

ARLEQUIN

A software for
population genetic
data analysis



Copyright © 1995-98, L. Excoffier.

ver 1.1

ARLEQUIN ver 1.1

A software for population genetic data analysis

Authors:

Stefan Schneider, Jean-Marc Kueffer, David Roessli, and Laurent Excoffier

Genetics and Biometry Laboratory
Dept. of Anthropology and Ecology
University of Geneva
CP 511
1211 Geneva 24
Switzerland

E-mail : arlequin@sc2a.unige.ch

URL: <http://anthropologie.unige.ch/arlequin>

Table of contents:

| | | |
|----------|--|-----------|
| 1 | Introduction | 6 |
| 1.1 | <i>Why Arlequin?</i> | 6 |
| 1.2 | <i>Arlequin philosophy</i> | 6 |
| 1.3 | <i>About this manual</i> | 6 |
| 1.4 | <i>Data types handled by Arlequin</i> | 7 |
| 1.4.1 | <i>DNA sequences</i> | 8 |
| 1.4.2 | <i>RFLP Data</i> | 8 |
| 1.4.3 | <i>Microsatellite data</i> | 8 |
| 1.4.4 | <i>Standard data</i> | 8 |
| 1.4.5 | <i>Allele frequency data</i> | 9 |
| 1.5 | <i>Methods implemented in Arlequin</i> | 9 |
| 1.6 | <i>System requirements</i> | 10 |
| 1.7 | <i>Installing and uninstalling Arlequin</i> | 10 |
| 1.8 | <i>List of files included in the Arlequin package</i> | 11 |
| 1.9 | <i>Arlequin limitations</i> | 12 |
| 1.10 | <i>How to cite Arlequin</i> | 12 |
| 1.11 | <i>Acknowledgements</i> | 12 |
| 1.12 | <i>Bug report and comments</i> | 12 |
| 1.13 | <i>How to get the last version of the Arlequin software?</i> | 12 |
| 1.14 | <i>What is new in version 1.1 compared to version 1.0</i> | 13 |
| 1.15 | <i>Forthcoming developments</i> | 13 |
| 1.16 | <i>Remaining problem</i> | 14 |
| 2 | Getting started | 15 |
| 2.1 | <i>Preparing input files</i> | 15 |
| 2.2 | <i>Loading project files into Arlequin</i> | 15 |
| 2.3 | <i>Selecting analyses to be performed on your data</i> | 15 |
| 2.4 | <i>Creating and using Setting Files</i> | 15 |
| 2.5 | <i>Performing the analyses</i> | 16 |
| 2.6 | <i>Stopping the computations</i> | 16 |
| 2.7 | <i>Consulting the results</i> | 16 |
| 3 | Input files | 17 |
| 3.1 | <i>Format of Arlequin input files</i> | 17 |
| 3.2 | <i>Project file structure</i> | 17 |
| 3.2.1 | <i>Profile section</i> | 17 |
| 3.2.2 | <i>Data section</i> | 19 |
| 3.2.2.1 | <i>Haplotype list (optional)</i> | 19 |
| 3.2.2.2 | <i>Distance matrix (optional)</i> | 20 |
| 3.2.2.3 | <i>Samples</i> | 21 |
| 3.2.2.4 | <i>Genetic structure</i> | 23 |
| 3.3 | <i>Eexample of an input file</i> | 24 |
| 3.4 | <i>Automatically creating the outline of a project file</i> | 25 |
| 3.5 | <i>Conversion of data files</i> | 26 |
| 3.6 | <i>Arlequin batch files</i> | 27 |
| 4 | Output files | 28 |
| 4.1 | <i>Result file</i> | 28 |

| | |
|---|-----------|
| 4.2 View your results in HTML browser | 28 |
| 4.3 Arlequin Log file | 28 |
| 4.4 Back-up file | 29 |
| 4.5 Linkage Disequilibrium Result File | 29 |
| 4.6 Variance components null distribution histograms | 29 |
| 5 Examples of input files | 30 |
| 5.1 Example of allele frequency data | 30 |
| 5.2 Example of standard data (Genotypic data, unknown gametic phase, recessive alleles) | 30 |
| 5.3 Example of DNA sequence data (Haplotypic) | 31 |
| 5.4 Example of microsatellite data (Genotypic) | 32 |
| 5.5 Example of RFLP data(Haplotypic) | 33 |
| 5.6 Example of standard data (Genotypic data, known gametic phase) | 34 |
| 6 Arlequin interface | 35 |
| 6.1 Menus | 35 |
| 6.1.1 File Menu | 35 |
| 6.1.2 Edit Menu | 35 |
| 6.1.3 Project Menu | 35 |
| 6.1.4 Setup Menu | 37 |
| 6.1.5 Special Menu | 37 |
| 6.1.6 Window Menu | 37 |
| 6.1.7 Help Menu | 38 |
| 6.2 Toolbar | 38 |
| 6.3 Status Bar | 39 |
| 6.4 Dialog boxes | 39 |
| 6.4.1 General Settings | 39 |
| 6.4.2 Diversity indices | 41 |
| 6.4.3 Neutrality tests | 43 |
| 6.4.4 Gametic disequilibrium | 44 |
| 6.4.5 Genetic structure | 46 |
| 6.4.6 Launch Pad | 48 |
| 7 Methodological outlines | 50 |
| 7.1 Intra-population level methods | 51 |
| 7.1.1 Standard diversity indices | 51 |
| 7.1.1.1 Gene diversity | 51 |
| 7.1.1.2 Number of usable loci | 51 |
| 7.1.1.3 Number of polymorphic sites (S) | 51 |
| 7.1.2 Molecular indices | 51 |
| 7.1.2.1 Mean number of pairwise differences (π) | 51 |
| 7.1.2.2 Nucleotide diversity or average gene diversity over L loci (RFLP and DNA data) | 52 |
| 7.1.2.3 Theta estimators | 52 |
| 7.1.2.3.1 Theta(Hom) | 52 |
| 7.1.2.3.2 Theta(S) | 53 |
| 7.1.2.3.3 Theta(k) | 53 |
| 7.1.2.3.4 Theta(π) | 54 |
| 7.1.2.4 Mismatch distribution | 54 |
| 7.1.2.5 Estimation of genetic distances between DNA sequences | 55 |
| 7.1.2.5.1 Pairwise difference | 56 |

| | |
|---|-----------|
| 7.1.2.5.2 Percentage difference | 56 |
| 7.1.2.5.3 Jukes and Cantor | 56 |
| 7.1.2.5.4 Kimura 2-parameters | 57 |
| 7.1.2.5.5 Tamura | 57 |
| 7.1.2.5.6 Tajima and Nei | 58 |
| 7.1.2.5.7 Tamura and Nei | 58 |
| 7.1.2.6 Estimation of genetic distances between RFLP haplotypes | 59 |
| 7.1.2.6.1 Number of pairwise difference | 59 |
| 7.1.2.6.2 Proportion of difference | 60 |
| 7.1.2.7 Estimation of distances between Microsatellite haplotypes | 60 |
| 7.1.2.7.1 No. of different alleles | 60 |
| 7.1.2.7.2 Sum of squared size difference | 60 |
| 7.1.2.8 Estimation of distances between Standard haplotypes | 61 |
| 7.1.2.8.1 Number of pairwise differences | 61 |
| 7.1.3 Haplotype frequency estimation | 61 |
| 7.1.3.1 Haplotypic data or Genotypic data with known Gametic phase | 61 |
| 7.1.3.2 Genotypic data with unknown Gametic phase | 61 |
| 7.1.4 Linkage disequilibrium between pairs of loci | 62 |
| 7.1.4.1 Exact test of linkage disequilibrium (haplotypic data) | 62 |
| 7.1.4.2 Likelihood ratio test of linkage disequilibrium (genotypic data, gametic phase unknown) | 64 |
| 7.1.4.3 Measures of gametic disequilibrium (haplotypic data) | 64 |
| 7.1.5 Hardy-Weinberg equilibrium. | 65 |
| 7.1.6 Neutrality tests. | 66 |
| 7.1.6.1 Ewens-Watterson homozygosity test | 66 |
| 7.1.6.2 Ewens-Watterson-Slatkin exact test | 66 |
| 7.1.6.3 Chakraborty's test of population amalgamation | 67 |
| 7.1.6.4 Tajima's test of selective neutrality | 67 |
| 7.2 <i>Inter-population level methods</i> | 67 |
| 7.2.1 Population genetic structure inferred by analysis of variance (AMOVA) | 67 |
| 7.2.1.1 Haplotypic data, one group of populations | 69 |
| 7.2.1.2 Haplotypic data, several groups of populations | 70 |
| 7.2.1.3 Genotypic data, one group of populations, no within- individual level | 70 |
| 7.2.1.4 Genotypic data, several groups of populations, no within- individual level | 71 |
| 7.2.1.5 Genotypic data, one population, within- individual level | 72 |
| 7.2.1.6 Genotypic data, one group of populations, within- individual level | 72 |
| 7.2.1.7 Genotypic data, several groups of populations, within- individual level | 73 |
| 7.2.2 Population pairwise genetic distances | 73 |
| 7.2.3 Exact tests of population differentiation | 75 |
| 8 Appendix | 76 |
| 8.1 <i>Overview of input file keywords</i> | 76 |
| 9 References | 79 |

1 INTRODUCTION

1.1 Why Arlequin?

Arlequin is the French translation of "Arlecchino", a famous character of the Italian "Commedia dell'Arte". As a character he has many aspects, but he has the ability to switch among them very easily according to its needs and to necessities. This polymorphic ability is symbolized by his colorful costume, from which the Arlequin icon was designed.

1.2 Arlequin philosophy

The goal of Arlequin is to provide the average user in population genetics with quite a large set of methods and statistical tests, in order to extract information on genetic and demographic features of a collection of population samples.

The graphical interface has been designed such as to allow the user to rapidly select the different analyses he wants to perform on his data. We felt important to be able to explore the data, to analyze several times the same data set from different perspectives, with different selected options.

The statistical tests implemented in Arlequin have been chosen such as to minimize hidden assumptions and to be as powerful as possible. Thus, they often take the form of either permutation tests or exact tests, with some exceptions.

Finally, we wanted Arlequin to be able to handle genetic data under many different forms, and to try to carry out the same types of analyses irrespective of the format of the data.

Because Arlequin has a rich set of features and many options, it means that the user has to spend some time in learning them. However, we hope that the learning curve will not be that steep.

Arlequin is made available free of charge, as long as we have enough local resources to support the development of the program.

1.3 About this manual

The main purpose of this manual is to allow you to use Arlequin on your own, in order to limit as far as possible e-mail exchange with us.

In this manual, we have tried to provide a description of

1. the data types handled by Arlequin
2. the way these data should be formatted before the analyses
3. the graphical interface
4. the impact of different options on the computations
5. methodological outlines describing which computations are actually performed by Arlequin.

Even though this manual contains the description of some theoretical aspects, it should not be considered as a textbook in basic population genetics. We strongly recommend you to consult the original references provided with the description of a given method if you are in doubt with any aspect of the analysis.

1.4 Data types handled by Arlequin

Arlequin can handle several types of data either in *haplotypic* or *genotypic* form. The basic data types are:

- DNA sequences
- RFLP data
- Microsatellite data
- Standard data
- Allele frequency data

By *haplotypic form* we mean that genetic data can be presented under the form of haplotypes (i.e. a combination of alleles at one or more loci). This haplotypic form can result from the analyses of haploid genomes (mtDNA, Y chromosome, prokaryotes), or from diploid genomes where the gametic phase could be inferred by one way or another. Note that allelic data are treated here as a single locus haplotype.

Ex 1: haplotypic RFLP data : 100110100101001010

Ex 2: haplotypic standard HLA data : DRB1*0101 DQB1*0102 DPB1*0201

By *genotypic form*, we mean that genetic data is presented under the form of diploid genotypes (i.e. a combination of pairs of alleles at one or more loci).

Ex1: genotypic DNA sequence data:

ACGGCATTTAAGCATGACATACGGATTGACA

ACGGGATTTTAGCATGACATTCGGATAGACA

Ex 2: genotypic Microsatellite data :

63 24 32

62 24 30

The gametic phase of a multi-locus genotype may be either known or unknown. If the gametic phase is known, the genotype can be considered as made up of two well-defined haplotypes. For genotypic data with unknown gametic phase, you can consider the two alleles present at each locus as codominant, or you can allow for the presence of a recessive allele. This gives finally four possible forms of genetic data:

- Haplotypic data,
- Genotypic data with known gametic phase,
- Genotypic data with unknown gametic phase (no recessive alleles)
- Genotypic data with unknown gametic phase (recessive alleles).

1.4.1 DNA sequences

DNA sequences of arbitrary length can be accommodated by Arlequin. Each nucleotide is considered as a distinct locus. The four nucleotides "C", "T", "A", "G" are considered as unambiguous alleles for each locus, and the "-" is used to indicate a deleted nucleotide. Usually the question mark "?" codes for an unknown nucleotide.

The following notation for ambiguous nucleotides are also recognized:

R: A/G (purine)

Y: C/T (pyrimidine)

M: A/C

W: A/T

S: C/G

K: G/T

B: C/G/T

D: A/G/T

H: A/C/T

V: A/C/G

N: A/C/G/T

1.4.2 RFLP Data

RFLP haplotypes of arbitrary length can be handled by Arlequin. Each restriction site is considered as a distinct locus. The presence of a restriction site should be coded as a "1", and its absence as a "0". The "-" character should be used to denote the deletion of a site, not its absence due to a point mutation.

1.4.3 Microsatellite data

The raw data consist here of the allelic state of one or an arbitrary number of microsatellite loci. For each locus, one should in principle provide the number of repeats of the microsatellite motif as the allelic definition, if one wants his data to be analyzed according to the step-wise mutation model (for the analysis of genetic structure). It may occur that the absolute number of repeats is unknown. If the difference in length between amplified products is the direct consequence of changes in repeat numbers, then the minimum length of the amplified product could serve as a reference, allowing to code the other alleles in terms of additional repeats as compared to this reference. If this strategy is impossible, then any other number could be used as an allelic code, but the step-wise mutation model could not be assumed for these data.

1.4.4 Standard data

Data for which the molecular basis of the polymorphism is not particularly defined, or when different alleles are considered as mutationally equidistant from each other. Standard data haplotypes are thus compared for their content at each locus, without taking special care about the nature of the alleles, which can be either similar or different. For instance, HLA data (human MHC) enters the category of standard data.

1.4.5 Allele frequency data

The raw data consist of only allele frequencies (mono-locus treatment), so that no haplotypic information is needed for such data. Population samples are then only compared for their allelic frequencies.

1.5 Methods implemented in Arlequin

The analyses Arlequin can perform on the data fall into two main categories: intra-population and inter-population methods. In the first category statistical information is extracted independently from each population, whereas in the second category, samples are compared to each other.

| <i>Intra-population methods:</i> | <i>Short description:</i> |
|---|---|
| Standard indices | Some diversity measures like the number of polymorphic sites, gene diversity. |
| Molecular diversity | Calculates several diversity indices like nucleotide diversity, different estimators of the population parameter θ . |
| Mismatch distribution | The distribution of the number of pairwise differences between haplotypes based on computed inter-haplotypic distances. |
| Haplotype frequency estimation | Estimates the frequency of haplotypes present in the population either by gene counting or by the maximum likelihood method, depending on the type of data (haplotypic or genotypic). |
| Linkage disequilibrium | Test of non-random association of alleles at different loci. |
| Hardy-Weinberg equilibrium | Test of non-random association of alleles within diploid individuals. |
| Tajima's neutrality test (infinite site model) | Test of the selective neutrality of a random sample of DNA sequences or RFLP haplotypes under the infinite site model. |
| Ewens-Watterson neutrality test (infinite allele model) | Tests of selective neutrality based on Ewens sampling theory under the infinite alleles model. |
| Chakraborty's amalgamation test (infinite allele model) | A test of selective neutrality and population homogeneity. This test can be used when sample heterogeneity is suspected. |
| <i>Inter-population methods:</i> | <i>Short description:</i> |
| Search for shared haplotypes between populations | Comparison of population samples for their haplotypic content. All the results are then summarized in a table. |
| AMOVA | Different hierarchical Analyses of MOlecular VAriance to evaluate the amount of population genetic structure. |
| Pairwise genetic distances | F_{ST} based genetic distances for short divergence time. |
| Exact test of population differentiation | Test of non-random distribution of haplotypes into population samples under the hypothesis of panmixia. |

1.6 System requirements

- 486DX CPU, or higher
- Windows 95, Windows NT, or Win 3.1x with Win32s installed
- 8 MB RAM or more
- At least 4Mb free hard disk space

1.7 Installing and uninstalling Arlequin

- *Win 3.1 installation*
 1. Because Arlequin is a pure 32-bit program, you need to first install the Win32s libraries on your system. You can retrieve them from our Arlequin web site (<http://anthropologie.unige.ch/arlequin>) or from some other web site. To install these libraries, download the Win32s zip file, unzip it in a temporary directory and launch the Win32s setup application. It will install a 32-bit subsystem in your windows directories. The unzipped temporary files can be removed after installation.
 2. Then, unzip the Arlequin compressed file in the directory of your choice (e.g. c:\programs\Arlequin\). The zip file contains 32 bit libraries, the Arlequin executable files, and the example files.
 3. Use the file manager and the program manager to put *Arlequin.exe* in the program group of your choice.
 4. To launch Arlequin, simply click twice on *Arlequin.exe* icon.
- *Win 3.1 uninstallation*

Simply delete the Arlequin directory.
- *Win95 and WinNT installation*

Unzip the Arlequin compressed file in a temporary directory. Launch the setup file and follow its instructions, step by step. The setup file will create a directory for Arlequin executable files, example files, put the required system libraries in the appropriate directories, and create shortcuts in the explorer structure of your choice.

In case of an Arlequin update, please first uninstall the obsolete version of Arlequin.
- *Win95 and WinNT uninstallation*

Use the Win95 *Settings / Add-Remove programs* feature, and select Arlequin 1.1. It will automatically remove all files previously installed during Arlequin 1.1 setup.

1.8 List of files included in the Arlequin package

| Files | Description | Required for Arlequin to run properly |
|----------------------------|---|---------------------------------------|
| Arlequin files | | |
| <i>arlequin.exe</i> | Arlequin executable file | ✓ |
| <i>arlequin.hlp</i> | Arlequin help file | |
| <i>arlequin.cnt</i> | Arlequin help file organizer | |
| <i>arlequin.ini</i> | A file containing the description of the last custom settings defined by the user | |
| <i>arlequin.log</i> | A file containing the last warning and error messages issued by Arlequin | |
| <i>readme11.txt</i> | A text file containing a description of the last release of Arlequin | |
| Example files | | |
| <i>batch\batch_ex.arb</i> | <i>microsat\2popmic.arb</i> | <i>haplfreq\hla_7pop.arb</i> |
| <i>batch\amova1.arb</i> | <i>microsat\2popmic.ars</i> | <i>haplfreq\hla_7pop.ars</i> |
| <i>batch\amova1.ars</i> | <i>microsat\micdipl.arb</i> | <i>amova\amovahap.arb</i> |
| <i>batch\amova1mat.dis</i> | <i>microsat\micdipl.ars</i> | <i>amova\amovahap.ars</i> |
| <i>batch\genotsta.arb</i> | <i>microsat\micdipl2.arb</i> | <i>amova\amovadis.arb</i> |
| <i>batch\genotsta.ars</i> | <i>microsat\micdipl2.ars</i> | <i>amova\amovadis.ars</i> |
| <i>batch\microsat.arb</i> | <i>dna\mtdna_hv1.arb</i> | <i>amova\56hapdef.txt</i> |
| <i>batch\microsat.ars</i> | <i>dna\mtdna_hv1.ars</i> | <i>amova\amovadis.dis</i> |
| <i>batch\missdata.arb</i> | <i>dna\nucl_div.arb</i> | <i>disequil\hwequil.arb</i> |
| <i>batch\missdata.ars</i> | <i>dna\nucl_div.ars</i> | <i>disequil\hwequil.ars</i> |
| <i>batch\phenohla.arb</i> | <i>neutrst\chak_tst.arb</i> | <i>disequil\ld_gen0.arb</i> |
| <i>batch\phenohla.ars</i> | <i>neutrst\chak_tst.ars</i> | <i>disequil\ld_gen0.ars</i> |
| <i>batch\relfreq.arb</i> | <i>neutrst\ew_watt.arb</i> | <i>disequil\ld_gen1.arb</i> |
| <i>batch\relfreq.ars</i> | <i>neutrst\ew_watt.ars</i> | <i>disequil\ld_gen1.ars</i> |
| <i>freqncy\cohen.arb</i> | | <i>disequil\ld_hap.arb</i> |
| <i>freqncy\cohen.ars</i> | | <i>disequil\ld_hap.ars</i> |
| System libraries | | |
| <i>owl501f.dll</i> | Dynamic link library installed in your system directory | ✓ |
| <i>bds501f.dll</i> | Dynamic link library installed in your system directory | ✓ |
| <i>cw3220.dll</i> | Dynamic link library installed in your system directory | ✓ |

1.9 Arlequin limitations

The amount of data that Arlequin can handle mostly depends on the memory available on your computer. However, a few parameters are limited to values within the range shown below.

| Portions of Arlequin concerned by the limitations | Limited parameter | Maximum value |
|--|---------------------------------|---------------|
| All | Number of population samples | 1000 |
| All | Number of groups of populations | 1000 |
| Ewens-Watterson and Chakraborty's neutrality tests | Sample size | 2000 |
| Ewens-Watterson and Chakraborty's neutrality tests | Number of haplotypes | 1000 |

1.10 How to cite Arlequin

Stefan Schneider, Jean-Marc Kueffer, David Roessli, and Laurent Excoffier (1997) Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

1.11 Acknowledgements

This program has been made possible by Swiss NSF grants No. 32-37821-93 and No 32.047053.96

Many thanks to:

André Langaney, Yannis Michalakis, Thierry Pun, Monty Slatkin, Peter Smouse, Alicia Sanchez-Mazas, Isabelle Dupanloup, Giorgio Bertorelle, Michele Belledi, Evelyne Heyer, Erika Bucheli, Alex Widmer, Philippe Jarne, Frédérique Viard, Peter de Knijff and all the beta-testers of Arlequin

1.12 Bug report and comments

Please report any bug through the bug report form available on

<http://anthropologie.unige.ch/arlequin/bug-report.html>

Other comments and suggestions will be also appreciated and can be communicated to us using the same web page.

1.13 How to get the last version of the Arlequin software?

Arlequin will be updated regularly and can be freely retrieved on

<http://anthropologie.unige.ch/arlequin>

1.14 What is new in version 1.1 compared to version 1.0

New features:

- Many bug corrections (see the list on our web site: <http://anthropologie.unige.ch/arlequin/buglist.html>).
- The input file is screened before being actually processed, removing unnecessary characters that lead to some errors in the previous version.
- Search for shared haplotypes between populations.
- All results concerning the analysis of a specific project are now output into a separate directory having the same name as the selected project, but with the [.res] extension.
- Implementation of an exact test of population differentiation, based on haplotype or genotype sample frequencies.
- The results can now be translated in an HTML file and consulted in any web browser.
- The creation of new project outlines has been made easier by the use of a special dialog box.
- Tables of expected haplotype frequencies under the hypotheses of linkage equilibrium and Hardy-Weinberg hypothesis are now included in the result file.
- The molecular diversity indices can now be calculated also for standard data.
- The matrix of pairwise distances used to compute the molecular diversity can optionally be printed in the result file.
- The probability of the observed Tajima's *D* value is calculated.
- It is now possible to import data from the GenePop version 3.0 format. In version 1.0, we considered only Genepop version 1.0.
- The results of the permutation tests done in AMOVA are now given in three forms: P (rand < obs.), P (rand = obs.) , P (rand ≤ obs.) for each statistic.
- The number of loci is now only limited by the available amount of memory. In version 1.0, the number of loci was limited to 500. See 1.8 for more details.
- Harpending's raggedness index is now computed on relative counts instead of absolute counts.
- Hardy-Weinberg exact test is made impossible with recessive data, contrary to what was done in ver 1.0.

1.15 Forthcoming developments

- Improved user interface.
- Treatment of pure dominant data (RAPD, AFLP).
- Incorporation of additional population genetics methods.
Suggestions are welcome, but we only have one life...
- Development of a portable C++ version of Arlequin with a Java graphical interface, to allow its deployment across different platforms.

1.16 Remaining problem

Arlequin build-in text editor cannot handle more than 32 KB of text under Windows 95 or Win 3.1x. Therefore, large files appears as if truncated. Under WinNT Arlequin windows can show the content of files of size up to 2 Mb, but it still cannot edit files larger than 32 kb.

Solution: Look at your result files with another text editor able to handle large data files. Even if the result file appears truncated, it is not when you open it in a more capable text editor. If you have an HTML browser (Netscape, Internet Explorer,...) installed on your machine, you can ask Arlequin to load the results into the browser by selecting the menu item *Window / View Results in HTML Browser*. The first time you do this, you are asked to locate the browser on your hard disk.

2 GETTING STARTED

The first thing to do before running Arlequin for the first time is certainly to read the manual or consult the help file. They will provide you with most of the information you are looking for. So, take some time to read them before you seriously start analyzing your data.

2.1 Preparing input files

The first step for the analysis of your data is to prepare an input data file for Arlequin. This input file is called here a *project file*. As Arlequin is quite a versatile program able to analyze several data types, you have to include some information about the properties of your data in the project file together with the raw data.

There are two ways to create Arlequin projects:

- 1) You can start from scratch and use a text editor to define your data using reserved keywords.
- 2) You can use Arlequin's "Project Outline Wizard" by selecting the menu function *Project / Build Project Outline...* . This calls a special dialog box where you can specify the type of project outline that should be build. Once created, the project outline is loaded into Arlequin. The name of the target file should have a "*.arp" extension (for ARlequin Project).

2.2 Loading project files into Arlequin

Once the project file is built, you must load it into Arlequin. You can do this either by activating the menu *Project / Open*, or by clicking on the Open project button on the toolbar. The Arlequin project files must have the *.arp extension. If your project file is not valid, Arlequin will open the Arlequin Log file to help you pointing out the problems. For each Arlequin session, Arlequin creates a log file called *arlequin.log*, where warnings and error messages are issued. The log file also keeps track of all the operations performed during an Arlequin session.

If your project file is valid, its main properties will be shown in the Project information window.

At this point, you just have to choose which analyses to perform on your data. The Project information window can be visible or invisible by selecting the menu item *Project / View Project Info*.

2.3 Selecting analyses to be performed on your data

The different analyses that can be performed on your data are selected using the launch pad dialog box. The options of the different analyses are set using special dialog boxes accessible by clicking on appropriate toolbar buttons, or by selecting the *Setup* menus. See the chapter on the description of the different dialog boxes.

2.4 Creating and using Setting Files

By settings we mean any alternative choice that can be made when using Arlequin. As you can choose different types of analyses, as well as different options for each of these analyses, all these choices can be saved into setting files. These files generally take the same name as the project files, but with the extension *.ars. Setting files can be created at any time of your work by activating the menu *Project / Settings / Save Settings*.

Alternatively, if you activate the menu *Project / Settings / Enable Save Settings on Close*, the analyses and the options that were validated when closing a project file will be automatically saved in a setting file. These setting files are convenient when you want to repeat some analyses done previously, or when you want to make different types of computations on several projects, as it is possible using batch files (see section 3.6) giving you considerable flexibility on the analyses you can perform, and avoiding tedious and repetitive mouse-clicks.

2.5 Performing the analyses

The selected analyses can be performed either by clicking on the Run button of the toolbar or by selecting the *Project / Run* menu. If an error occurs in the course of the execution, Arlequin will write diagnostic information in the log file. If the error is not too severe, Arlequin will open the log file for you. If there is a memory error, Arlequin will shut down itself. In the latter case, you should consult the Arlequin log file *before* launching a new analysis in order to get some information on where or at which stage of the execution the problem occurred. The file *Arlequin.log* is located in Arlequin original directory.

2.6 Stopping the computation

The computations can be stopped at any time by pressing the Stop button on the toolbar. However, note that the results may be incorrect if the computations did not terminate normally.

For very large project files, you may have to wait for a few seconds before the calculations are stopped.

2.7 Consulting the results

When the calculations are over, Arlequin will create a result directory, which has the same name as the project file, but with the **.res* extension. This directory contains all the result files, particularly the main result file with the same name as the project file, but with the **.arf* extension. The main result file is viewed in Arlequin's build-in editor.

Caution: If the result file is larger than 32,000 bytes, the file may appear as truncated under Win 95 or Win 3.1x. In this case you should close the result file window without saving it and open the relevant **.arf* file with a text editor allowing to view large files.

If you have an HTML browser (Netscape, Internet Explorer,...) installed on your machine, you can ask Arlequin to load the results into the browser, either by selecting the menu item *Window / View Results in HTML Browser*, or by pressing the corresponding button on the toolbar.

3 INPUT FILES

3.1 Format of Arlequin input files

Arlequin input files are also called project files. The project files contain both descriptions of the properties of the data, as well as the raw data themselves. The project file may also refer to one or more external data files.

Note that comments beginning by a "#" character can be put anywhere in the Arlequin project file.

Everything that follows the "#" character will be ignored until an end of line character.

3.2 Project file structure

Input files are structured into two main sections with additional subsections that must appear in the following order:

- | | |
|-----------------------|-------------|
| 1) Profile section | (mandatory) |
| 2) Data section | (mandatory) |
| 2a) Haplotype list | (optional) |
| 2b) Distance matrix | (optional) |
| 2c) Samples | (mandatory) |
| 2d) Genetic structure | (optional) |

We now describe the content of each (sub-) section in more detail.

3.2.1 Profile section

The properties of the data must be described in this section. The beginning of the profile section is indicated by the keyword [Profile] (within brackets).

One must also specify

- *The title of the current project* (used to describe the current analysis)

Notation: **Title=**

Possible value: Any string of characters within double quotes

Example: `Title="An analysis of haplotype frequencies in 2 populations"`

- *The number of samples or populations present in the current project*

Notation: **NbSamples =**

Possible values: Any integer number between 1 and 1000.

Example: `NbSamples =3`

- *The type of data to be analyzed.* Only one type of data is allowed per project

Notation: **DataType =**

Possible values: DNA, RFLP, MICROSAT, STANDARD and FREQUENCY

Example: `DataType = DNA`

- *If the current project deals with haplotypic or genotypic data*

Notation: **GenotypicData** =

Possible values: 0 (haplotypic data), 1 (genotypic data)

Example: `GenotypicData = 0`

One can also optionally specify

- *The character used to separate the alleles at different loci (the locus separator)*

Notation: **LocusSeparator** =

Possible values: WHITESPACE, TAB, NONE, or any character other than "#", or the character specifying missing data.

Example: `LocusSeparator = TAB`

Default value: WHITESPACE

- *If the gametic phase of genotypes is known*

Notation: **GameticPhase** =

Possible values: 0 (gametic phase not known), 1 (known gametic phase)

Example: `GameticPhase = 1`

Default value: 1

- *If the genotypic data present a recessive allele*

Notation: **RecessiveData** =

Possible values: 0 (co-dominant data), 1 (recessive data)

Example: `RecessiveData =1`

Default value: 0

- *The code for the recessive allele*

Notation: **RecessiveAllele** =

Possible values: Any string of characters within double quotes. This string can be explicitly used in the input file to indicate the occurrence of a recessive homozygote at one or several loci.

Example: `RecessiveAllele = "xxx"`

Default value: "null"

- *The character used to code for missing data*

Notation: **MissingData** =

Possible values: A character used to specify the code for missing data, entered between single or double quotes.

Example: `MissingData = '$'`

Default value: '?'

- *If haplotype or phenotype frequencies are entered as absolute or relative values*

Notation: **Frequency** =

Possible values: ABS (absolute values), REL (relative values: absolute values will be found by multiplying the relative frequencies by the sample sizes)

Example: `Frequency = ABS`

Default value: ABS

- *If a distance matrix needs to be computed from the original data, when calculating genetic structure indices*

Notation: **CompDistMatrix** =

Possible values: 0 (use the distance matrix specified in the DistanceMatrix sub-section), 1 (compute distance matrix from haplotypic information)

Example: `CompDistMatrix = 1`

Default value: 0

- *The number of significant digits for haplotype frequency outputs*

Notation: **FrequencyThreshold** =

Possible values: A real number between 1e-2 and 1e-7

Example: `FrequencyThreshold = 0.00001`

Default value: 1e-5

- *The convergence criterion for the EM algorithm used to estimate haplotype frequencies and linkage disequilibrium from genotypic data*

Notation: **EpsilonValue** =

Possible values: A real number between 1e-7 and 1e-12.

Example: `EpsilonValue = 1e-10`

Default value: 1e-7

3.2.2 Data section

This section contains the raw data to be analyzed. The beginning of the profile section is indicated by the keyword [Data] (within brackets).

It contains several sub-sections:

3.2.2.1 Haplotype list (optional)

In this sub-section, one can define a list of the haplotypes that are used for all samples. This section is most useful in order to avoid repeating the allelic content of the haplotypes present in the samples. For instance, it can be tedious to write a full sequence of several hundreds of nucleotides next to each haplotype in each sample. It is much easier to assign an identifier to a given DNA sequence in the haplotype list, and then use this identifier in the sample data section. This way Arlequin will know exactly the DNA sequences associated to each haplotype.

However, this section is optional. The haplotypes can be fully defined in the sample data section.

An identifier and a combination of alleles at different loci (one or more) describe a given haplotype. The locus separator defined in the profile section must separate each adjacent allele from each other.

It is also possible to have the definition of the haplotypes in an external file. Use the keyword EXTERN followed by the name of the file containing the definition of the haplotypes. Read Example 2 to see how to proceed. If the file "*hapl_file.hap*" contains exactly what is between the braces of Example 1, the two haplotype lists are equivalent.

Example 1:

```
[[HaplotypeDefinition]] #start the section of Haplotype definition
  HaplListName="list1" #give any name you wish to this list
  HaplList={
    h1 A T      #on each line, the name of the haplotype is
    h2 G C      # followed by its definition.
    h3 A G
    h4 A A
    h5 G G
  }
```

Example 2:

```
[[HaplotypeDefinition]] #start the section of Haplotype definition
  HaplListName="list1" #give any name you wish to this list
  HaplList = EXTERN "hapl_file.hap"
```

3.2.2.2 Distance matrix (optional)

Here, a matrix of genetic distances between haplotypes can be specified. This section is here to provide some compatibility with earlier WINAMOVA files. The distance matrix must be a lower diagonal with zeroes on the diagonal. This distance matrix will be used to compute the genetic structure specified in the genetic structure section. As specified in AMOVA, the elements of the matrix should be squared Euclidean distances. In practice, they are an evaluation of the number of mutational steps between pairs of haplotypes.

One also has to provide the labels of the haplotypes for which the distances are computed. The order of these labels must correspond to the order of rows and columns of the distance matrix. If a haplotype list is also provided in the project, the labels and their order should be the same as those given for the haplotype list.

Usually, it will be much more convenient to let Arlequin compute the distance matrix by itself.

It is also possible to have the definition of the distance matrix given in an external file. Use the keyword EXTERN followed by the name of the file containing the definition of the matrix. Read Example 2 to see how to proceed.

Example 1:

```
[[DistanceMatrix]] #start the distance matrix definition section
  MatrixName= "none" # name of the distance matrix
  MatrixSize= 4      # size = number of lines of the distance matrix
  MatrixData={
    h1 h2 h3 h4 # labels of the distance matrix (identifier of the
    0.00000 # haplotypes)
    2.00000 0.00000
    1.00000 2.00000 0.00000
```

```

        1.00000  2.00000  1.00000  0.00000
    }

```

Example2:

```

[[DistanceMatrix]]      #start the distance matrix definition section
  MatrixName= "none"    # name of the distance matrix
  MatrixSize= 4         # size = number of lines of the distance matrix
  MatrixData= EXTERN "mat_file.dis"

```

3.2.2.3 Samples

In this obligatory sub-section, one defines the haplotypic or genotypic content of the different samples to be analyzed.

Each sample definition begins by the keyword `SampleName` and ends after a `SampleData` has been defined.

One must specify:

- *A name for each sample*

Notation: **SampleName** =

Possible values: Any string of characters within quotes.

Example: `SampleName= "A first example of a sample name"`

Note: This name will be used in the Structure sub-section to identify the different samples, which are part of a given genetic structure to test.

- *The size of the sample*

Notation: **SampleSize** =

Possible values: Any integer value.

Example: `SampleSize=732`

Note: For haplotypic data, the sample size is equal to the haploid sample size. For genotypic data, the sample size should be equal to the number of diploid individuals present in the sample. When absolute frequencies are entered, the size of each sample will be checked against the sum of all haplotypic frequencies will check. If a discrepancy is found, a *Warning message* is issued in the log file, and the sample size is set to the sum of haplotype frequencies. When relative frequencies are specified, no such check is possible, and the sample size is used to convert relative frequencies to absolute frequencies.

- *The data itself*

Notation: **SampleData** =

Possible values: A list of haplotypes or genotypes and their frequencies as found in the sample, entered within braces

Example:

```

SampleData={
  id1 1  ACGGTGTCGA
  id2 2  ACGGTGTCAG
  id3 8  ACGGTGCCAA

```

```

id4 10 ACAGTGTCAA
id5 1  GCGGTGTCAA
}

```

Note: The last closing brace marks the end of the sample definition. A new sample definition begins with another keyword `SampleName`.

FREQUENCY data type:

If the data type is set to `FREQUENCY`, one must only specify for each haplotype its identifier (a string of characters without blanks) and its sample frequency (either relative or absolute). In this case the haplotype should not be defined.

Example:

```

SampleData={
  id1      1
  id2      2
  id3      8
  id4     10
  id5      1
}

```

Haplotypic data

For all data types except `FREQUENCY`, one must specify for each haplotype its identifier and its sample frequency. If no haplotype list has been defined earlier, one must also define here the allelic content of the haplotype. The haplotype identifier is used to establish a link between the haplotype and its allelic content maintained in a local database.

Once a haplotype has been defined, it needs not be defined again. However the allelic content of the same haplotype can also be defined several times. The different definitions of haplotypes with same identifier are checked for equality. If they are found identical, a warning is issued in the log file. If they are found to be different at some loci, an error is issued and the program stops, asking you to correct the error.

For complex haplotypes like very long DNA sequences, one can perfectly assign different identifiers to all sequences (each having thus an absolute frequency of 1), even if some sequences turn out to be similar to each other. If the option *Infer Haplotypes from Distance Matrix* is checked in the General Settings dialog box, Arlequin will check whether haplotypes are effectively different or not. This is a good precaution when one tests the selective neutrality of the sample using Ewens-Watterson or Chakraborty's tests, because these tests are based on the observed number of effectively different haplotypes.

Genotypic data

For each genotype, one must specify its identifier, its sample frequency, and its allelic content. Genotypic data can be entered either as a list of individuals, all having an absolute frequency of 1, or as a list of genotypes with different sample frequencies. During the computations, Arlequin will compare all genotypes to all others and recompute the genotype frequencies.

The allelic content of a genotype is entered on two separate lines in the form of two pseudo-haplotypes.

Examples:

1):

```

Id1 2  ACTCGGGTTCGCGCGC  # the first pseudo-haplotype
      ACTCGGGCTCACGCGC  # the second pseudo-haplotype

```

2)

```

my_id 4      0 0 1 1 0 1
           0 1 0 0 1 1

```

If the gametic phase is supposed to be known, the pseudo-haplotypes are treated as truly defined haplotypes.

If the gametic phase is not supposed to be known, only the allelic content of each locus is supposed to be known. In this case an equivalent definition of the upper phenotype would have been:

```

my_id 4      0 1 1 0 0 1
           0 0 0 1 1 1

```

3.2.2.4 Genetic structure

The hierarchical genetic structure of the samples is specified in this optional sub-section. It is possible to define groups of populations. This subsection starts with the keyword `[[Structure]]`. The definition of a genetic structure is only required for AMOVA analyses.

One must specify:

- *A name for the genetic structure*

Notation: **StructureName** =

Possible values: Any string of characters within quotes.

Example: `StructureName= "A first example of a genetic structure"`

Note: This name will be used to refer to the tested structure in the output files.

- *The number of groups defined in the structure*

Notation: **NbGroups** =

Possible values: Any integer value.

Example: `NbGroups = 5`

Note: If this value does not correspond to the number of defined groups, then calculations will not be possible, and an error message will be displayed.

- *If we add the individual level in the variance analysis*

Notation: **IndividualLevel** =

Possible values: 0 (no) or 1 (yes)

Example: `IndividualLevel = 0`

Note: Default value: 0. The value 1 is only possible with genotypic data.

- *The group definitions*

Notation: **Group** =

Possible values: A list containing the names of the samples belonging to the group, entered within braces. Repeat this for as many groups you have in your structure. It is of course not allowed to put the same population in different groups.

Example (NbGroups=2):

```
Group ={
    population1
    population2
    population3
}
Group ={
    population4
    population5
}
```

3.3 Example of an input file

The following small example is a project file containing four populations. The data type is STANDARD genotypic data with unknown gametic phase.

```
[Profile]
  Title="Fake HLA data"
  NbSamples=4
  GenotypicData=1
  GameticPhase=0
  DataType=STANDARD
  LocusSeparator=WHITESPACE
  MissingData='?'

[Data]

[[Samples]]
  SampleName="A sample of 6 Algerians"
  SampleSize=6
  SampleData={
    1  1  1104  0200
        0700  0301
    3  3  0302  0200
        1310  0402
    4  2  0402  0602
        1502  0602
  }
  SampleName="A sample of 11 Bulgarians"
  SampleSize=11
  SampleData={
    1  1  1103  0301
        0301  0200
    2  4  1101  0301
        0700  0200
    3  1  1500  0502
        0301  0200
    4  1  1103  0301
        1202  0301
    5  1  0301  0200
        1500  0601
    6  3  1600  0502
```

```

                1301    0603
    }
    SampleName="A sample of 12 Egyptians"
    SampleSize=12
    SampleData={
        1    2    1104    0301
           1600    0502
        3    1    1303    0301
           1101    0502
        4    3    1502    0601
           1500    0602
        6    1    1101    0301
           1101    0301
        8    4    1302    0502
           1101    0609
        9    1    1500    0302
           0402    0602
    }
    SampleName="A sample of 8 French"
    SampleSize=8
    SampleData={
        219    1    0301    0200
           0101    0501
        239    2    0301    0200
           0301    0200
        249    1    1302    0604
           1500    0602
        250    3    1401    0503
           1301    0603
        254    1    1302    0604
    }
}

[[Structure]]

    StructureName="My population structure"
    NbGroups=2
    Group={
        "A sample of 6 Algerians"
        "A sample of 12 Egyptians"
    }
    Group={
        "A sample of 11 Bulgarians"
        "A sample of 8 French"
    }
}

```

3.4 Automatically creating the outline of a project file

In order to help you setting up quickly a project file, Arlequin can create the outline of a project file for you. In order to do this, use the **Project outline wizard** dialog box by activating the *Project / Build Project Outline* menu. A special dialog box will appear, allowing to quickly define which type of data you have and some specificities of the data.

- **Data file**

Specify the name of the target file (the new Arlequin project). It should have the extension “.arp”.

- **Data type**

Specify which **type of data** you want to analyze (DNA, RFLP, Microsat, Standard, or Frequency).

Specify if the data is under **genotypic** or **haplotypic** form.

Specify if the **gametic phase** is known (for genotypic data only).

Specify if there are **recessive alleles** (for genotypic data only)

- **Controls**

Specify the number of population samples defined in the project

Choose a **locus separator**

Specify the character coding for **missing data**

Specify the **code for the recessive allele**

- **Optional sections**

Specify if you want to include a global **list of haplotypes**

Specify if you want to include a predefined **distance matrix**

Specify if you want to include a **group structure**

By pressing the OK button, an empty outline of a project file will be created for you. It will be automatically loaded in Arlequin, and you can then paste your sample data.

Note that if you have large amount so f data, you should note use Arlequin text editor, due to its inability to handle large data files

3.5 Conversion of data files

By selecting the menu *File / File Conversions*, it is possible to translate data files from one format to the other. This might be useful for users already having data files set up for other data software packages. It is also possible to convert Arlequin data files into other formats.

The currently recognized data formats are:

- Arlequin ver. 1.1
- Gene Pop ver. 3.0,
- Biosys ver.1.0,
- Phylip ver. 3.5
- Mega ver. 1.0
- Win Amova ver. 1.55.

The translation procedure is simple:

1. Select your source file with the upper left *Browse* button.
2. Select the format of the source data file, as well as that of the target file.
3. A default name is automatically given to the target file, but you can change the target file name with the upper right *Browse* button, or directly edit its name in the edit field.
4. The file conversion is started by pressing on the central button ">>>>".
5. In some cases, you might be asked for some additional information, for instance if input data is fractionated in several input files (like in WinAmova).

These conversion routines were done on the basis of the description of the input file format found in the user manuals of each of aforementioned programs. The tests done with the example files given with these programs worked fine. However, the original reading procedures of the other software packages may be more tolerant than our owns, and some data may be impossible to convert. Thus, some small corrections will need to be done by hand, and we apologize for that.

3.6 Arlequin batch files

A large number of data files can be analyzed one after the other using batch files.

A batch file (having usually the *.arb* extension) is simply a text file having on each line the name of the project file that should be analyzed. The number of data files to be analyzed can be arbitrary large.

By selecting the menu item *Project / Run batch file...*, you open a dialog box that allows you to set up the parameters for running your batch file.

You can either use the same options for all project files by selecting *Use identical settings for all projects*, or use the setting file associated with each project file by selecting *Use associated settings for each project*. In the first case, the same analyses will be performed on all project files listed in the batch file. In the second case, you can perform different computations on each project file listed in the batch file, giving you much more flexibility on what should be done. However, it implies that setting files have been prepared previously, recording the analyses needing to be performed on the data, as well as the options of these analyses.

If the associated project file does not exist, the current settings are used.

Note that the batch file, the project files, and the setting files should all be in the same folder.

4 OUTPUT FILES

The output files are now all located in a special sub-directory, having the same name as your project, but with the ".res" extension. This has been done to structure your result files according to different projects. For instance, if your project file is called my_file.arf, then the result files will be in a sub-directory called [my_file.res]

4.1 Result file

The file containing all the results of the analyses just performed. By default, it has the same name than the Arlequin input file, with the extension (.arf) for Arlequin Result File. This file is opened in a text window at the end of each run.

If the option *Project / Output Files / Append Results* is checked, the results of the current computations are appended to the one of previous calculations, otherwise the results of previous analyses are erased, and only the last results are overwritten in the result file.

4.2 View your results in HTML browser

For very large result files or result files containing the product of several analyses, it may be of practical interest to view the results in an HTML browser. This can be simply done by activating the menu *Window / View Results in HTML Browser*. This will trigger a translation of your result file (.arf) into html formatted files. These files will then be loaded into your Internet browser. The first time you use this option, you are asked to locate the browser on your hard disk. This can be done through the *General Settings* dialog box (see section 6.4.1). The location of your browser is then stored in *arlequin.ini*.

It is also possible to have an automatic generation of the result files in HTML format by checking the option *Project / Output Files / Generate HTML Result File*. This might be useful when running a batch file.

In the web browser, the main window is divided in three panes.

1. The **upper pane** lists all runs that have been output in the result file.
2. The **lower left pane** lists the inter-population analyses (Genetic structure, Shared haplotypes) or the population samples that have been analyzed in the run selected in the upper pane.
3. The **lower right pane**, shows the results concerning the selected item in the lower left pane.

The results can be consulted in html form, once they have been created, even when Arlequin is not active. The relevant files (*.htm) are located in the result sub-directory of your project, together with the conventional result files (*.arf). There are several html files in this directory, but you should only load the file having the same name as your project, but with the .htm extension. The other html files are called from this master file.

4.3 Arlequin Log file

A file where run-time *WARNINGS* and *ERRORS* encountered during any phases of the current Arlequin session are issued. By default, the file has the name *Arlequin.log*. You should consult this file if you observe any warning or error message in your result file. If Arlequin has crashed then consult *Arlequin.log* **before** running

Arlequin again. It will probably help you in finding where the problem was located. The content of *Arlequin.log* can be emptied in the *Special | Empty Log file* menu of the main window.

4.4 Back-up file

The name of a back-up file build from the data read in the input file. By default, it has the same name than the Arlequin input file, with the extension (*.bac*). It also allows you to check if your data have been read properly.

4.5 Linkage Disequilibrium Result File

This file contains the results of pairwise linkage disequilibrium tests between all pairs of loci. By default, it has the name LK_DIS.XL. As suggested by its extension, this file can be read with MS-Excel without modification. A tabulator separates the columns.

4.6 Variance components null distribution histograms

Specifies the name of an output file where the histograms of the variance component null distributions are output. By default, the name is set to AMO_HIST.XL. This tabulated text file can be read directly by MS-Excel, for a graphical output of the distributions.

All values of the permuted statistics are found in files, having the same name as the project file, with *.va*, *.vb*, *.vc* and *.vd* for σ_a^2 , σ_b^2 , σ_c^2 , and σ_d^2 , respectively.

5 EXAMPLES OF INPUT FILES

5.1 Example of allele frequency data

The following example is a file containing FREQUENCY data. The allelic composition of the individuals is not specified. The only information we have are the frequencies of the alleles.

```
[Profile]
  Title="Frequency data"
  NbSamples=2
  GenotypicData=0
  DataType=FREQUENCY
[Data]
  [[Samples]]
    SampleName="Population 1"
    SampleSize=16
    SampleData= {
      000 1
      001 3
      002 1
      003 7
      004 4
    }
    SampleName="Population 2"
    SampleSize=23
    SampleData= {
      000 3
      001 6
      002 2
      003 8
      004 4
    }
  }
```

5.2 Example of standard data (Genotypic data, unknown gametic phase, recessive alleles)

```
[Profile]
  Title="Genotypic Data, Phase Unknown, 5 HLA loci"
  NbSamples=1
  GenotypicData=0
  DataType=STANDARD
  LocusSeparator=WHITESPACE
  MissingData='?'
  GameticPhase=0
  RecessiveData=1
  RecessiveAllele="xxx"
  FrequencyThreshold=0.00001
  EpsilonValue=0.000000001
[Data]
  [[Samples]]
    SampleName="Population 1"
    SampleSize=63
    SampleData={
      MAN0102  12  A33  Cw10  B70  DR1304  DQ0301
              A33  Cw10  B7801  DR1304  DQ0302
      MAN0103  22  A33  Cw10  B70  DR1301  DQ0301
              A33  Cw10  B7801  DR1302  DQ0501
      MAN0108  23  A23  Cw6   B35  DR1102  DQ0301
              A29  Cw7   B57  DR1104  DQ0602
```

```

        MAN0109    6    A30    Cw4    B35    DR0801    xxx
                   A68    Cw4    B35    DR0801    xxx
    }

```

5.3 Example of DNA sequence data (Haplotypic)

```

[Profile]
  Title="An example of DNA sequence data"
  NbSamples=3
  GenotypicData=0
  DataType=DNA
  LocusSeparator=NONE
[Data]
  [[Samples]]
    SampleName="Population 1"
    SampleSize=6
    SampleData= {
      000    3    GACTCTCTACGTAGCATCCGATGACGATA
      001    1    GACTGTCTGCGTAGCATAACGACGACGATA
      002    2    GCCTGTCTGCGTAGCATAGGATGACGATA
    }
    SampleName="Population 2"
    SampleSize=8
    SampleData= {
      000    1    GACTCTCTACGTAGCATAACGATGACGATA
      001    1    GACTGTCTGCGTAGCATAACGATGACGATA
      002    1    GCCTGTCTGCGTAGCATAACGATGACGATA
      003    1    GCCTGTCTGCCTAGCATAACGATCACGATA
      004    1    GCCTGTCTGCGTACCATAACGATGACGATA
      005    1    GCCTGTCCGCGTAGCGTACGATGACGATA
      006    1    GCCCGTGTGCGTAGCATAACGATGGCGATA
      007    1    GCCTGTCTGCGTAGCATGCGACGACGATA
    }
    SampleName="Population 3"
    SampleSize=6
    SampleData= {
      023    1    GCCTGTCTGCGTAGCATAACGATGACGGTA
      024    1    GCCTGTCTGCGTAGCGTACGATGACGATA
      025    1    GCCTGTCTGCGTAGCATAACGATGACGATA
      026    1    GCCTGTCCGCGTAGCATAACGGTGACGGTA
      027    1    GCCTGTCTGCGTGGCATAACGATGACGATG
      028    1    GCCTGTCTGCGTAGCATAACGATGACGATA
    }
  }

```

5.4 Example of microsatellite data (Genotypic)

```

[Profile]
  Title="A small example of microsatellite data"
  NbSamples=4
  GenotypicData=1
  DataType=MICROSAT
  LocusSeparator=WHITESPACE
  CompDistMatrix=1
[Data]
  [[Samples]]
    SampleName="MICR1"
    SampleSize=28
    SampleData=
      {
        1      27      12 23
                13 22
        40     1       15 22
                13 22
      }
    SampleName="MICR2"
    SampleSize=59
    SampleData=
      {
        1      37      12 24
                12 22
        17     1       15 20
                13 22
        6      21      14 22
                14 23
      }
    SampleName="MICR3"
    SampleSize=30
    SampleData=
      {
        1      17      12 21
                13 22
        10     1       12 20
                13 23
        6      12      10 22
                12 22
      }
    SampleName="MICR4"
    SampleSize=16
    SampleData=
      {
        1      15      13 24
                13 23
        9      1       12 24
                13 23
      }
  [[Structure]]
    StructureName="Test microsat structure"
    NbGroups=2
    IndividualLevel=0
    Group={
      "MICR1"
      "MICR2"
    }
    Group={
      "MICR3"
      "MICR4"
    }
  }

```

5.5 Example of RFLP data(Haplotypic)

```
[Profile]
  Title="A small example of RFLP data: 3 populations"
  NbSamples=3
  GenotypicData=0
  DataType=RFLP
  LocusSeparator=WHITESPACE
  CompDistMatrix=1
  MissingData='?'
[Data]
  [[HaplotypeDefinition]]
    HaplListName="A fictive list of RFLP haplotypes"
    HaplList= {
      1      000011100111010011011001001011001101110100101101100
      2      100011100111010011011001001011001101110100101100100
      6      000011100111010010011001001011001101110100101101100
      7      100011100111010011011001001011001101110100101101100
      8      000011100111010011011001001001001101110100101101100
      11     000001100111011011011011001001011001101110100101111100
      12     000011100111010011011001101011001101110100101101100
      17     000011100111010011011001001011001100110100101101100
      22     000011100111011011011001001011001101110100101100100
      36     000011100111010011011001001010001100110100101101100
      37     000011100111011011011001001111001101110100101100100
      38     000111100111010011011001001011001101110100101101100
      40     000011100111000011011001001011001101110100101101100
      47     000011100111010011011001001011001101110100101100100
      139    000011100111010011011001001011001111110100101001110
      140    000011100111010011011001001011001101110100101100101
      141    000011100111010010011001000011001101110100101100100
    }
  [[Samples]]
    #1
    SampleName="pop 1"
    SampleSize=28
    SampleData= {
      1      27
      40     1
    }
    #2
    SampleName="pop 2"
    SampleSize=75
    SampleData= {
      1      37
      17     1
      6      21
      7      1
      2      1
      22     5
      11     2
      36     1
      139    1
      47     1
      140    1
      141    1
      37     1
      38     1
    }
    #3
```

```

    SampleName="pop 3"
    SampleSize=48
    SampleData=      {
        1      46
        8      1
        12     1
    }
[[Structure]]
    StructureName="A single group of 3 samples"
    NbGroups=1
    Group={
        "pop 1"
        "pop 2"
        "pop 3"
    }

```

5.6 Example of standard data (Genotypic data, known gametic phase)

```

[Profile]
    Title="An example of Genotypic data with known Gametic phase"
    NbSamples=3
    GenotypicData=1
    GameticPhase=1
    RecessiveData=0
    DataType=STANDARD
    LocusSeparator=WHITESPACE
[Data]
[[Samples]]
    SampleName="pop1"
    SampleSize=20
    SampleData=      {
        1      4      A  D
                          B  C
        3      5      A  B
                          A  A
        5      3      B  B
                          B  A
        7      8      D  C
                          D  C
    }
    SampleName="pop2"
    SampleSize=10
    SampleData=      {
        8      5      A  C
                          C  B
        9      5      B  C
                          D  B
    }
    SampleName="pop3"
    SampleSize=15
    SampleData=      {
        10     3      A  D
                          C  A
        4      12     A  C
                          B  B
    }

```

6 ARLEQUIN INTERFACE

6.1 Menus

6.1.1 File Menu

The menu by which you manipulate files input and output.

- New Open an empty text edit window. You can then save the content of the active window with *File / save as...*
- Open Open the selected file in Arlequin's text editor. You can then save any modification with *File / save*.

Note that the file is just opened in text edit mode. To open it as a project, and to make some analyses on it, use *Project / Open...*
- Close Close the active text edit window.
- Files conversions... This starts a dialog box that allows conversion between several formats. This is useful if you wish to convert data file used by other genetic data analysis programs into the Arlequin format.
- Save Save the content of the active text edit window with the current file name.
- Save As... Save the content of the active window in a file
- Exit Exit Arlequin

6.1.2 Edit Menu

A menu to access standard Edit command that can be performed on a text in a child window

- Undo Undo the last edit action.
- Cut Cut the highlighted text and put it into the clipboard.
- Copy Copy the highlighted text into the clipboard.
- Paste Paste the content of the clipboard.
- Clear All Delete the whole content of the active child window.
- Delete Delete the highlighted text without copying it into the clipboard.
- Find Find a given text string in the active child window.
- Replace Replace a given text string with another one in the active child window.
- Next Repeat the last Find action.

6.1.3 Project Menu

The menu that contains functions related to the manipulations of the project.

- Run Run the current project file with the chosen settings. The progress of the calculation is displayed on the bottom status line of Arlequin's window.
- Stop Stop the current run. For very large data files, the user might have to wait a little time before the program reacts. The results may be unreliable

- after an abnormal termination.
- Allow the opening and the loading of a project file. It can be edited by selecting *View Project File*. Some information appears in the Project information window, if the data file is valid. If the data file contains some errors, Arlequin's log file will be opened.
 - The dialog box that allows the project selection contains a history of the last ten selected projects. To open a new one, use the button browse. Clear list empties the list of recently selected projects. To open a listed project, highlight it in the list and press the open button, or simply double-click on it.
- Close the project file.
- Load a setting file. The setting files have usually the *.ars* extension. We recommend creating these setting files by using the *Save Settings* function. Only user familiar with Arlequin should try to create or modify these files with the text editor.
 - By loading a setting file, you will lose the settings that have been set up before. If you do not want to lose them you can save them with *Save Settings* under a different name.
- Save the current settings in a file, such as to be able to recall them for a later run.
- Reset all settings to original default values.
- If this menu item is checked, Arlequin looks for a setting file having the same name as the project file, but with the *.ars* extension, when a project is loaded. This associated setting file should be in the same folder as the project file.
 - Whether an associated setting file exists and whether it is used or not, is indicated in the project information window.
- If this menu item is checked, the actual settings are saved when you close the project or open a new one. The settings are saved in a file having the same name as the project file, but with the *.ars* extension.
- Run a series of Arlequin projects listed in a batch file. All projects will be executed one after the other with the current settings or with the settings specified in the associated setting files.
- Build the outline of a project file. You have to customize it to make it a valid project file. This is useful when creating new project files. The outline will be saved under the name you specify (the extension *.arp* is required).
- If this menu item is checked, the project information window is visible, otherwise it is invisible.
- A sub-menu for the specification of the format of output files.
- If this menu item is checked, the results of the actual computations are appended to those of previous computations done on this project. If it is unchecked, you will loose results of previous computations because the result file will only contain the results of the last run.
- If this menu item is checked, the inter-haplotypic distance matrix that might be computed (*DM_out.txt*) is not deleted, when you close the project. If you need this matrix for other computations (like phylogenetic reconstruction-programs), you should check this option.

- **Keep Null Distributions** If this option is checked, the null distributions of σ_a^2 , σ_b^2 , σ_c^2 , and σ_d^2 generated by an Amova analysis are written in files having the same name as the project file, but with the extensions *.va*, *.vb*, *.vc*, and *.vd*, respectively.
- **Generate Backup File** If this option is checked, a file having the project file name with the *.bac* extension is generated before the project is closed. This backup file contains the data of the original data file in a standardized format.
- **Generate HTML result File** If this option is checked, the results will be output in a series of HTML files that can be consulted in any web browser. The results of different runs can be accessed by dynamic links in separate pane windows of the browser.

6.1.4 Setup Menu

Set up the different dialog boxes. The dialog boxes allow one to choose which tasks to perform on the data, and to set up the options for each task. The dialog boxes are described in detail in paragraph 5 of this chapter.

- **General Settings** Opens the dialog box that contains the general settings. This dialog box is described in detail in section 6.4.1.
- **Diversity Indices** Open the dialog box that sets up the parameters for standard indices, molecular diversity, mismatch distribution, haplotype frequency estimation and search for shared haplotypes between populations. This dialog box is described in detail in section 6.4.2.
- **Neutrality Tests** Open the dialog box that sets up the parameters for selective neutrality tests. This dialog box is described in detail in section 6.4.3.
- **Tests of Disequilibrium** Open the dialog box that sets up the parameters for linkage disequilibrium test and Hardy-Weinberg equilibrium test. This dialog box is described in detail in section 6.4.4.
- **Genetic Structure** Open the dialog box that sets up the parameters for AMOVA, pairwise genetic distances and exact test of population differentiation. This dialog box is described in detail in section 6.4.5.
- **Launch Pad** Give an overview of the selected tasks. This dialog box is described in detail in section 6.4.6.

6.1.5 Special Menu

A menu to perform some special actions

- **Restore Tool Pad** By tearing off the tool pad, the user might close it by inadvertence. Selecting this option restores it at its original position.
- **Empty Log File** Empty the content of Arlequin's log-file. This can be useful if you have done several successive runs, and you want to trace more easily the cause of some problems during the execution of a specific run

6.1.6 Window Menu

The menu for handling windows and opening files related to the current project

- **View Project File** Open a window with the current active project file.
- **View Result file** Open a new window with the results of the last project run.
- **View Results in** This function translates the result file into HTML files that are then

- HTML Browser loaded in the internet web browser installed on your computer. The first time you use this option, you are asked to locate the browser on your hard disk. The location is then stored into *Arlequin.ini*.
- View Log file Open a window with the content of Arlequin.log, where Arlequin outputs warnings and error messages.
- View Distance Matrix If a distance matrix is computed, it is possible to view it in an edit window. If the file is too large, you can view the file *DM_out.txt* containing the distance matrix in a more capable text editor. If no matrix was calculated during the run, this item appears grayed.
- View group Structure If the project file contains a genetic structure, it can be viewed by selecting this function. However to modify the genetic structure, you have to edit the project file itself.
- Refresh Refresh the content of the project information window. Use it when for some reason the content is not readable.
- Cascade Cascade child windows.
- Tile Tile child windows.
- Arrange Icon Arrange Child window icons.
- Close All Close all child windows.

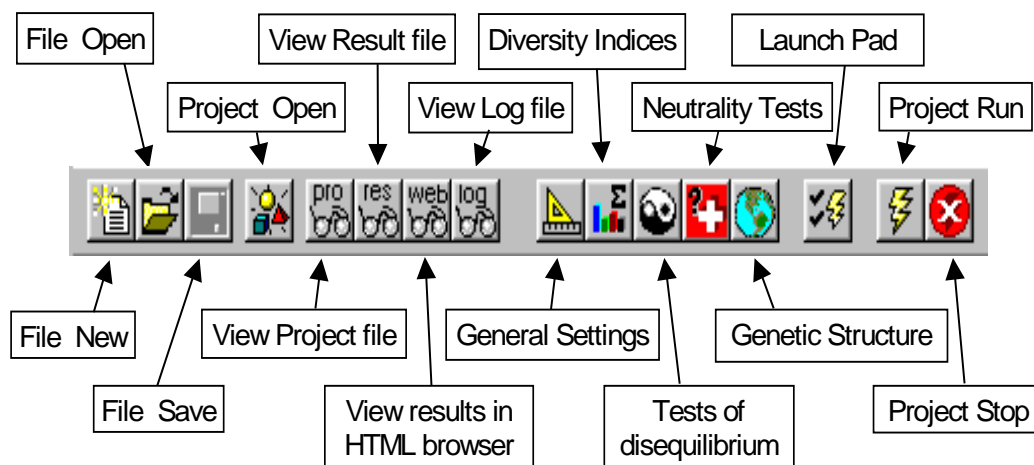
6.1.7 Help Menu

The menu to get access to the Help File System

- Content Open Arlequin Help File Welcome subject.
- Arlequin Overview Open Arlequin Help File Overview subject. It gives you an overview of Arlequin look and functionalities.
- About Some information about Arlequin, its authors, contact address and the Swiss NSF grants that supported its development.

6.2 Toolbar

Arlequin’s toolbar contains icons that are shortcuts to some commonly used menu items as shown below. Clicking on one of these icons is equivalent to activating the corresponding menu item.



6.3 Status Bar

A status bar is located at the bottom of the Arlequin program window. It issues brief information messages, or the progress of the computations, when the pointing device is located on Arlequin pane window.

6.4 Dialog boxes

Most of the tasks that Arlequin can perform are possible irrespective of the data type. Nevertheless, the testing procedure that might be used for performing a given task (e.g. testing linkage disequilibrium) may depend on the data type. The aim of this section is to give an overview of what happens in which situation and how to set up the numerous options in an optimal way.

The items that appear «grayed» in Arlequin's dialog boxes indicate that a given task is not possible in the current situation. For example, if you open a project containing haplotypic data, it is not possible to test Hardy-Weinberg equilibrium, and the task will appear as «grayed» in the dialog box. Or, for STANDARD data it is not possible to set up the transversion, transition, and deletion weights.

The way inter-haplotypic distances are calculated depends also on the data type. According to the situation, different lists of distance methods are presented in the dialog box.

Arlequin's interactive graphical user interface should prevent the user from selecting tasks impossible to perform, or from setting up parameters that are not taken into account in the analyses.

We do now describe in detail the six main dialog boxes that allow to select the options of the different possible analyses.

We have used the following symbols to specify which type of input was expected in the dialog boxes:

- [f] : parameter to be set in the dialog box as a floating number.
- [i] : parameter to be set in the dialog box as an integer.
- [b] : check box (two states: checked or unchecked).
- [m] : multiple selection radio buttons.
- [l] : List box, allowing the selection of an item in a downward scrolling list.
- [r] : read only setting, cannot be changed by the user.

6.4.1 General Settings

With this dialog box, the user can visualize some features of the data defined in the project file, and set up some additional general parameters

- **Project file** [r]: The name of the file containing the data to be analyzed.
- **Output files**: The files containing the results of the analyses generated by Arlequin.
 - **Main result file** [r]: The file containing all the results of the analyses just performed. It has the same name as the project file, with the extension (*.arf*) for Arlequin Result File. This file is opened in a text window at the end of each run.
 - **Backup file** [r]: The name of a back-up file built from the data read in the input file. It has the same name as the project, with the extension (*.bac*). It also allows you to check that your data have been read properly.

- **Log file** [r]: A file where run-time WARNINGS and ERRORS encountered during any phases of the current Arlequin session are issued. It has the name *arlequin.log*. You should consult this file if you observe any strange behavior of Arlequin. If Arlequin has crashed please consult this file located in Arlequin's installation directory **before** running Arlequin again. It will probably help you in finding where the problem was located. The file content can be emptied in the *Special / Empty Log file* menu of the main window.
- **Amova histograms** [r]: Specifies the name of an output file where the histograms of the variance component null distributions are output. By default, the name is set to *amo_hist.xl*. It is a tabulated text file which can be read directly by MS-Excel, for a graphical output of the distributions.
- **Linkage disequilibrium** [r]: The name of a tabulated text file where the results of pairwise linkage disequilibrium tests between all pairs of loci are output. By default, it has the name *lk_dis.xl*. As suggested by its extension, this file can be read with MS-Excel without modification.
- **Project information:**
 - **Project name**[r]: The title of the project as entered in the project.
 - **Locus separator**[r]: The character used to separate allelic information at adjacent loci.
 - **Missing data**[r]: The character used to represent missing data at any locus. By default, a question mark (?) is used for unknown alleles.
 - **Data type** [r]: Data type in the input file.
 - **Genotypic data** [r]: Specifies whether input data consist of diploid genotypic data or haplotypic data. For genotypic data, the diploid information of each genotype is entered on separate lines in the input file. The gametic phase of the genotype can be either assumed to be known or unknown. If the gametic phase is known, then the treatment of the data will be essentially similar to that of haplotypic data.
 - **Gametic phase** [r]: Specifies whether the gametic phase is known or unknown when the input file is made up of genotypic data.
- **Polymorphism control:**
 - **Allowed missing level per site** [f]: Specify the fraction of missing data allowed for any locus to be taken into account in the analyses. For instance, a level of 0.05 means that a locus with more than 5% of missing data will not be considered in any analysis. This option is especially useful when dealing with DNA data where different individuals have been sequenced for slightly different fragments. Setting a level of zero will force the analysis to consider only those sites that have been sequenced in all individuals. Alternatively, choosing a level of one means that all sites will be considered in the analyses, even if they have not been sequenced in any individual (not a very smart choice, however).
 - **Deletion weight** [f]: The weight given to deletions when comparing DNA or RFLP sequences.
 - **Transition weight** [f]: The weight given to transitions when comparing DNA sequences.
 - **Transversion weight** [f]: The weight given to transversions when comparing DNA sequences.

- **Infer haplotypes from distance matrix** [m] or **Use original haplotype definition** [m]: With the first option, similar haplotypes will be identified by computing a distance matrix based on the settings chosen above. Selecting the second option has the consequence that haplotypes are identified according to their original identifier.
- **Haplotype frequencies**: Some settings directly related to haplotype frequency estimation and output.
 - **Significant digits** [i]: The number of significant digits shown for the estimated haplotype frequencies in the result files.
 - **Epsilon value** [f]: The criterion used to stop the EM algorithm when estimating haplotype frequencies or linkage disequilibrium from genotypic data with unknown gametic phase (see section 7.1.3.2). The criterion is the difference in the sum of haplotypic frequency change between two successive iterations. The default value is 1e-7.
- **Internet browser location**
 - Use the **Browse** button to locate your Internet browser on your computer, if you want to be able to view your results in an HTML file.

6.4.2 Diversity indices

- **Standard diversity indices** [b]: Compute several common indices of diversity, like the number of alleles, the number of segregating loci, the heterozygosity level, etc. (see section 7.1.1).
- **Molecular diversity** [b]: Check box for computing several indices of diversity at the molecular level.
 - **Molecular distance** [l]: Choose the type of distance used when comparing haplotypes (see section 7.1.2.5 and below).
 - **Gamma a value** [f]: Set the value for the shape parameter of the gamma function, when selecting a distance allowing for unequal mutation rates among sites. This option is only valid for some distances computed between DNA sequences. Note that a value of zero deactivates here the Gamma correction of these distances, whereas in reality, a value of infinity would deactivate the Gamma correction procedure.
 - **Print distance matrix** [b]: If checked, the inter-haplotypic distance matrix used to evaluate the molecular diversity is printed in the result file.
 - **Theta(Hom)** [b]: An estimation of θ obtained from the observed homozygosity H (see section 7.1.2.3.1).
 - **Theta(S)** [b]: An estimation of θ obtained from the observed number of segregating site S (see section 7.1.2.3.2).
 - **Theta(k)** [b]: An estimation of θ obtained from the observed number of alleles k (see section 7.1.2.3.3).
 - **Theta(π)** [b]: An estimation of θ obtained from the mean number of pairwise differences $\hat{\pi}$ (see section 7.1.2.3.4).

- **Mismatch distribution** [b]: Compute the distribution of the observed differences between all pairs of haplotypes in the sample (see section 7.1.2.4). It also estimates parameters of a sudden demographic expansion according to the model presented in Rogers (1995) (see section 7.1.2.4).
 - Molecular distance [l]: Here we only allow one genetic distance: the mere number of observed differences between haplotypes.
- **Haplotype frequency information**

Depending on the data type, different methods are used to estimate the haplotypic frequencies.

Case a: Haplotypic data, or genotypic (diploid) data with known gametic phase

- **Gene frequency estimation** [b]: Estimate the maximum-likelihood haplotype frequencies from the observed data using a mere gene counting procedure
 - **Allele frequencies at all loci**: Estimate allele frequencies at all loci separately.

Case b: Genotypic data with unknown gametic phase

- **Haplotype frequency estimation** [b]: We estimate the maximum-likelihood (ML) haplotype frequencies from the observed data using an Expectation-Maximization (EM) algorithm for multi-locus genotypic data when the gametic phase is not known, or when recessive alleles are present (see section 7.1.3.2).
 - **No. of initial conditions** [i]: Set the number of random initial conditions from which the EM algorithm is started to repeatedly estimate haplotype frequencies. The haplotype frequencies globally maximizing the likelihood of the sample will be kept eventually. Figures of 100 or more are usually in order.
 - **No. of bootstrap replicates** [i]: Set the number of parametric bootstrap replicates of the EM estimation process on random samples generated from a fictive population having haplotype frequencies equal to previously estimated ML frequencies. This procedure is used to generate the standard deviation of haplotype frequencies. When set to zero, the standard deviations are not estimated.
 - **No. of initial conditions for bootstrap** [i]: Set the number of initial conditions for the bootstrap procedure. It may be smaller than the number of initial conditions set when estimating the haplotype frequencies, because the bootstrap replicates are quite time-consuming. Setting this number to small values is conservative, in the sense that it usually inflates the standard deviations.
 - **Maximum no. of iterations** [i]: Set the maximum number of iterations allowed in the EM algorithm. The iterative process will have at most this number of iterations, but may stop before if convergence has been reached. Here, convergence is reached when the sum of the differences between haplotypes frequencies between two successive iterations is smaller than the epsilon value defined in the general settings dialog box (section 6.4.1).

- **Recessive data** [b]: Specify whether a recessive allele is present. This option applies to all loci. The code for the recessive allele can be specified in the project file (see 3.2.1).
- **Compact haplotypes** [b]: Specify whether haplotypes can be compacted to get rid of monomorphic loci. This option just saves up memory and has no effect on the estimation procedure outcome.
- **Sub-haplotypes** [b]: Estimate haplotype frequencies for all haplotypes defined for all pairs of loci, as well as for all loci taken separately. This option can be quite time-consuming when the number of loci is large. The EM procedure is done with the same settings as those used for the haplotypic frequency estimation.
- **Search for shared haplotypes within and between populations** [b]: Look for haplotypes that are effectively similar after computing pairwise genetic distances according to the distance calculation settings in the « general settings » dialog box (6.4.1). For each pair of populations, the shared haplotypes will be printed out. Then will follow a table that contains, for every group of identified haplotypes, its absolute and relative frequency in each population. This task is only possible for haplotypic data.

6.4.3 Neutrality tests

Tests of selective neutrality, based either on the infinite-allele model or on the infinite-site model (see section 7.1.6).

- **Ewens-Watterson neutrality tests** [b]: Performs tests of selective neutrality based on Ewens sampling theory in a population at equilibrium (Ewens 1972). These tests are currently limited to sample sizes of 2000 genes or less and 1000 different alleles (haplotypes) or less.
 - **Ewens-Watterson homozygosity test**: This test, devised by Watterson (1978, 1986), is based on Ewens' sampling theory, but uses as a statistic the quantity F equal to the sum of squared allele frequencies, equivalent to the sample homozygosity in diploids (see section 7.1.6.1).
 - **Exact test based on Ewens' sampling theory**: In this test, devised by Slatkin (1994b, 1996), the probability of the observed sample is compared to that of a random neutral sample with same number of alleles and identical size. The probability of the sample selective neutrality is obtained as the proportion of random samples, which are less or equally probable than the observed sample.
 - **No. of random samples** [i]: Number of random samples to be generated for the two neutrality tests mentioned above. Values of several thousands are in order, and 16,000 permutations guarantee to have less than 1% difference with the exact probability in 99% of the cases (see Guo and Thomson 1992).
- **Chakraborty's test of population amalgamation** [b]: A test of selective neutrality and population homogeneity and equilibrium (Chakraborty, 1990). This test can be used when sample heterogeneity is suspected. It uses the observed homozygosity to estimate the population mutation parameter θ_{Hom} . The estimated value of this parameter is then used to compute the probability of observing k alleles or more in a neutral sample drawn from a stationary population. This test is based on Chakraborty's observation that the observed homozygosity is not very sensitive to population amalgamation or sample heterogeneity, whereas the number of observed (low frequency) alleles is more affected by this phenomenon.

- **Tajima's test of selective neutrality** [b]: This test described by Tajima (1989a, 1989b, 1993) compares two estimators of the population parameter θ , one being based on the number of segregating sites in the sample, and the other being based on the mean number of pairwise differences between haplotypes. Under the infinite-site model, both estimators should estimate the same quantity, but differences can arise under selection, population non-stationarity, or heterogeneity of mutation rates among sites (see section 7.1.6.4).

6.4.4 Gametic disequilibrium

- **Pairwise linkage disequilibrium** [b]: Test for the presence of significant association between pairs of loci.

This test can be done with all data types except FREQUENCY data type. The number of loci can be arbitrary, but if there are less than two polymorphic loci, there is no point performing this test.

Different approaches will be used depending on the data type:

Case a): Genotypic data with unknown gametic phase :

A procedure for testing the significance of the association between pairs of loci when the gametic phase is not known (see section 7.1.4.2). The likelihood of the sample under the hypothesis of no association between loci (linkage equilibrium) is compared to the likelihood of the sample when association is allowed (see Slatkin and Excoffier, 1996). The significance of the observed likelihood ratio is found by computing the null distribution of this ratio under the hypothesis of linkage equilibrium, using a permutation procedure.

- **No. of permutations** [i]: Number of random permuted samples to generate. Figures of several thousands are in order, and 16,000 permutations guarantee to have less than 1% difference with the exact probability in 99% of the cases (Guo and Thomson, 1992). A standard error for the estimated P -value is estimated using a system of batches (Guo and Thomson, 1992).
- **No. of initial conditions** [i]: Sets the number of random initial conditions from which the EM is started to repeatedly estimate the sample likelihood. The haplotype frequencies globally maximizing the sample likelihood will be eventually kept. Figures of 100 or more are in order.
- **Generate histogram and table** [b]: Generates an histogram of the number of loci with which each locus is in disequilibrium, and an s by s table (s being the number of polymorphic loci) summarizing the significant associations between pairs of loci. This table is generated for different levels of polymorphism, controlled by the value y : a locus is declared polymorphic if there are at least 2 alleles with y copies in the sample (Slatkin, 1994a). This is done because the exact test is more powerful at detecting departure from equilibrium for higher values of y (Slatkin 1994a).
- **Significance level** [f]: The level at which the test of linkage disequilibrium is considered significant for the output table.

Case b): Exact test of linkage disequilibrium

A test analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size (see section 7.1.4.1).

- **No. of steps in Markov chain** [i]: The maximum number of alternative tables to explore. Figures of 100,000 or more are in order. Larger values of the step number increases the precision of the P -value as well as its estimated standard deviation.
- **No. of dememorization steps** [i]: The number of steps to perform before beginning to compare the alternative table probabilities to that of the observed table. A few thousands steps are necessary to reach a random starting point corresponding to a table independent from the observed table.
- **Required precision on probability** [f]: The precision required on the inferred probability of linkage equilibrium. A system of batches (Guo and Thomson 1992) is used to constantly estimate the standard-deviation of the probability. The estimation process is stopped once the required precision has been reached, or once the maximal number of steps has been performed.
- **Generate histogram and table** [b]: Generates a histogram of the number of loci with which each locus is in disequilibrium, and an s by s table (s being the number of polymorphic loci) summarizing the significant associations between pairs of loci. This table is generated for different levels of polymorphism, controlled by the value y : a locus is declared polymorphic if there are at least 2 alleles with y copies in the sample (Slatkin, 1994a). This is done because the exact test is more powerful at detecting departure from equilibrium for higher values of y (Slatkin 1994a).
- **Significance level** [f]: The level at which the test of linkage disequilibrium is considered significant for the output table.
- **D and D' coefficients for all pairs of alleles** [b]:
See section 7.1.4.3
 1. D : The classical linkage disequilibrium coefficient measuring deviation from random association between alleles at different loci (Lewontin and Kojima, 1960) expressed as

$$D = p_{ij} - p_i p_j.$$
 2. D' : The linkage disequilibrium coefficient D standardized by the maximum value it can take (D_{\max}), given the allele frequencies (Lewontin 1964).
- **Hardy-Weinberg equilibrium** [b]: Test of the hypothesis that the observed diploid genotypes are the product of a random union of gametes. This test is only possible for genotypic data. Separate tests are carried out at each locus.

This test is analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size (see section 7.1.5). If the gametic phase is unknown the test is only possible locus by locus. For data with known gametic phase, it is also possible to test the association at the haplotypic level within individuals.
- **No. of steps in Markov chain** [i]: The maximum number of alternative tables to explore. Figures of 100,000 or more are in order.

- **No. of dememorization steps** [i]: The number of steps to perform before beginning to compare the alternative table probabilities to that of the observed table. A few thousands steps are necessary to reach a random starting point corresponding to a table independent from the observed table.
- **Required precision on probability** [f]: The precision required on the inferred probability of linkage equilibrium. A system of batches (Guo and Thomson 1992) is used to constantly estimate the standard-deviation of the probability. The estimation process is stopped once the required precision has been reached or once the maximal number of steps has been performed.
- **Test association at** [m]:
 1. **Locus level**: Test the association of alleles at each locus within individuals.
 2. **Haplotype level**: test the association of haplotypes within diploid individuals.
 3. **Both levels**: Test both 1 and 2.

6.4.5 Genetic structure

A dialog box to set up the options for the analysis of population genetic structure, and genetic distances between populations. The genetic structure is analyzed using an analysis of variance framework (Weir and Cockerham, 1984; Excoffier et al. 1992; Weir, 1996).

- **AMOVA** [b]: Analysis of MOlecular VAriance framework. Estimate genetic structure indices using information on the allelic content of haplotypes, as well as their frequencies (Excoffier et al. 1992). The information on the differences in allelic content between haplotypes is entered as a matrix of Euclidean squared distances. The significance of the variance components associated with the different possible levels of genetic structure (within individuals, within populations, within groups of populations, among groups) is tested using non-parametric permutation procedures (Excoffier et al. 1992). The type of permutations is different for each variance component (see section 7.2.1).

The number of hierarchical levels of the variance analysis and the kind of permutations that are done depend on the kind of data, the genetic structure that is tested, and the options the user might choose. All details will be given in section 7.2.1.

- **No. of permutations** [i]: Enter the number of permutations used to test the significance of variance components and fixation indices. A value of zero will not lead to any testing procedure. Values of several thousands are in order for a proper testing scheme, and 16 000 permutations guarantee to have less than 1% difference with the exact probability in 99% of the cases (Guo and Thomson 1992).

The number of permutations used by the program might be slightly larger. This is the consequence of subdivision of the total number of permutation in batches for estimating the standard error of the *P*-value.

Note that if several variance components need to be tested, the probability of each variance component will be estimated with this number of permutation. The distribution of the variance components is output into a tabulated text file called *amo_hist.xls*, which can be directly read into MS-EXCEL .

- **Include individual level for genotype data** [b]: Include the intra-individual variance component of genetic diversity, and its associated fixation indices. It thus takes into account the differences between

genes found within individuals. This is another way to test for global departure from Hardy-Weinberg equilibrium. The selection of this option is only possible for genotypic data with known gametic phase.

- **Compute population pairwise F_{ST} 's** [b]: Compute F_{ST} statistics for all pairs of populations.

Transformed pairwise F_{ST} 's can be used as short term genetic distances between populations (Reynolds et al. 1983; Slatkin, 1995).

The significance of the pairwise F_{ST} values is tested by permuting the haplotypes or individuals between the populations. See section 7.2.2 for more details on the output results (genetic distances and migration rates estimates between populations).

- **Test significance** [b]: Use a non-parametric permutation scheme to test for the significance of the derived genetic distances. Note that this procedure is quite time consuming when the number of populations is large.
- **No. of permutations** [i]: Enter the required number of permutations. If this number is set to zero, no testing procedure will be performed.
- **Use specified distance matrix** [m] or **Generate distance matrix** [m]: Select the first option to use a given matrix of Euclidean distances (computed by another program than Arlequin). The matrix can be put directly into the Arlequin project, or it can be put into a separate file whose name has to be defined in the project file. Select the second option to have the distance matrix generated directly by Arlequin, from the definition of the haplotypes. This matrix can be generated either for haplotypic data or genotypic data (Michalakis and Excoffier, 1996)
 - **Distance method** [l]: Select a distance method to compute the distances between haplotypes.
 - **Gamma a value** [f]: Set the value for the shape parameter a of the gamma function, when selecting a distance allowing for unequal mutation rates among sites. See the Molecular diversity section 7.1.2.5.
- **Compute conventional F -statistics** [b]: Estimate genetic structure indices using haplotype frequencies only, without taking into account their allelic content and the molecular differences between these haplotypes.

The distance matrix used for the calculations will simply have zeroes on its diagonal and ones elsewhere. This is equivalent to a conventional treatment based on *haplotype* frequencies, and different from a treatment averaged over all loci. By checking this, the three settings concerning the distance matrix are of course ignored.
- **Exact test of population differentiation** [b]: We test the hypothesis of random distribution of the individuals between pairs of populations as described in Raymond and Rousset (1995) and Goudet et al. (1996). This test is analogous to Fisher's exact test on a two-by-two contingency table, but extended to a contingency table of size two by (no. of haplotypes). We do also an exact differentiation test for all populations defined in the project by constructing a table of size (no. of populations) by (no. of haplotypes). (Raymond and Rousset, 1995).

- **No. of steps in Markov chain** [i]: The maximum number of alternative tables to explore. Figures of 100,000 or more are in order. Larger values of the step number increases the precision of the *P*-value as well as its estimated standard deviation.
- **No. of dememorization steps** [i]: The number of steps to perform before beginning to compare the alternative table probabilities to that of the observed table. A few thousands steps are necessary to reach a random starting point corresponding to a table independent from the observed table.
- **Required precision on probability** [f]: The precision required on the inferred probability of non-differentiation. A system of batches (Guo and Thomson, 1992) is used to constantly estimate the standard deviation of the probability. The estimation process is stopped once the required precision has been reached or once the maximal number of steps has been performed. Setting this value to zero ensures that the total amount of specified steps is performed.
- **Generate histogram and table** [b]: Generates a histogram of the number of populations which are significantly different from a given population, and a *s* by *s* table (*s* being the number of populations) summarizing the significant associations between pairs of populations. An association between two populations is considered as significant or not depending on the significance level specified below.
- **Significance level** [f]: The level at which the test of differentiation is considered significant for the output table. If the *P*-value is smaller than the *Significance level*, then the two populations are considered as significantly different.

6.4.6 Launch Pad

In this dialog box, users can rapidly select which tasks they want to perform on their data set. The options of each possible category can then be set by a series of other dialog boxes described below.

The population genetics methods of Arlequin have been grouped into four main categories:

- **Population genetic diversity indices:**

A series of diversity indices can be computed if these check-boxes are checked (see section 6.4.2).

- **Standard indices** [b]: Several common indices of diversity, like the number of alleles, the number of segregating loci, gene diversity.
- **Molecular diversity** [b]: Computes the sample molecular diversity, the sample nucleotide diversity for DNA data, and several estimators of the population mutation parameter θ .
- **Mismatch distribution** [b]: Computes the distribution of the number of pairwise differences.
- **Haplotype frequency estimation** [b]: Estimate haplotype frequencies and standard deviations.
- **Search for shared haplotypes between populations** [b]: Make a table of common haplotypes shared between populations.

- **Gametic disequilibrium:**

Possibility of testing for pairwise linkage equilibrium and Hardy-Weinberg equilibrium (see section 6.4.4).

- **Linkage disequilibrium** [b]: Test the linkage equilibrium hypothesis for all pairs of loci.

-
- **Hardy-Weinberg-disequilibrium** [b]: Test the Hardy-Weinberg equilibrium hypothesis at all loci and/or at the whole haplotype level. This test is only possible for genotypic data.
 - **Selective neutrality tests**: For details see section 6.4.3.
 - **Ewens-Watterson neutrality tests** [b]: The exact or the homozygosity tests of selective neutrality under the infinite allele model. These tests are not possible for microsatellite data
 - **Chakraborty's test** [b]: Chakraborty's (1990) test of neutrality and population subdivision under the infinite allele model.
 - **Tajima's D** [b]: Tajima's test of selective neutrality under the infinite-site model. This test is only meaningful on non-recombining DNA or RFLP data.
 - **Analysis of population genetic structure** : See section 6.4.5 for details
 - **AMOVA** [b]: Estimates and tests the amount of genetic structure among populations, using the Amova framework for mono- or multi-locus data.
 - **Pairwise genetic distances** [b]: Computes pairwise F_{ST} 's between all pairs of population samples.
 - **Exact test of population differentiation** [b]: Test if the haplotypic composition of the populations is significantly heterogeneous.

7 METHODOLOGICAL OUTLINES

The following table gives a rapid overview of the methods implemented in Arlequin. A ✓ indicates that the task corresponding to the table entry is possible. Some tasks are only possible or meaningful if there is no recessive data, and those cases are marked with a ✗

| | Data types | | | | | | | | | |
|--|--------------|----|---|----------|----|---|----------|----|---|-----------|
| | DNA and RFLP | | | Microsat | | | Standard | | | Frequency |
| Types of computations | G+ | G- | H | G+ | G- | H | G+ | G- | H | |
| Standard indices ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Molecular diversity ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Mismatch distribution | ✓ | | ✓ | ✓ | | ✓ | ✓ | | ✓ | |
| Haplotype frequency estimation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Linkage disequilibrium | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Hardy-Weinberg equilibrium ✗ | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ | | |
| Tajima's neutrality test | ✓ | | ✓ | | | | | | | |
| Ewens-Watterson neutrality tests | ✓ | | ✓ | | | | ✓ | | ✓ | ✓ |
| Chakraborty's amalgamation test | ✓ | | ✓ | | | | ✓ | | ✓ | ✓ |
| Search for shared haplotypes between samples | | | ✓ | | | ✓ | | | ✓ | ✓ |
| AMOVA ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pairwise genetic distances ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Exact test of population differentiation ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

G+: Genotypic data, gametic phase known,

G- : Genotypic data, gametic phase unknown,

H : Haplotypic data.

7.1 Intra-population level methods

7.1.1 Standard diversity indices

7.1.1.1 Gene diversity

Equivalent to the expected heterozygosity for diploid data. It is defined as the probability that two randomly chosen haplotypes are different in the sample. Gene diversity and its sampling variance are estimated as

$$\hat{H} = \frac{n}{n-1} \left(1 - \sum_{i=1}^k p_i^2\right)$$

$$V(\hat{H}) = \frac{2}{n(n-1)} \left\{ 2(n-2) \left[\sum_{i=1}^k p_i^3 - \left(\sum_{i=1}^k p_i^2\right)^2 \right] + \sum_{i=1}^k p_i^2 - \left(\sum_{i=1}^k p_i^2\right)^2 \right\},$$

where n is the number of gene copies in the sample, k is the number of haplotypes, and p_i is the sample frequency of the i -th haplotype.

Reference:

Nei, 1987, p.180.

7.1.1.2 Number of usable loci

Number of loci that present less than a specified amount of missing data. The maximum amount of missing data must be specified in the *General Settings* dialog box.

7.1.1.3 Number of polymorphic sites (S)

Number of usable loci that present more than one allele per locus.

7.1.2 Molecular indices

7.1.2.1 Mean number of pairwise differences (π)

Mean number of differences between all pairs of haplotypes in the sample. It is given by

$$\hat{\pi} = \sum_{i=1}^k \sum_{j<i} p_i p_j \hat{d}_{ij},$$

where \hat{d}_{ij} is an estimate of the number of mutations having occurred since the divergence of haplotypes i and j , k is the number of haplotypes, and p_i is the frequency of haplotype i . The total variance (over the

stochastic and the sampling process), assuming no recombination between sites and selective neutrality, is obtained as

$$V(\hat{\pi}) = \frac{3n(n+1)\hat{\pi} + 2(n^2 + n + 3)\hat{\pi}^2}{11(n^2 - 7n + 6)}. \quad (\text{Tajima, 1991})$$

Note that similar formulas are also used for *Microsat* and *Standard* data, even though the underlying assumptions of the model may be violated.

References:

Tajima, 1983

Tajima, 1991

7.1.2.2 Nucleotide diversity or average gene diversity over L loci (*RFLP* and *DNA* data)

It is the probability that two randomly chosen homologous nucleotides are different. It is equivalent to the gene diversity at the nucleotide level.

$$\hat{\pi}_n = \frac{\sum_{i=1}^k \sum_{j<i} p_i p_j \hat{d}_{ij}}{L}$$

$$V(\hat{\pi}_n) = \frac{n+1}{3(n-1)L} \hat{\pi}_n + \frac{2(n^2 + n + 3)}{9n(n-1)} \hat{\pi}_n^2$$

Note that similar formulas are used for computing the average gene diversity over L loci for *Microsat* and *Standard* data, assuming no recombination and selective neutrality. As above, one should be aware that these assumption may not hold for these data types.

References:

Tajima, 1983

Nei, 1987, p. 257

7.1.2.3 Theta estimators

Several methods are used to estimate the population parameter $\theta = 2Mu$, where M is equal to $2N$ for diploid populations of size N , or equal to N for haploid populations, and u is the overall mutation rate at the haplotype level.

7.1.2.3.1 Theta(Hom)

The expected homozygosity in a population at equilibrium between drift and mutation is usually given by

$$H = \frac{1}{\theta + 1}.$$

However, Zouros (1979) has shown that this estimator was an overestimate when estimated from a single or a few loci. Although he gave no closed form solution, Chakraborty and Weiss (1991) proposed to iteratively solve the following relationship between the expectation of $\hat{\theta}_H$ and the unknown parameter θ

$$E(\hat{\theta}_H) = \theta \left(1 + \frac{2(1+\theta)}{(2+\theta)(3+\theta)} \right) \quad (\text{Zouros, 1979})$$

starting with a first estimate of $\hat{\theta}_H$ of $(1-H)/H$, and equating it to its expectation.

Chakraborty and Weiss (1991) give an approximate formula for the standard error of $\hat{\theta}_H$ as

$$\text{s.d.}(\hat{\theta}_H) \approx \frac{(2+\theta)^2(3+\theta)^2 \text{s.d.}(H)}{H^2(1+\theta)[(2+\theta)(3+\theta)(4+\theta) + 10(2+\theta) + 4]}$$

where $\text{s.d.}(H)$ is the standard error of H given in section 7.1.1.1.

7.1.2.3.2 Theta(S)

$\hat{\theta}_S$ is estimated from the infinite-site equilibrium relationship (Watterson, 1975) between the number of segregating sites (S), the sample size (n) and θ for a sample of non-recombining DNA:

$$\theta = \frac{S}{a_1}$$

where

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

The variance of $\hat{\theta}_S$ is obtained as

$$V(\hat{\theta}_S) = \frac{a_1^2 S + a_2 S^2}{a_1^2 (a_1^2 + a_2)}, \quad (\text{Tajima, 1989})$$

where

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

7.1.2.3.3 Theta(k)

$\hat{\theta}_k$ is estimated from the infinite-allele equilibrium relationship (Ewens, 1972) between the expected number of alleles (k), the sample size (n) and θ :

$$E(k) = \theta \sum_{i=0}^{n-1} \frac{1}{\theta + i}$$

Instead of the variance of $\hat{\theta}_k$, we give the limits ($\hat{\theta}_0$ and $\hat{\theta}_1$) of a 95% confidence interval around $\hat{\theta}_k$, obtained from Ewens (1972)

$$\Pr(\text{less than } k \text{ alleles} | \theta = \theta_0) = 0.025$$

$$\Pr(\text{more than } k \text{ alleles} | \theta = \theta_1) = 0.025,$$

These probabilities are obtained by summing up the probabilities of observing k' alleles ($k'=0, \dots, k$), obtained as (Ewens, 1972)

$$\Pr(K = k | \theta) = \frac{|S_n^k| \theta^k}{S_n(\theta)}$$

where $|S_n^k|$ is a Stirling number of the first kind (see Abramovitz and Stegun, 1970), and $S_n(\theta)$ is defined as $\theta(\theta+1)(\theta+2)\dots(\theta+n-1)$.

7.1.2.3.4 Theta (π)

$\hat{\theta}_\pi$ is estimated from the infinite-site equilibrium relationship between the mean number of pairwise differences ($\hat{\pi}$) and theta (θ):

$$E(\hat{\pi}) = \theta, \quad (\text{Tajima, 1983})$$

and its variance $V(\hat{\pi})$ is given in section 7.1.1.1.

7.1.2.4 Mismatch distribution

It is the distribution of the observed number of differences between pairs of haplotypes. This distribution is usually multimodal in samples drawn from populations at demographic equilibrium, as it reflects the highly stochastic shape of gene trees, but it is usually unimodal in populations having passed through a recent demographic expansion (Rogers and Harpending, 1992; Hudson and Slatkin, 1991).

If one assumes that a stationary haploid population at equilibrium has suddenly passed τ generations ago from a population size of N_0 to N_1 , then the probability of observing S differences between two randomly chosen non-recombining haplotypes is given by

$$F_S(\tau, \theta_0, \theta_1) = F_S(\theta_1) + \exp\left(-\tau \frac{\theta_1 + 1}{\theta_1}\right) \sum_{j=0}^S \frac{\tau^j}{j!} [F_{S-j}(\theta_0) - F_{S-j}(\theta_1)], \quad (\text{Li, 1977})$$

where $F_S(\theta) = \frac{\theta^S}{(\theta+1)^{S+1}}$ is the probability of observing two random haplotypes with S differences in a stationary population (Watterson, 1975), $\theta_0 = 2uN_0$, $\theta_1 = 2uN_1$, $\tau = 2ut$, and u is the mutation rate for the whole haplotype.

Rogers (1995) has simplified the above equation, by assuming that $\theta_1 \rightarrow \infty$, implying there are no coalescent events after the expansion, which is only reasonable if the expansion size is large. With this simplifying assumption, it is possible to derive the moment estimators of the time to the expansion (τ) and the mutation parameter θ_0 , as

$$\begin{aligned} \hat{\theta}_0 &= \sqrt{v-m} \\ \hat{\tau} &= m - \hat{\theta}_0 \end{aligned} \quad , \quad (\text{Rogers, 1995})$$

where m and v are the mean and the variance of the observed mismatch distribution, respectively. These estimators can then be used to plot $F_S(\tau, \theta_0, \infty)$ values. Note, however, that this estimation cannot be done if the variance of the mismatch is smaller than the mean.

A simple chi-square test of goodness of fit can be performed to judge whether the observed distribution fits with the predicted expansion scenario.

For convenience, we also compute the raggedness index of the observed distribution defined by Harpending (1994) as

$$r = \sum_{i=1}^{d+1} (x_i - x_{i-1})^2 ,$$

where d is the maximum number of observed differences between haplotypes, and the x 's are the observed relative frequencies of the mismatch classes. This index takes larger values for multimodal distributions commonly found in a stationary population than for unimodal and smoother distributions typical of expanding populations.

7.1.2.5 Estimation of genetic distances between DNA sequences

Definitions:

- | | |
|-------------------|---|
| <i>L</i> : | Number of loci |
| Gamma correction: | This correction is proposed when the mutation rates cannot be assumed as uniform for all sites. It had been originally proposed for mutation rates among amino acids (Uzell and Corbin, 1971), but it seems also to be the case of the control region of human mtDNA (Wakeley, 1993). In such a case, a Gamma distribution of mutation rates is often assumed. The shape of this distribution (the unevenness of the mutation rates) is mainly controlled by a parameter α , which is the inverse of the coefficient of variation of the mutation rate. The smaller the α coefficient, the more uneven the mutation rates. A uniform |

| | |
|----------|---|
| | mutation rate corresponds to the case where a is equal to infinity. |
| n_d : | Number of observed substitutions between two DNA sequences |
| n_s : | Number of observed transitions between two DNA sequences |
| n_v : | Number of observed transversions between two DNA sequences |
| ω | G+C ratio, computed on all the DNA sequences of a given sample |

7.1.2.5.1 Pairwise difference

Outputs the number of loci for which two haplotypes are different

$$\hat{d} = n_d$$

$$V(\hat{d}) = \hat{d}(L - \hat{d}) / L$$

7.1.2.5.2 Percentage difference

Outputs the percentage of loci for which two haplotypes are different

$$\hat{d} = n_d / L$$

$$V(\hat{d}) = \hat{d}(1 - \hat{d}) / L$$

7.1.2.5.3 Jukes and Cantor

Outputs a corrected percentage of nucleotides for which two haplotypes are different.

The correction allows for multiple substitutions per site since the most recent common ancestor of the two DNA sequences. The correction also assumes that the rate of nucleotide substitution is identical for all 4 nucleotides A, C, G and T.

$$\hat{p} = n_d / L$$

$$\hat{d} = -\frac{3}{4} \log\left(1 - \frac{4}{3} \hat{p}\right)$$

$$V(\hat{d}) = \frac{\hat{p}(1 - \hat{p})}{\left(1 - \frac{4}{3} \hat{p}\right)^2 L}$$

Gamma correction:

$$\hat{d} = -\frac{3}{4} a \left[\left(1 - \frac{4}{3} p\right)^{-1/a} - 1 \right]$$

$$V(\hat{d}) = p(1 - p) \left[\left(1 - \frac{4}{3} p\right)^{-2(1/a+1)} \right] / L$$

References:

- Jukes and Cantor 1969
- Jin and Nei 1990
- Kumar et al. 1993

7.1.2.5.4 Kimura 2-parameters

Outputs a corrected percentage of nucleotides for which two haplotypes are different.

The correction also allows for multiple substitutions per site, but takes into account different substitution rates between transitions and transversions. The transition-transversion ratio is estimated from the data.

$$\hat{P} = \frac{n_s}{L}, \quad \hat{Q} = \frac{n_v}{L}$$

$$c_1 = (1 - 2\hat{P} - \hat{Q})^{-(1/a+1)}, c_2 = (1 - 2\hat{Q})^{-(1/a+1)}, c_3 = \frac{c_1 + c_2}{2}$$

$$\hat{d} = \frac{1}{2} \log(1 - 2\hat{P} - \hat{Q}) - \frac{1}{4} \log(1 - 2\hat{Q})$$

$$V(\hat{d}) = \frac{c_1^2 \hat{P} + c_3^2 \hat{Q} - (c_1 \hat{P} + c_3 \hat{Q})^2}{L}$$

Gamma correction:

$$\hat{d} = \frac{a}{2} \left[(1 - 2\hat{P} - \hat{Q})^{-1/a} + \frac{1}{2} (1 - 2\hat{Q})^{-1/a} - \frac{3}{2} \right]$$

$$V(\hat{d}) = \frac{c_1^2 \hat{P} + c_3^2 \hat{Q} - (c_1 \hat{P} + c_3 \hat{Q})^2}{L}$$

References:

Kimura (1980)

Jin and Nei 1990

7.1.2.5.5 Tamura

Outputs a corrected percentage of nucleotides for which two haplotypes are different.

The correction is an extension of Kimura 2-parameters method, allowing for unequal nucleotide frequencies.

The transition-transversion ratios, as well as the overall nucleotide frequencies are computed from the original data.

$$\hat{P} = \frac{n_s}{L}, \quad \hat{Q} = \frac{n_v}{L}$$

$$c_1 = \frac{1}{1 - \frac{\hat{P}}{2\omega(1-\omega)}}, \quad c_2 = \frac{1}{1 - 2\hat{Q}}, \quad c_3 = 2\omega(1-\omega)(c_1 - c_2) + c_2$$

$$\hat{d} = -2\omega(1-\omega) \log\left(1 - \frac{\hat{P}}{2\omega(1-\omega)} - \hat{Q}\right) - \frac{1}{2} (1 - 2\omega(1-\omega)) \log(1 - 2\hat{Q})$$

$$V(\hat{d}) = \frac{c_1^2 \hat{P} + c_3^2 \hat{Q} - (c_1 \hat{P} + c_3 \hat{Q})^2}{L}$$

References:

Tamura, 1992,

Kumar et al. 1993

7.1.2.5.6 Tajima and Nei

Outputs a corrected percentage of nucleotides for which two haplotypes are different.

The correction is an extension of Jukes and Cantor method, allowing for unequal nucleotide frequencies. The overall nucleotide frequencies are computed from the data.

$$\hat{p} = \frac{n_d}{L}, \quad b = \frac{1}{2} \left(1 - \sum_{i=1}^4 g_i^2 + \frac{\hat{p}^2}{c} \right), \quad c = \sum_{i=1}^3 \sum_{j=i+1}^4 \frac{x_{ij}^2}{2g_i g_j},$$

where the g 's are the four nucleotide frequencies, and x_{ij} is the relative frequency of the nucleotide pair i and j .

$$\hat{d} = -b \log \left(1 - \frac{\hat{p}}{b} \right)$$

$$V(\hat{d}) = \frac{\hat{p}(1 - \hat{p})}{\left(1 - \frac{\hat{p}}{b} \right)^2 L}$$

References:

Tajima and Nei, 1984,

Kumar et al. 1993

7.1.2.5.7 Tamura and Nei

Outputs a corrected percentage of nucleotides for which two haplotypes are different.

Like Kimura 2-parameters, and Tajima and Nei distances, the correction allows for different transversion and transition rates, but a distinction is also made between transition rates between purines and between pyrimidines.

$$c_1 = \frac{2g_A g_G}{g_R}, \quad c_2 = \frac{2g_C g_T}{g_Y}, \quad c_3 = \frac{2g_A g_G g_R}{2g_A g_G g_R - g_R^2 \hat{P}_1 - g_A g_G \hat{Q}}$$

$$c_4 = \frac{2g_T g_C g_Y}{2g_T g_C g_Y - g_Y^2 \hat{P}_2 - g_T g_C \hat{Q}}$$

$$c_5 = \frac{2g_A^2 g_G^2}{g_R (2g_A g_G g_R - g_R^2 \hat{P}_1 - g_A g_G \hat{Q})}$$

$$+ \frac{2g_T^2 g_C^2}{g_Y (2g_T g_C g_Y - g_Y^2 \hat{P}_2 - g_T g_C \hat{Q})}$$

$$+ \frac{g_R^2(g_T^2 + g_C^2) + g_Y^2(g_A^2 + g_G^2)}{2g_R^2g_Y^2 - g_Rg_YQ}$$

$$\hat{P}_1 = n_s(A \leftrightarrow G), \quad \hat{P}_2 = n_s(C \leftrightarrow T), \quad \hat{Q} = \frac{n_s}{n_d}$$

$$\hat{d} = -c_1 \log\left(1 - \frac{\hat{P}_1}{c_1} - \frac{\hat{Q}}{2g_R}\right) - c_2 \log\left(1 - \frac{\hat{P}_2}{c_2} - \frac{\hat{Q}}{2g_Y}\right)$$

$$- 2(g_Rg_Y - c_1g_Y - c_2g_R) \log\left(1 - \frac{Q}{2g_Rg_Y}\right)$$

$$V(\hat{d}) = \frac{c_3^2\hat{P}_1 + c_4^2\hat{P}_2 + c_5^2\hat{Q} - (c_3\hat{P}_1 + c_4\hat{P}_2 + c_5\hat{Q})^2}{L}$$

Gamma correction:

$$\hat{d} = 2a \left[c_1 \left(1 - \frac{\hat{P}_1}{c_1} - \frac{\hat{Q}}{2g_R}\right)^{-1/a} + c_2 \left(1 - \frac{\hat{P}_2}{c_2} - \frac{\hat{Q}}{2g_Y}\right)^{-1/a} \right. \\ \left. + \left(g_Rg_Y - \frac{g_Y}{c_1} - \frac{g_R}{c_2}\right) \left(1 - \frac{\hat{Q}}{2g_Rg_Y}\right)^{-1/a} - 2g_Ag_G - 2g_Tg_C - 2g_Rg_Y \right]$$

$$V(\hat{d}) = \frac{c_3^2\hat{P}_1 + c_4^2\hat{P}_2 + c_5^2\hat{Q} - (c_3\hat{P}_1 + c_4\hat{P}_2 + c_5\hat{Q})^2}{L}$$

References:

Tamura and Nei, 1994,

Kumar et al. 1993

7.1.2.6 Estimation of genetic distances between RFLP haplotypes

7.1.2.6.1 Number of pairwise difference

We simply count the number of different alleles between two RFLP haplotypes.

$$\hat{d}_{xy} = \sum_{i=1}^L \delta_{xy}(i)$$

where $\delta_{xy}(i)$ is the Kronecker function, equal to 1 if the alleles of the i -th locus are identical for both haplotypes, and equal to 0 otherwise.

When estimating genetic structure indices, this choice amounts at estimating weighted F_{ST} statistics over all loci (Weir and Cockerham, 1984; Michalakis and Excoffier, 1996).

7.1.2.6.2 Proportion of difference

We simply count the proportion of loci that are different between two RFLP haplotypes.

$$\hat{d}_{xy} = \frac{1}{L} \sum_{i=1}^L \delta_{xy}(i)$$

where $\delta_{xy}(i)$ is the Kronecker function, equal to 1 if the alleles of the i -th locus are identical for both haplotypes, and equal to 0 otherwise.

When estimating genetic structure indices, this choice will lead to exactly the same results as the number of pairwise differences.

7.1.2.7 Estimation of distances between Microsatellite haplotypes

7.1.2.7.1 No. of different alleles

We simply count the number of different alleles between two haplotypes.

$$\hat{d}_{xy} = \sum_{i=1}^L \delta_{xy}(i)$$

where $\delta_{xy}(i)$ is the Kronecker function, equal to 1 if the alleles of the i -th locus are identical for both haplotypes, and equal to 0 otherwise.

When estimating genetic structure indices, this choice amounts at estimating weighted F_{ST} statistics over all loci (Weir and Cockerham, 1984; Michalakis and Excoffier, 1996).

7.1.2.7.2 Sum of squared size difference

Counts the sum of the squared number of repeat difference between two haplotypes (Slatkin, 1995).

$$\hat{d}_{xy} = \sum_{i=1}^L (a_{xi} - a_{yi})^2 ,$$

where a_{xi} is the number of repeats of the microsatellite for the i -th locus.

When estimating genetic structure indices, this choice amounts at estimating an analog of Slatkin's R_{ST} (1995) (see Michalakis and Excoffier, 1996, as well as Rousset, 1996, for details on the relationship between F_{ST} and R_{ST}).

7.1.2.8 Estimation of distances between Standard haplotypes

7.1.2.8.1 Number of pairwise differences

Simply counts the number of different alleles between two haplotypes.

$$\hat{d}_{xy} = \sum_{i=1}^L \delta_{xy}(i)$$

where $\delta_{xy}(i)$ is the Kronecker function, equal to 1 if the alleles of the i -th locus are identical for both haplotypes, and equal to 0 otherwise.

When estimating genetic structure indices, this choice amounts at estimating weighted F_{ST} statistics over all loci (Weir and Cockerham, 1984; Michalakis and Excoffier, 1996).

7.1.3 Haplotype frequency estimation

7.1.3.1 Haplotypic data or Genotypic data with known Gametic phase

If haplotype i is observed x_i times in a sample containing n gene copies, then its estimated frequency (\hat{p}_i) is given by

$$\hat{p}_i = \frac{x_i}{n},$$

whereas an unbiased estimate of its sampling variance is given by

$$V(p_i) = \frac{\hat{p}_i(1 - \hat{p}_i)}{n - 1}.$$

7.1.3.2 Genotypic data with unknown Gametic phase

Maximum-likelihood haplotype frequencies are computed using an Expectation-Maximization (EM) algorithm (see e.g. Dempster et al. 1977; Excoffier and Slatkin, 1995; Lange, 1997; Weir, 1996). This procedure is an iterative process aiming at obtaining maximum-likelihood estimates of haplotype frequencies from multi-locus genotype data when the gametic phase is unknown (phenotypic data). In this case, a simple gene counting is not possible because several genotypes are possible for individuals heterozygote at more than one locus.

Therefore, a slightly more elaborate procedure is needed.

The likelihood of the sample (the probability of the observed data \mathbf{D} , given the haplotype frequencies - \mathbf{p}) is given by

$$L(\mathbf{D} | \mathbf{p}) = \sum_{i=1}^n \prod_{j=1}^{g_i} G_{ij},$$

where the sum is over all n individuals of the sample, and the product is over all possible genotypes of those individuals, and $G_{ij} = 2p_i p_j$, if $i \neq j$ or $G_{ij} = p_i^2$, if $i = j$.

The principle of the EM algorithm is the following:

1. Start with arbitrary (random) estimates of haplotype frequencies.
2. Use these estimates to compute expected genotype frequencies for each phenotype, assuming Hardy-Weinberg equilibrium (The E-step).
3. The relative genotype frequencies are used as weights for their two constituting haplotypes in a gene counting procedure leading to new estimates of haplotype frequencies (The M-step).
4. Repeat steps 2-3, until the haplotype frequencies reach equilibrium (do not change more than a predefined epsilon value).

Dempster et al (1977) have shown that the likelihood of the sample could only grow after each step of the EM algorithm. However, there is no guarantee that the resulting haplotype frequencies are maximum likelihood estimates. They can be just local optimal values. In fact, there is no obvious way to be sure that the resulting frequencies are those that globally maximize the likelihood of the data. This would need a complete evaluation of the likelihood for all possible genotype configurations of the sample. In order to check that the final frequencies are putative maximum likelihood estimates, one has generally to repeat the EM algorithm from many different starting points (many different initial haplotype frequencies). Several runs may give different final frequencies, suggesting the presence of several "peaks" in the likelihood surface, but one has to choose the solution that has the largest likelihood. It may also arise that several distinct peaks have the same likelihood, meaning that different haplotypic compositions explain equally well the observed data. At this point, there is no way to choose among the alternative solutions from a likelihood point of view. Some external information should be provided to make a decision.

Standard deviations of the haplotype frequencies are estimated by a parametric bootstrap procedure (see e.g. Rice, 1995), generating random samples from a population assumed to have haplotype frequencies equal to their maximum-likelihood values. For each bootstrap replicate, we apply the EM algorithm to get new maximum-likelihood haplotype frequencies. The standard deviation of each haplotype frequency is then estimated from the resulting distribution of haplotype frequencies. Note however that this procedure is quite computer intensive.

7.1.4 Linkage disequilibrium between pairs of loci

Depending on whether the haplotypic composition of the sample is known or not, we have implemented two different ways to test for the presence of pairwise linkage disequilibrium between loci.

We describe in detail below how the two tests are done.

7.1.4.1 Exact test of linkage disequilibrium (haplotypic data)

This test is an extension of Fisher exact probability test on contingency tables (Slatkin, 1994a). A contingency table is first built. The $k_1 \times k_2$ entries of the table are the observed haplotype frequencies (absolute values), with k_1 and k_2 being the number of alleles at locus 1 and 2, respectively. The test consists in obtaining the

probability of finding a table with the same marginal totals and which has a probability equal or less than the observed table. Under the null-hypothesis of no association between the two tested loci, the probability of the observed table is

$$L_0 = \frac{n!}{\prod_{i,j} n_{ij}} \prod_i (n_{i*} / n)^{n_{i*}} \prod_i (n_{*i} / n)^{n_{*i}},$$

where the n_{ij} 's denote the count of the haplotypes that have the i -th allele at the first locus and the j -th allele at the second locus, n_{i*} is the overall frequency of the i -th allele at the first locus ($i=1, \dots, k_1$) and n_{*i} is the count of the i -th allele at the second locus ($i=1, \dots, k_2$).

Instead of enumerating all possible contingency tables, a Markov chain is used to efficiently explore the space of all possible tables. This Markov chain consists in a random walk in the space of all contingency tables. It is done in such a way that the probability to visit a particular table corresponds to its actual probability under the null hypothesis of linkage equilibrium. A particular table is modified according to the following rules (see also Guo and Thompson, 1992; or Raymond and Rousset, 1995) :

1. We select in the table two distinct lines i_1, i_2 and two distinct columns j_1, j_2 at random.
2. The new table is obtained by decreasing the counts of the cells (i_1, j_1) and (i_2, j_2) and increasing the counts of the cells (i_1, j_2) and (i_2, j_1) by one unit. This leaves the marginal allele counts n_i unchanged.
3. The switch to the new table is accepted with a probability equal to

$$R = \frac{L_1}{L_0} = \frac{(n_{i_1, j_2} + 1)(n_{i_2, j_1} + 1)}{n_{i_1, j_1} n_{i_2, j_2}},$$

where R is just the ratio of the probabilities of the two tables.

The steps 1-3 are done a large number of times to explore a large amount of the space of all possible contingency tables having identical marginal counts. In order to start from a random initial position in the Markov chain, the chain is explored for a pre-defined number of steps (the dememorization phase) before the probabilities of the switched tables are compared to that of the initial table. The number of dememorization steps should be enough (some thousands) such as to allow the Markov chain to "forget" its initial state, and make it independent from its starting point. The P -value of the test is then taken as the proportion of the visited tables having a probability smaller or equal to the observed contingency table.

A standard error on P is estimated by subdividing the total amount of required steps into B batches (see Guo and Thompson, 1992, p. 367). A P -value is calculated separately for each batch. Let us denote it by P_i ($i=1, \dots, B$).

The estimated standard error is then calculated as

$$s.d.(P) = \sqrt{\frac{\sum_{i=1}^B (P - P_i)^2}{B(B-1)}}.$$

The process is stopped as soon as the estimated standard deviation is smaller than a pre-defined value specified by the user.

7.1.4.2 Likelihood ratio test of linkage disequilibrium (genotypic data, gametic phase unknown)

For genotypic data where the haplotypic phase is unknown, the test based on the Markov chain described above is not possible because the haplotypic composition of the sample is unknown, and is just estimated. Therefore, linkage disequilibrium between a pair of loci is tested for genotypic data using a likelihood-ratio test, whose empirical distribution is obtained by a permutation procedure (Slatkin and Excoffier, 1996). The likelihood of the data assuming linkage equilibrium (L_{H^*}) is computed by using the fact that, under this hypothesis, the haplotype frequencies are obtained as the product of the allele frequencies. The likelihood of the data *not* assuming linkage equilibrium (L_H) is obtained by applying the EM algorithm to estimate haplotype frequencies. The likelihood-ratio statistic given by

$$S = -2 \log\left(\frac{L_{H^*}}{L_H}\right)$$

should in principle follow a Chi-square distribution, with $(k_1-1)(k_2-1)$ degrees of freedom, but it is not always the case in small samples with large number of alleles per locus. In order to better approximate the underlying distribution of the likelihood-ratio statistic under the null hypothesis of linkage equilibrium, we use the following permutation procedure:

1. Permute the alleles between individuals at one locus only.
2. Re-estimate the likelihood of the data L_H by the EM algorithm. Note that L_{H^*} is unaffected by the permutation procedure.
3. Repeat steps 1-2 a large number of times to get the null distribution of L_H , and therefore the null distribution of S .

Note that this test of linkage disequilibrium assumes Hardy-Weinberg proportions of genotypes, and the rejection of the test could be also due to departure from Hardy-Weinberg equilibrium (see Excoffier and Slatkin, 1998)

7.1.4.3 Measures of gametic disequilibrium (haplotypic data)

- **D and D' coefficients:**

1. *D*: The classical linkage disequilibrium coefficient measuring deviation from random association between alleles at different loci (Lewontin and Kojima, 1960) is expressed as

$$D_{ij} = p_{ij} - p_i p_j,$$

where p_{ij} is the frequency of the haplotype having allele i at the first locus and allele j at the second locus, and p_i and p_j are the frequencies of alleles i and j , respectively.

2. D'_{ij} : The linkage disequilibrium coefficient D_{ij} standardized by the maximum value it can take

($D_{ij,max}$), given the allele frequencies (Lewontin 1964), as

$$D'_{ij} = \frac{D_{ij}}{D_{ij,max}},$$

where $D_{ij,max}$ takes one of the following values:

$$\min(p_i p_j, (1-p_i)(1-p_j)) \quad \text{if } D_{ij} < 0$$

$$\min((1-p_i)p_j, p_i(1-p_j)) \quad \text{if } D_{ij} > 0$$

7.1.5 Hardy-Weinberg equilibrium.

To detect significant departure from Hardy-Weinberg equilibrium, we follow the procedure described in Guo and Thompson (1992) using a test analogous to Fisher's exact test on a two-by-two contingency table, but extended to a triangular contingency table of arbitrary size. The test is done using a modified version of the Markov-chain random walk algorithm described Guo and Thomson (1992). The modified version gives the same results than the original one, but is more efficient from a computational point of view.

This test is obviously only possible for genotypic data. If the gametic phase is unknown, the test is only possible for each locus separately. For data with known gametic phase, it is also possible to test for the non random association of haplotypes into individuals. Note that this test assumes that the allele frequencies are given. Therefore, this test is not possible for data with recessive alleles, as in this case the allele frequencies need to be estimated.

A contingency table is first built. The $k \times k$ entries of the table are the observed allele frequencies and k is the number of alleles. Using the same notations as in section 8.2.2, the probability to observe the table under the null-hypothesis of no association is given by Levene (1949)

$$L_0 = \frac{n! \prod_{i=1}^k n_{i*}!}{(2n)! \prod_{i=1}^k \prod_{j=1}^i n_{ij}!} 2^H,$$

where H is the number of heterozygote individuals.

Much like it was done for the test of linkage disequilibrium, we explore alternative contingency tables having same marginal counts. In order to create a new contingency table from an existing one, we select two distinct lines i_1, i_2 and two distinct columns j_1, j_2 at random. The new table is obtained by decreasing the counts of the cells (i_1, j_1) (i_2, j_2) and increasing the counts of the cells (i_1, j_2) (i_2, j_1) by one unit. This leaves the allele counts n_i unchanged. The switch to the new table is accepted with a probability R equal to :

$$\begin{aligned}
1. \quad R &= \frac{L_{n+1}}{L_n} = \frac{n_{i_1 j_1} n_{i_2 j_2}}{(n_{i_1 j_2} + 1)(n_{i_2 j_1} + 1)} \frac{(1 + \delta_{i_1 j_1})(1 + \delta_{i_2 j_2})}{(1 + \delta_{i_1 j_2})(1 + \delta_{i_2 j_1})}, \text{ if } i_1 \neq j_1 \text{ or } i_2 \neq j_2 \\
2. \quad R &= \frac{L_{n+1}}{L_n} = \frac{n_{i_1 j_1} n_{i_2 j_2}}{(n_{i_1 j_2} + 1)(n_{i_2 j_1} + 2)} \frac{4}{1}, \text{ if } i_1 = j_1 \text{ and } i_2 = j_2 \\
3. \quad R &= \frac{L_{n+1}}{L_n} = \frac{n_{i_1 j_1} (n_{i_2 j_2} - 1)}{(n_{i_1 j_2} + 1)(n_{i_2 j_1} + 1)} \frac{1}{4}, \text{ if } i_1 = j_2 \text{ and } i_2 = j_1
\end{aligned}$$

As usual δ denotes the Kronecker function. R is just the ratio of the probabilities of the two tables. The switch to the new table is accepted if R is larger than 1.

The P -value of the test is the proportion of the visited tables having a probability smaller or equal to the observed (initial) contingency table. The standard error on the P -value is estimated like in the case of linkage disequilibrium using a system of batches (see section 7.1.4.1).

7.1.6 Neutrality tests.

7.1.6.1 Ewens-Watterson homozygosity test

This test is based on *Ewens* (1972) sampling theory of neutral alleles. *Watterson* (1978) has shown that the distribution of selectively neutral haplotype frequencies could be conveniently summarized by the sum of haplotype (allele) frequencies (F), equivalent to the expected homozygosity for diploids. This test can be performed equally well on diploid or haploid data, as the test statistic is not used for its biological meaning, but just as a way to summarize the allelic frequency distribution. The null distribution of F is generated by simulating random neutral samples having the same number of genes and the same number of haplotypes using the algorithm of *Stewart* (1977). The probability of observing random samples with F values identical or smaller than the original sample is recorded. This test is currently limited to sample sizes of 2000 genes or less and 1000 different alleles (haplotypes) or less. It can be used to test the hypothesis of selective neutrality and population equilibrium against either balancing selection or the presence of advantageous alleles.

7.1.6.2 Ewens-Watterson-Slatkin exact test

This test is essentially similar to that of *Watterson* (1978) test, but instead of using F as a summary statistic, it compares the probabilities of the random samples to that of the observed sample (*Slatkin* 1994b, 1996). The probability of obtaining a random sample having a probability smaller or equal to the observed sample is recorded. The results are in general very close to those of *Watterson's* homozygosity test. Note that the random samples are generated as explained for the Ewens-Watterson homozygosity test.

7.1.6.3 Chakraborty's test of population amalgamation

This test is also based on the infinite-allele model, and on Ewens (1972) sampling theory of neutral alleles. By simulation, Chakraborty (1990) has noticed that the number of alleles in a heterogeneous sample (drawn from a population resulting from the amalgamation of previously isolated populations) was larger than the number of alleles expected in a homogeneous neutral sample. He also noticed that the homozygosity of the sample was less sensitive to the amalgamation and therefore proposed to use the mutation parameter inferred from the homozygosity ($\hat{\theta}_{Hom}$) (see section 7.1.2.3.1) to compute the probability of observing a random neutral sample with a number of alleles similar or larger than the observed value ($\Pr(K \geq k_{obs})$) (see section 7.1.2.3.3 to see how this probability can be computed). It is an approximation of the conditional probability of observing some number of alleles given the observed homozygosity.

7.1.6.4 Tajima's test of selective neutrality

Tajima's (1989a) test is based on the infinite-site model without recombination, appropriate for short DNA sequences or RFLP haplotypes. It compares two estimators of the mutation parameter theta ($\theta = 2Mu$, with $M=2N$ in diploid populations or $M=N$ in haploid populations of effective size N). The test statistic D is then defined as

$$D = \frac{\hat{\theta}_{\pi} - \hat{\theta}_S}{\sqrt{\text{Var}(\hat{\theta}_{\pi} - \hat{\theta}_S)}}$$

where $\hat{\theta}_{\pi} = \hat{\pi}$ and $\hat{\theta}_S = S / \sum_{i=0}^{n-1} (1/i)$, and S is the number of segregating sites in the sample. The limits of confidence intervals around D may be found in Table 2 of Tajima's paper (Tajima 1989a) for different sample sizes. The P -value of the observed D under the hypothesis of population equilibrium and selective neutrality is computed here assuming a beta-distribution limited by minimum and maximum possible D values (see Tajima 1989a, p.589). Note that departure from the confidence interval can be due to factors other than selective effects, like population expansion, bottleneck, or heterogeneity of mutation rates (see Tajima, 1993; Aris-Brosou and Excoffier, 1996; or Tajima 1996, for further details).

7.2 Inter-population level methods

7.2.1 Population genetic structure inferred by analysis of variance (AMOVA)

The genetic structure of population is investigated here by an analysis of variance framework, as initially defined by Cockerham (1969, 1973), and extended by others (see e.g. Weir and Cockerham, 1984; Long 1986). The Analysis of Molecular Variance approach used in Arlequin (AMOVA, Excoffier et al. 1992) is essentially similar to other approaches based on analyses of variance of gene frequencies, but it takes into account the number of mutations between molecular haplotypes (which first need to be evaluated).

By defining groups of populations, the user defines a particular genetic structure that will be tested (see the input file notations for more details). A hierarchical analysis of variance partitions the total variance into components due to intra-individual differences, inter-individual differences, and/or inter-population differences. See also Weir (1996), for detailed treatments of hierarchical analyses. The variance components (σ_i^2 's) are used to compute fixation indices, as originally defined by Wright (1951, 1965), in terms of inbreeding coefficients, or later in terms of coalescent times by Slatkin (1991).

Formally, in the haploid case, we assume that the i -th haplotype frequency vector from the j -th population in the k -th group is a linear equation of the form

$$\mathbf{x}_{ijk} = \mathbf{x} + \mathbf{a}_k + \mathbf{b}_{jk} + \mathbf{c}_{ijk}.$$

The vector \mathbf{x} is the unknown expectation of \mathbf{x}_{ijk} , averaged over the whole study. The effects are \mathbf{a} for group, \mathbf{b} for population, and \mathbf{c} for haplotypes within a population within a group, assumed to be additive, random, independent, and to have the associated variance components σ_a^2 , σ_b^2 , and σ_c^2 , respectively. The total molecular variance (σ^2) is the sum of variances due to differences among haplotypes within a population (σ_c^2), the sum of variances due to differences among haplotypes in different populations within a group (σ_b^2), and those due to differences among the G populations (σ_a^2). The same framework could be extended to additional hierarchical levels, such as to accommodate, for instance, the variance component due to differences between haplotypes within diploid individuals.

Note that in the case of a simple hierarchical genetic structure consisting of haploid individuals in populations, the implemented form of the algorithm leads to a fixation index F_{ST} which is absolutely identical to the weighted average F -statistic over loci, $\hat{\theta}_w$, defined by Weir and Cockerham (1984) (see Michalakis and Excoffier 1996 for a formal proof). In terms of inbreeding coefficients and coalescence times, this F_{ST} can be expressed as

$$F_{ST} = \frac{f_0 - f_1}{1 - f_1} = \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1}, \quad (\text{Slatkin, 1991})$$

where f_0 is the probability of identity by descent of two different genes drawn from the same population, f_1 is the probability of identity by descent of two genes drawn from two different populations, \bar{t}_1 is the mean coalescence times of two genes drawn from two different populations, and \bar{t}_0 is the mean coalescence time of two genes drawn from the same population.

The significance of the fixation indices is tested using a non-parametric permutation approach described in Excoffier et al. (1992), consisting in permuting haplotypes, individuals, or populations, among individuals, populations, or groups of populations. After each permutation round, we recompute all statistics to get their

null distribution. Depending on the tested statistic and the given hierarchical design, different types of permutations are performed. Under this procedure, the normality assumption usual in analysis of variance tests is no longer necessary, nor is it necessary to assume equality of variance among populations or groups of populations. A large number of permutations (1,000 or more) is necessary to obtain some accuracy on the final probability.

We have implemented here 6 different types of hierarchical AMOVA. The number of hierarchical levels varies from two to four. In each of the situations, we describe the way the total sum of squares is partitioned, how the variance components and the associated F -statistics are obtained, and which permutation schemes are used for the significance test.

Before enumerating all the possible situations, we introduce some notations:

| | | |
|--------------|---|--|
| $SSD(T)$ | : | Total sum of squared deviations. |
| $SSD(AG)$ | : | Sum of squared deviations Among Groups of populations. |
| $SSD(AP)$ | : | Sum of squared deviations Among Populations. |
| $SSD(AI)$ | : | Sum of squared deviations Among Individuals. |
| $SSD(WP)$ | : | Sum of squared deviations Within Populations. |
| $SSD(WI)$ | : | Sum of squared deviations Within Individuals. |
| $SSD(AP/WG)$ | : | Sum of squared deviations Among Populations, Within Groups. |
| $SSD(AI/WP)$ | : | Sum of squared deviations Among Individuals, Within Populations. |
| G | : | Number of groups in the structure. |
| P | : | Total number of populations. |
| N | : | Total number of individuals for genotypic data or total number of gene copies for haplotypic data. |
| N_p | : | Number of individuals in population p for genotypic data or total number of gene copies in population p for haplotypic data. |
| N_g | : | Number of individuals in group g for genotypic data or total number of gene copies in group g for haplotypic data.. |

7.2.1.1 Haplotypic data, one group of populations

| Source of variation | Degrees of freedom | Sum of squares (SSD) | Variance component |
|---------------------|--------------------|----------------------|----------------------------|
| Among Populations | $P - 1$ | $SSD(AP)$ | $n\sigma_a^2 + \sigma_b^2$ |
| Within Populations | $N - P$ | $SSD(WP)$ | σ_b^2 |
| Total | $N - 1$ | $SSD(T)$ | σ_T^2 |

Where n and F_{ST} are defined by

$$n = \frac{N - \sum \frac{N_p^2}{N}}{P - 1},$$

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2}.$$

- We test σ_a^2 and F_{ST} by permuting haplotypes among populations.

7.2.1.2 Haplotypic data, several groups of populations

| Source of variation | Degrees of freedom | Sum of squares (SSD) | Variance components |
|--------------------------------------|--------------------|----------------------|---|
| Among Groups | $G - 1$ | SSD(AG) | $n''\sigma_a^2 + n'\sigma_b^2 + \sigma_c^2$ |
| Among Populations / Within Groups | $P - G$ | SSD(AP/WG) | $n\sigma_b^2 + \sigma_c^2$ |
| Within Populations | $N - P$ | SSD(WP) | σ_c^2 |
| Total: | $N - 1$ | SSD(T) | σ_T^2 |

Where the n 's and the F -statistics are defined by:

$$S_G = \sum_{g \in G} \sum_{p \in g} \frac{N_p^2}{N_g}, \quad n = \frac{N - S_G}{P - G},$$

$$n' = \frac{S_G - \sum_{p \in P} \frac{N_p^2}{N}}{G - 1}, \quad n'' = \frac{N - \sum_{g \in G} \frac{N_g^2}{N}}{G - 1}$$

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2}, \quad F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2} \quad \text{and} \quad F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2}$$

- We test σ_c^2 and F_{ST} by permuting haplotypes among populations among groups.
- We test σ_b^2 and F_{SC} by permuting haplotypes among populations within groups.
- We test σ_a^2 and F_{CT} by permuting populations among groups.

7.2.1.3 Genotypic data, one group of populations, no within- individual level

| Source of variation | Degrees of freedom | Sum of squares (SSD) | Variance components |
|---------------------|--------------------|----------------------|----------------------------|
| Among Populations | $P - 1$ | SSD(AP) | $n\sigma_a^2 + \sigma_b^2$ |
| Within Populations | $2N - P$ | SSD(WP) | σ_b^2 |
| Total: | $2N - 1$ | SSD(T) | σ_T^2 |

Where n and F_{ST} are defined by

$$n = \frac{2N - \sum_P \frac{2N^2_p}{N}}{P - 1},$$

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2}.$$

If the gametic phase is known:

- We test σ_a^2 and F_{ST} by permuting haplotypes among populations.

If the gametic phase is unknown:

- We test σ_a^2 and F_{ST} by permuting individual genotypes among populations.

7.2.1.4 Genotypic data, several groups of populations, no within- individual level

| Source of Variation | Degrees of freedom | Sum of squares (SSD) | Variance components |
|--------------------------------------|--------------------|----------------------|---|
| Among Groups | $G - 1$ | SSD(AG) | $n''\sigma_a^2 + n'\sigma_b^2 + \sigma_c^2$ |
| Among Populations / Within Groups | $P - G$ | SSD(AP/WG) | $n\sigma_b^2 + \sigma_c^2$ |
| Within Populations | $2N - P$ | SSD(WP) | σ_c^2 |
| Total: | $2N - 1$ | SSD(T) | σ_T^2 |

Where the n 's and the F -statistics are defined by:

$$S_G = \sum_{g \in G} \sum_{p \in g} \frac{2N^2_p}{N_g}, \quad n = \frac{2N - S_G}{P - G},$$

$$n' = \frac{S_G - \sum_{p \in P} \frac{2N^2_p}{N}}{G - 1}, \quad n'' = \frac{2N - \sum_{g \in G} \frac{2N^2_g}{N}}{G - 1},$$

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2}, \quad F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} \quad \text{and} \quad F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}.$$

If the gametic phase is known:

- We test σ_c^2 and F_{ST} by permuting haplotypes among populations and among groups.
- We test σ_b^2 and F_{SC} by permuting haplotypes among populations but within groups.

If the gametic phase is not known:

- We test σ_c^2 and F_{ST} by permuting individual genotypes among populations and among groups.
- We test σ_b^2 and F_{SC} by permuting individual genotypes among populations but within groups.

In all cases:

- We test σ_a^2 and F_{CT} by permuting whole populations among groups.

7.2.1.5 Genotypic data, one population, within- individual level

| Source of variation | Degrees of freedom | Sum of squares (SSD) | Variance component |
|---------------------|--------------------|----------------------|----------------------------|
| Among Individuals | $N - 1$ | SSD(AI) | $2\sigma_a^2 + \sigma_b^2$ |
| Within Individuals | N | SSD(WI) | σ_b^2 |
| Total: | $2N - 1$ | SSD(T) | σ_T^2 |

Where F_{IS} is defined as:

$$F_{IS} = \frac{\sigma_a^2}{\sigma_T^2}.$$

- We test σ_a^2 and F_{IS} by permuting haplotypes among individuals.

7.2.1.6 Genotypic data, one group of populations, within- individual level

| Source of Variation | Degrees of freedom | Sum of squares (SSD) | Variance component |
|---|--------------------|----------------------|--|
| Among Populations | $P - 1$ | SSD(AP) | $n\sigma_a^2 + 2\sigma_b^2 + \sigma_c^2$ |
| Among Individuals / Within Populations | $N - P$ | SSD(AI/WP) | $2\sigma_b^2 + \sigma_c^2$ |
| Within Individuals | N | SSD(WI) | σ_c^2 |
| Total | $2N - 1$ | SSD(T) | σ_T^2 |

Where n and the F -statistics are defined by:

$$n = \frac{2N - \sum_{p \in P} \frac{2N^2}{N}}{P - 1}$$

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2}, \quad F_{IT} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} \quad \text{and} \quad F_{IS} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}.$$

- We test σ_c^2 and F_{IT} by permuting haplotypes among individuals among populations.

- We test σ_b^2 and F_{IS} by permuting haplotypes among individuals within populations.
- We test σ_a^2 and F_{ST} by permuting individual genotypes among populations.

7.2.1.7 Genotypic data, several groups of populations, within- individual level

| Source of Variation: | Degrees of freedom | Sum of squares (SSD) | Variance component |
|---|--------------------|----------------------|--|
| Among Groups | $G - 1$ | SSD(AG) | $n'\sigma_a^2 + n'\sigma_b^2 + 2\sigma_c^2 + \sigma_d^2$ |
| Among Populations / Within Groups | $P - G$ | SSD(AP/WG) | $n\sigma_b^2 + 2\sigma_c^2 + \sigma_d^2$ |
| Among Individuals / Within Populations | $N - P$ | SSD(AI/WP) | $2\sigma_c^2 + \sigma_d^2$ |
| Within Individuals | N | SSD(WI) | σ_d^2 |
| Total: | $2N - 1$ | SSD(T) | σ_T^2 |

Where the n 's and the F -statistics are defined by:

$$n = \frac{2N - \sum_{g \in G} \sum_{p \in g} \frac{2N_p^2}{N_g}}{P - G}, \quad n' = \frac{\sum_{g \in G} \frac{(N - N_g)}{N_g} \sum_{p \in g} 2N_p^2}{N(G - 1)}, \quad n'' = \frac{\sum_{g \in G} 2N_g^2}{N(G - 1)}$$

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2}, \quad F_{IT} = \frac{\sigma_a^2 + \sigma_b^2 + \sigma_c^2}{\sigma_T^2}, \quad F_{IS} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_d^2} \quad \text{and} \quad F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2 + \sigma_d^2}.$$

- We test σ_d^2 and F_{IT} by permuting haplotypes among populations and among groups.
- We test σ_c^2 and F_{IS} by permuting haplotypes among individuals within populations.
- We test σ_b^2 and F_{SC} by permuting individual genotypes among populations but within groups.
- We test σ_a^2 and F_{CT} by permuting populations among groups.

7.2.2 Population pairwise genetic distances

The pairwise F_{ST} s can be used as short-term genetic distances between populations, with the application of a slight transformation to linearize the distance with population divergence time (Reynolds et al. 1983; Slatkin, 1995).

The pairwise F_{ST} values are given in the form of a matrix.

The null distribution of pairwise F_{ST} values under the hypothesis of no difference between the populations is obtained by permuting haplotypes between populations. The P -value of the test is the proportion of permutations leading to a F_{ST} value larger or equal to the observed one. The P -values are also given in matrix form.

Three other matrices are computed from the F_{ST} values:

- A matrix of coancestry coefficients (Reynolds et al. 1983):

Since F_{ST} between pairs of stationary haploid populations of size N having diverged t generations ago varies approximately as

$$F_{ST} = 1 - \left(1 - \frac{1}{N}\right)^t \approx 1 - e^{-t/N}$$

The genetic distance $D = -\log(1 - F_{ST})$ is thus approximately proportional to t/N for short divergence times.

- A matrix of Slatkin linearized F_{ST} 's (Slatkin 1995):

Slatkin considers a simple demographic model where two haploid populations of size N have diverged τ generations ago from a population of identical size. These two populations have remained isolated ever since, without exchanging any migrants. Under such conditions, F_{ST} can be expressed in terms of the coalescence times \bar{t}_1 , which is the mean coalescence time of two genes drawn from two different populations, and \bar{t}_0 which is the mean coalescence time of two genes drawn from the same population.

Using the analysis of variance approach, the F_{ST} 's are expressed as

$$F_{ST} = \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1} \quad (\text{Slatkin, 1991, 1995})$$

Because, \bar{t}_0 is equal to N generations (see e.g. Hudson, 1990), and \bar{t}_1 is equal to $\tau + N$ generations, the above expression reduces to

$$F_{ST} = \frac{\tau}{\tau + N}.$$

Therefore, the ratio $D = F_{ST} / (1 - F_{ST})$ is equal to τ / N , and is therefore proportional to the divergence time between the two populations.

- A matrix of M values ($M = Nm$ for haploid populations, $M = 2Nm$ for diploid populations).

This matrix is computed under very different assumptions than the two previous matrices. Assume that two populations of size N exchange a fraction m of migrants each generation, and that the mutation rate u is negligible as compared to the migration rate m . In this case, we have the following simple relationship at equilibrium between migration and drift,

$$F_{ST} = \frac{1}{2M + 1}$$

Therefore, M , which is the absolute number of migrants exchanged between the two populations, can be estimated by

$$M = \frac{1 - F_{ST}}{2F_{ST}}$$

7.2.3 Exact tests of population differentiation

We test the hypothesis of a random distribution of k different haplotypes or genotypes among r populations as described in Raymond and Rousset (1995). This test is analogous to Fisher's exact test on a 2×2 contingency table extended to a $r \times k$ contingency table. All potential states of the contingency table are explored with a Markov chain similar to that described for the case of the linkage disequilibrium test (section 7.1.4.1). During this random walk between the states of the Markov chain, we estimate the probability of observing a table less or equally likely than the observed sample configuration under the null hypothesis of panmixia.

For haplotypic data, the table is built using sample haplotype frequencies (Raymond and Rousset 1995).

For genotypic data with unknown gametic phase, the contingency table is built from sample genotype frequencies (Goudet et al. 1996).

As it was done previously, an estimation of the error on the P -value is done by partitioning the total number of steps into a given number of batches (see section 7.1.4.1).

8 APPENDIX

8.1 Overview of input file keywords

| Keywords | Description | Possible values |
|------------------------|---|---|
| [Profile] | | |
| Title | A title describing the present analysis | A string of alphanumeric characters within double quotes |
| NbSamples | The number of different samples listed in the data file | A positive integer larger than zero |
| DataType | The type of data to be analyzed (only one type of data per project file is allowed) | STANDARD, DNA, RFLP, MICROSAT, FREQUENCY |
| GenotypicData | Specifies if genotypic or gametic data is available | 0 (haplotypic data), 1 (genotypic data) |
| LocusSeparator | The character used to separate adjacent loci | WHITESPACE, TAB, NONE, or any character other than "#", or the character specifying missing data Default: WHITESPACE |
| GameticPhase | Specifies if the gametic phase is known (for genotypic data only) | 0 (gametic phase not known), 1 (known gametic phase) Default: 1 |
| RecessiveData | Specifies whether recessive alleles are present at all loci (for genotypic data) | 0 (co-dominant data), 1 (recessive data) Default: 0 |
| RecessiveAllele | Specifies the code for the recessive allele | Any string within quotation marks This string can be explicitly used in the input file to indicate the occurrence of a recessive homozygote at one or several loci. Default: "null" |
| MissingData | A character used to specify the code for missing data | "?" or any character within quotes, other than those previously used Default: "?" |
| Frequency | Specifies the format of haplotype frequencies | ABS (absolute values), REL (relative values: absolute values will be found by multiplying the relative frequencies by the sample sizes) Default: ABS |

| | | |
|---------------------------|--|---|
| CompDistMatrix | Specifies if the distance matrix has to be computed from the data | 0 (use any specified distance matrix), 1 (compute distance matrix from haplotypic information) Default: 0 |
| FrequencyThreshold | The minimum frequency a haplotype has to reach for being listed in any output file | A real number between 1e-2 and 1e-7. Default: 1e-5 |
| EpsilonValue | The EM algorithm convergence criterion. (For advanced users only) | A real number between 1e-7 and 1e-12. Default: 1e-7 |

| Keywords | Description | Possible values |
|--|---|---|
| [Data] | | |
| [[HaplotypeDefinition]] (facultative section) | | |
| HaplListName | The name of a haplotype definition list | A string within quotation marks |
| HaplList | The list of haplotypes listed within braces ({...}) | A series of haplotype definitions given on separate lines for each haplotype. Each haplotype is defined by a haplotype label and a combination of alleles at different loci. The Keyword EXTERN followed by a string within quotation marks may be used to specify that a given haplotype list is in a different file |

| Keywords | Description | Possible values |
|---|---|---|
| [Data] | | |
| [[DistanceMatrix]] (facultative section) | | |
| MatrixName | The name of the distance matrix | A string within quotation marks |
| MatrixSize | The size of the matrix | A positive integer larger than zero (corresponding to the number of haplotypes listed in the haplotype list) |
| LabelPosition | Specifies whether haplotypes labels are entered by row or by column | ROW (the haplotype labels will be entered consecutively on one or several lines, within the MatrixData segment, before the distance matrix elements), COLUMN (the haplotype labels will be entered as the first column of each row of the distance matrix itself) |
| MatrixData | The matrix data itself listed within braces ({...}) | The matrix data will be entered as a format-free lower-diagonal matrix. The haplotype labels can be either entered consecutively on one or several lines (if LabelPosition=ROW), or entered at the first column of each row (if labelPosition=COLUMN). The special keyword EXTERN may be used followed by a file name within quotation marks, stating that the data must be read in an another file |

| Keywords | Description | Possible values |
|--------------------|---|---|
| [Data] | | |
| [[Samples]] | | |
| SampleName | The name of the sample. This keyword is used to mark the beginning of a sample definition | A string within quotation marks |
| SampleSize | Specifies the sample size | An integer larger than zero. For haplotypic data, it must specify the number of gene copies in the sample. For genotypic data, it must specify the number of individuals in the sample. |
| SampleData | The sample data listed within braces ({...}) | The keyword EXTERN may be used followed by a file name within quotation marks, stating that the data must be read in a separate file. The SampleData keyword ends a sample definition |

| Keywords | Description | Possible values |
|------------------------|--|---|
| [Data] | | |
| [[Structure]] | | |
| | (facultative section) | |
| StructureName | The name of a given genetic structure to test | A string of characters within quotation marks |
| NbGroups | The number of groups of populations | An integer larger than zero |
| IndividualLevel | Specifies whether the level of genetic variability within individuals has to be taken into account (for genotypic data only) | 0 (the component of variance due to differences between haplotypes within individuals will be ignored) 1 (the component of variance due to differences between haplotypes within individuals, and its associated statistics will be computed) |
| Group | The definition of a group of samples, identified by their SampleName listed within braces ({...}) | A series of strings within quotation marks all enclosed within braces, and, if desired, on separate lines |

9 REFERENCES

- Abramovitz, M., and I. A. Stegun, 1970 Handbook of Mathematical Functions. Dover, New York.
- Aris-Brosou, S., and L. Excoffier, 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.* 13: 494-504.
- Cavalli-Sforza, L. L., and W. F. Bodmer, 1971 The Genetics of Human Populations. W.H. Freeman and Co., San Francisco, CA.
- Chakraborty, R. 1990 Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47:87-94.
- Chakraborty, R., and K. M. Weiss, 1991 Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *Am. J. Hum. Genet.* 86: 497-506.
- Cockerham, C. C., 1969 Variance of gene frequencies. *Evolution* 23: 72-83.
- Cockerham, C. C., 1973 Analysis of gene frequencies. *Genetics* 74: 679-700.
- Dempster, A., N. Laird and D. Rubin, 1977 Maximum likelihood estimation from incomplete data via the EM algorithm. *J Roy Statist Soc* 39: 1-38.
- Efron, B. 1982 The Jackknife, the Bootstrap and other Resampling Plans. Regional Conference Series in Applied Mathematics, Philadelphia:.
- Ewens, W.J. 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87-112.
- Ewens, W.J. 1977 Population genetics theory in relation to the neutralist-selectionist controversy. In: *Advances in human genetics*, edited by Harris, H. and Hirschhorn, K. New York: Plenum Press, p. 67-134.
- Excoffier, L., Smouse, P., and Quattro, J. 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479-491.
- Excoffier, L. and M. Slatkin. 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12:921-927
- Excoffier, L., and M. Slatkin, 1998 Incorporating genotypes of relatives into a test of linkage disequilibrium. *Am. J. Hum. Genet.* (January issue)
- Goudet, J., M. Raymond, T. de Meertis and F. Rousset, 1996 Testing differentiation in diploid populations. *Genetics* 144: 1933-1940.
- Guo, S. and Thompson, E. 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48:361-372.
- Harpending, R. C., 1994 Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum. Biol.* 66: 591-600.

- Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by Futuyama, and J. D. Antonovics. Oxford University Press, New York.
- Jukes, T. and Cantor, C. 1969 Evolution of protein molecules. In: *Mammalian Protein Metabolism*, edited by Munro HN, New York:Academic press, p. 21-132.
- Kimura, M. 1980 A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.
- Kumar, S., Tamura, K., and M. Nei. 1993 MEGA, Molecular Evolutionary Genetic Analysis ver 1.0. The Pennsylvania State University, University Park, PA 16802.
- Lange, K., 1997 *Mathematical and Statistical Methods for Genetic Analysis*. Springer, New York.
- Levene H. (1949). On a matching problem arising in genetics. *Annals of Mathematical Statistics* 20, 91-94.
- Lewontin, R. C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49-67.
- Lewontin, R. C., and K. Kojima. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* 14: 450-472.
- Long, J. C., 1986 The allelic correlation structure of Gainj and Kalam speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* 112: 629-647.
- Michalakis, Y. and Excoffier, L. , 1996 A generic estimation of population subdivision using distances between alleles with special reference to microsatellite loci. *Genetics* 142:1061-1064.
- Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY, USA.
- Raymond M. and F. Rousset. 1994 GenePop. ver 3.0. Institut des Sciences de l'Evolution. Université de Montpellier, France.
- Raymond M. and F. Rousset. 1995 An exact test for population differentiation. *Evolution* 49:1280-1283.
- Reynolds, J., Weir, B.S., and Cockerham, C.C. 1983 Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105:767-779.
- Rice, J.A. 1995 *Mathematical Statistics and Data Analysis*. 2nd ed. Duxbury Press: Belmont, CA
- Rogers, A., 1995 Genetic evidence for a Pleistocene population explosion. *Evolution* 49: 608-615.
- Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9: 552-569.
- Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* 142: 1357-1362.
- Slatkin, M., 1991 Inbreeding coefficients and coalescence times. *Genet. Res. Camb.* 58: 167-175.
- Slatkin, M. 1994a Linkage disequilibrium in growing and stable populations. *Genetics* 137:331-336.
- Slatkin, M. 1994b An exact test for neutrality based on the Ewens sampling distribution. *Genet. Res.* 64(1):71-74.

- Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.
- Slatkin, M., 1996 A correction to the exact test based on the Ewens sampling distribution. *Genet. Res.* 68: 259-260.
- Slatkin, M. and Excoffier, L. 1996 Testing for linkage disequilibrium in genotypic data using the EM algorithm. *Heredity* 76:377-383.
- Stewart, F. M. 1977 Computer algorithm for obtaining a random set of allele frequencies for a locus in an equilibrium population. *Genetics* 86:482-483.
- Strobeck, K., 1987 Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. *Genetics* 117: 149-153.
- Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460.
- Tajima, F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595,.
- Tajima, F. 1989b. The effect of change in population size on DNA polymorphism. *Genetics* 123:597-601,.
- Tajima, F. 1993. Measurement of DNA polymorphism. In: *Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology*, edited by Takahata, N. and Clark, A.G., Tokyo, Sunderland, MA:Japan Scientific Societies Press, Sinauer Associates, Inc., p. 37-59.
- Tajima, F. and Nei, M. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1:269-285.
- Tajima, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* 143: 1457-1465.
- Tamura, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.* 9: 678-687.
- Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10: 512-526.
- Uzell, T., and K. W. Corbin, 1971 Fitting discrete probability distribution to evolutionary events. *Science* 172: 1089-1096.
- Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. *Theor.Popul.Biol.* 7: 256-276.
- Watterson, G. 1978. The homozygosity test of neutrality. *Genetics* 88:405-417
- Watterson, G. A., 1986 The homozygosity test after a change in population size. *genetics* 112: 899-907.
- Weir, B. S., 1996 *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Assoc., Inc., Sunderland, MA, USA.

- Weir, B.S. and Cockerham, C.C. 1984 Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370.
- Wright, S., 1951 The genetical structure of populations. *Ann.Eugen.* 15: 323-354.
- Wright, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. *Evol* 19: 395-420.
- Zouros, E., 1979 Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics* 92: 623-646.