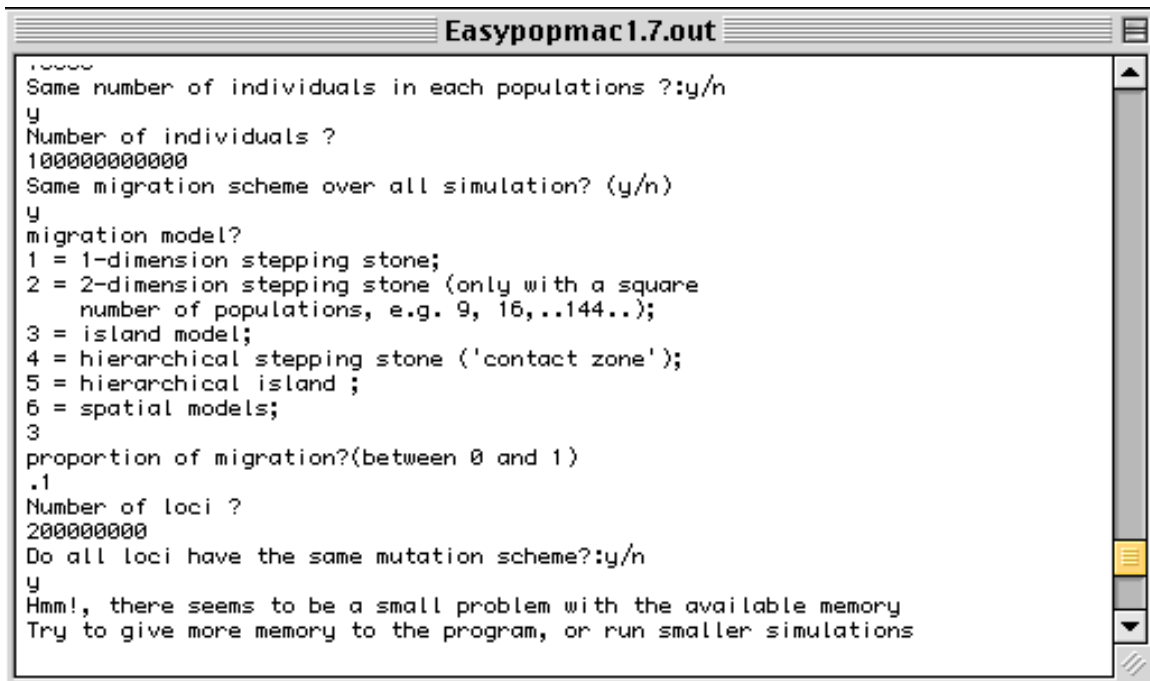
Fig.1. Interface window at opening of EASYPOP

EASYPOP is easy to install. Just put it where you want in your computer. The program will start by a double click on the icon. A window will open where all parameters for the simulation have to be specified (fig. 1). As the simulation runs, the interface window will give the replicate currently running and the number of generations left. At the end, the message "finished" will be displayed. The application can be quit at this moment. The user will then be asked if this window containing the simulation parameters should be saved (The result files are automatically saved!). All output files will be stored in the folder where EASYPOP has been running.

3. Input

3.1 Available memory

When running very large simulations (with a huge number of populations, individuals or loci), it can happen that the available memory of the computer is not sufficient. In this case, the error message displayed in fig. 2 will appear.



```

Easypopmac1.7.out
10000
Same number of individuals in each populations ?:y/n
y
Number of individuals ?
10000000000
Same migration scheme over all simulation? (y/n)
y
migration model?
1 = 1-dimension stepping stone;
2 = 2-dimension stepping stone (only with a square
   number of populations, e.g. 9, 16,..144..);
3 = island model;
4 = hierarchical stepping stone ('contact zone');
5 = hierarchical island ;
6 = spatial models;
3
proportion of migration?(between 0 and 1)
.1
Number of loci ?
200000000
Do all loci have the same mutation scheme?:y/n
y
Hmm!, there seems to be a small problem with the available memory
Try to give more memory to the program, or run smaller simulations

```

Fig.2. Error message when the memory requested for the simulations exceeds the available memory of the computer.

3.2 Simulation parameters

In the following section, the parameters that must be entered are explained. These parameters are the only inputs EASYPOP needs to fulfil the simulations. There is no extra data file needed.

3.2.1 Haploids

In this case, a proportion of recombination (sexual reproduction) between zero and one can be chosen. For organisms reproducing only clonally, recombination equals zero. In the output files .dat and .gen genotypes are coded as diploids (the same allele is repeated twice). This allows the direct analysis by the programs FSTAT and GENPOP. With haploids, most quantities given in the .equ file are meaningless (H_o , H_s , F_{is}). Nevertheless, H_t and F_{ST} are still appropriate measures of respectively gene diversity and the Wahlund effect.

3.2.2 Diploids (one sex)

Selfing: When using diploids “with one sex” (hermaphrodites), the user is asked the following question: “Random mating?:y/n”. The option random mating corresponds to a probability of selfing of $1/n_s$ where n_s stands for the number of individuals in population s . If the user answers “n” for no, he will be asked if there is any proportion of clonal reproduction (between zero and one). If the population does not reproduce completely clonally, the proportion of selfing (between 0 and 1) of individuals not born as clones can additionally be set.

3.2.3 Diploids (two sexes)

Social structure: When simulating data with two sexes, different social structures are proposed: random mating, monogamy and polygyny.

Random mating is the basic promiscuous mating system, with each male in a subpopulation having the same probability to be the father of any individual born in this patch.

Polygyny: In complete polygyny, only one male (the alpha male) is the father of all offspring born in the patch. A proportion of mating achieved by subordinate males can be selected. In this case mating performed by subordinates are assigned at random between all males of the patch (except the alpha male).

Monogamy: Under complete monogamy, offspring, which share the same mother, share also the same father. Only individuals involved in a pair do reproduce, if there are for instance ten females and one male per subpopulations, there will still be only one reproducing pair in each population. Some proportion of extra pair mating can be selected. Each time this type of mating occurs, all males, both those forming a pair and the unpaired males (the supernumerary males if the number of males exceeds number of females) have the same probability to be the father of the newly generated individual. Extrapair mating are also the only way for unpaired females (when number of females exceeds number of males) to participate to reproduction.

3.2.4 Haplodiploids

This option allows simulating haplodiploid organisms as social hymenopterans. The number of diploid queens, haploid males and diploid workers (non-reproductive females) can be chosen for each population. For social hymenopterans (simulations with workers), a population will represent a colony. In the output files .dat and .gen, the first lines represent the queens genotypes, the following the workers genotypes and finally the males. If the males are haploid, their genotypes are coded as diploids (the same allele is duplicated). This allows the analysis by FSTAT and GENPOP. As most quantities given in the file .equ are meaningless for haploid genotypes, heterozygosities and F -statistics are given only for females (queens and workers). The current version of EASYPOP allows only for single mated queens. It is further important to take into account that the different genotypes given in the output files belong to the same generation (ie queens, workers and males are brothers and sisters and not mothers, fathers and daughters).

3.2.5 Number of populations

The number of population has been arbitrarily limited to 10'000 in this version. When choosing the number of populations keep in mind the migration model you intend to use afterwards. For instance a two dimensional stepping-stone will work only with a square number of populations (e.g. 9, 16, 25...10'000).

3.2.6 Number of individuals

Any number of individuals can be selected for each population. Different numbers of males and females can be chosen in each subpopulation. This number of individuals (males and females) will be kept constant during all subsequent generations.

3.2.7 Migration

Migration in EASYPOP is zygotic and not gametic. In other words, individuals and not their gametes migrate. The difference between gametic and zygotic migration is however negligible in most cases (Nagylaki 1983). Migration rates are probabilities that an individual moves in a population different from the population where he has been born. In hierarchical models, the program will ask for different migration rates (e.g. within groups and between groups). As these migration rates are probabilities, the sum must not exceed unity. Different migration models are proposed:

- **One-dimensional stepping stone:** In this model the population are ordered along a line. m is the migration rate. In the n^{th} population a proportion $m/2$ will migrate in population $n-1$ and $m/2$ in $n+1$. The proportion of individuals staying at home is given as $(1-m)$. Migration in the borders leading outside the stepping stone is neglected.
- **Two-dimensional stepping stone:** This model is the extension of the previous one to two dimensions. As in the previous model migration events which lead outside the grid are neglected. Each population (excepted those on the border of the grid) are connected to 4 other populations. The proportion of individuals staying at home is given as $(1-m)$. Under this migration model, the populations are ordered as presented in figure 3 for a hypothetical simulation with 100 populations.

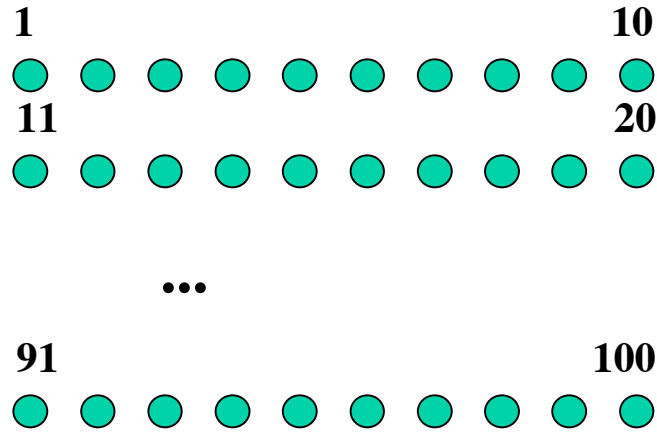


Fig.3. Ordering of populations in the two dimensional stepping stone for a hypothetical simulation with 100 populations

- **Island model:** This is Wright's classical island model. Any migrant has the same probability to reach any of the s subpopulations. The probability to migrate to any of the subpopulations is therefore $m(1/s-1)$. The probability for not migrating is given as $1-m$.
- **Hierarchical island model:** This model is an extension of the classical island model at two hierarchical levels. There are different groups of islands with a migration rate (m_1) within the group of island. In addition there is migration between islands (m_2). If migration occurs between groups of islands the probability of reaching any group is the same (independently of the size of the group). The probability of migrating for any individual is m_1+m_2 . As $1-(m_1+m_2)$ is the probability for not migrating, (m_1+m_2) must be smaller than one.
- **Hierarchical stepping stone:** This model is based on two stepping stone models, which are in contact at their border. Each stepping stone has a migration rate m_1 within, and there is a migration rate m_2 between the last population of group 1 and the first population of group 2. This model has been written to explore contact zones.
- **Spatial models:** This model allows working with isolation by distance models. First it needs the number of dimensions of the space. Then all population coordinates must be given. Distances between populations are then computed as $d_{i,j}$. The migration follows a negative exponential and is fitted for mean dispersal. Migration rate from population i to population j is then computed as $\exp^{-r*(d_{i,j})}$ with $r=1/\text{mean dispersal}$.

3.2.8 Same migration scheme over all simulation?

This option allows changing the migration scheme in a given simulation. It is possible to change both the migration model and migration rates. If the user decides to run simulations with two distinct migration schemes, he will be asked how many generations will be run under the first migration scheme. This option allows for instance letting populations reach equilibrium under panmixia, before imposing any model of population structure. This option also permits simulating population size reductions. Let for example run for n_1 generations an island model of 100 populations with 100 individuals, with a migration rate of 0.99. This will be equivalent to a panmictic population of 10000 individuals. Then select an island model (or any other) with no migration at all and let it go for n_2 generations. You will get 100 populations, which are the product of an immediate bottleneck of a factor 100. Variation in size of several populations can be simulated by the same rationale while using the hierarchical island model.

3.2.9 Linkage

From version 1.8 on, EASYPOP allows simulating linked loci in diploids (and for females in haplodiploids). If the simulation comprises more than one locus, the program first asks: "Free recombination between loci?:y/n". If "y" is selected the loci will behave completely independently. If "n" is selected, there will be some linkage between loci, the user is then further asked:

"Recombination rate between adjacent loci (ie between locus n and n+1)

The recombination rate must be comprised between 0.0 and 0.5 "

A recombination rate of 0.0 means that the loci are fully linked, and thus behave together as a single locus. A recombination rate of 0.5 is equivalent to free recombination. All values in between will represent various linkage intensities between adjacent pairs of loci. The recombination rate represents the probability that the allele at the next locus is randomly drawn.

3.2.10 Mutation models

Different mutation models are implemented, and are especially designed to take into account microsatellite loci. Different mutation models, mutation rates and number of possible allelic states can be selected for each individual locus.

- **Number of possible allelic states:** Most theoretical studies assume an infinite number of allelic states to avoid any homoplasy. If it facilitates analysis, this is however extremely unrealistic. In EASYPOP, the maximum possible number of allelic states must be stated. The highest number is 999. However care should be taken when planning to analyse data with GENEPOP, since this latter program does not accept more than 99 alleles. The latest versions of FSTAT (from 2.8) are able to deal with up to 999 alleles.
- **KAM:** In this model, mutation will generate a new allele of any possible allelic state at random. It is a special case of the infinite allele model limited to a finite number of allelic states.
- **SSM:** This is the single step mutation model limited to a finite number of allelic states. Mutation will depend on its previous allelic state; the new allele size is either increased or decreased by one with a probability of 50%. Mutations reaching outside the border will be neglected.
- **Mixed model SSM/KAM:** In this model, the probability that a mutation event follows a KAM process must be stated. For each mutation, a random number is drawn and tested against the proportion of KAM, then the mutation follows either a KAM or a SSM process.
- **Mixed model SSM/SSM2:** This model follows the same rationale than the previous one except that here KAM mutations are replaced by mutations increasing or decreasing the allelic state by two steps. Mutations reaching outside the border of possible allele sizes are set at the limit.

3.2.11 Starting variability

There are two options available for the starting population. Either, the user chooses “minimum variability” and in the initial generation all individuals share the same allele defined as the number of possible allelic states divided by two. The option “maximum

variability” will randomly distribute alleles drawn from all the possible allelic states to the genotypes of the first generation.

3.2.12 Partial simulation results

It is possible to choose between complete sampling of the simulation or a subset in the files .dat and .gen. Indeed, it can be useful to run simulations with a large number of populations and individuals, but only analyse a subsample. When selecting only some of the populations of the simulation, EASYPOP selects the **first ones**. The number of individuals, which can be subsampled is the same in each population. If the number of individuals selected for analysis is inferior to the real number in any population, EASYPOP automatically makes the correction by giving the actual number of individuals. The calculation of the file .equ is still computed on the complete dataset.

3.2.13 Retrieving the complete genealogy

From version 1.8.2 onwards, it is possible to retrieve the complete genealogy of the simulation. This can be very useful as it allows for computing expectations for identities by descent (e.g. inbreeding coefficients for any individuals, or pair-wise relatedness). These expectations can then be compared to estimates based on the genetic data obtained from the same simulations. The message on the screen is as follows:

“Do you want a file giving the complete pedigree of the simulation?: y/n
Please notice that this file can be very huge and will slow down simulations)”

The complete genealogy is stored in a file with “pdg” if the user chooses to retrieve the pedigree. The pedigree file has as many columns as individuals in the simulation times the number of generations so that a simulation of 100 individuals over 1,000 generations would have 100,000 lines. The file has nine columns. The first three columns give the information for each individual, respectively the generation, the subpopulation of origin and the label of the individual. The following three columns give information for the first parent (the mother for dioecious organisms), and the last three for the second parent.

Again these are respectively the generation ($x-1$ in this case) , the population and the individual's label. Each individual appears only once as a new individual within columns 1-3, but can occur several times in the last six columns as a parent. For dioecious simulations, The first slots in a subpopulation are always allocated to females and the following to males. For example in a simulation with a subpopulation including five males and five females, females would be labeled 1-5 and males 6-10.

3.2.13 Replicates

It is possible to compute up to 99 replicates. Since the seed of the random generator are modified for each replicate, they provide independent realisations of the same expectation.

4. Output files

There are four result files provided by the program. They are characterised by their different extensions (dat, gen, equ and pdg). Different replicates are recognisable by their associated number. For example a hypothetical simulation called “test_” with two replicates would provide six files (test_001.dat, test_002.dat, test_001.gen...). The use of the different files is given below:

- **Extension dat:** This file can be directly analysed with FSTAT (Goudet 1995). When working with two sexes, the first lines for each population are the female genotypes. The allelic phase is conserved, the first alleles of each single locus genotype correspond to chromosome 1 and the following to chromosome 2. Haploid genotypes are coded as diploids. This allows direct analysis by the program FSTAT. For additional information, see FSTAT user guide. Recent versions of this software allow analysing batches of replicates.
- **Extension gen:** This file can be analysed by GENPOP (Raymond & Rousset 1995), it is also recognized by the software ARLEQUIN (Schneider *et al.* 1997). The extension (.gen) must be eliminated before (just rename the file without extension). When working with two sexes, the first lines for each population are the female genotypes. The allelic phase is conserved, the first alleles of each single locus genotype correspond to chromosome 1 and the following to chromosome 2. Haploid genotypes are coded as diploids. For additional information, see GENPOP or ARLEQUIN user guides.
- **Extension equ** Number of alleles, observed heterozygosity, Nei & Chesser’s (1983) unbiased H_s , H_t and F-statistics are given in this file for each 10th generation. This allows following the value of these quantities over time. In hierarchical migration models (hierarchical stepping-stone and hierarchical island model), the subscript s of H_s , F_s and F_{st} refers to the population. For haplodiploids, heterozygosities and F-statistics are computed on females only (queens and workers).

- **Extension pdg:** From version 1.8.2 onwards, it is possible to retrieve the complete genealogy of the simulation. As this file rapidly becomes very large even for reasonable parameters, it is optional.

5. General problems

5.1 How to get EASYPOP

OSX and PC versions of EASYPOP can be downloaded with its user guide from the following website:

<http://www.unil.ch/dee>

5.2 Debugging and troubleshooting

EASYPOP has been debugged as carefully as possible by testing simulation results against analytical expectations in all cases where they could be computed. If you suspect the presence of a bug, please feel free to contact me. Over time I have also developed several versions featuring options not included in the “official” version.

5.3 How to cite EASYPOP

Please cite EASYPOP as follows:

Balloux, F., EASYPOP (version 1.7), (2001) A computer program for the simulation of population genetics. *J. Heredity* **92**: 301-302.

6. References

Goudet, J., 1995. FSTAT: A computer program to calculate F-statistics. *J. Heredity* **86**: 485-486.

L'Ecuyer, P. 1988. Efficient and portable Random Number Generators. *Commun ACM* **31**: 147-157.

Nagylaki, T. 1983. The robustness of neutral models of geographical variation. *Theor. Popul. Biol.* **24**: 268-294.

Nei, M. & R. K. Chesser 1983. Estimation of fixation indices and gene diversities. *Ann Hum. Genet.* **47**: 253-259.

Raymond, M., & Rousset, F. 1995 GENEPOP: population genetics software for exact tests and ecumenicism. *J. Heredity* **86**: 248-249.

Schneider, S., J.-M. Kueffer, D. Roessli & L. Excoffier 1997. Aequin ver 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

7. Acknowledgements

It is a pleasure to thank Jean Lehman and Max Reuter who introduced me to programming in 1998. I also thank Jérôme Goudet, Eric Petit and Thierry De Meeûs for the time they invested into testing earlier versions of EASYPOP.