Arlequin (version 3.0): An integrated software package for population genetics data analysis

Laurent Excoffier, Guillaume Laval, Stefan Schneider

Computational and Molecular Population Genetics Lab, , Zoological Institute, University of Berne, Baltzerstrasse 6, 3012 Berne, Switzerland

Abstract: Arlequin ver 3.0 is a software package integrating several basic and advanced methods for population genetics data analysis, like the computation of standard genetic diversity indices, the estimation of allele and haplotype frequencies, tests of departure from linkage equilibrium, departure from selective neutrality and demographic equilibrium, estimation or parameters from past population expansions, and thorough analyses of population subdivision under the AMOVA framework. Arlequin 3 introduces a completely new graphical interface written in C++, a more robust semantic analysis of input files, and two new methods: a Bayesian estimation of gametic phase from multi-locus genotypes, and an estimation of the parameters of an instantaneous spatial expansion from DNA sequence polymorphism. Arlequin can handle several data types like DNA sequences, microsatellite data, or standard multi-locus genotypes. A Windows version of the software is freely available on http://cmpg.unibe.ch/software/arlequin3.
Keywords: Computer package, population genetics, genetic data analysis, AMOVA, EM algorithm, gametic phase estimation, spatial expansion.

Introduction

Most genetic studies on non-model organisms require a description of the pattern of diversity within and between populations, based on a variety of markers often including mitochondrial DNA (mtDNA) sequences and microsatellites. The genetic data are processed to extract information on the mating system, the extent of population subdivision, the past demography of the population, or on departure from selective neutrality at some loci. A series of computer packages have been developed in the last 10 years to assist researchers in performing basic population genetics analyses like Arlequin2 (Schneider et al. 2000), DNASP (Rozas et al. 2003), FSTAT (Goudet 1995), GENEPOP (Raymond and Rousset 1995b), or GENETIX (Belkhir et al. 2004). These programs have been widely used in the molecular ecology and conservation genetics community (Labate 2000; Luikart and England 1999; Schnabel et al. 1998). Among these, Arlequin is a very versatile (though not universal) program, and complements the other programs listed above. It can handle several data types like RFLPs, DNA sequences, microsatellite data, allele frequencies, or standard multi-locus genotypes, while allowing the user to carry out the same types of analyses irrespective of the data types.

We present here the version 3 of Arlequin with additional methods extending its capacities for the handling of unphased multi-locus genotypes and for the estimation of parameters of a spatial expansion. Note that these new developments are mainly implementations of new methodologies developed in our lab. We believe these methods will be useful to the research community, but we do not claim that alternative methods implemented by other groups in other programs are inadequate. A new graphical interface has been developed to provide a better integration of the different analyses into a common framework, and an easier exploration of the data by performing a wide variety of analyses with different settings. The tight coupling of Arlequin with the simulation programs SIMCOAL2 (Laval and Excoffier 2004) and SPLATCHE (Currat et al. 2004) should also make it useful to describe patterns of genetic diversity under complex evolutionary scenarios.

Methods implemented in Arlequin

Arlequin provides methods to analyse patterns of genetic diversity within and between population samples.

Intra-population methods

- Computation of different standard genetic indices, like the number of segregating sites, the number of dif-

Correspondence: Laurent Excoffier, Tel: +41 31 631 30 31, Fax: +41 31 631 48 88 Email: laurent.excoffier@zoo.unibe.ch

ferent alleles, the heterozygosity, the base composition of DNA sequences, gene diversity, or the population effective size N_e scaled by the mutation rate μ as $\theta=4N_e u$.

- Maximum-likelihood estimation of allele and haplotype frequencies via the EM algorithm (Excoffier and Slatkin 1995).
- Estimation of the gametic phase from multilocus genotypes via the Excoffier-Laval-Balding (ELB) algorithm (Excoffier et al. 2003).
- Estimation of the parameters of a demographic (Rogers and Harpending 1992; Schneider and Excoffier 1999) or a spatial (Excoffier 2004; Ray et al. 2003) expansion, from the mismatch distribution computed on DNA sequences.
- Calculation of several measures of linkage disequilibrium (LD) like D, D', or r^2 (Hedrick 1987), and test of non-random association of alleles at different loci when the gametic phase is known (Weir 1996) or unknown (Slatkin and Excoffier 1996).
- Exact test of departure from Hardy-Weinberg equilibrium (Guo and Thompson 1992).
- Computation of Tajima's D (Tajima 1989) and Fu's F_s (Fu 1997) statistics, and test of their significance by coalescent simulations (Hudson 1990; Nordborg 2003) under the infinite-site model.
- Tests of selective neutrality under the infinite-alleles model, like the Ewens-Watterson test (Slatkin 1996; Watterson 1978), and Chakraborty's amalgamation test (Chakraborty 1990).

Inter-population methods

- Search for shared haplotypes between populations
- Analysis of population subdivision under the AMOVA framework (Excoffier 2003; Excoffier et al. 1992), with three hierarchical levels: genes within individuals, individuals within demes, demes within groups of demes. Computation of F-statistics like the local inbreeding coefficient F_{IS} or the index of population differentiation F_{ST}.
- Computation of genetic distances between populations related to the pairwise F_{ST} index (Gaggiotti and Excoffier 2000; Reynolds et al. 1983; Slatkin 1995).
- Exact test of population differentiation (Goudet et al. 1996; Raymond and Rousset 1995a).
- A simple assignment test of individual genotypes to populations according to their likelihood (Paetkau et al. 1997).
- Computation of correlations or partial correlations between a set of 2 or 3 distance matrices (Mantel test: Smouse et al. 1986)

New features in Arlequin 3

- Version 3 of Arlequin integrates the core computational routines and the interface in a single program written in C++ for the Windows environment. The interface has been entirely redesigned to provide better usability.
- Incorporation of two new methods to estimate gametic phase and haplotype frequencies:
  - ◊ The ELB algorithm (Excoffier et al. 2003) is a pseudo-Bayesian approach aiming at reconstructing the gametic phase of multi-locus genotypes, and the estimation of the haplotype frequencies are a by-product of this process. Phase updates are made on the basis of a window of neighbouring loci, and the window size varies according to the local level of linkage disequilibrium.
  - ◊ The EM zipper algorithm, which is an extension of the EM algorithm for estimating haplotype frequencies (Excoffier and Slatkin 1995), aims at estimating the haplotype frequencies in unphased multi-locus genotypes. The estimation of the gametic phases are a by-product of this process. It proceeds by adding loci one at a time and progressively extending the length of the reconstructed haplo-

types. With this method, Arlequin does not need to build all possible genotypes for each individual like in the conventional EM algorithm, but it only considers the genotypes whose sub-haplotypes have non-null estimated frequencies. It can thus handle a much larger number of polymorphic sites than the strict EM algorithm. It also gives final haplotype frequencies that often have a higher likelihood than those estimated under the strict EM algorithm, due to the difficulty in exploring the space of all possible genotypes when the number of polymorphic loci in the sample is large. Note that this version of the EM algorithm is equivalent to that implemented in the SNPHAP program by David Clayton fully described on http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt, and whose efficiency for inferring gametic phase has been favorably evaluated (Adkins 2004).

- Estimation of the parameters of a spatial expansion (age of the expansion and deme size scaled by the mutation rate, as well as the number of migrants exchanged between neighbouring demes) from the patterns of polymorphism in a sample of DNA sequences. The estimation is based on a simple model of instantaneous and infinite range expansion, where some time ago, a single deme instantaneously colonized an infinite number of demes subsequently interconnected by migration (as under an infinite-island model) (Excoffier 2004). The parameters are obtained by a least-square approach maximizing the fit between the observed and expected distribution of pairwise differences (the mismatch distribution) computed on DNA sequences. Confidence intervals of the estimates are obtained under a parametric bootstrap approach involving the simulation of an instantaneous expansion under a coalescent framework.
- Estimation of confidence intervals for *F*-statistics estimated under the AMOVA framework by bootstrapping over loci for multi-locus data. A minimum of 8 loci are necessary for the computation of these confidence intervals.
- A completely rewritten and more robust input file parsing procedure, giving more precise information on the location of potential syntax and format errors in input files.
- Use of the ELB algorithm described above to generate samples of phased multi-locus genotypes, which allows one to analyse unphased multi-locus genotype data as if the phase was known. The phased data sets are output in Arlequin projects that can be analysed in a batch mode to obtain the distribution of statistics taking phase uncertainty into account.
- New output files fully compatible with modern web browsers.

Availability

A Windows executable version Arlequin ver 3 can be freely downloaded on http://cmpg.unibe.ch/software/arlequin3, together with an up-to-date user manual in Adobe Acrobat PDF format incorporating more technical details on the methods used in Arlequin 3, as well as several example files.

Acknowledgements

This work was partially made possible thanks to a Swiss NSF grant No 31-56755.99 to LE.

References

Adkins RM. 2004. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genet.* 5**:** 22.

Belkhir K, Borsa P, Chikhi L et al. 2004. GENETIX 4.05, logiciel sous Windows pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier.

Chakraborty R. 1990. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am J Hum Genet.* 47**:** 87-94.

Currat M, Ray N and Excoffier L. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol.* 4**:** 139-142.

Excoffier L. 2003. Analysis of Population Subdivision. In Balding D Bishop M, and Cannings C, eds. Handbook of Statistical Genetics, 2nd Edition. New York: John Wiley & Sons, Ltd. p 713-750.

Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol.* 13**:** 853-864.

Excoffier L, Laval G and Balding D. 2003. Gametic phase estimation over large genomic regions using an adaptive window approach. *Mol Ecol.* 1**:** 7-19.

Excoffier L and Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 12**:** 921-927.

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.

Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. *Genetics.* 147**:** 915-925.

Gaggiotti O and Excoffier L. 2000. A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proceedings of the Royal Society London B.* 267**:** 81-87.

Goudet J. 1995. Fstat version 1.2: a computer program to calculate F-statistics. *J Heredity.* 86**:** 485-486.

Goudet J, Raymond M, de Meeüs T et al. 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.

Hedrick P. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics.* 117**:** 331-3412.

Hudson RR. 1990. Gene genealogies and the coalescent process. In Futuyma DJ and Antonovics JD, eds. Oxford Surveys in Evolutionary Biology. New York: Oxford University Press. p 1-44.

Labate JA. 2000. Software for Population Genetic Analyses of Molecular Marker Data. *Crop Sci.* 40**:** 1521-1528.

Laval G and Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics.* 20**:** 2485-2487.

Luikart G and England PR. 1999. Statistical analysis of microsatellite DNA data. *Trends Ecol Evol.* 14**:** 253-256.

Nordborg M. 2003. Coalescent Theory. In Balding D Bishop M, and Cannings C, eds. Handbook of Statistical Genetics, 2nd edition. New York: John Wiley & Sons Ltd. p 602-635.

Paetkau D, Waits LP, Clarkson PL et al. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics.* 147**:** 1943-1957.

Ray N, Currat M and Excoffier L. 2003. Intra-Deme Molecular Diversity in Spatially Expanding Populations. *Mol. Biol. Evol.* 20**:** 76-86.

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

Raymond M and Rousset F. 1995b. GENEPOP Version 1.2: Population genetics software for exat tests and ecumenicism. *J Heredity.* 248-249.

Reynolds J, Weir BS and Cockerham CC. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics.* 105**:** 767-779.

Rogers AR and Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol.* 9**:** 552-569.

Rozas J, Sanchez-DelBarrio JC, Messeguer X et al. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics.* 19**:** 2496-2497.

Schnabel A, Beerli P, Estoup A et al. 1998. A guide to software packages for data analysis in molecular ecology. In Carvalho G, eds. Advances in Molecular Ecology. Amsterdam: IOS Press. pp 291-303.

Schneider S and Excoffier L. 1999. Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: Application to human mitochondrial DNA. *Genetics.* 152**:** 1079-1089.

Schneider S, Roessli D and Excoffier L. 2000. Arlequin: a software for population genetics data analysis. User manual ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva.

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.

Smouse PE, Long JC and Sokal RR. 1986. Multiple regression and correlation extensions of the Mantel Test of matrix correspondence. *Syst Zool.* 35**:** 627-632.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 123**:** 585-595.

Watterson G. 1978. The homozygosity test of neutrality. *Genetics.* 88**:** 405-417.

Weir BS. 1996. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Assoc., Inc.: Sunderland, MA, USA.