Instructions of how to use PGEToolbox

Installation and Getting Started

PGEToolbox can be downloaded from http://bioinformatics.org/pgetoolbox/

Step by step instruction:

1. Download 'PGEToolbox.zip' or 'PGEToolbox.tar.gz' file;
2. Unzip the files in a local directory keeping the structure of sub-directories. A
3. directory called `pgetoolbox` will be created;
4. Start up Matlab;
5. Add the local directory `./pgetoolbox` to the Matlab path;
6. From Matlab run command `PGEGUI`. This should bring up the main menu GUI of the toolbox.

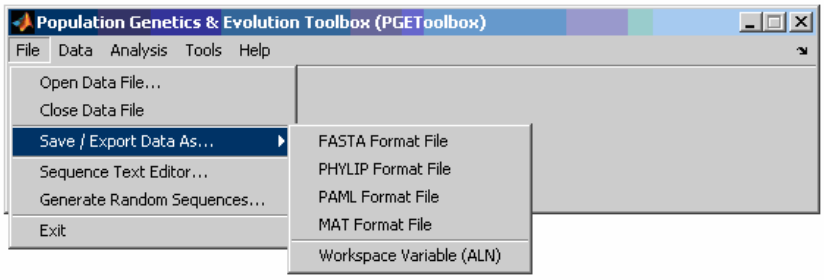
Tutorial and Help System

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 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.

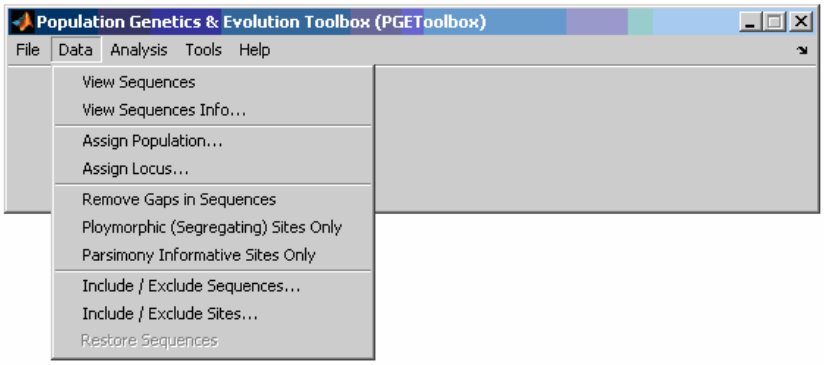
Supplementary Figures

Figure S1 - PGEToolbox GUI
(A) File submenu; (B) Data submenu; (C) Analysis submenu; and (D) Tools submenu

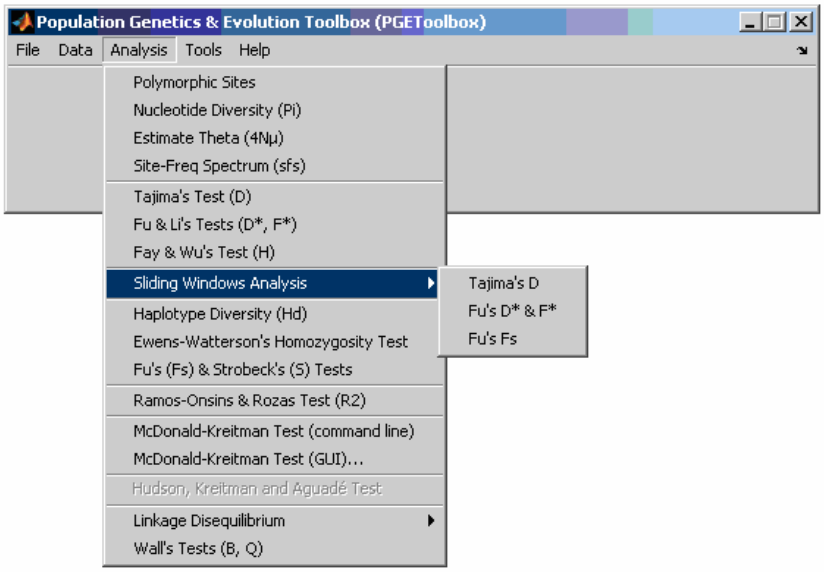
(A)


Population Genetics & Evolution Toolbox (PGEToolbox)
File
Data
Analysis
Tools
Help
Open Data File...
Close Data File
Save / Export Data As...
Sequence Text Editor...
Generate Random Sequences...
Exit
FASTA Format File
PHYLIP Format File
PAML Format File
MAT Format File
Workspace Variable (ALN)

(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
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)
Tajima's D
Fu's D* & F*
Fu's Fs

(D)

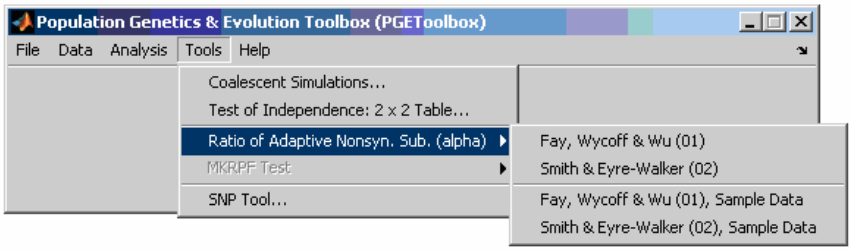

Population Genetics & Evolution Toolbox (PGEToolbox)
File
Data
Analysis
Tools
Help
Coalescent Simulations...
Test of Independence: 2 x 2 Table...
Ratio of Adaptive Nonsyn. Sub. (alpha)
MKRPF Test
SNP Tool...
Fay, Wycoff & Wu (01)
Smith & Eyre-Walker (02)
Fay, Wycoff & Wu (01), Sample Data
Smith & Eyre-Walker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

Figure 2 - Example output from function estimatetheta

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

```
========================================
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.

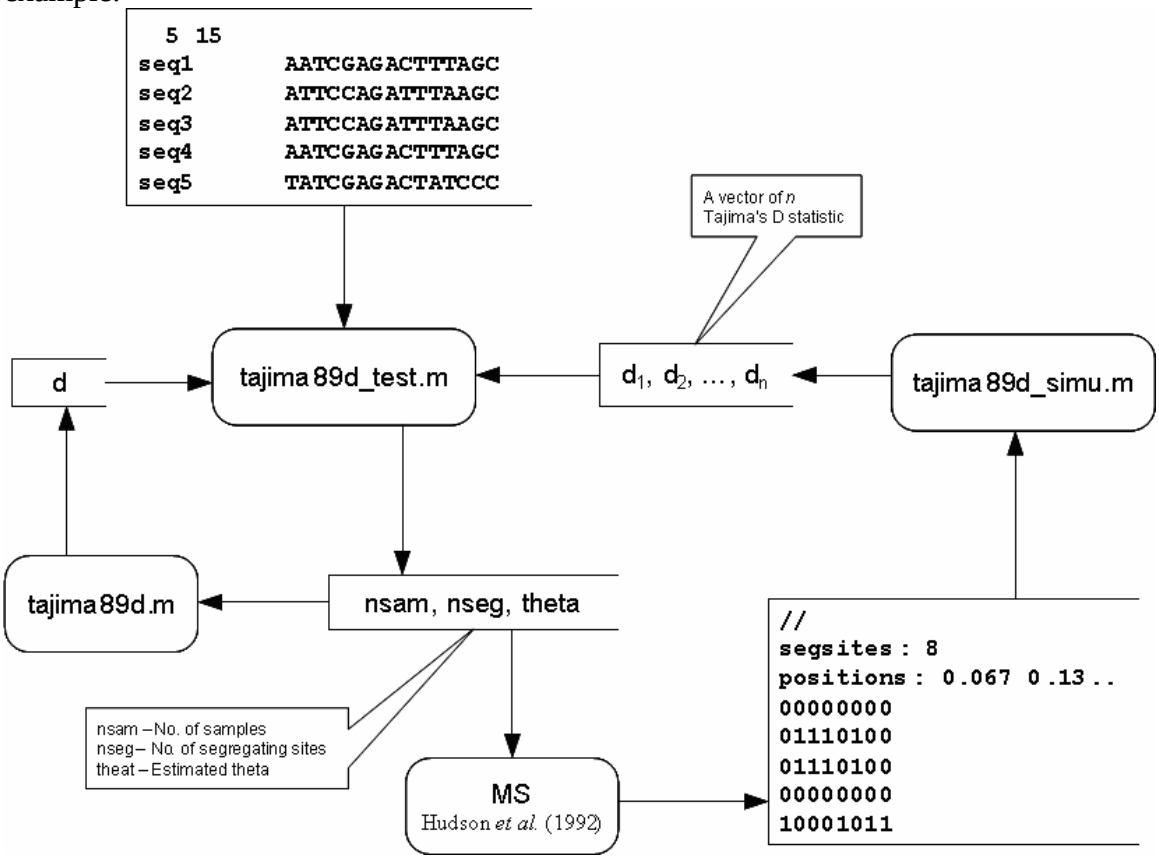


Figure 4 - Coalescent simulation dialog and histogram of result

(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.

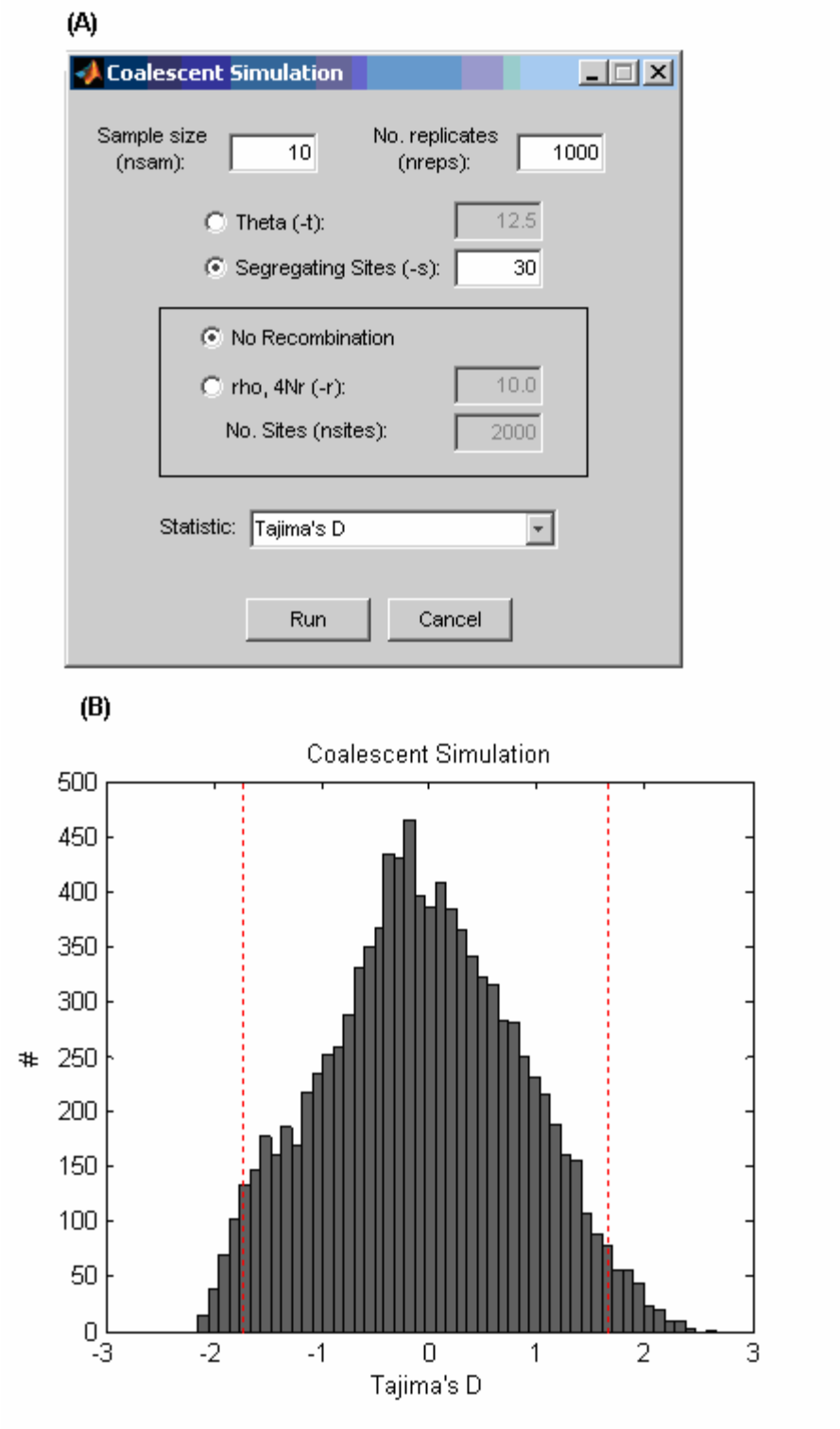
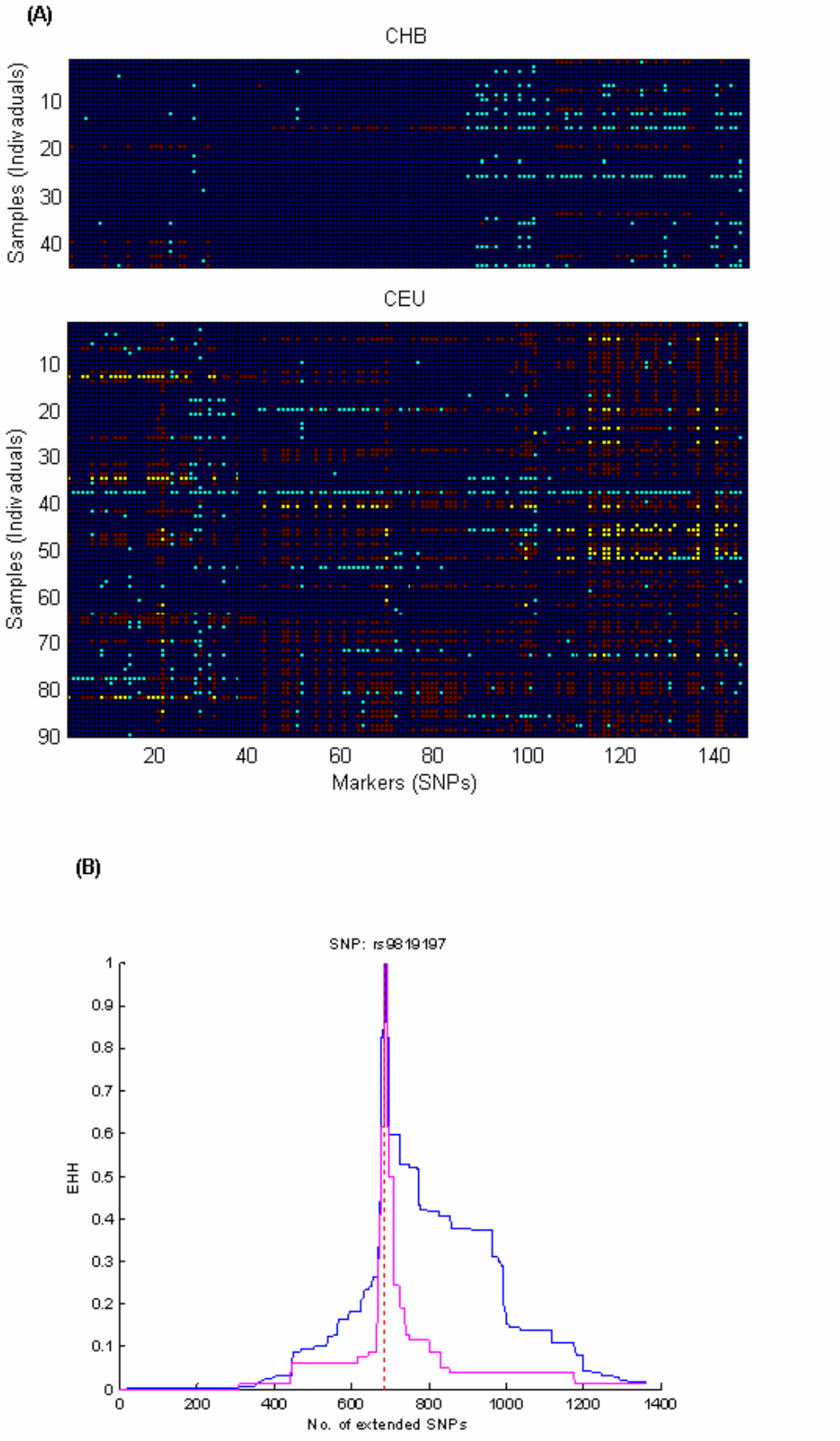


Figure 5 - Example results from SNP-related functions
(A) Visual genotype (VG) view. In each panel, a graphical representation of genotypes is shown for the CHB (Chinese individuals from Beijing) and CEU (CEPH trios from Utah) samples. Rows correspond to individuals and columns denote SNPs. For each 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)
CHB
Samples (Individuals)
10
20
30
40
CEU
Samples (Individuals)
10
20
30
40
50
60
70
80
90
20
40
60
80
100
120
140
Markers (SNPs)
(B)
SNP: rs9819197
EHH
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
200
400
600
800
1000
1200
1400
No. of extended SNPs