PGEToolbox: a Matlab Toolbox for Population Genetics and Evolution

James J. Cai[1]*

[1]Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA

*Corresponding author

Email addresses:
JJC: jamescai@stanford.edu

Abstract

Background

Matlab is a high-performance language for technical computing, integrating computation, visualization, and programming in an easy-to-use environment. Its usefulness has been increasingly appreciated by biologists. In population genetics studies, a large quantity of genetic diversity data can now be produced at unprecedented rate. Analysing these data at the population level and at the genomic level is vital for understanding evolutionary processes. However, few Matlab functions are freely available for data analysis in evolutionary population genetics.

Results

PGEToolbox is a Matlab-based software package for analysis of polymorphism and divergence data for population genetics and evolution. It estimates several basic statistics of DNA sequence variation and carries out statistical tests of selective neutrality under the infinite alleles model, such as Tajima's D test, Fu & Li's tests and Fay & Wu's H test. The significance of tests is determined from the distribution of the statistics obtained by coalescent simulation. The toolbox performs McDonald-Kreitman test (and several extensions). PGEToolbox also contains functions for handling SNP (Single Nucleotide Polymorphism) genotype and haplotype data. Open-source PGEToolbox can be easily extended or tailored for specific tasks, and scaled up for large data sets.

Conclusions

PGEToolbox is a useful tool that can aid in exploration, interpretation and visualization of data in population genetics and evolution. For academic uses, PGEToolbox is available free of charge at http://bioinformatics.org/pgetoolbox.

Background

Assessing genetic diversity within populations is vital for understanding the nature of the 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. Recent advent of cost-efficient, large-scale genotyping techniques has greatly facilitated the assessment of genetic diversity within population. Massive computations are often involved in analysing these genetic data.

Matlab as a high-performance language for technical computing has been increasingly appreciated by biologist for data manipulation and analysis. The Mathworks, Inc. has recently upgraded its bioinformatics toolbox by including more functions. MBEToolbox, the open-source toolbox for molecular biology and evolution has been widely used since its release [1]. However, to my knowledge, few functions are available for population genetics.

PGEToolbox (from **P**opulation **G**enetics and **E**volution) is a software package for data analysis in molecular population genetics under Matlab. It assists population

geneticists in many ways from manipulating data to performing statistical tests. It contains functions to manipulate polymorphism and/or divergence data and compute many population genetic statistics (Table 1). It allows users to test the departure from selective neutrality with a number of established tests. The significance of results can be evaluated via coalescent simulations. It also provides tests based on comparison of polymorphism and divergence between species, which has been an effective strategy for testing population genetic hypotheses on the causes of variation. In addition, PGEToolbox includes a SNP analysis tool called `snptool` to provide the most frequently used functions related to SNP analysis. Open source and sophisticated graphic function are major advantages of PGEToolbox over similar packages like DnaSP [2] or libsequence [3], which are either proprietarily developed or being lack of graphic function.

Implements

Data Type and File Format

PGEToolbox supports three types of data - DNA sequences, SNP genotype data and phased haplotype data. For DNA sequences, it can read/write alignments in FASTA or Phylip formats. The lengths of the sequences are limited only by the size of the memory. For SNP, it reads genotype information from the HapMap [4, 5] and Perlegen [6] projects. For phased haplotype data, it recognizes files from the HapMap. The functions handling DNA sequences data and SNP data are separate, allowing the user to carry out the many same types of analyses irrespective of the data types. All functions working with SNP data are named with the prefix '`snp_`'. The native Matlab MAT-file is a binary file format which allows any kind of information about individual sequences (or SNPs) and sets of sequences (or SNPs) to be saved.

System and Implementation

PGEToolbox is developed and tested in Matlab version 6.5 (R13) under Microsoft Windows. It is then deployed into versions for UNIX platforms (GNU/Linux and Solaris) and Macintosh platforms. PGEToolbox has been designed with several considerations in mind: batch-ability or scalability, extendibility and usability. As a result, PGEToolbox can be easily set up as scripts (calling one function after another) to perform an entire job in an unattended "batch mode". It is straightforward to apply to large data set. In many case, implementing functions in the Matlab framework greatly reduced the complexity of the original implementation in other languages and allows users to debug or add new functions much more easily than before. The software is under an open source license which allows others to extend and re-use components, allows inter-operation via an open and published interfaces, and can reduce duplication of effort within the community. Graphic user interfaces (GUIs) are useful in hiding the complexity of the computations from the user. PGEToolbox contains many simple yet efficient menu- or dialog- driven GUIs. These GUIs were developed by using GUIDE in Matlab. The main access of functions is through `PGEGUI`, which brings up the menu-driven interface with four major drop-down menu items: File, Data, Analysis and Tools (Figure 1). It aids the usage of the most frequently required functions so that users do not have to run any scripts or functions from the Matlab command line in most cases.

Tutorial and Help System

PGEToolbox provides a comprehensive tutorial and help system. Command `PGEDEMO` brings up slideshow-style demos for sequence-based and SNP-related analysis. Command `help pgetoolbox` lists the name of functions and brief introduction of those functions in the command line window. Command `help` or `edit` followed by a function name gives the usage description or the source code of the function. PGEToolbox website (http://bioinformatics.org/pgetoolbox) contains a step-by-step tutorial and documentation of functions. `PGEGUI` contains a help menu, in which user can bring up Matlab build-in help browser. The version checker under the same menu allows user to check for the latest updates.

Results and Discussion

Polymorphism Statistics

The calculation of polymorphism statistics is a routine task in molecular population genetics. PGEToolbox calculates several polymorphism statistics as a routine task in molecular population genetics, such as polymorphic sites, number of segregating sites, site-frequency spectrum, nucleotide diversity [7] and its sampling variance [8, equation 10.7], Fay's θ_H statistic [9], and linkage disequilibrium (LD) like D, D', or r2 [10] from sequences. For the population mutation parameter, $\theta = 4N\mu$ (where N is the effective population size and μ is the per-locus mutation rate per generation), PGEToolbox computes several common estimates of θ, including the number of segregating sites, θ_W, [11], the mean pairwise difference between nucleotide sequences, θ_π [8], Fay's θ_H [9], and Ewens' θ [12]. An example output from function `estimatetheta` is given in Figure 2.

Neutrality Tests

PGEToolbox conducts several statistical tests for detecting departures from neutrality. These tests include Ewens-Watterson's homozygosity test [13, 14], Tajima's *D* test [15], Fu & Li's *D**-, *F**- tests [16], Strobeck's *S* statistic [17], Wall's *B*- & *Q*-tests [18], Fay & Wu's *H*-test [9], Watterson's homozygosity test of neutrality [14], and Kelly's *ZnS* test [19]. PGEToolbox also offers tests for detecting population growth, Fu's *F*s test [20] and the *R*2 test [21]. For most important tests, like Tajima's D test, three functions are evaluated to complete the whole test. The three functions are `tajima89d_test`, `tajima89d_simu` and `tajima89d`. The first function `tajima89d_test` takes sequences as input to evaluate parameters: θ_π and θ_W, required for calculating the statistics *D*. The second function `tajima89d_simu` generates multiple samples using coalescent simulation (see below) so that the significance of test can be evaluated. Both functions finally call a common procedure deployed in the third function `tajima89d`, which takes merely the necessary parameters θ_π and θ_W to compute *D* directly. Such a design allows user to start with sequence data, simulated data or direct parameter to compute a statistic. Figure 3 illustrates the relationship between functions and input and output information, using `tajima89d` and related functions as an example.

Coalescent Simulations

Testing of the significance of computed statistics like Tajima's *D* requires generate parametric bootstrap samples from a wide variety of neutral models using a coalescent approach and an infinitely many sites model of mutation. Hudson [22]'s program `ms` has been incorporated as a Matlab MEX-function (the C interface to Matlab) to do coalescent simulations [23], giving PGEToolbox extensive capabilities in coalescent-based analyses. Simulations can be conducted for different parameter combinations. A dialog interface for coalescent simulation called `coalsimdlg` was developed to assist users in setting up those parameters (Figure 4).

McDonald-Kreitman Test and Derivatives

PGEToolbox provides methods to analyze patterns of genetic diversity within and between population samples – the McDonald-Kreitman (MK) test [24] and extensions. Functions are available to count the numbers of synonymous (Ds) and nonsynonymous (Dn) divergences, and the numbers of synonymous (Ps) and nonsynonymous (Pn) polymorphisms. The MK test can be initiated from the command-line function `mktestcmd` or from another function called `mktestgui`, which 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 method of Fay, Wyckoff & Wu [25] and the method of Smith & Eyre-Walker's [26].

SNP Tool

PGEToolbox provides many functions to explore the frequency and distribution of SNPs. The interfaces of those SNP-related functions is `snptool`. Depends on user's SNP data type, `snptool` adjusts its menu to provide relevant functions or commands for either genoetype or haplotype data. A haplotype here refers to a set of SNPs found to be statistically associated on a single chromatid. A genotype is distinct from a haplotype because an individual's genotype may not uniquely define that individual's haplotype.

`snptool` opens genotype data file in the formats specified by HapMap or Perlegen. It also can retrieve genotype data from the HapMap and Perlegen databases over the Internet. `snptool` computes several statistics from genotype data, including observed and predicted heterozygosity, minor allele frequency (MAF), *p*-value for Hardy-Weinberg equilibrium test, allele frequency and genotype frequency. It calculates composite likelihood [27], Tajima's *D* [15] and Fay & Wu's *H* [9] statistics for SNPs with frequencies. Display and interpretation of large genotype data sets can be simplified by using `snptool`'s graphical display, such as the pie chart of allele and genotype frequencies among populations for a given SNP, plot for relative position of SNPs on chromosome, and the visual genotype (VG) view. The VG view presents complete raw datasets of individuals' genotype data (Figure 5A). `snptool` uses the expectation-maximization (EM) algorithm to estimate probabilities of haplotypes and calculates LD statistics, such as, D, D' and R between pairs of SNPs.

In haplotype data mode, `snptool` reads the HapMap haplotype data file or retrieves the file from HapMap database directly.

When phased haplotype data is given, `haptool` calculates extended haplotype homozygosity (EHH) [28] and the integrated haplotype score (iHS) [29] for selected core haplotypes (Figure 5B). Both EHH and iHS are based on haplotype homozygosity (HH), which effectively measures LD for more than 2 SNPs. EHH calculate HH in a stepwise manner to see how LD breaks down with increasing distance to a specified core region. HH is evaluated as:

$$HH = \frac{\sum p_i^2 - 1/n}{1 - 1/n}$$

where p_i is the relative haplotype frequency and n the sample size. The variance of HH is estimated according to Nei [30]. iHS is based on the differential levels of EHH surrounding a positively selected allele compared to the background allele at the same position. An extreme positive iHS score (iHS > 2) means that haplotypes on the ancestral allele background are longer compared to derived allele background. An extreme negative iHS score (iHS < -2) means that the haplotypes on the derived allele background are longer compared to the haplotypes associated with the ancestral allele [29]. Both haplotype-based tests have been increasingly used for detecting recent selection [28, 29]. EHH is powerful for detecting a rapid increase in the frequency of an advantageous mutation under recent selection [28, 31]. iHS detected up to 20% of candidate genes under recent positive selection are in the fifth percentile of gene-based |iHS| scores [29].

Taken together, `snptool` provides both genotype and haplotype base tests. The genotype-based tests (usually summarizing data information into site frequency spectrum) likely have higher power to identify selection where the advantageous allele is approaching fixation or completed selective sweeps. The haplotype-based statistics are most suitable for identifying recent, incomplete, and/or ongoing selection.

Conclusions

The usefulness of Matlab as a power and convenient scientific computation has been increasingly appreciated by biologists. PGEToolbox is the first Matlab toolbox dedicated to analysis of polymorphism data in molecular population genetics. It provides the open-source framework containing most frequently used functions ready to be scaled up for massive computations. These functions, together with the user friendly interface, should allow PGEToolbox to gain its popularity within the research community.

Availability and requirements

Project name: PGEToolbox

Project web page: http://bioinformatics.org/pgetoolbox

Operating system: Platform independent

Programming language: Matlab 6.5 or higher

Other requirements: MBEToolbox

License: GPL

Any restrictions on use by non-academics: License needed

Authors' contributions

JJC designed and implemented the software and wrote the manuscript.

Acknowledgements

JJC thanks Dmitri Petrov, Mike Macpherson, Felicity Jones and Frank Chen at the Department of Biological Sciences, Stanford University, Gavin Huttley, Peter Maxwell, Ray Sammut and Helen Lindsay of the Centre for Bioinformation Science (CBiS) at the Australian National University for valuable technical discussions.

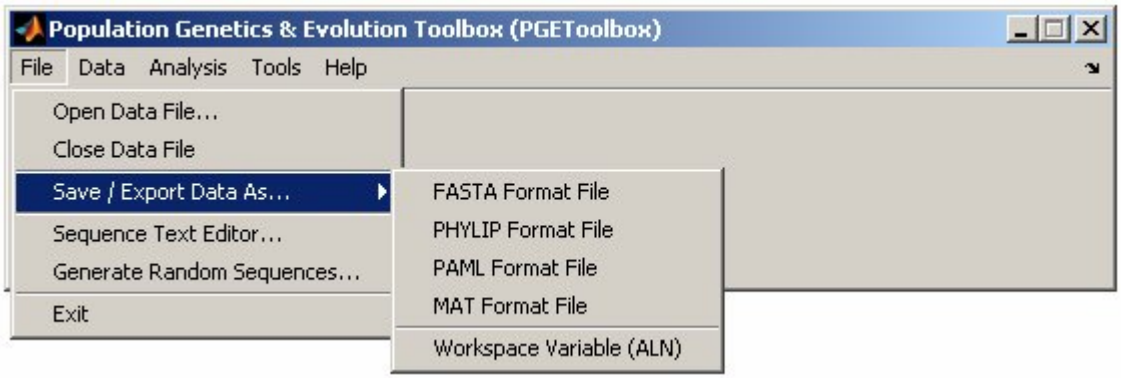
References

Figures

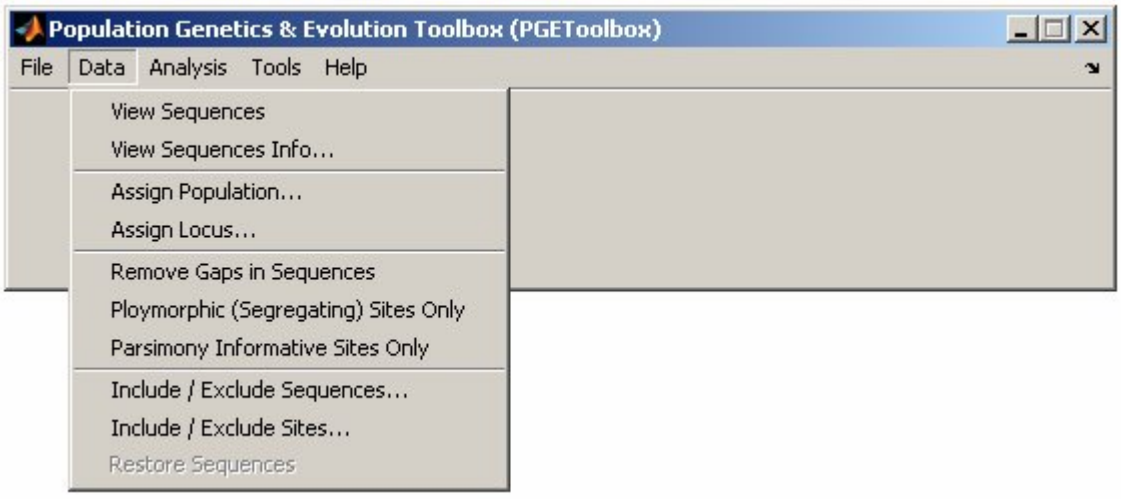
Figure 1 - PGEToolbox GUI, PGEGUI

(A) File submenu; (B) Data submenu; (C) Analysis submenu; and (D) Tools submenu

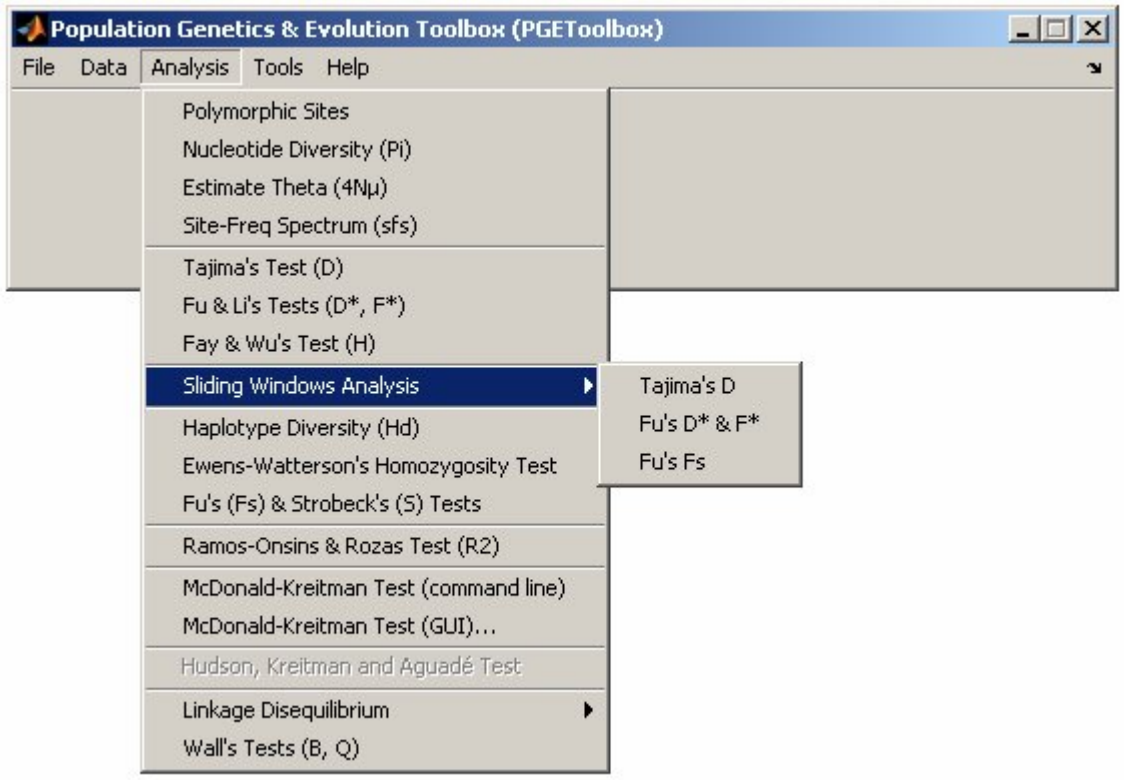
(A)



(B)


Population Genetics & Evolution Toolbox (PGEToolbox)
File
Data
Analysis
Tools
Help
View Sequences
View Sequences Info...
Assign Population...
Assign Locus...
Remove Gaps in Sequences
Ploymorphic (Segregating) Sites Only
Parsimony Informative Sites Only
Include / Exclude Sequences...
Include / Exclude Sites...
Restore Sequences

(C)


Population Genetics & Evolution Toolbox (PGEToolbox)
File
Data
Analysis
Tools
Help
Polymorphic Sites
Nucleotide Diversity (Pi)
Estimate Theta (4Nμ)
Site-Freq Spectrum (sfs)
Tajima's Test (D)
Fu & Li's Tests (D*, F*)
Fay & Wu's Test (H)
Sliding Windows Analysis
Tajima's D
Fu's D* & F*
Fu's Fs
Haplotype Diversity (Hd)
Ewens-Watterson's Homozygosity Test
Fu's (Fs) & Strobeck's (S) Tests
Ramos-Onsins & Rozas Test (R2)
McDonald-Kreitman Test (command line)
McDonald-Kreitman Test (GUI)...
Hudson, Kreitman and Aguadé Test
Linkage Disequilibrium
Wall's Tests (B, Q)

(D)

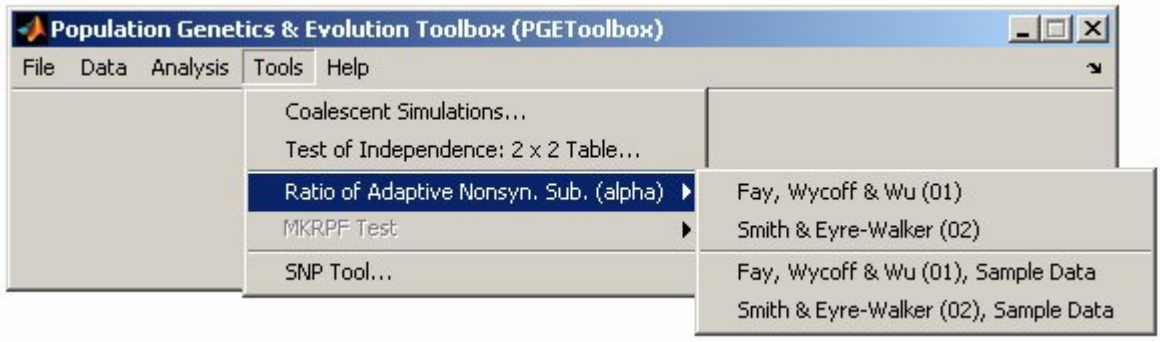

Population Genetics & Evolution Toolbox (PGEToolbox)
File
Data
Analysis
Tools
Help
Coalescent Simulations...
Test of Independence: 2 x 2 Table...
Ratio of Adaptive Nonsyn. Sub. (alpha)
MKRPF Test
SNP Tool...
Fay, Wycoff & Wu (01)
Smith & Eyre-Walker (02)
Fay, Wycoff & Wu (01), Sample Data
Smith & Eyre-Walker (02), Sample Data

Figure 2 - Example output from function estimatetheta

The polymorphism sequences were randomly generated. The results are consistent with those from DnaSP [2].

```
========================================
Estimation of theta (i.e., 4Nu)
========================================
Theta-W (from Eta*) = 3.065693, var = 3.981063 (per sequence)
                    = 0.005135 (per site)
Theta-W (from S**)  = 3.065693, var = 3.981063 (per sequence)
                    = 0.005135 (per site)
Theta-Pi***         = 2.333333, var = 2.903704 (per sequence)
                    = 0.003908 (per site)

*   Eta is the total number of mutations
**  S is number of segregating sites
*** Theta-Pi equals nucleotide diversity
========================================
```

Figure 3 - Relationship among calculation, simulation and testing functions.

Using functions `tajima89d`, `tajima89d_simu` and `tajima89d_test` as example.

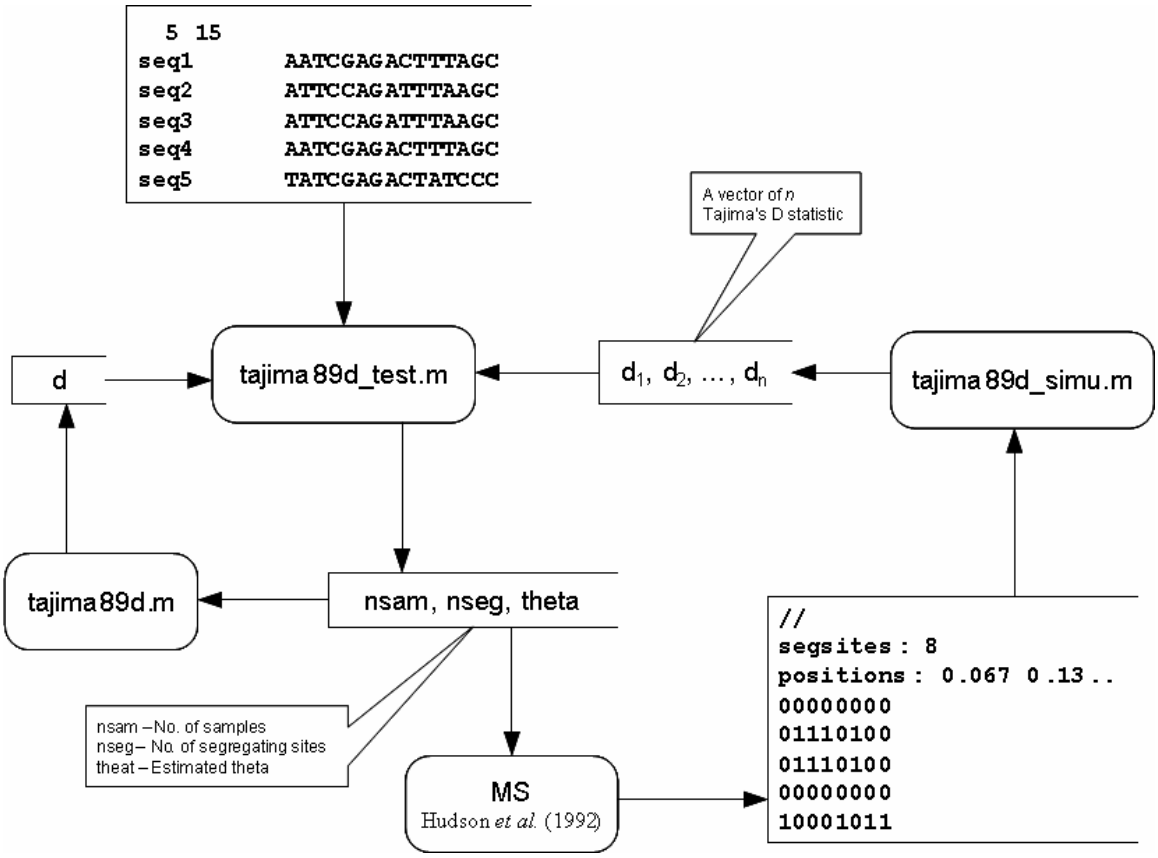
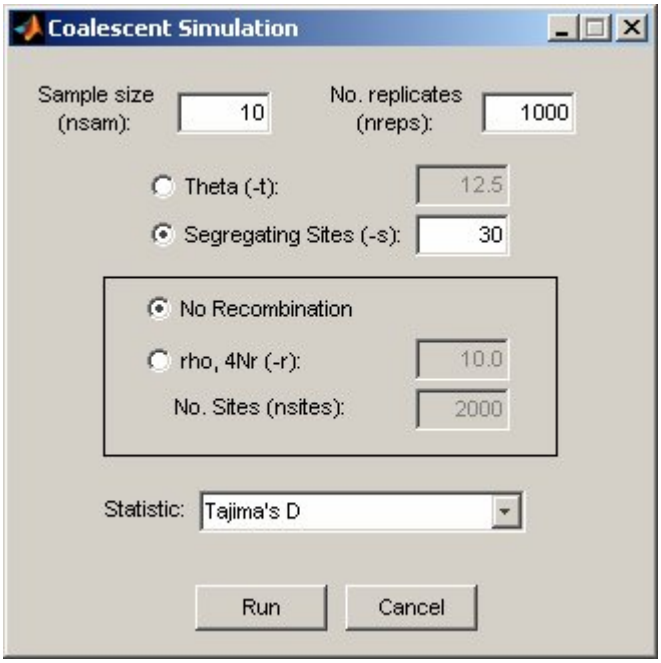


Figure 4 - Coalescent simulation dialog and histogram of result

(A) Dialog of parameter input for coalescent simulation. Simulated data will be generated by using 30, the observed number of segregating sites in the sample and

under the conservative assumption of no recombination. (B) Histogram of estimated statistics (here, Tajima's *D*) from simulated replicates.

(A)



(B)

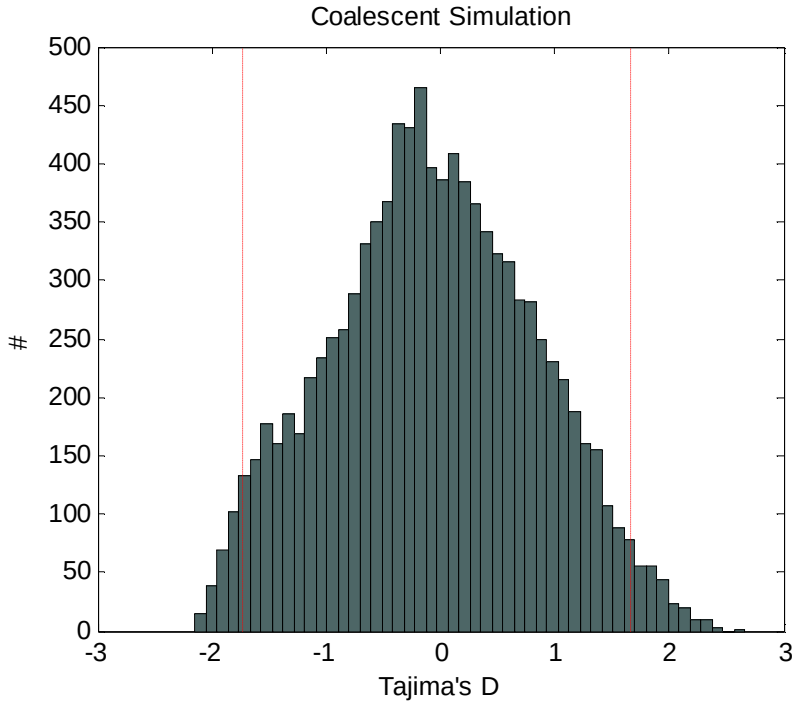
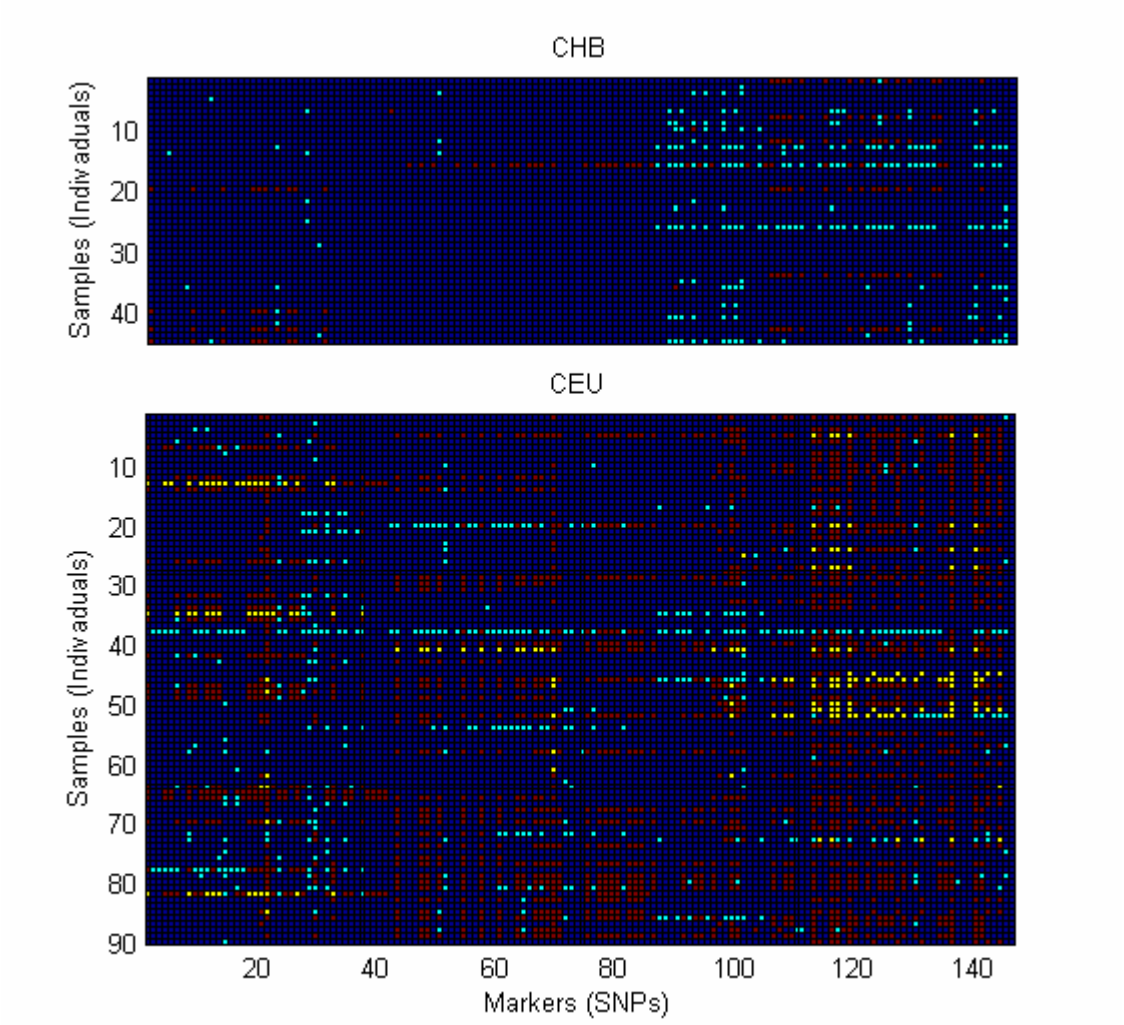


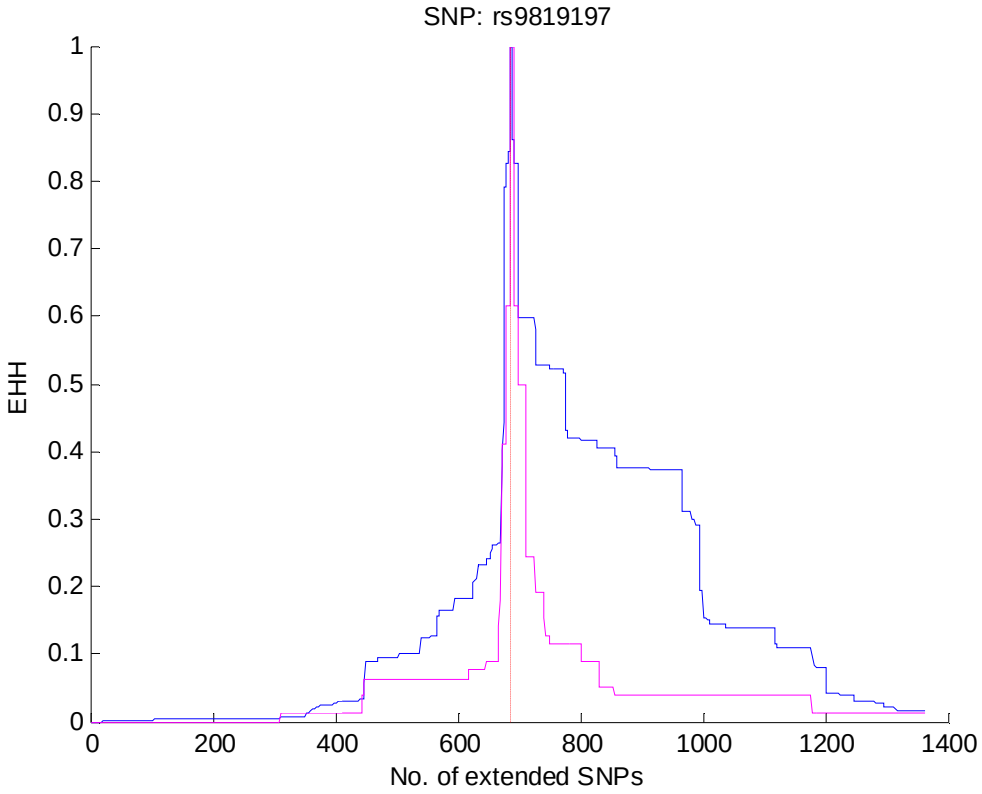
Figure 5 - Example results from SNP-related functions

(A) Visual genotype (VG) view. In each panel, a graphical representation of genotypes is shown for the CHB (Chinese individuals from Beijing) and CEU (CEPH trios from Utah) samples. Rows correspond to individuals and columns denote SNPs. For each SNP, blue, yellow, and red boxes indicate whether the individual is homozygous for the common allele, heterozygous, or homozygous for the rare allele, respectively. Cyan boxes indicate missing data. The SNPs are from human locus EDAR, in which strong signature of positive selection has been identified in the CHB sample [32]. (B) Plot of EHH for two core haplotypes of the single SNP, rs9819197, with haplotype data for HapMap CEU population. Red dash line indicates the position of the core SNP.

(A)



(B)



Tables

Table 1 - Major statistic and tests implemented in PGEToolbox

Statistic or Test	Reference
Watterson's theta, θ_W	[11]
Nucleotide diversity (π), θ_π	[8]
Theta *H*, θ_H	[9]
Ewens' θ	[12]
Tajima's *D*	[15]
Fu and Li's *D*, F**	[16]
Walls *B, Q*	[18]
Watterson's homozygosity test of neutrality	[14]
Kelly's *ZnS* test	[19]
Fu's *Fs* test	[20]
R2 test	[21]
Number of haplotypes	[33]
Haplotype diversity	[33]
Fu's *Fs* test	[20]
McDonald-Kreitman test	[24]
Proportion of positively selected amino-acid substitutions, α	[25, 26]
Extended haplotype homozygosity (EHH)	[28]

Integrated haplotype score (iHS)	[29]

Additional files

Additional file 1 – Using PGEToolbox
This is a step-by-step tutorial file lead the first time user to go through major step in installing and using the toolbox.

Additional file 2 – Comparison of Running Numerical Results
This is a comparison of numerical results between PGEToolbox and two related softwares, DnaSP [2] and NeutralityTest (http://www.hgc.sph.uth.tmc.edu/neutrality_test). The results are from several testing datasets. The comparison shows PGEToolbox is accurate under all these testing conditions.

References

1. Cai JJ, Smith DK, Xia X, Yuen KY: **MBEToolbox: a MATLAB toolbox for sequence data analysis in molecular biology and evolution**. *BMC Bioinformatics* 2005, **6**(1):64.
2. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods**. *Bioinformatics* 2003, **19**(18):2496-2497.
3. Thornton K: **Libsequence: a C++ class library for evolutionary genetic analysis**. *Bioinformatics* 2003, **19**(17):2325-2327.
4. International-HapMap-Consortium: **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.
5. International-HapMap-Consortium: **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.
6. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations**. *Science* 2005, **307**(5712):1072-1079.
7. Nei M, Miller JC: **A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data**. *Genetics* 1990, **125**(4):873-879.
8. Nei M: **Molecular evolutionary genetics**. New York: Columbia University Press; 1987.
9. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection**. *Genetics* 2000, **155**(3):1405-1413.
10. Hedrick PW: **Gametic disequilibrium measures: proceed with caution**. *Genetics* 1987, **117**(2):331-341.

11. Watterson GA: **On the number of segregating sites in genetical models without recombination**. *Theor Popul Biol* 1975, **7**(2):256-276.
12. Ewens WJ: **Mathematical population genetics**, 2nd edn. New York: Springer; 2004.
13. Ewens WJ: **The sampling theory of selectively neutral alleles**. *Theor Popul Biol* 1972, **3**(1):87-112.
14. Watterson GA: **The homozygosity test of neutrality**. *Genetics* 1978, **88**(2):405-417.
15. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism**. *Genetics* 1989, **123**(3):585-595.
16. Fu YX, Li WH: **Statistical tests of neutrality of mutations**. *Genetics* 1993, **133**(3):693-709.
17. Strobeck C: **Average Number of Nucleotide Differences in a Sample from a Single Subpopulation: A Test for Population Subdivision**. *Genetics* 1987, **117**(1):149-153.
18. Wall JD: **Recombination and the power of statistical tests of neutrality**. *Genet Res Camb* 1999, **74**:65–79.
19. Kelly JK: **A test of neutrality based on interlocus associations**. *Genetics* 1997, **146**(3):1197-1206.
20. Fu YX: **Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection**. *Genetics* 1997, **147**(2):915-925.
21. Ramos-Onsins SE, Rozas J: **Statistical properties of new neutrality tests against population growth**. *Mol Biol Evol* 2002, **19**(12):2092-2100.
22. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation**. *Bioinformatics* 2002, **18**(2):337-338.
23. Hudson R: **Gene genealogies and the coalescent process**. In: *Oxford surveys in evolutionary biology.* Edited by Futuyma D, Antonovics J, vol. 7. New York: Oxford University Press; 1990: 1-44.
24. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila**. *Nature* 1991, **351**(6328):652-654.
25. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome**. *Genetics* 2001, **158**(3):1227-1234.
26. Smith NG, Eyre-Walker A: **Adaptive protein evolution in Drosophila**. *Nature* 2002, **415**(6875):1022-1024.
27. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data**. *Genome Res* 2005, **15**(11):1566-1575.
28. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ *et al*: **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 2002, **419**(6909):832-837.
29. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome**. *PLoS Biol* 2006, **4**(3):e72.
30. Nei M: **Molecular population genetics and evolution**. American Elsevier Pub. Co., 1975.
31. Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K: **Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection**. *Am J Hum Genet* 2004, **74**(6):1198-1208.

32. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches**. *Genome Res* 2006, **16**(8):980-989.
33. Depaulis F, Veuille M: **Neutrality tests based on the distribution of haplotypes under an infinite-site model**. *Mol Biol Evol* 1998, **15**(12):1788-1790.