Journal of Heredity 2008:99(4):438–440
doi:10.1093/jhered/esm127
Advance Access publication February 29, 2008

Computer Note

PGEToolbox: A Matlab Toolbox for Population Genetics and Evolution

JAMES J. CAI

From the Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305.

Address correspondence to James J. Cai at the address above, or e-mail: jamescai@stanford.edu.

Assessing genetic diversity within populations is vital for understanding the nature of evolutionary processes at the molecular level. PGEToolbox is a Matlab-based open-sourced software package for data analysis in population genetics. The main features of this software are as follows: 1) capability for handling both DNA sequence polymorphisms and single nucleotide polymorphisms (SNPs), which include genotype and haplotype data; 2) exhaustive population genetic analyses and neutrality tests based on the coalescent theory; 3) extendibility and scalability for complex and large genome-wide datasets; 4) simple yet effective graphic user interfaces and sophisticated visualization of data and results. For academic uses, PGEToolbox is available free of charge at http://bioinformatics.org/pgetoolbox.

Assessing genetic diversity is vital for understanding the nature of evolutionary processes at the molecular level. Over many years, powerful methods have been developed to analyze genetic data to elucidate the influence of mutation, random genetic drift, migration, and natural selection on genetic diversity. Dedicated computer programs implementing these methods become essential for extracting embedded information. The recent advent of cost-efficient genotyping techniques has greatly facilitated the assessment of genetic diversity within a population. Consequently, massive computations are often required to analyze the large-scale genetic data obtained.

PGEToolbox (from Population Genetics and Evolution Toolbox) is a software package written in Matlab for data analysis in molecular population genetics. As a high-performance language for technical computing, Matlab has been increasingly appreciated by biologists for data analysis (e.g., Cai et al. 2005). However, few Matlab functions are available for analyzing genetic sequence variation data (including DNA sequence polymorphisms and single nucleotide polymorphisms [SNPs]). PGEToolbox has been developed to fill this gap, providing functions for data manipulation, population genetic statistics calculation, and neutrality tests. Major statistics and tests implemented in PGEToolbox for DNA sequence polymorphisms are given in Table 1.

The majority of its functions can be achieved with existing software, but PGEToolbox has several advantages. Existing software can be categorized as either program or library. The former includes, for example, DnaSP (Rozas et al. 2003), Genepop (Raymond and Rousset 1995), and Arlequin (Schneider et al. 2000). The later includes the libsequence (Thornton 2003), the PopGen module of the BioPerl project, the PAL project (Drummond and Strimmer 2001), and the PopGenLib library of the Bio++ project (Dutheil et al. 2006). Compared with libraries, programs are usually more user friendly because of the developed interfaces and outputs, but they lack flexibility in terms of code customization. PGEToolbox stands in the continuum between program and library. It is highly scalable because it can be easily set up as scripts (calling one function after another) to perform an entire job in an unattended batch mode. Furthermore, it contains simple yet efficient menu-driven and dialog-driven graphic user interfaces, which hide the complexity of the computations from the end users. PGEToolbox is under an open source license, which allows others to extend and reuse components, enables interoperation via an open and published interfaces, and reduces duplication of effort within the community. It is running under all 3 major operation systems—Microsoft Windows, UNIX, and Macintosh. Although PGEToolbox requires a Matlab-running environment, the dependency has been minimized: PGEToolbox is independent from any other Matlab toolboxes and back compatible with the earlier version of Matlab (version 6.5). PGEToolbox might be compatible with free alternatives to Matlab as Octave (http://www.gnu.org/software/octave/) and Scilab (http://www.scilab.org/) given some revisions. The more recent version of Matlab includes some code and memory efficiency features, such as the single-precision arithmetic, the memory mapping function, and the support for 64-bit platforms. When these features are used, PGEToolbox should be able to handle modern genetics datasets on the scale of thousands of individuals and hundreds of thousands of sequences or SNPs.

The calculation of sequence polymorphism statistics is a routine task in molecular population genetics. To do this, PGEToolbox first reads a DNA sequence alignment in FASTA or Phylip format. Several polymorphism statistics, such as the number of segregating sites, site-frequency spectrum (SFS), and the nucleotide diversity, can then be calculated. For the population mutation parameter, $\theta = 4$

Table 1. Major statistics and tests implemented in PGEToolbox for DNA sequence data

Statistic or test	Reference
Watterson's theta, θ_W	Watterson (1975)
Nucleotide diversity (π), θ_π	Nei (1987)
Theta *H*, θ_H	Fay and Wu (2000)
Tajima's *D* test	Tajima (1989)
Fu and Li's *D** and *F** tests	Fu and Li (1993)
Fay and Wu's *H* test	Fay and Wu (2000); Zeng et al. (2006)
Wall's *B* and *Q* tests	Wall (1999)
Watterson's homozygosity test	Watterson (1978)
Kelly's *ZnS* test	Kelly (1997)
Fu's *F* test	Fu (1997)
*R*2 test	Ramos-Onsins and Rozas (2002)
Haplotype diversity	Depaulis and Veuille (1998)
Minimum number of recombination events, R_m	Hudson and Kaplan (1985)
MK test	McDonald and Kreitman (1991)
Proportion of positively selected amino acid substitutions, α	Fay et al. (2001); Smith and Eyre-Walker (2002)
EHH	Sabeti et al. (2002)
iHS	Voight et al. (2006)
F_{ST}	Weir (1996); Weir and Cockerham (1983)

$N\mu$ (where N is the effective population size and μ is the per-locus mutation rate per generation), PGEToolbox calculates several common estimates, including the number of segregating sites, θ_W, (Watterson 1975), the mean pairwise difference between nucleotide sequences, θ_π (Nei 1987), and Fay's θ_H (Fay and Wu 2000). PGEToolbox conducts several neutrality tests, such as Tajima's *D* test (Tajima 1989), Fu and Li's *D** and *F** tests (Fu and Li 1993), Strobeck's *S* statistic (Strobeck 1987), Wall's *B* and *Q* tests (Wall 1999), Fay and Wu's *H* test (Fay and Wu 2000) (where *H* statistic is normalized according to Zeng et al. [2006]), Ewens–Watterson homozygosity test (Ewens 1972; Watterson 1978), and Kelly's *ZnS* test (Kelly 1997). Testing of the significance of these statistics requires generating bootstrap samples from a neutral model using a coalescent approach. To do this in Matlab, PGEToolbox incorporates the program ms, which is originally written by Hudson (2002) in C, into a MEX-function (the C interface to Matlab). Such a code migration supplies PGEToolbox with extensive capabilities in coalescent simulations with various parameter settings. The dialog called `coalsimdlg` has been developed to assist users in setting up these parameters. The function `fst_weir` calculates Weir's formulation of Wright's F_{ST} (Weir and Cockerham 1983; Weir 1996). PGEToolbox can also analyze patterns of genetic diversity within and between population samples by using the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) and its 2 extensions (Fay et al. 2001; Smith and Eyre-Walker 2002). In doing so, 2 functions count the numbers of synonymous (D_s) and nonsynonymous (D_n) divergences, and the numbers of synonymous (P_s) and nonsynonymous (P_n) polymorphisms. The MK test can be initiated from 2 functions: the command-line function `mktestcmd` and mktestgui that invokes a pop-up dialog of 2 × 2 contingency table. The function `sewfww` estimates the average proportion of amino acid substitutions driven by positive selection by using the methods of Fay et al. (2001), and Smith and Eyre-Walker (2002). Finally, it is important to note that current version PGEToolbox deletes all sites with missing data that may lead to the loss of information.

PGEToolbox provides a variety of functions for SNP analysis. The graphic user interface of these SNP-related functions is `snptool`, which can adjust its menu for genotype or haplotype SNP data according to users' choice. In the genotype data mode, `snptool` opens input file in the format specified by either HapMap (International HapMap Consortium 2003) or Perlegen (Hinds et al. 2005) projects. Alternatively, `snptool` can download SNP data for all 4 HapMap populations: Yoruba from Ibadan, Nigeria, Japanese from Tokyo, Chinese Han from Beijing, and CEPH individuals from Utah (with northern and western European ancestry). After reading the data, `snptool` computes the observed and predicted heterozygosities, the minor allele frequency, the *P* value of the Hardy–Weinberg equilibrium test, the allele and genotype frequency, the composite likelihood (Nielsen et al. 2005), Tajima's *D* (1989), and Fay and Wu's *H* (2000). A warning message displays in case the "folded" SFS are mistakenly provided by users to calculate Fay and Wu's *H* which requires the "unfolded" SFS (whose ancestral alleles are usually inferred via parsimony using an outgroup). The `snptool` uses the expectation-maximization algorithm to estimate the probabilities of haplotypes and calculates linkage disequilibrium (LD) statistics, such as *D*, *D*′, and *R*, between pairs of SNPs. The `snptool` displays results and datasets graphically. Some examples include a pie chart displaying SNP allele and genotype frequencies of 4 HapMap populations, a plot of relative positions of SNPs on the chromosome, and a visual genotype view (via function `snp_vgview`) presenting complete raw dataset of individuals' genotypes. In the haplotype mode, `snptool` calculates the haplotype diversity (Depaulis and Veuille 1998), the haplosimilarity score (Hanchard et al. 2006), the chromosome segment homozygosity (Hayes et al. 2003), the minimum number of recombination events (R_m) (Hudson and Kaplan 1985), the extended haplotype homozygosity (EHH) (Sabeti et al. 2002), and the integrated haplotype score (iHS) (Voight et al. 2006). EHH and iHS are 2 particularly powerful statistics which have been used for detecting recent selection (Sabeti et al. 2002; Voight et al. 2006). EHH is based on the statistic haplotype homozygosity, $\mathrm{HH}=\left(\sum p_i^2 - 1/n\right)/(1-1/n)$, where p_i is the relative haplotype frequency and n the sample size. For selected core haplotypes, EHH calculates HH in a stepwise manner to determine how LD breaks down with increasing distance to a specified core region. The iHS is based on the differential levels of EHH surrounding an allele compared with the background allele at the same position.

In summary, I show the usefulness of Matlab as a powerful and convenient scientific computation language in molecular population genetics. PGEToolbox is a powerful and flexible Matlab toolbox dedicated to the analysis of DNA sequence polymorphisms and SNPs. The current version implements a number of algorithms, methods, and tools and is ready to be tailored or extended to specific tasks and scaled up for exhaustive exploratory analyses of genome-wide data.

Acknowledgments

J.J.C. thanks Dmitri Petrov and Mike Macpherson for very helpful discussions and the anonymous reviewers who provided useful comments.

References

Cai JJ, Smith DK, Xia X, Yuen KY. 2005. MBEToolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution. BMC Bioinformatics. 6:64.

Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol. 15:1788–1790.

Drummond A, Strimmer K. 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics. 17:662–663.

Dutheil J, Gaillard S, Bazin E, Glemin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinformatics. 7:188.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. Theor Popul Biol. 3:87–112.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics. 158:1227–1234.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 147:915–925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet. 78:153–159.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13:635–643.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science. 307:1072–1079.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111:147–164.

International HapMap Consortium, 2003. The International HapMap Project. Nature. 426:789–796.

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics. 146:1197–1206.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351:652–654.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. Mol Biol Evol. 19:2092–2100.

Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered. 86:248–249.

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

Sabeti PC, Reich DE, Higgins JM, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva. http://lgb.unige.ch/arlequin/

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024.

Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 117:149–153.

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

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 19:2325–2327.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wall JD. 1999. Recombination and the power of statistical tests of neutrality. Genet Res. 74:65–79.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.

Watterson GA. 1978. The homozygosity test of neutrality. Genetics. 88:405–417.

Weir BS. 1996. Genetic data analysis II: methods for discrete population genetic data. Sunderland (MA): Sinauer Associates.

Weir BS, Cockerham CC. 1983. Estimating F-Statistics for the analysis of population structure. Evolution. 38:1358–1370.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 174:1431–1439.

Received August 1, 2008
Accepted December 7, 2008

Corresponding Editor: Robert Wayne