REVIEWS

Computer programs for population genetics data analysis: a survival guide

Laurent Excoffier and Gerald Heckel

Abstract | The analysis of genetic diversity within species is vital for understanding evolutionary processes at the population level and at the genomic level. A large quantity of data can now be produced at an unprecedented rate, requiring the use of dedicated computer programs to extract all embedded information. Several statistical packages have been recently developed, which offer a panel of standard and more sophisticated analyses. We describe here the functionalities, special features and assumptions of more than 20 such programs, indicate how they can interoperate, and discuss new directions that could lead to improved software and analyses.

Linkage disequilibrium
The non-random association of alleles at different loci.

Gametic phase
In a diploid individual, it represents the original allelic combinations that an individual received from its parents. It is therefore a particular association of alleles at different loci on the same chromosome, which is often unknown.

Selective neutrality
Null model of evolution that assumes that all the alleles observed at given locus are functionally equivalent.

Computational and Molecular Population Genetics (CMPG) Laboratory, Zoological Institute, University of Berne, Baltzerstrasse 6, 3012 Berne, Switzerland.
Correspondence to L.E.
e-mail:
laurent.excoffier@zoo.unibe.ch
doi:10.1038/nrg1904
Published online
22 August 2006

Recent population genetics methods can provide accurate information on the past demography of a population, which is necessary to correctly interpret patterns of linkage disequilibrium[1], recognize regions of the genome that are under selection[2,3] or help to develop good conservation strategies and priorities[4]. At the same time, the advent of cost-efficient, large-scale genotyping techniques has greatly facilitated the assessment of genetic diversity within populations. Powerful new methods have been developed to analyse these genetic data, sometimes relying on massive computations. These methods are implemented in various software packages and programs, which have grown in number tremendously in the past few years. Although many computer programs in population genetics have been successful in hiding the complexity of the computations from the user, they often rely on assumptions that are crucial for a correct interpretation of the results. There is therefore a clear need to help researchers to navigate this complicated field in order to promote the informed use of these programs.

In this review, we describe some of the most widely used computer programs in population genetics, as well as a series of more specialized programs that implement new and advanced methodologies. We promote the view that the proper analysis of any population genetics data set requires the use of several approaches, beginning with the examination of the basic properties of the data, followed by various more specialized analyses, which will probably be implemented in several programs. So, unlike previous reviews of population genetics software (for examples, see REFS 5,6), we do not compare the value of the programs, nor assess their relative performance and accuracy. We instead briefly describe their principles, the statistics they compute, the assumptions they make and some of their limitations.

Even though we have included many programs in this review, the list is by far not exhaustive. We have chosen, for instance, to leave aside packages that deal with phylogenetic analysis, parentage analysis, gametic phase estimation or linkage analysis, because they have been the subject of other reviews or books (for examples, see REFS 6–11), whereas no overview seems to exist for population-genetics software. Nevertheless, we have tried to include a wide range of applications, implemented in more than 20 programs that have been selected on the basis of their generality, usability, interoperability and unique features, to present a set of programs that carry out classical as well as recent and more sophisticated analyses. They therefore cover the estimation of basic descriptive statistics of genetic diversity within and between populations, tests of random mating, linkage equilibrium and selective neutrality, detection of new immigrants and admixed individuals, as well as the inference of various demographic parameters, such as population size, population divergence time and migration rates.

Because users usually need to analyse the same data set with several programs, we describe the input formats used by different programs, and pay special attention to their interoperability, as an important factor that limits the use of a particular program is often the need to reformat the raw data for that particular purpose. We also discuss some limitations and common problems associated with the use of the current software and the ways in which they could be improved, and indicate potential new directions for the development of future packages.

© 2006 Nature Publishing Group

Computer programs included in this review

The programs included in this review are listed in TABLE 1, with some of their basic properties, and in BOX 1. They are all freely downloadable from the internet. We have grouped the reviewed programs in three categories. Multi-purpose programs compute basic statistics that describe the genetic diversity within and between populations, as well as a few more elaborate or specialized analyses, which are highlighted in TABLES 2–4 and in BOX 1 (additional resources are shown in BOX 2). Individual-centred programs represent a recent development of population genetics:

Table 1 | **List of population genetics programs examined in this review**

Name	Version	Platform	Graphical interface	Accepted data type	Handled data format	References
Multi-purpose packages						
Arlequin	3.01	Win	Yes	DNA, SNP, STR, MULT, FREQ	Specific, GENEPOP	49
DnaSP	4.10	Win	Yes	DNA, SNP	In — MEGA, NEXUS, FASTA, PHYLIP; out — MEGA, NEXUS, FASTA, PHYLIP, Arlequin	50
FSTAT	2.93	Win	Yes	STR, MULT	Specific, GENPOP	51
GDA	1.1	Win	Yes	AFLP, MULT	In — NEXUS, BIOSYS, GeneStrut; out — NEXUS, BIOSYS, GeneStrut, GENESTAT-PC, SAS	See Boxes 1,2
Genepop	3.4	DOS	No	STR, MULT	Specific	52
GENETIX	4.05	Win	Yes	MULT	In — specific, FSTAT, Genepop; out — specific, FSTAT, Genepop, BIOSYS, Arlequin	See Box 1
MEGA	3.1	Win	Yes	DNA, DIST	In — specific, CLUSTAL, NEXUS, PHYLIP, GCG, FASTA, NBRF/PIR, MSF, IG; out — specific, PHYLIP, NEXUS	53
MSA	4.0	DOS, MacOS, Linux	No	STR, MULT	In — EXCEL; out — Genepop, MSVAR, Structure, Arlequin, Migrate	54
SPAGeDi	1.2	DOS	No	STR, MULT	Specific, FSTAT, Genepop	55
Individual-centred programs						
BayesAss+	1.3	Win, MacOS, Linux	Yes	MULT	Specific, IMMANC	56
BAPS	3.2	Win	Yes	MULT	Specific, Genepop	57
GeneClass	2.0g	Win	Yes	MULT	Genepop, FSTAT, GENETIX	58
Geneland	1.05	R	No	MULT	Specific	46,47
NewHybrids	1.1b3	Win, Linux	Yes	MULT	Specific	59
Structure	2.1	Java	Yes	MULT	Specific	60,61
Specialized programs						
BATWING	–	DOS, MacOS, Linux	No	STR, SNP	Specific	62
COLONISE	1.0	Win	Yes	MULT	Specific	63
FDIST2	2.0	DOS, Linux	No	DNA, STR, MULT	Specific	31
Hickory	1.0	Win, Linux	Yes	AFLP, RAPD, MULT	NEXUS	See Boxes 1,2
IM	–	DOS, MacOS	No	DNA, STR, hapSTR	Specific	37
LAMARC	2.0.2	DOS, MacOS, Linux	No	DNA, SNP, STR	Specific, PHYLIP, Migrate	See Box 1
Migrate	2.1.3	DOS, MacOS, Linux	No	DNA, SNP, STR, MULT	Specific, PHYLIP	13
MSVAR	0.4.1.b	DOS, Linux	No	STR	Specific	64
Conversion programs						
Convert	1.3	Win	Yes		In — EXCEL, Genepop; out — GDA, Genepop, Arlequin, Popgene, MICROSAT, PHYLIP, Structure	65
Formatomatic	0.2	Java	Yes		In — Genepop; out — Genepop, Arlequin, IMMANC	See Box 1

AFLP, amplified fragment length polymorphism (dominant markers); DNA, DNA-sequence data — usually, the infinite-site model of mutation is assumed; DIST, distance matrix, used as input for drawing phylogenetic trees or performing Mantel tests; FREQ, frequency data; hapSTR, linked SNP and STR (short tandem repeat) markers; MULT, multi-allelic markers, for which no particular mutation model is assumed; RAPD, random amplified polymorphic DNA (dominant markers); SNP, here one assumes that a single mutation occurred in the ancestry of all genes at that locus, creating only two alleles; STR, also called microsatellites, where a ladder (stepwise) mutation model is assumed; Win, Windows.

© 2006 Nature Publishing Group

here, the main focus of the analysis is on individuals and the very recent history of a population. The last group includes more specialized programs, generally intended to infer some population parameters under a specific evolutionary scenario. Most of these programs use a Bayesian framework for parameter inference[12],

Box 1 | Online links to programs and resources

Some freely downloadable computer programs and packages for analysing population genetics data, and related programs and resources. See BOX 2 for links to additional resources.

Multi-purpose packages

Arlequin	http://cmpg.unibe.ch/software/arlequin3/
DnaSP	http://www.ub.es/dnasp/
FSTAT	http://www2.unil.ch/popgen/softwares/fstat.htm
GDA	http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php
Genepop	http://ftp.cefe.cnrs.fr/PC/MSDOS/GENEPOP
GENETIX	http://www.univ-montp2.fr/~genetix/genetix/genetix.htm
MEGA	http://www.megasoftware.net/
MSA	http://i122server.vu-wien.ac.at/MSA/MSA_download.html
SPAGeDi	http://www.ulb.ac.be/sciences/ecoevol/spagedi.html

Individual-based programs

BayesAss+	http://www.rannala.org/labpages/software.html
BAPS	http://www.rni.helsinki.fi/~jic/bapspage.html
GeneClass	http://www.montpellier.inra.fr/URLB/index.html
Geneland	http://www.inapg.inra.fr/ens_rech/mathinfo/personnel/guillot/Geneland.html
NewHybrids	http://ib.berkeley.edu/labs/slatkin/eriq/software/software.htm
Structure	http://pritch.bsd.uchicago.edu/software/structure2_1.html

Specialized programs

BATWING	http://www.mas.ncl.ac.uk/~nijw/
COLONISE	http://www-leca.ujf-grenoble.fr/logiciels.htm
FDIST2	http://www.rubic.rdg.ac.uk/~mab/software.html
Hickory	http://darwin.eeb.uconn.edu/hickory/hickory.html
IM	http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM
LAMARC	http://evolution.gs.washington.edu/lamarc/lamarc_prog.html
Migrate	http://popgen.csit.fsu.edu/
MSVAR	http://www.rubic.rdg.ac.uk/~mab/software.html

Conversion programs

Convert	http://www.agriculture.purdue.edu/fnr/html/faculty/Rhodes/Students%20and%20Staff/glaubitz/software.htm
Formatomatic	http://taylor0.biology.ucla.edu/~manoukis/Pub_programs/Formatomatic/

XML specifications

BioPAX	http://www.biopax.org/
MAGE-ML	http://www.mged.org/Workgroups/MAGE
SBML	http://sbml.org/index.psp

R resources

HIERFSTAT	http://www2.unil.ch/popgen/softwares/hierfstat.htm
R-project	http://www.r-project.org/
Statistical Genetics Resources	http://cran.au.r-project.org/src/contrib/Views/Genetics.html

Front-ends for command-line programs

CBSU	http://cbsuapps.tc.cornell.edu/index.aspx
Genepop on the web	http://wbiomed.curtin.edu.au/genepop
SNAP	http://www.cals.ncsu.edu/plantpath/people/faculty/carbone/workbench.html

Individual programs not reviewed here

IMMANC	http://www.rannala.org/labpages/software.html
MESQUITE	http://mesquiteproject.org/Mesquite_Folder/docs/mesquite/manual.html
MR BAYES 3.1	http://mrbayes.csit.fsu.edu/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
STRUCTURAMA	http://www.structurama.org/
TFPGA	http://www.marksgeneticsoftware.net/tfpga.htm

HIERFSTAT, an R-package to compute and test hierarchical *F*-statistics; CBSU, Computational Biology Service Unit Web Computing Resources; SNAP, a workbench management tool for evolutionary population genetic analysis; MESQUITE, a modular system for evolutionary analysis, version 1.1; STRUCTURAMA, a program for inferring population structure from genetic data; TFPGA, Tools for Population Genetic Analyses.

Bayesian
Inference framework, based on the work of Thomas Bayes (1702–1761), in which the posterior probability of a parameter depends explicitly on its prior probability, reflecting some previous belief about this parameter.

© 2006 Nature Publishing Group

which facilitates the incorporation of some prior knowledge on the parameters of interest, often leading to improved estimations (for an example, see REF. 13).

In this review, we have tested only the last compiled Windows version of the programs, even though most of the applications are available on different platforms (TABLE 1), either as precompiled executable programs or as a source code to allow compilation for specific machines. All the programs are accompanied by some documentation and most of them with sample files, allowing one to get a feel for their execution speed, the presentation of the results in output files and the range of computations they can perform.

We have tested the programs with the provided sample files and some additional data sets, and in most cases they ran flawlessly. The quality of the documentation that accompanies the programs was generally good. Most authors have taken care to include help files, ranging from a simple read-me file (for example, MSVAR and Genepop), help files under windows (for example, MEGA, FSTAT and DnaSP), PDF documentation (for example, MSA and BATWING), and a simple HTML help file (for example, GENETIX

Table 2 | **Program functionalities and assumptions: multi-purpose packages**

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues*
Arlequin	Computes indices of genetic diversity, *F*-statistics and genetic distances between populations; exact test of HWE, LD and population differentiation; tests selective neutrality within populations; Mantel test; estimates gametic phase from multilocus genotypes; estimates demographic parameters from mismatch distribution	Hierarchical analysis of genetic structure based on the AMOVA framework, with up to 3 levels; tests LD without specifying gametic phase; analysis of multiple files in batch mode	Moment, least-square, ML or Bayesian estimators; most tests are either non-parametric, exact or based on coalescent simulations	Recessive alleles only supported when estimating haplotype frequencies
DnaSP	DNA sequences only; computes LD indices and several pairwise distances between populations; estimates demographic parameters from mismatch distributions; tests selective neutrality at the intra- and interspecific level	Convenient DNA-sequence browser; definition of working subsets of sites; detection of recombination and gene conversion; possibility to compute several statistics in a sliding window	Moment or least-square estimators; coalescent simulation module for confidence intervals of statistics	
FSTAT	Computes basic indices of genetic diversity, allelic richness and *F*-statistics; test of HWE; multiple regression analysis; Mantel test	Test of genotypic disequilibrium between loci for unphased diploid data; compares various statistics among groups of samples; tests sex-biased dispersal based on summary statistics; batch mode	Moment estimators; tests and confidence intervals are based on resampling techniques	
GDA	Computes basic indices of genetic diversity, LD indices and *F*-statistics	Hierarchical analysis of genetic variance with up to 4 levels; detects private alleles	Moment estimators; most tests are exact or based on resampling techniques	
Genepop	Computes basic indices of genetic diversity and *F*-statistics; exact test of HWE and LD; Mantel test	Web interface for remote computations; estimates the number of migrants exchanged between populations based on rare alleles	Moment estimators; most tests are either based on MCMC or resampling techniques	Not updated for some time
GENETIX	Computes basic indices of genetic diversity, LD indices, *F*-statistics and genetic distances between populations; Mantel test	Factorial correspondence analysis; synthetic measure of LD between multi-allelic loci; partitions LD owing to population structure; convenient data editor	Moment-based estimators; tests and confidence intervals are based on resampling techniques	French version only
MEGA	DNA sequences only; computes basic indices of nucleotide diversity, evolutionary distances between sequences and populations; computes phylogenetic trees; tests selective neutrality	Can handle DNA or protein sequences and distance matrices; powerful sequence-data viewer; tree explorer for the visualization of phylogenetic trees with bootstrap support	Moment-based estimators; tests are based on resampling techniques	
MSA	Handles large STR data sets; computes standard genetic diversity indices and several genetic distances between populations	Computes $\delta\mu^2$; input file created in a spreadsheet, and conversion into several other formats	Moment-based estimators	Assumes diploid data
SPAGeDi	Computes genetic distances between populations, as well as several indices of inbreeding, kinship and relatedness at the individual level; focuses on relationships between genetic and geographical distances	Specific handling of spatially explicit genotypic data; estimates gene dispersal from patterns of isolation by distance; polyploid, diploid dominant or co-dominant markers and haploid markers	Moment-based estimators; tests are based on resampling techniques	

*Specific assumptions exist for all implemented methodologies. AMOVA, analysis of molecular variance (this is an inference framework for *F*-statistics that is derived from molecular data, and is based on conventional analysis of variance); HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo method; ML, maximum likelihood; STR, short tandem repeats; $\delta\mu^2$, the genetic distance between populations that is specially designed to handle microsatellite (STR) data.

© 2006 Nature Publishing Group

Short tandem repeat (or microsatellite)
A class of repetitive DNA that is made up of repeats that are 2–5 nucleotides in length. The number of these repeats is usually extremely variable in a population.

and GeneClass), to tutorials or detailed user manuals (for example, Arlequin, LAMARC, Migrate and Structure). A description of the underlying methodology is sometimes included in the documentation (for example, Arlequin and Migrate), but most programs simply refer to the paper in which the methodology was originally described.

It is good practice for users to take the time to carefully read the provided instructions before running the program, as well as the original papers describing the implemented methodologies. This approach is necessary for a sound interpretation of the results, as the real job of the user starts when the job of the programmer ends — that is, in interpreting the results. Users should therefore understand the theoretical aspects of the programs they are using, to avoid interpretation errors or using inconsistent settings when running the programs.

Properties and assumptions of different packages

Notwithstanding their functionalities, the programs differ from each other in several aspects, such as the types of marker they can handle, the way in which raw data are formatted and how users select the details of the computations to be performed. These properties are important to consider when choosing the most appropriate program, as detailed below.

Supported data types. The data types supported by the different programs are shown in TABLE 1. Some programs have been designed to deal with a given type of data, such as DNA sequences (DnaSP and MEGA) or short tandem repeat (STR) data (for example, BATWING, MSA and MSVAR), whereas others can handle several types of marker. Among these programs, one should be careful to distinguish between those that have routines specially dedicated to the handling of specific markers (such as Arlequin, BATWING, FDIST2, FSTAT, GDA, Genepop, IM, LAMARC, Migrate and SPAGeDI) and those that support different data types because they do not model mutations explicitly. If no mutation model is specified, then it is implied that the program assumes that only genetic drift and/or migration are responsible for the observed differences between populations. This assumption is probably legitimate when applied to related populations, but less so when comparing the genetic diversity that has developed between populations that have diverged for hundreds or thousands of generations; in these circumstances the mutation process cannot be ignored. It is worth noting that no individual-centred program explicitly incorporates mutations, even though this feature is important when dealing with STR markers that are analysed in highly divergent populations (for example, as in REF. 14), and

Table 3 | **Program functionalities and assumptions: individual-centred programs**

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues
BayesAss+	Estimates recent migration rates between populations from multilocus genotype data	Estimates each individual's immigrant ancestry, the generation in which immigration occurred, and inbreeding levels within populations	MCMC, Bayesian	Assumes co-dominant unlinked markers, and sampling of source populations of the immigrants; allows for missing data
BAPS	Assigns individuals to genetic clusters by either considering them as immigrants (mixture analysis) or as descendents from immigrants (admixture analysis)	Estimates the number of genetic clusters; provides the proportion of the genome of each individual that can be assigned to the inferred clusters (admixture analysis)	Bayesian	Assumes HWE within clusters and unlinked markers; partially uses information on the sampling origin of the individuals
GeneClass	Detects immigrants from multilocus genotypes, assignment of individuals to populations	Assesses whether a given genotype can be excluded from a given population	Bayesian, likelihood	Assumes HWE within populations; assignment to sampled populations only; no attempt to reconstruct virtual populations
Geneland	R package to detect population subdivisions that explicitly take into account the spatial position of sampled multilocus genotypes; computes *F*-statistics between inferred virtual populations	Determines the best number of subdivisions, and assigns geo-referenced individuals to a subdivision; provides graphical output of the spatial distribution of the subdivisions	MCMC, Bayesian	Assumes HWE and no linkage within subdivisions; immigrant genes are supposed to be present only in new immigrants
NewHybrids	Specifically designed for the study of a single hybrid population; genes of hybrid individuals can come from only two parental populations	Computes the posterior probability that individuals fall into different hybrid or pure parental categories	MCMC, Bayesian	Assumes HWE within parental populations and independent co-dominant diploid markers
Structure	Detects the underlying genetic structure among a set of individuals genotyped at multiple markers; can detect new immigrants or individuals whose ancestors were immigrants (admixture analysis)	Computes the proportion of the genome of an individual originating from the different inferred populations (admixture analysis); reports genetic distances between inferred virtual populations and the ancestral populations	MCMC, Bayesian	Assumes HWE within clusters; can model LD due to admixture; assumes diploid data (haploid data possible); sequential analysis for different potential number of virtual populations

HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo; RJ-MCMC, reversible-jump MCMC.

© 2006 Nature Publishing Group

Table 4 | **Program functionalities and assumptions: specialized programs**

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues
BATWING	Estimation of the past demography of one or more populations based on multilocus genotypes	Estimates the relative sizes of the examined populations and the ancestral population; reports posterior distributions for the divergence times, the beginning of population growth and the exponential growth rate	MCMC, Bayesian	Assumes HWE within populations, no migration after divergence; growth is assumed to have started simultaneously in all populations
COLONISE	Designed to study the pattern of colonization events having occurred in the history of populations based on multilocus genotypes	Integrates genetic and non-genetic information into a hierarchical model, to estimate the contributions of source populations to new colonies; provides the posterior probabilities of various models	RJ-MCMC, Bayesian	Assumes that all potential source populations are sampled, and that colonizers are either new immigrants or F1
FDIST2	Detects outlier loci, potentially due to positive or balancing selection, in samples from a subdivided population	Reports the expected relationship between genetic diversity within (heterozygosity) and between (F_{ST}) populations, obtained under a simple structured coalescent model.	Coalescent simulations and moment-based estimators	Assumes an infinite- or finite-island model of migration; tested loci are assumed to be unlinked
Hickory	Bayesian estimation of *F*-statistics in samples from a subdivided population genotyped at dominant or co-dominant markers	Reports the posterior distribution of inbreeding coefficients and F_{ST}; possibility to compare the posterior distribution of *F*-statistics from different data sets	MCMC, Bayesian	Little power to estimate local inbreeding coefficients with dominant markers
IM	Estimates the divergence time and the migrations having occurred in the ancestry of two populations, which might have grown exponentially since their split	Reports the posterior distributions of the ancestral population size, the divergence time, the relative initial population sizes, the growth rates and potentially asymmetrical migration rates between populations	MCMC, Bayesian	Assumes that there are no other populations exchanging migrants with the sampled populations, no linkage between loci, and no recombination within loci; individual genotypes are entered as haplotype data
LAMARC	Estimates the past demographic history of a series of populations using unlinked or partially linked markers	Estimates simultaneously immigration rates, average recombination rate, current population sizes and exponential growth rates	MCMC, ML, Bayesian	Assumes a stable migration structure, constant exponential growth or decline and uniform recombination rates
Migrate	Estimates the effective population sizes and immigration rates in a series of populations, using unlinked markers only	Reports the ML estimators or the posterior distribution of the effective sizes and immigration rates between populations	MCMC, ML, Bayesian	Assumes constant migration rates over time, and that migration occurred only between sampled populations
MSVAR	Estimates the past demographic history of a population analysed at STR loci; mostly applied to the detection of population expansions or bottlenecks	Reports the posterior distribution of TMRCA, ancestral population size, growth rate and time of onset of growth	MCMC, Bayesian	Assumes unlinked loci and no immigrants in the population; the presence of many immigrants can result in a signal of recent bottleneck (M. Beaumont, personal communication)

HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo; ML, maximum likelihood; STR, short tandem repeats; TMRCA, time to the most recent common ancestor.

in situations in which homoplasic mutations are likely to accumulate[15–18]. Departures from this assumption have not been studied, and their effect on the estimation process is largely unknown. On the other hand, most of the specialized programs that aim to infer demographic parameters assume a specific mutation model for STRs, DNA sequences or SNPs, which is explicitly simulated in an estimation procedure based on the coalescent process.

Homoplasic mutations
Mutations that lead to identical character states (identity-in-state) despite having occurred by different evolutionary processes.

Input files and communication between programs. Unfortunately, most of the programs use a specific data-file format, but several offer the possibility to read or write data from, or to, other file formats (TABLE 1). It should therefore be possible to format one's own data in a given format and use conversion tools to analyse it with different programs. This possibility is essential to avoid being limited to the analyses provided in a single program, and therefore to perform various analyses on a given data set without having to reformat it manually. Such exchange pathways are shown in FIG. 1. In this figure, we have identified three programs that could be considered as starting points to format input data files. Two of these programs (Convert and Formatomatic) are specialized conversion utilities, and can create input files for several other programs; the third, MSA, incorporates specific methodologies for the analysis of STR markers, but can also directly convert data into several interesting formats such as Migrate, MSVAR and Structure. Note that MS-Excel or another spreadsheet software can be

© 2006 Nature Publishing Group

Box 2 | **Additional resources for population genetics data analysis**

An Alphabetic List of Genetic Analysis Software	http://linkage.rockefeller.edu/soft/list1.html
Genetic Software Forum	http://www.rannala.org/gsf
libsequence	http://molpopgen.org/software/libsequence_html/libsequence.htm
Nexus Class Library (version 2.0)	http://hydrodictyon.eeb.uconn.edu/ncl/
Population Genetics Links	http://www.geocities.com/CapeCanaveral/Lab/4709/popgen.htm
PyPop	http://allele5.biol.berkeley.edu/pypop/index.html
Resources for Ecology, Evolutionary Biology, Systematics, and Conservation Biology	http://darwin.eeb.uconn.edu/links/index.php
Software for Population Genetic Analyses	http://www.biology.lsu.edu/general/software.html

libsequence, a C++ class library for evolutionary genetic analysis; PyPop, Python for Population Genetics.

conveniently used to generate MSA and Convert input files. FIGURE 1 also reveals that the Genepop format has a central role in the interactions between programs, as many other programs can directly read it or produce data files in this format. This is because this program was among the first integrated population genetics packages to be made available. Although its functionalities are now available in several other regularly updated programs, its file format remains a standard in population genetics analyses.

It follows that some specific or manual formatting is necessary for only a few specialized-purpose or individual-based programs (BATWING, COLONISE, FDIST2, Geneland, IM and NewHybrids). Note, however, that file conversions might be impeded by the continuous update of some programs and their input formats. For instance, Arlequin was able to read early MEGA files (version 1), but cannot read the format of the current version 3. The authors of the programs would need to constantly update their input and output routines to remain compatible with the latest releases of the other programs, which might be difficult to achieve in practice.

Graphical interface programs versus command-line programs. Several programs have user-friendly and sophisticated graphical interfaces, allowing users to easily choose the types of analysis to be performed and to set up computation parameters (Arlequin, BAPS, BayesAss+, COLONISE, DnaSP, FSTAT, GDA, GeneClass, GENETIX, MEGA and Structure), to examine some properties of the data and to select loci, individuals or populations to include in the analyses (DnaSP, MEGA and GENETIX), or to directly edit data files (GENETIX and Structure). Some programs also allow a graphical representation and/or analysis of the results (BAPS, COLONISE, DnaSP, Geneland, GENETIX, NewHybrids, MEGA and Structure). Note, however, that only a few programs (BayesAss+, Formatomatic, Migrate, NewHybrids and Structure) offer a graphical interface on platforms other than Windows, which can make them more accessible and appealing to some users. Genepop is special in this respect, as a web-based front-end has been developed (see Genepop on the web in BOX 1) which allows remote computations to be performed from any machine.

The other programs are simple command-line executable files, where the settings of the computations can often be specified in a separate text file or on a command line. Despite being less user friendly than graphical packages, command-line programs have the advantage that they can be easily launched in parallel on computer clusters, or serially on a single computer. They can therefore be used to automatically analyse a large number of input files, such as those resulting from large genomic studies or simulations. Note that some graphical packages can be run alternatively on the command line (Structure, Hickory and GDA) or incorporate a batch mode to analyse a series of files at once (Arlequin, FSTAT and LAMARC). SPAGeDi and Genepop are particular cases because computational settings need to be chosen at run-time through a text-based menu. This last feature is unfortunate, as it prevents the automatic execution of these programs from batch files under Windows or from shell scripts under Linux. Note that source code is available for most command-line programs, allowing advanced users to recompile programs for other platforms.

Computation time. Owing to the speed of current processors, the computation of descriptive statistics and their confidence intervals through resampling techniques is extremely fast, not exceeding a few minutes. Only programs that require an explicit simulation of demographic and mutational processes need a substantial computational time (BATWING, IM, Migrate, MSVAR and LAMARC), which can sometimes extend over several days, or even weeks. This should not be a problem in most cases, especially as data generation usually takes much longer than their analysis. However, Bayesian or maximum-likelihood estimations based on Markov chain Monte Carlo (MCMC) methods often require several consecutive runs to be performed to check that the chains have converged and that parameter space has been correctly explored (BOX 3). This need for multiple runs can considerably extend computing time if the program cannot be simultaneously run on different computers. However, users should always privilege result accuracy and robustness over execution speed.

Functionalities of multi-purpose packages

The functionalities and main features of the different programs are summarized in TABLE 2, together with a description of the inference framework they use, and their assumptions.

Coalescent (theory)
A theory that describes the structure of the genealogy of a sample of genes from present time to their most recent common ancestor. For neutral genes, this genealogy is extremely variable but only depends on the past demography (deme sizes and immigration rates) of the population.

Maximum-likelihood estimation
Inference technique in which the estimated parameters of a model are those that maximize the probability of the data under that model.

© 2006 Nature Publishing Group

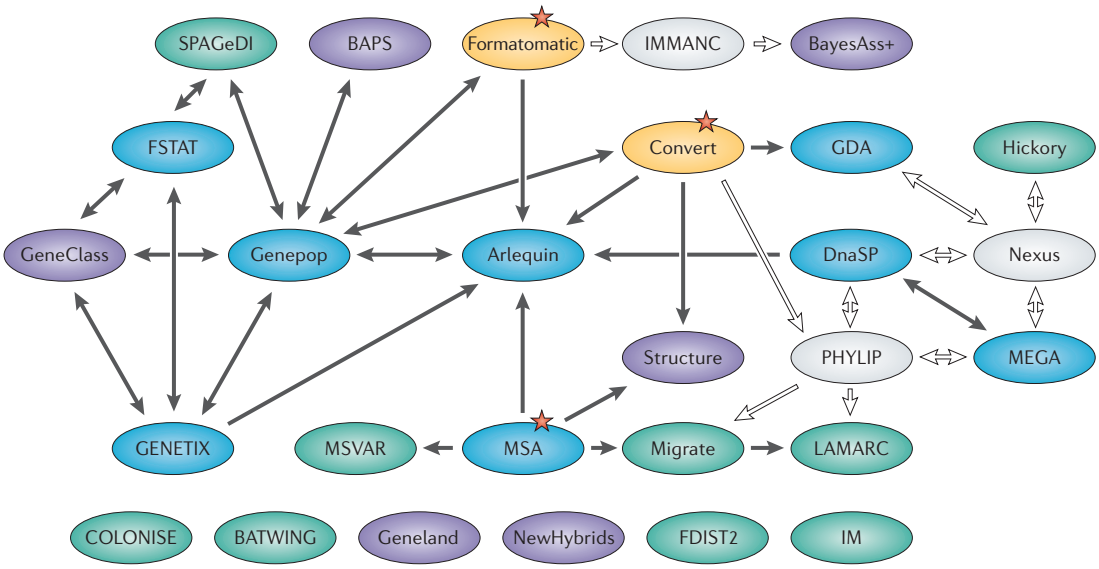


Figure 1 | **Flow chart of possible data exchange between different population genetics programs.** Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.

Descriptive statistics. Most multi-purpose packages compute basic indices of genetic diversity within populations, such as degree of heterozygosity, number of alleles or number of polymorphic loci. Additionally, FSTAT offers the possibility to compute allelic richness, which represents the number of alleles standardized to the smallest sample size in the study, whereas Arlequin and GDA provide different ways to detect alleles that are private for (that is, specific to) some populations. Tests of Hardy–Weinberg equilibrium (HWE) at all loci, and of linkage disequilibrium between pairs of loci, are also frequently available. Although several packages (DnaSP, FSTAT, GDA, GENETIX and MEGA) can define subsets of loci, populations or individuals on which to compute various statistics, DnaSP is the only one that can display statistics along a sliding window, the size and shifting increment of which can be specified. This feature is extremely useful for locating chromosomal segments with unusual patterns of diversity.

Population comparisons and genetic structure. Various programs and approaches allow users to describe the genetic relationships between a set of populations and their genetic structure. Populations can be compared by means of various genetic-distance measures (in Arlequin, DnaSP, GDA, GENETIX, MEGA, MSA and SPAGeDi), which can be used to produce phylogenetic trees of populations; for example, in MEGA. The analysis of population subdivision through *F*-statistics is also quite standard, although only Arlequin and GDA provide the possibility to perform hierarchical analyses of genetic variance. Indices of genetic diversity within populations can also be compared between groups of populations — in FSTAT this is done by means of a resampling technique, whereas in DnaSP this comparison occurs through a coalescent simulation module, which can provide confidence intervals for several statistics in an isolated and constant-sized population. Genetic and non-genetic data such as geographical distances can be compared through Mantel tests, which are provided in Arlequin, FSTAT, GENETIX, and Genepop; SPAGeDi implements several methods to compare genetic diversity with geography at both population and individual levels, allowing the user to obtain estimates of dispersal patterns. SPAGeDi is also the only program reviewed here that computes several indices of relatedness between individuals. Exact tests of population differentiation[19] are provided in Arlequin and Genepop.

Demographic inference and neutrality tests. Analysis of past population expansions that is based on DNA-sequence diversity can be performed in Arlequin and DnaSP by means of mismatch distribution analysis. These two programs, as well as MEGA, also provide routines to perform tests of selective neutrality and population equilibrium that are based on DNA-sequence diversity within populations, assuming the infinite-sites model. MEGA and DnaSP can also perform other tests of selection on DNA sequences, comparing the genetic diversity within and between species, whereas Arlequin provides additional neutrality tests for multi-allelic loci under an infinite-allele mutation model.

Hardy–Weinberg equilibrium
(HWE). Fit between the observed frequencies of the different genotype categories and the frequencies that are expected under random mating in an ideal population. Departure from HWE can also be due to selection, migration or hidden population subdivision.

***F*-statistics**
Statistics that measure the correlation between genes drawn at different levels of a (hierarchically) subdivided population. This correlation is influenced by several evolutionary forces, such as mutation and migration, but it was originally designed to measure how far populations had gone in the process of fixation owing to genetic drift.

Hierarchical analyses of genetic variance
Analysis in which genetic diversity is hierarchically organized, with subunits nested in larger units (for example, genes in diploid individuals drawn from demes belonging to a subdivided population).

Mantel test
Test designed to measure the association between the elements of two matrices, by taking into account the autocorrelation that exists between the elements of each matrix. It is often used to test for a significant association between genetic and geographical distances.

Mismatch distribution
The distribution of the number of differences (mismatches) between pairs of DNA sequences in a sample. The exact shape of this distribution is affected by the past demography of a population.

Infinite-sites model
A mutation model according to which each new mutation occurs at a site that has not mutated before. This model was originally developed for protein- and DNA-sequence evolution, and is obviously related to the infinite allele model.

© 2006 Nature Publishing Group

Assumptions. The assumptions made by the multi-purpose packages are difficult to summarize, because they implement various computations that are based on different methodologies, all relying on different premises. The computations of most summary statistics (for example, LD indices such as *D* or *D′* (REF. 20)) or tests of selective neutrality (for example, based on Tajima's *D* (REF. 21)) that are computed from diploid samples of DNA sequences assume that the gametic phase is known. This is rarely the case, as the phase is now more commonly reconstructed using statistical methods (for example, see REF. 6), some of them being implemented in Arlequin. Note that Arlequin and FSTAT implement tests of LD that do not require phase information. The effect of possible errors in reconstructed gametic phase on demographic parameter estimation is, however, still largely unknown. Obtaining such an understanding would require dedicated studies, or the development of new methodologies that take into account phase uncertainties in the estimation procedure.

Functionalities of individual-centred packages

Detecting recent immigrants. All individual-centred programs aim to detect immigrants among samples analysed at various multi-allelic markers, using the fact that these immigrants will present different multilocus genotypes than expected for native individuals (see TABLE 3 for a list of individual-centred programs). Some of them attempt to allocate individuals to predefined populations (GeneClass and BayesAss+) or to 'virtual' populations (BAPS, Geneland, NewHybrids and Structure), for which allele frequencies are also iteratively estimated. These last programs therefore attempt to reconstruct the underlying population genetic structure, in the sense that they try to define the number of 'subpopulations' from which the sampled individuals were drawn, and to attribute the individuals to these reconstructed populations. BAPS and Geneland begin by determining the optimal number of virtual populations or 'clusters', and then allocate individuals to these clusters, whereas Structure performs this allocation sequentially for different numbers of clusters, and then flags the number of clusters with the highest likelihood, which might not always be optimal[22]. Structure and BAPS can estimate the fraction of the individual's genome that originates from the different clusters, whereas the other programs either assume that individuals with foreign genes have just arrived (GeneClass and Geneland) or that immigrants arrived earlier and mixed with locals (NewHybrids and BayesAss+). BayesAss+ estimates local inbreeding levels within populations, and a linkage map between markers can be provided to Structure. GeneClass allocates individuals to known populations, but offers the possibility to flag individuals whose genotype cannot be allocated to any sampled population. Geneland offers the interesting option to explicitly use information on the spatial location of the sampled individuals, and so to infer the spatial and genetic structure of the population (BOX 4).

Infinite-allele mutation model
A mutation model according to which each new mutation produces an allele that has not previously existed.

Summary statistics
In the current genetic context, these are descriptive statistics summarizing the pattern of genetic diversity, such as the level of heterozygosity or the number of alleles per locus.

D
A measure of linkage disequilibrium defined as the difference between the frequency of a two-locus haplotype and the product of the frequencies of its constituent alleles ($D_{ij} = p_{ij} - p_i p_j$).

D′
A standardized version of *D* that is obtained by dividing *D* by its maximum possible value given the allele frequencies ($D' = D/D_{max}$).

Tajima's *D*
Statistic used in a selective neutrality test to decide whether the mean number of differences between pairs of DNA sequences is compatible with the observed number of segregating sites in a sample.

Likelihood (of a model)
The probability of the data under a given model defined by a particular set of parameter values.

Joint posterior distribution
When a model is defined by more than one parameter, it is the posterior distribution of all possible combinations of parameter values.

Box 3 | The Markov chain Monte Carlo technique

The Markov chain Monte Carlo (MCMC) technique (for example, see REF. 45) is often used to estimate the joint posterior distribution of a set of parameters without having to explore the whole parameter space. As can be seen in panel **a**, which represents a likelihood surface, most of the parameter space might have a very low likelihood (flat surface), whereas a limited portion of the space will have a much higher (and therefore more interesting) likelihood (raised surface). So, instead of naively and exhaustively exploring the parameter space, as indicated by the dark dots in panel **b**, MCMC better explores the parameter space by concentrating on the high-likelihood portion of the space, starting from a random point shown by an arrow in panel **c**. After computing the likelihood for this initial state, a new state is chosen in its vicinity, and its likelihood is also evaluated. Depending on the ratio of these likelihoods (weighted by the ratios of the prior probabilities of these states and the probability of moving between these states), the new state is accepted or not. By repeating this process for a sufficiently long time, one hopes to efficiently explore the space of parameters and get a sample of parameter values that is fully representative of the true posterior distribution, with a concentration of the sampled values in the space having the highest likelihoods.

The quality of the results is usually influenced by many factors, including the starting point, the length of the chain and the way we modify the parameter values between successive states. The way to remove the influence of the starting position is to let the chain run for some time before beginning to sample points (a 'burn-in' period). The number of burn-in steps should be large enough to allow the chain to reach the interesting portion of the surface, which is sometimes difficult to assess. The way in which transitions between states are performed is also crucial for a good exploration of the parameter space and to avoid being stuck in a particular region. Several programs can have several chains running simultaneously (for example, IM and Migrate) to better explore the likelihood surface.

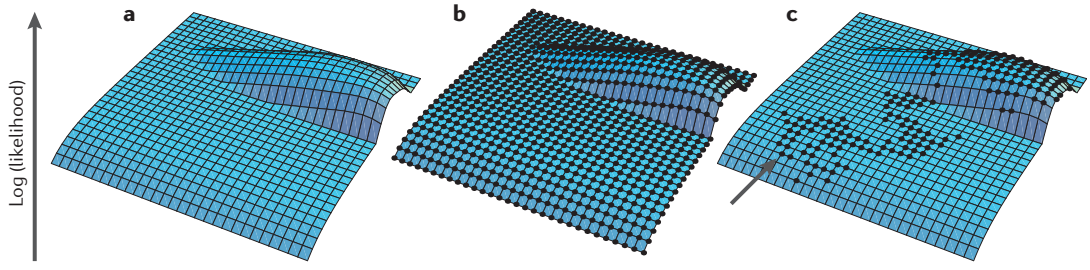


Image courtesy of Peter Beerli, Florida State University, USA.

© 2006 Nature Publishing Group

Assumptions. All individual-centred programs assume HWE within 'populations' (except BayesAss+) and assume that loci are unlinked (except Structure). NewHybrids further assumes that there are only two parental populations, and so aims at specifically identifying different categories of hybrid individuals (F1, F2 and backcrosses). BayesAss+ assumes that all the source populations of immigrants have been sampled, and estimates for each potential source population whether the individuals are local residents.

Functionalities of specialized programs

The specialized programs we review here have been developed recently to infer demographic parameters such as effective population size or migration rates from genetic data (see the list of programs in TABLE 4). Exceptions are: FDIST2, which attempts to detect loci under selection from genome scans; Hickory, which has been initially designed to estimate *F*-statistics from dominant data; and COLONISE, which tries to identify the origin of the founders of a new population.

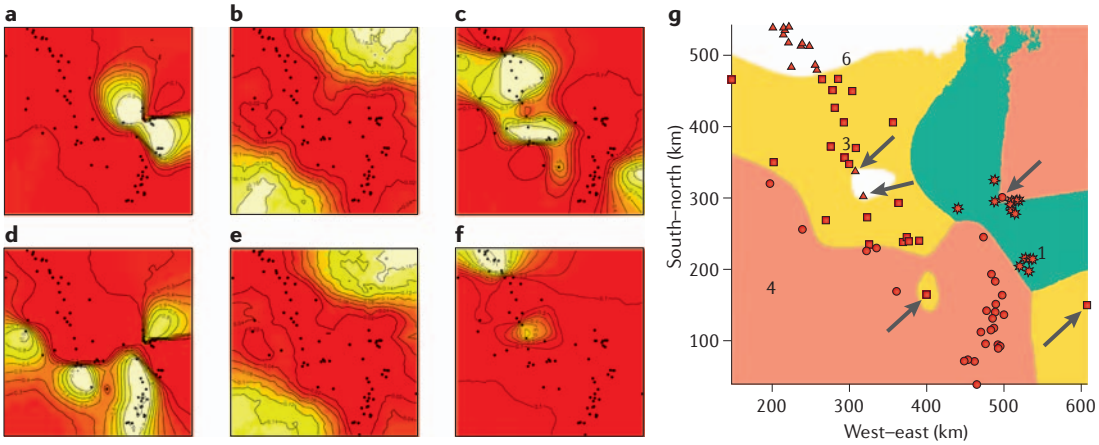
Demographic-inference programs. These programs allow users to estimate population size or migration rates between two populations that diverged some time ago (IM), or under a finite-island model between an arbitrary number of populations (LAMARC and Migrate). Note, however, that these programs do not directly estimate demographic parameters, but only the product of the demographic parameters by the mutation rate, which determines the observed pattern of genetic diversity. Note also that the functionalities of Migrate version 1.7 were incorporated into the LAMARC package, but we have reviewed Migrate separately; this is because this program has continued to evolve (to version 2.1) and now differs from LAMARC in several aspects, including a different Bayesian estimation framework, speed, the possibility of parallel program execution, output format (P. Beerli, personal communication), and it provides more feedback to users. On the other hand, the LAMARC package incorporates several features that are not present in Migrate, such as the estimation of recombination rates or past population expansions.

MSVAR concentrates on a single population analysed at STR markers to detect signs of recent expansions or contractions. The IM program extends a previous approach[23], and aims to simultaneously estimate divergence time and immigration rates between two populations. Whereas there is little information in allele frequencies to distinguish population divergence from ongoing gene flow[24,25], molecular data are more informative[26], and IM can handle various markers, including hapSTRs, which are a combination of SNPs linked to an STR locus[27]. BATWING has been originally developed

Box 4 | **Analysis of population genetic structure and geographical subdivision using Geneland**

Individual-centred analysis programs aim to detect new immigrants in populations, using the fact that these individuals will have different allele frequencies than native individuals at many loci. In most approaches, populations are predefined by the user and/or the spatial structure of sampling locations is ignored.

A novel approach, implemented in the software Geneland[46,47], takes the sampling location explicitly into account when individuals are allocated to 'virtual' populations. It attempts to infer the spatial and genetic structure of these populations. The approach is exemplified with data from a study on the wolverine (*Gulo gulo*), a medium-sized, highly mobile carnivore of the Northern hemisphere[47]. Eighty-nine individuals were genotyped at ten microsatellite loci[48] and analysed for cryptic genetic structure and migration patterns. Panels **a**–**f** show maps of the study area with the relative posterior probability of belonging to each of six inferred populations: black dots represent the geographical position of sampled individuals, and lighter colour reflects a higher posterior probability. Panel **g** shows a synthetic map of the mode of the posterior probability distribution for each pixel belonging to each wolverine population. Large character numbers (1, 3, 4 and 6) indicate population labels. Populations 2 and 5 are not shown because they represent inferred ghost populations without individuals, a by-product of the algorithm (see REF. 47 for details). Arrows indicate putative migrants. For example, two migrants from population 6 (represented by triangles) are found within population 3 (represented by squares), and one migrant from population 3 is found within population 4 (represented by circles). (Individuals in population 1 are represented by stars.) The figure is adapted with permission from REF. 47 © (2005) The Genetics Society of America.



Effective population size
The size of a virtual, randomly mating, stationary and isolated population that would have the same amount and type of polymorphisms as the population under study.

Finite-island model
A conceptual model for gene flow under which a finite number of demes exchange migrants with each other. The spatial location of the populations is not specified, and the constituent demes are usually assumed to have the same size and to exchange migrants at the same rate.

© 2006 Nature Publishing Group

Table 5 | **List of computer programs suited for a given analysis and genetic marker**

	Multi-allelic markers*	STR	Dominant markers (AFLP)	SNP	DNA sequences
Descriptive statistics	Arlequin, FSTAT, GDA, Genepop, GENETIX, MSA, SPAGeDi, Hickory		SPAGeDi		Arlequin, DnaSP, MEGA
Linkage disequilibrium	Arlequin, FSTAT, GDA, Genepop, GENETIX, Structure				
Analysis of population subdivision	Arlequin, FSTAT, GDA, Genepop, GENETIX, MSA, SPAGeDi, Hickory, Structure, BAPS, Geneland	Arlequin, FSTAT, GDA, Genepop, MSA, SPAGeDi	Hickory		Arlequin, DnaSP, MEGA
Detection of new immigrants: known populations	BayesAss+, GeneClass				
Detection of new immigrants: inferred populations	BAPS, NewHybrids, Structure, Geneland	BATWING, IM, LAMARC, MSVAR			
Demographic expansion or decline		BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	Arlequin, DnaSP, IM, LAMARC, Migrate
Population size	Migrate	BATWING, IM		BATWING, LAMARC, Migrate	IM, LAMARC, Migrate
Divergence time	Arlequin, FSTAT, GDA, Genepop, GENETIX	BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	DnaSP, IM, LAMARC, Migrate
Migration rates	Arlequin, FSTAT, Genepop, BayesAss+, COLONISE, Migrate	BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	DnaSP, IM, LAMARC, Migrate
Neutrality tests	Arlequin, FDIST2				Arlequin, DnaSP, MEGA
Spatially explicit analyses	SPAGeDi, Geneland, COLONISE				

*By multi-allelic markers, we mean loci for which no specific mutation model is assumed, or for which mutations can be neglected. In the latter case, computations are based on allele frequencies only. Otherwise, specific mutation models are assumed by the different programs. Note that DNA sequence, STR and SNP allele frequencies, as well as nucleotide frequencies, can also be used by several packages to estimate descriptive statistics and linkage disequilibrium, and to detect new immigrants. AFLP, amplified fragment length polymorphism; STR, short tandem repeat.

to estimate the size of a population from multilocus STR data, but it has now been extended to accommodate several populations that diverged from each other by a series of fissions, the times of which are also estimated, and a specific model for SNPs has also been incorporated. Finally, COLONISE is an extension of mixture analysis to estimate the contribution of different populations to new colonies by using genetic and non-genetic information. The originality of the approach is that it is based on reversible-jump MCMC (RJ-MCMC[28]) — an extension of the MCMC approach that allows the chain to jump between models that include different types of information (for example, population densities and geographical distance to the new colony). The RJ-MCMC approach also provides the posterior probability of the alternative models, and therefore allows the user to identify factors that might have a role in colonization processes[29,30].

Detecting loci under selection. FDIST2 is a program to detect loci of which the genetic diversity within (heterozygosity) and between populations (F_{ST}) does not conform to the prediction of a simple infinite or finite-island model obtained by coalescent simulations. Although this migration model is not very realistic, simulations have shown it to be quite robust[31], and this approach performs as well as a more sophisticated Bayesian approach[32]. However, loci under balancing selection seem more difficult to detect than those under positive selection with this approach, especially if populations are not very divergent[32].

Assumptions. As mentioned earlier, demographic-inference programs are powerful, but they assume a specific population and mutation model, which needs to be explicitly laid out and understood by the users. Any departure from the initial assumptions is likely to affect the results. In most programs, population sizes are assumed to either be constant (Migrate) or potentially exponentially growing (IM, BATWING, LAMARC and MSVAR). Inferred migrations rates are also assumed to have been constant over time in IM and Migrate, and all programs assume that there are no unsampled populations that send migrants to those that are sampled. Such populations are often referred to as 'ghost populations' (see REFS 33,34): they affect estimations of migration rates among sampled populations[33,34] and could mimic bottleneck effects (for example, in MSVAR; M. Beaumont, personal communication).

How to choose the right software

From our previous description of the properties of the different programs, it should be clear that similar computations can be performed by several packages,

F_{ST}
A measure of the level of population genetic differentiation, which usually reflects the proportion of total genetic variability that is due to the net differences between populations (see *F*-statistics).

Balancing selection
A form of natural selection that maintains polymorphism within populations.

© 2006 Nature Publishing Group

and that a given program can often be used for different purposes. TABLE 5 shows the programs that could be used to perform a given task depending on the type of marker at hand. Although most programs have been developed for multi-allelic markers, it should be underlined that in the absence of temporal sampling, sound inference of demographic parameters such as population growth can only be performed if a mutation model is defined, and that multi-allelic markers are not suited for such estimations. Moreover, several types of analysis reported under the multi-allelic markers column in TABLE 5 can be performed on STR, SNP and DNA sequences; these studies include the computation of basic summary statistics, studies of linkage disequilibrium or the detection of new immigrants. However, in these cases computations are only based on allele and/or genotype frequencies; mutations are neglected, which might often bias the results.

Common problems when analysing data

The interpretation of population genetics data analyses is still more an art than a process that can be fully automated. Major pitfalls in the analysis of genetic data are not due to computer programs, which almost always do what they are supposed to, but rather to them being used on inappropriate data or not correctly set up.

Forcing data to programs. An inadequacy between the available data and a computer program arises from the complexity of the biological world. In a perfect world, research teams would be able to develop analysis tools to address their specific problem, but in practice they have to make their data fit the available tools, leading to obvious discrepancies between the initial goals and the results. For example, many programs assume that genetic markers are either fully linked or completely unlinked, whereas many data sets consist of partially linked markers, or they assume that markers are co-dominant, while some researchers use dominant markers (for example, AFLPs). The treatment of diploid data where gametic phase has been inferred as phase-known data, or the presence of unsampled ghost populations[33,34], are other cases in which computationally correct results might be erroneously interpreted. Also, likelihood-based programs designed for diploid data might give incorrect point estimates and support intervals if they are applied to haploid data sets that have been made diploid by a mere duplication (diploidization) of the haploid genotypes, to fit the specifications of the input file.

Inadequate use of the programs. The non-optimal use of existing software occurs when a program is not correctly parameterized. For instance, the settings of MCMC-based demographic-inference packages often need to be fine-tuned (for example, with respect to chain length, number of chain states to sample or number of chains to run) to provide accurate results, and the choice of appropriate prior probabilities for parameters is crucial in Bayesian programs.

Ascertainment bias. A more difficult issue arises in cases of ascertainment bias. For instance, the use of non-random SNPs can lead to biased demographic parameters or alter tests of selective sweep[35,36]. Another example of ascertainment bias would consist of the elimination of nuclear loci that show evidence for recombination to make them fit models with no recombination, which might lead to an overrepresentation of genome regions with low diversity[37].

Uninformed use of programs. We cannot recommend strongly enough to read the program documentation carefully, as well as the most relevant theoretical papers on the implemented methodologies. Unfortunately, this is rarely done. User-friendly programs with an attractive graphical interface are especially dangerous in this respect, because they seem extremely easy to use, and some results can be obtained by a few mouse clicks and the use of default settings. Additional information on the best use of a program can sometimes be found on the internet, where news groups or discussion forums allow a direct discussion between users and programmers (see, for example, the Genetic Software Forum web site). Users must therefore make sure that the program was designed for their data, and that the results are reproducible. In particular, programs based on the MCMC method (BOX 3) often need to be run multiple times with different settings, and users should check that similar results are obtained over different runs. In Bayesian analyses, it is also good practice to compare the posterior and the prior distributions to check whether data are informative about the parameters[13].

Future directions for software development

We believe that there are two series of measures that could improve the analysis of population genetics data: a better use of existing software (as discussed in the previous section) and the development of improved programs. As noted above, current programs have been developed under a restrictive set of assumptions concerning mutation and demographic models; this is because parameter inference becomes rapidly difficult under more realistic models. However, some approximate Bayesian methods, which rely on summary statistics instead of likelihoods, have been developed recently to deal with more complex models[38,39]. This methodology can, in principle, be applied to any model that can be simulated[40], and so has the potential to open up the development of programs that estimate the specific parameters in which empiricists are interested. More work is needed to assess the statistical properties of these methods under complex models.

Nevertheless, current programs offer a vast panel of analyses that are rarely completely explored, owing to communication problems between programs or perhaps owing to a lack of information on what is available. FIGURE 1 shows that data are indeed exchangeable between most programs through a few intermediate steps, but users who are unaware of this fact might hesitate to recode their data to have access to other methodologies, and might become stuck with a single or a few packages. Improved communicability between programs could be achieved by defining an exchangeable format for population genetics data, or by adopting

AFLP
Amplified fragment length polymorphism. A method for the selective PCR amplification of anonymous, dominant DNA polymorphisms using restriction enzymes and DNA linkers.

Ascertainment bias
Systematic bias introduced by the criteria used to select individuals and/or genetic markers to be analysed (for example, choosing SNPs with heterozygosity that is higher than a given threshold).

Selective sweep
Drastic reduction of the genetic diversity along a chromosomal segment as a consequence of the fixation of an advantageous mutation by selection in that region.

© 2006 Nature Publishing Group

a pre-existing format that would be general enough to deal with various markers and different ploidy levels. Programmers would therefore either need to adopt it as their default format, or develop new routines to read from and write to it, instead of having to constantly update their conversion routines with two or more other formats. Such data-exchange formats already exist in other areas of biology (SBML, systems biology markup language; BioPAX, a data-exchange format for biological pathway data; MAGE-ML, microarray gene expression markup language). The data-exchange flow chart between programs would therefore be considerably simplified compared with FIG. 1, and people could easily switch between programs. Another advantage of having a single data format would be that the different programs could be considered as different modules for population genetics analysis, and future programs could concentrate on implementing new methodologies instead of replicating previous work.

A logical extension of a standard format would therefore be the development of standard and re-usable program components. Recently, several population genetics packages, such as Geneland reviewed above, have been developed under the R software environment (see, for example, REFS 41,42, and the R genetic resources in BOX 1). With their integration to R, these packages have the advantage of being platform independent, but they might not be adapted for computationally demanding applications such as making demographic inferences, because they need to be interpreted and therefore run much more slowly than compiled programs. This approach nevertheless has the merit of showing that there is a community of scientists willing to develop and distribute software components for a common platform. Alternatively, some modularity between compiled programs could be achieved by developing front-ends to existing programs under a common interface. Such a web interface has been made for Genepop (see also Genepop on the web in BOX 1), and others are being developed for evolutionarily related computations to be run on a remote computer or cluster[43] (see also the CBSU Web Computing Interface web site). The scientific community would greatly benefit from a collaborative effort to develop a library of modular and potentially re-usable population genetics algorithms, functions and programs. Such a development is certainly a great endeavour, but it would be facilitated if scientific funding agencies, which have spent enormous sums on the production of population genetics data, were less reluctant to invest in developing the tools necessary for their analysis.

Conclusions

Various computer programs are available for population genetics analyses of different types of molecular marker, ranging from the computation of descriptive statistics to more specific and model-based analyses that aim to reconstruct the past demography of a set of populations, or detect loci under selection. These programs are based on many documented assumptions that need to be integrated into the interpretative framework of the users. Most data analyses will require the use of more than one program, and should start with generalist packages to uncover the basic properties of the data and be followed by the use of specialized methodologies to address more specific questions, a process that often requires many file conversions. The development of a missing exchangeable data format for population genetics analysis would allow users to access a wider range of already implemented methodologies, and avoid them being limited to a restricted set of programs. It would also be beneficial to programmers, who could concentrate on the development of new methodologies, which could be considered as modules of a meta-package that regroups all programs compatible with this standard format. An international collaborative effort in this direction should be seen as complementary to current population genomics projects (for example, see REF. 44).

1. Schaffner, S. F. *et al*. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
2. Akey, J. M. *et al*. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286 (2004).
3. Williamson, S. H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci. USA* **102**, 7882–7887 (2005).
One of the first and more elaborate attempts to correct for the effect of past demography when inferring patterns of selection at the sequence level.
4. Fernandez, J., Villanueva, B., Pong-Wong, R. & Toro, M. A. Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics* **170**, 1313–1321 (2005).
5. Labate, J. A. Software for population genetics analyses of molecular marker data. *Crop Sci.* **40**, 1521–1528 (2000).
6. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
7. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, Sunderland, 2003).
8. Knowles, L. L. The burgeoning field of statistical phylogeography. *J. Evol. Biol.* **17**, 1–10 (2004).
9. Morrison, D. A. Networks in phylogenetic analysis: new tools for population biology. *Int. J. Parasitol.* **35**, 567–582 (2005).
10. Jones, A. G. & Ardren, W. R. Methods of parentage analysis in natural populations. *Mol. Ecol.* **12**, 2511–2523 (2003).
11. Dudbridge, F. A survey of current software for linkage analysis. *Hum. Genomics* **1**, 63–65 (2003).
12. Beaumont, M. A. & Rannala, B. The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261 (2004).
A necessary read presenting a broad overview of the use of Bayesian approaches in population genetics.
13. Beerli, P. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341–345 (2006).
14. Rosenberg, N. A. *et al*. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
15. Valdes, A. M., Slatkin, M. & Freimer, N. B. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737–749 (1993).
16. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (1995).
17. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463–471 (1995).
18. Balloux, F., Brunner, H., Lugon-Moulin, N., Hausser, J. & Goudet, J. Microsatellites can be misleading: an empirical and simulation study. *Evolution Int. J. Org. Evolution* **54**, 1414–1422 (2000).
19. Raymond, M. & Rousset, F. An exact test for population differentiation. *Evolution* **49**, 1280–1283 (1995).
20. Lewontin, R. C. The interaction of selection and linkage. II. Optimum models. *Genetics* **50**, 757–782 (1964).
21. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
22. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software Structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
23. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
24. Nielsen, R. & Slatkin, M. Likelihood analysis of ongoing gene flow and historical association. *Evolution Int. J. Org. Evolution* **54**, 44–50 (2000).
25. Felsenstein, J. How can we infer geography and history from gene frequencies? *J. Theor. Biol.* **96**, 9–20 (1982).
26. Wakeley, J. Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Popul. Biol.* **49**, 369–386 (1996).

© 2006 Nature Publishing Group

27. Mountain, J. L. *et al.* SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res.* **12**, 1766–1772 (2002).
28. Brooks, S. & Giudici, P. in *Bayesian Statistics* (eds Bernardo, J., Berger, J., Dawid, A. P. & Smith, A. F. M.) 733–742 (Oxford Univ. Press, Oxford, 1999).
29. Gaggiotti, O. E. *et al.* Patterns of colonization in a metapopulation of grey seals. *Nature* **416**, 424–427 (2002).
One of the first applications of RJ-MCMC in population and conservation genetics. The method allowed the authors to integrate non-genetic data such as demographic or environmental information directly in the inference process.
30. Gaggiotti, O. E., Brooks, S. P., Amos, W. & Harwood, J. Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Mol. Ecol.* **13**, 811–825 (2004).
31. Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B* **263**, 1619–1626 (1996).
The first description of how genome scans that were performed in several populations can be used to detect loci under selection.
32. Beaumont, M. A. & Balding, D. J. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13**, 969–980 (2004).
33. Slatkin, M. Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Mol. Ecol.* **14**, 67–73 (2005).
34. Beerli, P. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.* **13**, 827–836 (2004).
35. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genomics* **1**, 218–224 (2004).
A lucid description of the effect of SNP ascertainment bias on parameter inference and ways to correct it.
36. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
37. Hey, J. & Nielsen, R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760 (2004).
A comprehensive presentation of the elaborate methodology underlying the IM program, which can simultaneously estimate gene flow and divergence time between two populations.
38. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
A fundamental paper presenting the principles of ABC computations. It shows how genetic simulations can be used to accurately estimate parameters of arbitrarily complex demographic models, for which the likelihood is impossible or too costly to compute.
39. Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15324–15328 (2003).
40. Excoffier, L., Estoup, A. & Cornuet, J.-M. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727–1738 (2005).
41. Zhao, J. H. & Tan, Q. Integrated analysis of genetic data with R. *Hum. Genomics* **2**, 258–265 (2006).
42. Goudet, J. Hierfstat, a package for R to compute and test hierarchical *F*-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
43. Price, E. W. & Carbone, I. SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics* **21**, 402–404 (2005).
44. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
45. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice* (Chapman and Hall/CRC, London, 1996).
46. Guillot, G., Mortier, F. & Estoup, A. Geneland: a computer package for landscape genetics. *Mol. Ecol. Notes* **5**, 712–715 (2005).
47. Guillot, G., Estoup, A., Mortier, F. & Cosson, J. F. A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280 (2005).
An extension of the Structure approach that explicitly uses spatial information to infer the genetic structure of populations and to detect recent immigrants.
48. Cegelski, C. C., Waits, L. P. & Anderson, N. J. Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Mol. Ecol.* **12**, 2907–2918 (2003).
49. Excoffier, L., Laval, G. & Schneider, S. Arlequin ver. 3. 0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50 (2005).
50. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497 (2003).
51. Goudet, J. FSTAT (version 1.2): a computer program to calculate *F*-statistics. *J. Hered.* **86**, 485–486 (1995).
52. Raymond, M. & Rousset, F. Genepop (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249 (1995).
53. Kumar, S., Tamura, K. & Nei, M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5**, 150–163 (2004).
54. Dieringer, D. & Schlötterer, C. Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169 (2003).
55. Hardy, O. J. & Vekemans, X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**, 618–620 (2002).
56. Wilson, G. A. & Rannala, B. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191 (2003).
57. Corander, J., Waldmann, P., Marttinen, P. & Sillanpaa, M. J. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363–2369 (2004).
58. Piry, S. *et al.* GeneClass2: A software for genetic assignment and first- generation migrant detection. *J. Hered.* **95**, 536–539 (2004).
59. Anderson, E. C. & Thompson, E. A. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229 (2002).
A powerful Bayesian method that uses multilocus genotype information to identify the different types of hybrid individual present in a population.
60. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
A highly influential and innovative paper that uses multilocus genotype information to assign individuals to populations, and to identify recent immigrants and admixed individuals.
61. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
62. Wilson, I. J., Weale, M. E. & Balding, D. J. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. A* **166**, 155–188 (2003).
63. Foll, M. & Gaggiotti, O. E. Colonise: a computer program to study colonization processes in metapopulations. *Mol. Ecol. Notes* **5**, 705–707 (2005).
64. Beaumont, M. A. Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013–2029 (1999).
65. Glaubitz, J. C. Convert: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol. Ecol. Notes* **4**, 309–310 (2004).

Acknowledgements
We are grateful to P. Beerli for providing an illustration from Migrate's manual. We also thank him, as well as O. Gaggiotti, J. Goudet and A. Estoup, for suggestions and comments on an early version of this manuscript. We are indebted to three reviewers for their comments. We apologize to the authors of programs which, owing to space constraints, we have not been able to cover here. The work in L.E.'s laboratory is partially supported by a grant from the Swiss National Science Foundation.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Berne's CMPG (Computational and Molecular Population Genetics) programs: http://cmpg.unibe.ch/software.htm
Bob Griffith's GeneTree program: http://www.stats.ox.ac.uk/~griff/software.html
Bruce Rannala's programs: http://www.rannala.org/labpages/software.html
Computational and Molecular Population Genetics Laboratory homepage: http://cmpg.unibe.ch
Gil McVean's programs: http://www.stats.ox.ac.uk/~mcvean/
Giorgio Bertorelle's programs: http://web.unife.it/progetti/genetica/Giorgio/giorgio_soft.html
Ian Wilson's programs: http://www.mas.ncl.ac.uk/~nijw/
Jerôme Goudet's programs: http://www2.unil.ch/popgen/softwares/
Jinliang Wang's programs: http://www.zoo.cam.ac.uk/ioz/software.htm
Jody Hey's programs: http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm
Jonathan Pritchard's programs: http://pritch.bsd.uchicago.edu/software.html
Kent Holsinger's programs: http://darwin.eeb.uconn.edu/#software
Louis Bernatchez's laboratory links and programs: http://www.bio.ulaval.ca/louisbernatchez/links.htm#soft
Mark Beaumont's programs: http://www.rubic.rdg.ac.uk/~mab/software.html
Matthew Stephens's programs: http://www.stat.washington.edu/stephens/software.html
Montgomery Slakin's programs: http://ib.berkeley.edu/labs/slatkin/software.html
Montpellier's CBGP (Centre de Biologie et de Gestion des Populations) programs: http://www.montpellier.inra.fr/URLB/
Montpellier's Genome Populations Interactions Adaptation laboratory's programs: http://www.genetix.univ-montp2.fr/index.htm#programmes
Noah Rosenberg's programs: http://rosenberglab.bioinformatics.med.umich.edu/software.html
Oxford Evolutionary Biology Group's programs: http://evolve.zoo.ox.ac.uk/software.html
Oxford's Mathematical Genetics programs: http://www.stats.ox.ac.uk/mathgen/software.html
Peter Andolfatto's programs: http://www.biology.ucsd.edu/labs/andolfatto/programs/programs.html
Rasmus Nielsen's programs: http://www.binf.ku.dk/~rasmus/webpage/programs.html
Richard Hudson's programs: http://home.uchicago.edu/~rhudson1/source.html
Ziheng Yang's programs: http://abacus.gene.ucl.ac.uk/
Access to this links box is available online.

© 2006 Nature Publishing Group