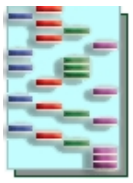


DnaSP

DNA Sequence Polymorphism

Version: 6.10.03
November 30, 2017

Table of contents

Contents

DnaSP Version: 6.10.03 Help Contents
November 30, 2017

Running DnaSP, press **F1** to view the context-sensitive help.

Introduction

Summary

References: DnaSP v1 DnaSP v2 DnaSP v3 DnaSP v4 DnaSP v5

Population genetics is a branch of the evolutionary biology that tries to determine the level and distribution of genetic polymorphism in natural populations and also to detect the evolutionary forces (mutation, migration, selection and drift) that could determine the pattern of genetic variation observed in natural populations. Ideally, the best way to quantify genetic variation in natural populations should be by comparison of DNA sequences (Kreitman 1983). However, although the methodology for DNA sequencing is available since 1977 (Maxam and Gilbert 1977 Sanger et al. 1977), until 1990 the use of DNA sequence data had had little impact on population genetics. This is because the effort (in terms of both money and time) required to obtain DNA sequence data from a relative large number of alleles was substantial.

The introduction of the polymerase chain reaction (PCR) (Saiki et al. 1985; 1988) which allows direct sequencing of PCR products and avoids, therefore, their cloning, has changed the situation. Undoubtedly this has produced a revolutionary change in population genetics. Although, at present, population studies at the DNA sequence level are still scarce and primarily carried out in Drosophila (for example: McDonald and Kreitman 1991 Schaeffer and Miller 1993 Rozas and Aguadé 1994), they will certainly increase in the future.

The DnaSP (DNA Sequence Polymorphism) is a software addressed to molecular population geneticists and can compute several measures of DNA sequence variation within and between populations in noncoding, in synonymous or in nonsynonymous sites; gene flow, gene conversion (Betrán et al. 1997), recombination and linkage disequilibrium parameters. In addition, DnaSP performs some neutrality tests: the Hudson, Kreitman and Aguadé (1987), the Tajima (1989), McDonald and Kreitman 1991; and the Fu and Li (1993) tests. DnaSP takes advantage of the Microsoft Windows capabilities, so that it can handle a large number of sequences of thousands of nucleotides each on a microcomputer. Furthermore, DnaSP can easily exchange data with other programs, for example, programs to perform multiple sequence alignments, phylogenetic tree analysis, or statistical analysis.

What DnaSP can do

What DnaSP can do:

References & Abstracts: DnaSP v1 DnaSP v2 DnaSP v3 DnaSP v4 DnaSP v5 DnaSP 2009 (chapter book) DnaSP v6

DnaSP, DNA sequence polymorphism, is an interactive computer program for the analysis of DNA polymorphism from nucleotide sequence data. The program calculates several measures of DNA sequence variation within and between populations (with or without the sliding window method) in noncoding, synonymous or nonsynonymous sites; linkage disequilibrium, recombination, gene flow and gene conversion parameters; and also computes some neutrality tests, Fu and Li's, Hudson, Kreitman and Aguadé's, McDonald and Kreitman, and Tajima's tests. DnaSP can also conduct computer simulations based on the coalescent process. The input data file is a multiple sequence alignment (MSA), or (new in version 6) a Multi-MSA file format.

What DnaSP can not do:

DnaSP can not align sequences. There are some available programs that can do this. For example, you can perform the multiple alignment with CLUSTAL W (Thompson et al. 1994), MAFFT, T-Coffee, Muscle, or many other tools. These programs produces outputs (multiple aligned sequences; MSA) in different formats that can be read by DnaSP.

DnaSP can not make phylogenetic inferences or manipulate trees. There are many programs to do this, for example, MacClade (Maddison and Maddison 1992), MEGA (Kumar et al. 1994), PHYLIP (Felsenstein 1993), PAUP (Swofford 1991), RAxML or MrBayes. Nevertheless, the input formats used by DnaSP (FASTA, MEGA, NBRF/PIR, NEXUS and PHYLIP format) are also recognized for some of them.

DNA sequences can not be edited or manipulated by DnaSP. You can do this by using, for example, MacClade (Maddison and Maddison 1992) or SeqApp / SeqPup programs (Gilbert 1996).

DnaSP can not directly analyze diploid genetic information (for instance, SNPs data from diploid genomic regions). If you are using diploid unphase data, you can reconstruct the phase using the Open Unphase/Genotype Data module, or use the Unfold a FASTA File with Ambiguity Codes Convert a FASTA File with Ambiguity Codes to Ns modules.

System Requirements

System requirements

DnaSP has been mainly written in Visual Basic.NET (Microsoft), and it runs on an IBM-compatible PC under 32 or 64-bit MS Windows.
The software can run in Microsoft Windows, versions Vista/7/8/10

Limitations (Open Data File command (A single MSA data file)

DnaSP has been successfully tested with data files as long as 120 Mbp (for instance, 30 DNA sequences of 4 Mbp each) in a Windows-based computer with 4 Gb of RAM memory.
Maximum number of nucleotides per sequence: Depends on the available memory (> 3,000,000 nt).
Maximum number of sequences: 32767

The grid control cannot display more than 16351 rows or 5448 columns. Therefore, for the sliding window option, the maximum number of rows of results is 16351. Hence, the maximum number of polymorphic sites (linkage disequilibrium module) or of sequences (synonymous and nonsynonymous module) that can be analyzed and displayed on the screen is 181 (the total number of pairwise comparisons is: 181*180 / 2 = 16290). Although DnaSP will not display the results of these analyses on the screen, the results could be saved in a file.

Both the number and length of the sequences that can be handled by DnaSP mainly depend on the available memory. Nevertheless, DnaSP is able to use all RAM memory available in a computer, both the conventional and the extended memory. DnaSP can also use virtual memory (it can use the hard disk space as memory, although in this case the computation speed will be much lower than when using RAM). Thus, the program can handle large numbers of sequences of up to thousands nucleotides each.

For large data sets, the user can use the Multi-MSA Analysis (All Positions) or the Multi-MSA Analysis (SNP Positions) options.

DnaSP under Linux and Macintosh

DnaSP can also be run on Apple Macintosh platforms (using VirtualBox, VMWare Fusion, Parallels Desktop), Linux-Unix-based operating systems (using VirtualBox, VMWare or Wine). See www.ub.edu/dnasp/DnaSP_OSv6.html
Using virtual machines or emulators, the computation speed of the program will decrease.

File Menu / Input and Output

File Menu

See Also: Input Data Files Output

This menu has (among others) the following commands:

Open Data File
This command allows you to open the data file. The command displays the standard Windows directory dialog box in which you may locate files.

Close Data File
Use this command if you wish to close the active data file.

Save/Export Data As
Use this command to save the changes made in the active data file or to export (translate) the active data file from one file format to another (note: the data file exported will not contain the excluded sequences; see the Include / Exclude Sequences command). The command displays the standard Windows directory dialog box where you may choose where to place the file.
This command also allows you to generate an Arlequin project file or a Roehl Data File (see the Haplotype Data File command).

Update NEXUS Data File
Use this command to update the information of the opened NEXUS Data File. The command is enabled for non NEXUS Data Files or if there are some excluded sequences.

Options for Saving (NEXUS format)
You can use this command to specify some options about saving or exporting NEXUS files:
- Saving in an interleaved format. The number of nucleotides of each interleaved block.
- To indicate the type of nucleotide sequences (DNA or RNA).
- To indicate the type of line delimiter:
  - IBM-PC or compatible: CR + LF (ASCII 13 & ASCII 10).
  - Macintosh: CR (ASCII 13).
  - Unix systems: LF (ASCII 10).
- To indicate the version of NEXUS file format:
  - Old version (used by MacClade 3.04 or older)
  - New version, NEXUS version 1 (used by MacClade 3.05 or later)
- To indicate the symbol used for:
  - missing data, alignment gap, and identical site (matching character).

Send All Output to File
Use this command to send all generated output (except graphs) in a file. The command displays the standard Windows directory dialog box where you may choose where to place the file.

Close Output File
Use this command if you wish to close the output file.

Save Current Output
Use this command to save the output (of the last analysis) in a file. The command displays the standard

Windows directory dialog box where you may choose where to place the file.

Page Setup
The command displays the standard Windows Page Setup dialog box where you may change various printer settings, for example, the default printer, paper size, orientation, etc.

Print Output.
Use this command to print the output on the default printer.

Files 1, 2, 3, 4
Lists the four most recently used Data Files.

Exit
This command ends the current DnaSP session.

Shortcut Keys
Open Data File CTRL+O
Close File CTRL+W
Save Output CTRL+S
Print Output CTRL+P
Exit CTRL+X

Input Data Files Formats

Input Data Files Formats

DnaSP can read the many types of multiple sequence alignment (MSA) data file formats:

Standard analysis (standard Open Data File command)
FASTA,
MEGA (Kumar et al. 1994),
NBRF/PIR (Sidman et al. 1988),
NEXUS (Maddison et al. 1997),
PHYLIP (Felsenstein 1993),
HapMap3 Phased Haplotypes

In all cases one or more homologous nucleotide sequences should be included in just one file (ASCII file). The sequences must be aligned (i. e. the sequences must have the same length). Nucleotide sequences should be entered using the letters A, T (or U), C or G (in lower case, upper case, or any mixture of lower and upper case).

DnaSP allows you to analyze a subset of sites of the data file (this option is useful for the analysis of particular regions of the data file, for example, when analyzing exonic and intronic regions separately), or to carry out analyses in a subset of sequences of the data file (see the Include / Exclude Sequences command).

Analysis using the Multi-MSA Data File Analysis (All sites; SNP positions) commands
DnaSP can read some standard RADseq-like multi-MSA file formats, including *.alleles and *.loci generated by pyRAD (Eaton 2014), *.fa generated by STACKS (Catchen et al., 2011) softwares, as well as the *.vcf, generated by many genome-based projects (Danecek et al. 2011).

Analysis using the Haplotype Frequency Data File Analysis command

DnaSP can read *.arp file formats (from Arlequin; Schneider et al. 2000) that include DNA-Haplotype information with their absolute frequencies. See the Haplotype Frequency Data File Analysis command.

Analysis using the Unphase/Genotype Data File command

To use this option DnaSP requires that the DNA sequences (including unphase or genotype information from diploid individuals) be formatted in FASTA format (see FASTA). This format is the standard FASTA format but including the IUPAC nucleotide ambiguity codes to represent heterozygous sites. See the Unphase/Genotype Data File command.

Data File Examples

You can found examples of all data file types in the folder:

```
Program Files (x86)/DnaSP v6/Examples
```

Tip:

Computational Speed. To increase the computational speed using a FASTA format, you can use the new Multi-MSA Data File Analysis (All Sites), including information of a single region. This module accept the format named *.loci (Eaton 2014), which is identical to a MSA in FASTA format, with the exception that file ends with the symbol '//'. The above example in *.loci format would have the following form:

```
>seq_1
ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTTTTTTTTTTCATGTG
>seq_2
ATATAC--GGATA---TTACA----AGAATCTATGTCTGCTTTCTTTTTCATGTG
>seq_3
ATATACGGGGATA---TTATA----AGAATGTGTGTGTTTTTTTTTTCATGTG
>seq_4
ATATACGGGGATA---GTAGT----AAAATGTGTGTGTTTTTTTTTTCATGTG
//
```

FASTA Format

FASTA Format

See Also: Input Data Files

DnaSP can recognize FASTA (*.fas) data file formats (also called Person format). FASTA file format must begin with the symbol '>' in the first line of the file; the sequence name is the first word after that symbol. Additional characters in this line are considered to be comments. The sequence data starts in the second line. Nucleotide data can be written in one or more lines.
DnaSP only recognize non-interleaved FASTA data files.

Special characters

Blank spaces, Tabs, and Carriage returns are ignored (i. e. they can be used to separate blocks of nucleotides). By default DnaSP uses the following symbols:
the hyphen character '-' to specify an alignment gap;
the dot character '.' to specify that the nucleotide in this site is identical to that in the same site of the first sequence (i.e. identical site or matching symbol);
the symbols '?', 'N', 'n' to designate missing data.

Sequence name

The sequence name can be up to 20 characters. Blank spaces and tabs are not allowed (underlines should

be used to indicate a blank space).

Example of FASTA Format

```
>seq_1 [comment -optional-]
ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
>seq_2 [comment -optional-]
ATATAC--GGATA---TTACA----AGAATCTATGTCTGCTTTCTTTTTCATGTG
>seq_3
ATATACGGGGATA---TTATA----AGAATGTGTGTGTGTTTTTTTTTTCATGTG
>seq_4
ATATACGGGGATA---GTAGT----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
```

Tip:
Computational Speed. To increase the computational speed using a FASTA format, you can use the new Multi-MSA Data File Analysis (All Sites), including information of a single region. This module accept the format named *.loci (Eaton 2014), which is identical to a MSA in FASTA format, with the exception that file ends with the symbol '//'. The above example in *.loci format would have the following form:

```
>seq_1
ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
>seq_2
ATATAC--GGATA---TTACA----AGAATCTATGTCTGCTTTCTTTTTCATGTG
>seq_3
ATATACGGGGATA---TTATA----AGAATGTGTGTGTGTTTTTTTTTTCATGTG
>seq_4
ATATACGGGGATA---GTAGT----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
//
```

MEGA Format

MEGA Format

See Also: Input Data Files Kumar et al. 1994

DnaSP can recognize interleaved and non-interleaved MEGA formats (*.meg). MEGA formats must contain the identifier #MEGA in the first line of the file. The second line must start with the word TITLE: followed by some comments (if any) on the data (comments within the sequences must be contained by a pair of double quotation marks: “comment“). The sequence data starts in the third line. The sequence name is the text after the character '#' until the first Blank space, Tab or Carriage return. The nucleotide sequence is written in one or more lines after the sequence name, until the next sequence name that also starts with the symbol '#' (see the MEGA user manual).

Special characters

Blank spaces, Tabs, and Carriage returns are ignored (i. e. they can be used to separate blocks of nucleotides). By default DnaSP uses the following symbols: the hyphen character '-' to specify an alignment gap; the dot character '.' to specify that the nucleotide in this site is identical to that in the same site of the first sequence (i.e. identical site or matching symbol); the symbols '?', 'N', 'n' to designate missing data. Nevertheless, these symbols can be changed in the dialog box that appears when opening a data file.

Sequence name

The sequence name can be up to 20 characters. Blank spaces and tabs are not allowed (underlines should be used to indicate a blank space).

Example of MEGA Format

```
#MEGA
TITLE: 4 sequences (55 nucleotides). File: EX##N1.MEG
#seq_1
ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTTTTTTTTTTCATGTG
#seq_2
.......--..A........C......G...C.A...C..C...C..........
#seq_3
..........A........T......G............................
#seq_4
..........A.....G...T..................................
```

NBRF/PIR Format

NBRF/PIR Format

See Also: Input Data Files Sidman et al. 1988

In the NBRF/PIR files (*.meg, *.pir), the sequence names are placed immediately after the identifier >DL; . The next line is used for comments. The nucleotide sequence is written in the next line (in one or more lines) and is ended with the symbol '*' . The file must contain nucleotide sequences in a noninterleaved form.

Sequence data

Blank spaces, Tabs, and Carriage returns are ignored (i. e. they can be used to separate blocks of nucleotides). The hyphen character '-' must be used to specify an alignment gap. The dot character '.' can be used to specify that the nucleotide in this site is identical to that in the same site of the first sequence. The symbols '?', 'N', 'n' could be used to designate missing data. No other symbols are allowed.

Sequence name

The sequence name can be up to 20 characters. Blank spaces and tabs are not allowed (underlines should be used to indicate a blank space).

Example of NBRF/PIR Format

```
>DL;seq_1
Comment on seq 1 (example file: EX##N1.NBR).
ATATACGGGG TTA---TTAG A----AAAAT GTGTGTGTGT TTTTTTTTTC ATGTG*
>DL;seq_2
Comment: seq 2
ATATAC--GG ATA---TTAC A----AGAAT CTATGTCTGC TTTCTTTTTC ATGTG*
>DL;seq_3
Comment: seq 3
ATATACGGGG ATA---TTAT A----AGAAT GTGTGTGTGT TTTTTTTTTC ATGTG*
>DL;seq_4
Comment: seq 4
ATATACGGGG ATA---GTAG T----AAAAT GTGTGTGTGT TTTTTTTTTC ATGTG*
```

NEXUS Format

NEXUS Format

See Also: Input Data Files Maddison et al. 1997

DnaSP can read NEXUS (*.nex) file formats. These files are standard text files that have been designed (Maddison et al. 1997) to store systematic data. DnaSP can read NEXUS files (both old and new versions, Maddison et al. 1997) containing DNA or RNA sequence data. The file can contain one or more sequences; in the later case, the homologous nucleotide sequences must be aligned (i. e. the sequences must have the same length).

Nucleotide sequences should be entered using the letters A, T (or U), C or G (in lower case, upper case, or any mixture of lower and upper case). Blank spaces and Tabs are ignored (i. e. they can be used to separate blocks of nucleotides). Carriage returns are also ignored in non-interleaved file formats.

Alignment gap symbol

The symbol used to designate an alignment gap should be indicated by the subcommand GAP:
For example, GAP=- indicates that the hyphen character '-' should be used to specify an alignment gap.
Default symbol: -

Identical site (matching character) symbol

The symbol used to designate that the nucleotide in a site is identical to that in the same site of the first sequence should be indicated by the subcommand MATCHCHAR:
For example, MATCHCHAR=.
Default symbol: .

Missing data symbol

The symbol used to designate missing data should be indicated by the subcommand MISSING:
For example, MISSING=?
Default symbol: ?

Note: the following symbols are not allowed in the subcommands GAP, MISSING, and MATCHCHAR:
The white space, and ( ) [ ] { } / \ , ; : = * ' " ` < >
(see Maddison et al. 1997).
Moreover, these subcommands cannot share the same symbol.

Sequence name

There is no limit for the sequence name length; nevertheless, DnaSP will only display the first 20 characters. Blank spaces and tabs are not allowed (underlines should be used to indicate a blank space).

Interleaved format

NEXUS files can contain nucleotide sequences with interleaved and non-interleaved formats. The former format must be indicated by the subcommand INTERLEAVE

NEXUS blocks

NEXUS blocks must end with the command END;. DnaSP will read the following NEXUS blocks (see Maddison et al. 1997):

DATA, TAXA, CHARACTERS blocks. These blocks contain information about the taxa and the molecular sequence data.

SETS block. That block allows the user to store information of groups of sequences, characters, taxa, etc. DnaSP only uses the TaxSet command. This block contains information about groups of sequences.
NOTE: See also Define Sequence Sets.

CODONS block. This block contains information about the genetic code, and about the regions of the

sequence that are noncoding, or protein coding regions.
NOTE: See also Assign Coding Regions.

CODONUSAGE block. This is a private NEXUS that contains information about the specific table of Preferred and Unpreferred codons that will be used in the Preferred and Unpreferred Synonymous Substitutions analysis. There are 8 predefined tables; nevertheless, the user can define their own table.
Subcommands:

- Pref*: subcommand. Includes the preferred codons.
- Unknown: subcommand. Includes codons of unknown preference nature.

NOTE: See also the Data Menu. See also the NEXUS Format Example 1.

DNASP block. This is a private NEXUS block that contains information about:
i) the chromosomal location of the DNA region:
CHROMOSOMALLOCATION= command. There are 8 predefined chromosomal locations:

- Autosome
- Xchromosome
- Ychromosome
- Zchromosome
- Wchromosome
- prokaryotic
- mitochondrial
- chloroplast

ii) or the organism's genomic type:
GENOME= command. There are 2 predefined genomic types:

- Diploid
- Haploid

NOTE: See also the Data Menu

Example of NEXUS version1 Format

```
#NEXUS
[This is an example of the new NEXUS file format, NEXUS version 1. This is the
version used by MacClade 3.05 or later. File: EX##new1.nex]

BEGIN TAXA;
DIMENSIONS NTAX=4;
TAXLABELS
seq_1
seq_2
seq_3
seq_4;
END;

BEGIN CHARACTERS;
DIMENSIONS NCHAR=55;
FORMAT DATATYPE=DNA MISSING=? GAP=- MATCHCHAR=. INTERLEAVE ;
MATRIX
seq_1 ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGT
seq_2 .........A.......C.-----.G...C.A...C..C
seq_3 ..........A.......T.-----.G.............
seq_4 ..........A..---G...T----...............
seq_1 TTTTTTTTTCATGTG
seq_2 ...C...........
```

```
seq_3 ..............
seq_4 ..............
;
END;

BEGIN SETS;
TaxSet Barcelona = 1-2;
TaxSet Girona = 3;
TaxSet Catalunya = 1-3;
TaxSet Outgroup = 4;
END;

BEGIN CODONS;
CODONPOSSET * UNTITLED =
N: 1 2 6-26 51-55,
1: 3 27-48\3,
2: 4 28-49\3,
3: 5 29-50\3;
CODESET * UNTITLED = Universal: all ;
END;

BEGIN CODONUSAGE;
PREFUNPREFCODONS GENETICCODE=Universal Drosophila_melanogaster =
PREF*: UUC UCC UCG
UAC UGC CUC CUG
CCC CAC CAG CGC
AUC ACC AAC AAG
AGC GUC GUG GCC
GAC GAG GGC;
END;

BEGIN DNASP;
CHROMOSOMALLOCATION= Autosome;
GENOME= Diploid;
END;
```

Example of NEXUS (old version) Format

```
#NEXUS
[This is an example of the Old NEXUS File Format used by MacClade 3.0 File:
EX##old1.nex]

BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=55;
FORMAT MISSING=? GAP=- DATATYPE=DNA ;
MATRIX
seq_1 ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
seq_2 ATATAC--GGATA---TTACA----AGAATCTATGTCTGCTTTCTTTTTCATGTG
seq_3 ATATACGGGGATA---TTATA----AGAATGTGTGTGTGTTTTTTTTTTCATGTG
seq_4 ATATACGGGGATA---GTAGT----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
;
END;

BEGIN CODONS;
CODPOSSET UNTITLED =
1: 3 27 30 33 36 39 42 45 48,
2: 4 28 31 34 37 40 43 46 49,
3: 5 29 32 35 38 41 44 47 50;
GENCODE UNIVNUC
;
```

```
END;
```

PHYLIP Format

NEXUS Format

See Also: Input Data Files Felsenstein 1993

DnaSP can recognize interleaved and non-interleaved PHYLIP (*.phy) formats. PHYLIP formats must contain two integers in the first line of the file: the first number indicates the number of sequences in the data file, while the second indicates the total number of sites. The sequence data starts in the second line. The sequence name can be up to 10 characters. The nucleotide sequence starts immediately (position 11). Nucleotide data can be written in one or more lines.
In PHYLIP interleaved formats, the sequence name must be indicate only in the first block.

Special characters

Blank spaces, Tabs, and Carriage returns are ignored (i. e. they can be used to separate blocks of nucleotides). By default DnaSP uses the following symbols:
the hyphen character '-' to specify an alignment gap;
the dot character '.' to specify that the nucleotide in this site is identical to that in the same site of the first sequence (i.e. identical site or matching symbol);
the symbols '?', 'N', 'n' to designate missing data.

Sequence name

The sequence name can be up to 10 characters. Blank spaces are allowed.

Example of PHYLIP Format

```
4 55
seq_1      ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
secuencia2ATATAC--GGATA---TTACA----AGAATCTATGTCTGCTTTCTTTTTCATGTG
DmelanogasATATACGGGGATA---TTATA----AGAATGTGTGTGTGTTTTTTTTTTCATGTG
seq_4      ATATACGGGGATA---GTAGT----AAAATGTGTGTGTGTTTTTTTTTTCATGTG
```

HapMap3 Phased Haplotypes Format

HapMap3 Phased Haplotypes Format

See Also: Input Data Files

DnaSP can recognize HapMap3 Phased Haplotypes (*.phased) file formats (phased haplotypes generated in the third HapMap phase). HapMap3 Phased Haplotypes format is a space-separated file with phased SNP information (haplotype information).
In the below example, the HapMap3 file contains 3 individuals (in total 6 chromosomes -or haplotypes-) with 9 positions (8 polymorphic and 1 monomorphic).

First row

```
rsID position_b36 NA19028_A NA19028_B NA19031_A NA19031_B NA19035_A NA19035_B
```

The first row must contain –separated by spaces- two any strings (in the above example rsID and position_b36) followed by the haplotypes IDs (the IDs must end with "_A" or "_B").

In the example, NA19035_A and NA19035_B correspond to the two haplotypes IDs from individual NA19035.

Following rows

rs28832292 18095260 C T T T T T

The first column is the SNP ID (rs28832292) and the second column the physical position in the reference chromosome (18095260). The subsequent columns contain the 6 nucleotide variants (from position 18095260). For instance, the nucleotide variants of the chromosomes NA19028_A and NA19028_B in the 18095260 position are a C and a T, respectively.

Special characters

Double-spaces and tabs are treated as a single spaces.
Other symbols than A, C, G, T, U, N, ? or - are not accepted.

Note

DnaSP will export any data file to the HapMap3 format including only polymorphic sites (but also positions with gaps/missing data).

Very Important Note

Since this format might not contain all the monomorphic sites, statistics based on the physical distance, or in the total number of positions (i.e., per-site genetic distances like p, K, nucleotide divergence, D_{xy}, D_a, etc) will be incorrect.

Example of HapMap3 Phased Haplotypes Format

```
rsID position_b36 NA19028_A NA19028_B NA19031_A NA19031_B NA19035_A NA19035_B
rs28832292 18095260 C T T T T T
rs28439049 18136371 A A A A A A
rs28505894 18179985 C C T C C C
rs35630207 18206177 C C C A C C
rs28842485 18325726 A A C A A A
rs4633700 18357066 G G C G G G
rs2300680 18398549 G G C G G G
rs28620789 18520261 A A A C A A
rs28841911 18534123 T C T T T C
```

Arlequin File Format

Arlequin File Format

See Also: Input Data Files Haplotype Frequency Data File Analysis
References: Excoffier and Lischer 2010

DnaSP can read Arlequin *.arp (Arlequin project) data files with DNA sequence information (haplotype) and their frequency, of a single locus (genomic region). The DNA sequence data must be aligned; this file, therefore, store information of a single MSA.

The structure of Arlequin data file are well described in their manual: http://cmpg.unibe.ch/software/arlequin35/Arlequin35.html). In this data file, everything (except in the Structure section) following the "#" character (until the end of the line) are comments (font color in black).

The sections of the Arlequin data file required for DnaSP are the following:

```
[Profile]
      GenotypicData=0          #Haplotypic data
      DataType=DNA             #DNA sequence data
      NbSamples=XX             #XX is the number of samples or populations; each sample may include the DNA sequence data from several individuals

[Data]
      [[Samples]]
            SampleName="Sample_of_Hospitalet"          #The name of the sample/population
            SampleSize=YY1     #YY1 is the total number of individuals of the sample/population
            SampleData= {
id1   Z1    ATCCCTCCTCCTTCTCGGT
id2   Z2    ATGCCTCCTCCTTCTCGGT
id3   Z3    ATCCTTCCTCCTTCTCGGT
}           #There are 3 different haplotypes (id1, id2 and id3) with frequencies Z1, Z2 and Z3 (Z1+Z2+Z3 = YY1)

            SampleName="Sample_of_Alella"          #The name of the sample/population
            SampleSize=YY2     #YY2 is the total number of individuals of the sample/population
            SampleData= {
id1   W1    ATCCCTCCTCCTTCTCGGT
id5   W2    TTCCCTCCTCCTTCTCGGT
id6   W3    ATCCCTCCTCCTTCTCGGG
}           #There are 3 different haplotypes (id1, id5 and id6) with frequencies W1, W2 and W3 (W1+W2+W3 = YY2)

...   #Information of the following samples, until the population number XX

      [[Structure]]      #optional section. It allows to define hierarchical groups of samples/populations. The user can also define this information by means of a *.SG.txt file (see below)
            NbGroups=2   #In this example, there are two groups of samples

            #Barcelona   #In this section, the "#" character precedes the group name
            Group={
                  "Sample_of_Hospitalet"
                  "Sample_of_ElPrat"
                  "Sample_of_Premia"
                  "Sample_of_Alella"
            }

            #Valencia
            Group={
                  "Sample_of_ElSaler"
                  "Sample_of_Alcira"
                  "Sample_of_Brujasot"
            }
```

Multi-MSA Formats

Multi-MSA File Formats

See Also: Input Data Files Multi-MSA Data File Analysis
***References*:** Eaton 2014 Catchen et al. 2011 Danecek et al. 2011

DnaSP can read several Multi-MSA (Multiple Sequence Alignment) file formats. A Multi-MSA format is a single data file containing DNA sequence data of several (1..>50,000) different genomic regions (different MSAs). Example of these data files includes the VCF (Variant Call Format) file formats (Danecek et al. 2011), as well as those data files generated by some popular programs for analyzing RADseq-like data, such as pyRAD (Eaton 2014) and STACKS (Catchen et al., 2011).

Multi-MSA file formats

In particular DnaSP can read the following Multi-MSA file formats:

*.fa and *.alleles File formats generated by stacks (Catchen et al., 2011) and pyRAD (Eaton 2014), respectively. DNA sequence data from a diploid organism, where the two alleles of each individual are separated (phased data)

*.loci File format generated by pyRAD (Eaton 2014), a format very similar to that of FASTA. Stores DNA sequence data of a single sequence per individual (phased data). DnaSP will consider any ambiguity code as a sequencing error (equivalent to an 'N'). If the *.loci file contain true diploid data (that is, if the ambiguity codes represent true IUPAC nucleotide ambiguity codes), the user should use the pyRAD software to obtain the corresponding *.alleles data file.

*.VCF File format under the VCF (Variant Call Format; Danecek et al. 2011) specifications. This format can store meta-information of DNA sequence variation and:
the genome status (haploid, diploid, triploid, etc),
genotype data,
the nucleotide variants of variable positions (the VCF format can also store monomorphic positions)
and phased (using the '|' symbol) or unphased ('/' symbol) sequence status (see the VCF specifications).
The structure of the VCF data file format is well described in its manual: https://samtools.github.io/hts-specs/VCFv4.3.pdf, https://samtools.github.io/hts-specs

*.gVCF The gVCF (Genomic Variant Call Format) file format (https://software.broadinstitute.org/gatk/documentation/article.php?id=4017) is a kind of the VCF (Danecek et al. 2011), which includes information of the state of all positions (variant and non-variant positions). [Not implemented yet]

All of these Multi-MSA formats can store DNA sequence information of a several regions of the genome. These regions can differ in the length (typically less than 1000bp for a RADseq experiment), and can also differ in the number of individuals surveyed (not all individuals should be sequenced in all regions). The empirical data could be obtained from some RRL (Reduced-Representation Libraries) approaches such as RADseq.
See also the Multi-MSA Data File Analysis section to know what types of analysis the program can perform.

Multi-MSA -based data files & analyses

DnaSP can read up to three different text files, each one including different information.

Multi-MSA files The above mentioned data files, i.e., data files with the DNA sequence information (and in some cases some general features of the genomic regions)

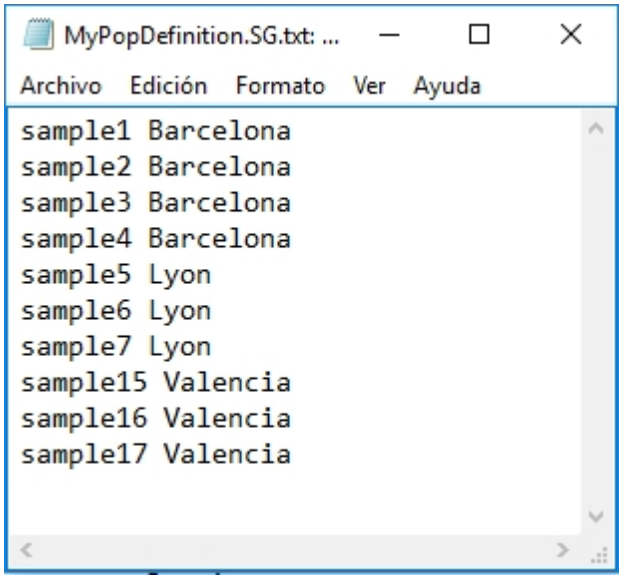
*.SG.txt the SampleToGroups data file that specifies which samples belong to which group (population, species, outgroup, etc).

*.GFF the GFF3 (Generic Feature Format, version 3), the data file storing genomic features of each MSA (such as exon, intron, etc). [Not implemented yet]

Definition of hierarchical groups of individuals or populations (*.SG.txt)

The assignment of samples (individuals) to a particular group (that might represent a population, species, outgroup, etc) must be done in a separate file. If provided (is optional) DnaSP will be able to perform analyses separately within or between groups (e.g., between populations). This feature is equivalent to the Define Sequence Sets command (used for the standard 1-locus analyses).
The structure of *.SG.txt text file is the same as that used by STACKS (a population map file; Catchen et al., 2011). This text file has two columns separated by a tab (or blanc space). The first column includes

information of the Sample (sample/individual name) and the second column for the Group (population; an upper hierarchical category) name.



Be careful, if you are using *.fa files (as shown in the example below, obtained from the STACKS manual), the sample ID (name of the individual) is the second value reported in the file, the value in green (eg, Sample_934, Sample_935 or Sample_936; in the example).

```
>CLocus_10056_Sample_934_Locus_12529_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_934_Locus_12529_Allele_1 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGTTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_935_Locus_13271_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_935_Locus_13271_Allele_1 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGTTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_936_Locus_12636_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
```

Assignment of genomic features (exon, intron, etc) of each MSA. (a GFF3 file format: *.GFF). Optional Data File. Not implemented yet

If this information is provided DnaSP will be able to analyze some statistics separately for different gene regions (exonic, intronic, synonymous changes, non-synonymous changes, etc). DnaSP read the standard GFF3 file, where the seqid value of the GFF file (column 1) must be the same that the RegionName values (i.e., the region/gene/scaffold ID).

Note. It is not necessary provide information (GFF values) for all regions included in the Multi-MSA file.

Tips:

Computational Speed. You can use the new module Multi-MSA Data File Analysis (All Sites) to increase the computational speed in the analysis of a single MSA. For that you should use a data file in FASTA format. Since the *.loci format (Eaton 2014) is nearly identical to that of FASTA format, with the exception that the *.loci format ends with the symbols '//', you only need to include these symbols in your FASTA file. FASTA format.

Multiple Data Files Analysis (Batch Mode)

Multiple Data Files Analysis

This module allows the user to read and analyze -at once- multiple data files (see Input Data Files) sequentially (as a Batch mode). Each data file (an MSA) can contain different number of sequences, or represent different genomic regions. The software can compute a number of measures of the extent of DNA polymorphism (DNA Polymorphism option) or DNA Polymorphism and Divergence (DNA Polymorphism/Divergence option). For the latter the user should define which (only one) sequence is the outgroup (the first or the last sequence of each MSA); the rest of DNA sequences are considered as the ingroup (intraspecfic data).

Analysis

DNA will conduct the following analyses:

1. DNA Polymorphism

1.1 GC content

- G+Cn, G+C content at noncoding positions.
- G+Cc, G+C content at coding positions.
- G+Ctot, G+C content in the complete genomic region.

1.2 Haplotype/Nucleotide Diversity

- The number of Segregating Sites, S
- The total number of mutations, Eta
- The number of haplotypes, Hap (Nei 1987, p. 259).
- Haplotype (gene) diversity (Hd), and its sampling variance (VarHd) (Nei 1987).
- Nucleotide diversity, Pi (π) (Nei 1987), and its sampling variance (not implemented yet; VarPi) (Nei 1987, equation 10.7).
- The average number of nucleotide differences, k (aka ThetaK) (Tajima 1983).
- Watterson theta per site (ThetaWattNuc) from Eta (η) or from S (Watterson 1975; Nei 1987).
- Watterson theta per gene -sequence (ThetaWatt) from Eta (η) or from S (Watterson 1975; Nei 1987).
- ZnS statistic (Kelly 1997, equation 3).

1.3 Neutrality tests

- Tajima's D (Tajima 1989), and its statistical significance.
- Fu and Li's D* (Fu and Li 1993; computed for biallelic positions), and its statistical significance.
- Fu and Li's F* (Fu and Li 1993, Achaz 2009; computed for biallelic positions), and its statistical significance.
- Achaz's Y* (Achaz 2008, equation 21; computed for biallelic positions).
- Fu's Fs (Fu 1997, equation 1).
- Ramos-Onsins and Rozas R_2 (Ramos-Onsins and Rozas 2002).

2. DNA Polymorphism/Divergence

In addition of the DNA Polymorphism statistics (1.1, 1.2 and 1.3), DnaSP will also compute:

- K(JC), average number of substitutions per site (using the Jukes and Cantor correction).
- Fu and Li's D (Fu and Li 1993; computed for biallelic positions), and its statistical significance.
- Fu and Li's F (Fu and Li 1993, Achaz 2009; computed for biallelic positions), and its statistical significance.
- Fay and Wu's Hn (normalized) (Fay and Wu 2000, Zeng et al. 2006; computed for biallelic positions).
- Zeng et al. E (Zeng et al. 2006, equation 13; computed for biallelic positions).
- Achaz's Y (Achaz 2008, equation 21; computed for biallelic positions).

Output

The results are saved on text files, *.MF.out (results for DNA polymorphism) and *.MFd.out (results for DNA Polymorphism/Divergence), with tab-separated values. These files are ready to be read by any spreadsheet application (such as Excel).

Multiple Data Files Analyses and The Coalescent (n-loci | 1-pop)

The output (*.MF.out and *.MFd.out) can also be used as input (in a new session of DnaSP) for the Coalescent Simulations (n-loci | 1-pop) module. Under this module DnaSP can compute the CI and P-values values of many statistics under the Coalescent process.

More information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Linkage Disequilibrium Tajima's Test, etc.

Abbreviations:

n.d., not determined (not implemented yet).
n.a., not available.
n.s., not significant.

Tips:

Computational Speed. You might consider to use the old DnaSP version (you can execute both versions in your computer) www.ub.edu/dnasp/indexDnaSPv5, or use the new Multi-MSA Data File Analysis (All Sites) utilizing a *.SG.txt file.

Multi-MSA Data File Analysis (All Positions; RADseq data)

Multi-MSA Data File Analysis (All Positions; RADseq Data)

See Also: Multi-MSA Formats Multi-MSA Analysis (SNP Positions)
***References*:** Eaton 2014 Catchen et al. 2011 Danecek et al. 2011

This module allows the user to read and analyze a Multi-MSA data file, containing full DNA Sequence data (monomorphic and polymorphic positions). That is a single data file containing DNA sequence data of several (1..>50,000) different genomic regions (different MSAs). Example of these data file includes those files generated by some popular programs for pre-processing and assembling RADSeq-like data, such as pyRAD (Eaton 2014) and STACKS (Catchen et al., 2011). See also the Multi-MSA Format section. The information of the regions (MSA) included in the Multi-MSA file can differ both in the number of positions (although typically less than 1000bp for a RADSeq data file), and in the number of individuals analyzed (not all individuals need to be sequenced in all regions). Example of these data files also includes the gVCF (Genomic Variant Call Format) file format (https://software.broadinstitute.org/gatk/documentation/article.php?id=4017), a kind of the VCF (Danecek et al. 2011), which also includes information of the monomorphic (non-variant) positions.

The present version of DnaSP can read the following file formats formats:
*.fa (generated by STACKS; Catchen et al., 2011). Diploid data (the two alleles of each individual are separated; phased data)
*.alleles (generated by pyRAD; Eaton 2014). Diploid data (the two alleles of each individual are separated; phased data)
*.loci (generated by pyRAD; Eaton 2014). This format contain information from a single sequence per individual, and DnaSP will consider that is a phased data. Therefore, DnaSP will consider any ambiguity

code as a sequencing error (equivalent to an 'N'). If the data file contain true diploid-genotype data (that is, if the ambiguity codes represent true IUPAC nucleotide ambiguity codes), the user should use the pyRAD software to obtain the corresponding *.alleles data file.
*.gvcf (generated by many genome-based projects; Danecek et al. 2011). This format store meta-information of DNA sequence variation, including the state of all positions (variant and non-variant positions). [not implemented yet]

Polymorphism and Divergence

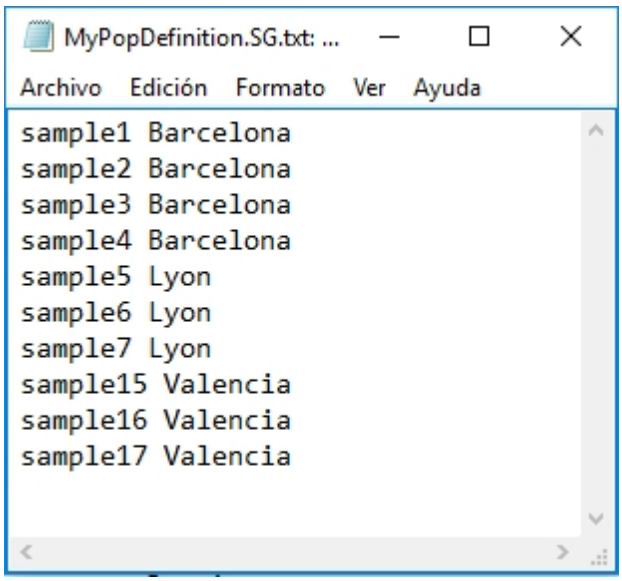
Multi-MSA Data File Analysis (All Positions) --Polymorphism and Divergence

See Also: Multi-MSA Data File Analysis (SNP Positions)
***References*:** Eaton 2014 Catchen et al. 2011 Danecek et al. 2011

This module allows the user to analyze the levels and patterns of nucleotide variation from a Multi-MSA data file, containing DNA sequence data from a single or from several populations, and including information from all positions (monomorphic and variable). DnaSP can estimates a number of measures of the extent of DNA polymorphism, the levels of heterozygosity per individual (data files including genotype information), the amount of DNA Divergence between/among populations and the levels of gene flow. For the analysis, DnaSP requires a Multi-MSA data file, that is a data file containing DNA sequence data from (per example) a RADSeq-based experiment. See the Multi-MSA Format and the Multi-MSA Data File Analysis (All Positions) sections.

Population Assignment File (Definition of hierarchical groups of individuals or populations; *.SG.txt)

The assignment of samples (individuals) to a particular group (that might represent a population, species, outgroup, etc) must be done in a separate file. If provided (is optional) DnaSP will be able to perform analyses separately within or between groups (e.g., between populations). This feature is equivalent to the Define Sequence Sets command (used for the standard 1-locus analyses).
The structure of *.SG.txt text file is the same as that used by STACKS (a population map file; Catchen et al., 2011). This text file has two columns separated by a tab (or blanc space). The first column includes information of the Sample (sample/individual name) and the second column for the Group (population; an upper hierarchical category) name.



Be careful, if you are using *.fa files (as shown in the example below, obtained from the STACKS manual),

the sample ID (name of the individual) is the second value reported in the file, the value in green (e.g., Sample_934, Sample_935 or Sample_936; in the example).

```
>CLocus_10056_Sample_934_Locus_12529_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_934_Locus_12529_Allele_1 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGTTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_935_Locus_13271_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_935_Locus_13271_Allele_1 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGTTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
>CLocus_10056_Sample_936_Locus_12636_Allele_0 [groupI, 49712]
TGCAGGCCCCAGGCCACGCCGTCTGCGGCAGCGCTGGAAGGAGGCGGTGGAGGAGGCGGCCAACGGCTCCCTGCCCCAGAAGGCCGAGTTCACCG
```

Analysis

DnaSP conducts different types of analysis separately for each region/MSA: DNA Polymorphism within populations and between populations, and gene flow analysis. The analyses can differ in function of the information provided in the input data file. If the input data file contains full (monomorphic and variable positions) phased DNA sequence data (*.fa , *.alleles and *.loci), the output is the following:

1. DNA Polymorphism (within populations or within species)

1.1 GC content [*.RAD.out]

- G+Ctot, G+C content in the complete genomic region.

1.2 Haplotype/Nucleotide Diversity [*.RAD.out]

- The number of segregating sites, S
- The total number of mutations, Eta
- The total number of heterozygous positions (only for data files with genotype information).
- The number of haplotypes, Hap (Nei 1987, p. 259).
- Haplotype (gene) diversity (Hd), and its sampling variance(VarHd) (Nei 1987).
- Nucleotide diversity, Pi (π) (Nei 1987).
- The average number of nucleotide differences, k (aka ThetaK) (Tajima 1983).
- Watterson theta per site (ThetaWattNuc) from Eta (η) or from S (Watterson 1975; Nei 1987).
- Watterson theta per gene -sequence (ThetaWatt) from Eta (η) or from S (Watterson 1975; Nei 1987).
- ZnS statistic (Kelly 1997, equation 3).

1.3 Neutrality tests [*.RAD.out]

- Tajima's D (Tajima 1989).
- Fu and Li's D* (Fu and Li 1993; computed for biallelic positions).
- Fu and Li's F* (Fu and Li 1993, Achaz 2009; computed for biallelic positions).
- Achaz's Y* (Achaz 2008, equation 21; computed for biallelic positions).
- Fu's Fs (Fu 1997, equation 1).
- Ramos-Onsins and Rozas R_2 (Ramos-Onsins and Rozas 2002).

1.4 Heterozygosity (within individuals) -if the data file contains genotype information [*.RAD.Hetz.out]

- The observed heterozygosity positions at a particular individual (across all loci-MSAs).
- The total number of sites sequenced at a particular individual (across all loci).
- The net number of sites analyzed at a particular individual (across all loci); i.e., the total number of sites excluding alignment gaps and missing data.
- Ho, the heterozygosity per site at a particular individual (across all loci).

- The total number of loci sequenced (and analyzed) at a particular individual.

2. DNA Divergence among populations (and gene flow) [*.RAD.Btw.out; *.RAD.GFlow.out; *.RAD.PW.out]

2.1 For a given population pairwise comparison (for each region/MSA) [*.RAD.Btw.out]

- The sample size in population 1 (S_size1) and in population 2 (S_size2).
- The total number of net sites (total number of positions excluding missing data and alignment gaps).
- The total number of segregating sites in population 1 (S1), in population 2 (S2), or in the total sample (populations 1 plus population 2) (ST).
- The total number of mutations in the total sample (EtaT).
- The total number of fixed differences (mutations) between populations (Fix), segregating only in population 1 or population 2 (M1 and M2, respectively), or shared between populations (MSh).
- The number of haplotypes in population 1, population 2 and in the total sample (H1, H2 and HT, respectively).
- The haplotype diversity in population 1, population 2 and in the total sample (Hd1, Hd2 and HdT, respectively).
- The average number of nucleotide differences in population 1, population 2 and in the total sample (k1, k2 and kT, respectively).
- The average number of nucleotide differences between population 1 and population 2 (kxy).
- The nucleotide diversity in population 1, population 2 and in the total sample (Pi1, Pi2 and PiT, respectively).
- The average number of nucleotide substitutions per site between population 1 and population 2 (Dxy).
- The net number of nucleotide substitutions per site between population 1 and population 2 (Da).
- The Hs and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a; eq. 2); see also note 1 below.
- The Ks and Kst nucleotide-based statistics (Hudson et al., 1992a, eq. 10; eq. 9).
- The Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 36).
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).

2.2 Gene flow among populations [*.RAD.Btw.out; *.RAD.GFlow.out]

- The Hst haplotype-based statistic among populations (Hudson et al., 1992a, eq. 2); see also note 1 below.
- The Nm(Hst) parameter estimated from Hst.
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).
- The Nm(Fst) parameter estimated from Fst.
- The Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 36).
- The Nm(Nst) parameter estimated from Nst.

DnaSP estimates the gene flow levels (across all loci/MSA), as the average the Fst (or Hst, Nst) over all loci.

The estimates of Nm (from the Fst -Hst or Nst values) are based on the island model of population structure (Wright 1951):

Haploids (Mitochondrial, Bacterial, Virus): Nm = (1 - Fst)/2Fst
Diploids (autosome): Nm = (1 - Fst)/4Fst
Diploids (X-chromosome): Nm = (1 - Fst)/3Fst
Diploids (Y-chromosome): Nm = (1 - Fst)/Fst
Triploids (autosome): Nm = (1 - Fst)/6Fst
Tetraploids (autosome): Nm = (1 - Fst)/8Fst

2.3 Genetic differentiation among all populations (for each region/MSA), and gene flow

[*.RAD.GFlow.out]

- The Hs, Ht and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a, 3b and 2, respectively); see also note 1 below.
- The Fw and Fb values. The Fw and Fb are the same than the Vw and Vb (Lynch and Crease 1990), but without applying the Jukes and Cantor correction. The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3). Fst = Fb / (Fw + Fb)
- The Vw, Vb and Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 3, 15 and 36, respectively). Nst = Vb / (Vw + Vb)

DnaSP estimates the gene flow levels (across all loci/MSA) as in point 2.2 above.

2.4 Pairwise genetic distances (differentation or Fst-related values) among populations [*.RAD.PW.out]

- The Dxy, Da, Hst, Nst and Fst values, for any population pair (represented as a semi-matrix).

Output

The results are saved on different text files with tab-separated values. These files are ready to be read by any spreadsheet application (such as Excel).

*.RAD.out -The results of the DNA polymorphism analysis across all loci

*.RAD.PopName1.out -The results of the DNA polymorphism analysis across all loci, for the PopName1 population

*.RAD.Hetz.out -The results of the heterozygosity values within each individual

*.RAD.Btw.out -The results of the DNA Divergence between populations and the Gene Flow estimates

*.RAD.GFlow.out -The results of the Genetic Differentiation among populations (for each region/MSA) and the Gene Flow estimates

*.RAD.PW.out -The pairwise genetic distances among populations

Note

DnaSP computes the Hs, Ht and Hst statistics (Hudson et al., 1992a) using the weighting factors recommended in page 144 (Hudson et al. 1992a); that is using the n-2 correction (only for cases where all populations have samples sizes greater than 2).

Computational issues/limitations

For sample sizes higher than 4000, DnaSP computes the Fu and Li's D*, Fu and Li's D* and Achaz Y* using a bootstrapping (approximate) algorithm instead of the analytical equations given in Achaz (2008 and 2009).

The ZnS statistic is not computed if the number of segregating sites (S) in a particular MSA/region is higher than 1000.

More information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Linkage Disequilibrium Tajima's Test, etc.

Abbreviations:

n.a., not available.
n.s., not significant.
n.d., not determined.

Multi-MSA Data File Analysis (SNP Positions; *.vcf)

Multi-MSA Data File Analysis (SNP Positions; *.vcf)

See Also: Multi-MSA Formats Multi-MSA Analysis (All Positions)
***References*:** Danecek et al. 2011

This module allows the user to read and analyze a Multi-MSA data file, containing SNP data (information from just variable positions; without the state of the monomorphic positions). That is a single data file containing DNA sequence data of several (1..>50,000) different genomic regions (different MSAs). Example of these data files includes the VCF (Variant Call Format) file format (Danecek et al. 2011). See also the Multi-MSA Format section. The information of the regions (MSA) included in the Multi-MSA file can differ both in the number of variable positions, and in the number of individuals analyzed (not all individuals need to be sequenced in all regions).

The present version of DnaSP can read and analyze the following file format:
*.vcf (generated by many genome-based projects; Danecek et al. 2011). This format store meta-information of DNA sequence variation. DnaSP can read and interpret VCF files including information from different ploidy levels (haploid, diploid, triploid, tetraploid), and using both phased or unphased data.

Polymorphism and Divergence

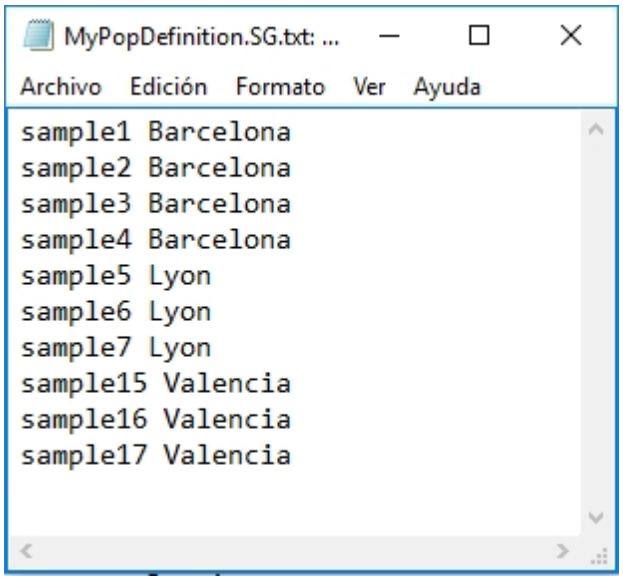
Multi-MSA Data File Analysis (SNP Positions) --Polymorphism and Divergence

See Also: Multi-MSA Data File Analysis (All Positions)
***References*:** Catchen et al. 2011 Danecek et al. 2011

This module allows the user to analyze the levels and patterns of nucleotide variation from a Multi-MSA data file, containing SNP information from a single or from several populations (a *.vcf file). DnaSP can estimates a number of measures of the extent of DNA polymorphism, the levels of heterozygosity per individual (data files including genotype information), the amount of DNA Divergence between/among populations and the levels of gene flow. For the analysis, DnaSP requires a Multi-MSA data file, that is a data file containing SNP information from, per example, a RADSeq-based experiment. See the Multi-MSA Format and the Multi-MSA Data File Analysis (SNP Positions) sections

Population Assignment File (Definition of hierarchical groups of individuals or populations; *.SG.txt)

The assignment of samples (individuals) to a particular group (that might represent a population, species, outgroup, etc) must be done in a separate file. If provided (is optional) DnaSP will be able to perform analyses separately within or between groups (e.g., between populations). This feature is equivalent to the Define Sequence Sets command (used for the standard 1-locus analyses).
The structure of *.SG.txt text file is the same as that used by STACKS (a population map file; Catchen et al., 2011). This text file has two columns separated by a tab (or blanc space). The first column includes information of the Sample (sample/individual name) and the second column for the Group (population; an upper hierarchical category) name.



Analysis

DnaSP conducts different types of analysis separately for each region/MSA: DNA Polymorphism within populations and between populations, and gene flow analysis. The analyses can differ in function of the information provided in the input data file (a *.vcf format), which can incorporate different kind of information (phased/unphased; genotype; population definition in a *.SG.txt file, etc).

1. DNA Polymorphism (within populations or within species)

1.1 Summary of the information provided in each scaffold/MSA [*.VCF.out]

- TotalPos, total number of variable positions included in each scaffold.
- FilteredQual, total number of non-analyzed positions since they do not pass the quality filter (values other than 'PASS' or '.' in the FILTER field of the VCF file).
- FilteredIndels, total number of non-analyzed positions since they include indel variation.
- FilteredOthers, other positions non-analyzed: monomorphic positions; multiple replacements in a given position (eg. REF=AG, ALT=GT).
- SegSites, total number of variable positions analyzed (TotalPos -FilteredQual -FilteredIndels -FilteredOthers).

1.2 Summary of the SNP variation and neutrality tests [*.VCF.out]

- NetSegSites. The net number of segregating sites. All SegSites excluding those positions with missing data in any individual (of the scaffold/MSA).
- Pos1. Scaffold coordinate of the first NetSegSites analyzed.
- Pos2. Scaffold coordinate of the last NetSegSites analyzed.
- NetSites. The net number of positions analyzed (Pos2 -Pos1 -SegSites +NetSegSites +1).
- Sample_Size. Sample size for the particular scaffold/MSA. This value does not include individuals who have missing data in all SegSites positions of a particular MSA.
- Eta. The total number of mutations (η).
- Heterozigosity. The total number of heterozygous positions from the SegSites positions [for genotype data only].
- Hap. The number of haplotypes (Nei 1987, p. 259) [for phased data only].
- Hd. Haplotype (gene) diversity (Nei 1987) [for phased data only].
- VarHd. Sampling variance of the Hd (Nei 1987) [for phased data only].

- ThetaK. The average number of nucleotide differences (aka k) (Tajima 1983).
- Pi. Nucleotide diversity (π) (Nei 1987). DnaSP computes Pi as: ThetaK/NetSites [for NetSites>1 only].
- ThetaWatt. Watterson theta per gene -sequence, from Eta or from S (Watterson 1975; Nei 1987).
- ZnS statistic (Kelly 1997, equation 3) [for phased data only].

1.3 Neutrality tests [*.VCF.out]

- TajimaD. The Tajima's D statistic (Tajima 1989). The statistic can be computed from Eta or from S.
- FuLiD*. The Fu and Li's D* (Fu and Li 1993; computed for biallelic positions).
- FuLiF*. The Fu and Li's F* (Fu and Li 1993, Achaz 2009; computed for biallelic positions).
- AchazY*. The Achaz's Y* (Achaz 2008, equation 21; computed for biallelic positions).
- FuFs The Fu's Fs (Fu 1997, equation 1) [for phased data only].
- Ramos-Onsins_Rozas's_R2. The Ramos-Onsins and Rozas R2 (Ramos-Onsins and Rozas 2002) [for phased data only].

1.4 Heterozygosity (within individuals) -for genotype data only [*.VCF.Hetz.out]

- The observed heterozygosity positions at a particular individual (across all loci-MSAs).
- The total number of SegSites surveyed at a particular individual (across all loci).
- The NetSegSites analyzed at a particular individual (across all loci).
- The total number of MSA analyzed in a particular individual.

2. DNA Divergence among populations (and gene flow) [*.VCF.Btw.out; *.VCF.GFlow.out]

2.1 For a given population pairwise comparison (for each region/MSA) [*.VCF.Btw.out]

- The sample size in population 1 (S_size1) and in population 2 (S_size2).
- The total number of segregating sites in population 1 (S1), in population 2 (S2), or in the total sample (populations 1 plus population 2) (ST).
- The total number of mutations in the total sample (EtaT).
- The total number of fixed differences (mutations) between populations (Fix), segregating only in population 1 or population 2 (M1 and M2, respectively), or shared between populations (MSh).
- The number of haplotypes in population 1, population 2 and in the total sample (H1, H2 and HT, respectively). [for phased data only].
- The haplotype diversity in population 1, population 2 and in the total sample (Hd1, Hd2 and HdT, respectively). [for phased data only].
- The average number of nucleotide differences in population 1, population 2 and in the total sample (k1, k2 and kT, respectively).
- The average number of nucleotide differences between population 1 and population 2 (kxy).
- The net number of nucleotide differences between population 1 and population 2 (ka). This statistic is equivalent to the Da statistic (that is, the net number of nucleotide substitutions per site between population 1 and population 2) but per sequence (per region), not per site.
- The average number of nucleotide substitutions per site between population 1 and population 2 (Dxy). The NetSites values are computed as in Pi (π). Dxy = kxy / NetSites
- The net number of nucleotide substitutions per site between population 1 and population 2 (Da). The NetSites values are computed as in Pi (π). Da = ka / NetSites
- The Hs and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a; eq. 2); see also note 1 below. [for phased data only].
- The Ks and Kst nucleotide-based statistics (Hudson et al., 1992a, eq. 10; eq. 9).
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).

2.2 Gene Flow among populations [*.VCF.Btw.out; *.VCF.GFlow.out]

- The Hst haplotype-based statistic among populations (Hudson et al., 1992a, eq. 2); see also note 1 below. [for phased data only].
- The Nm(Hst) parameter estimated from Hst. [for phased data only].
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).
- The Nm(Fst) parameter estimated from Fst.

DnaSP estimates the gene flow levels (across all loci/MSA), as the average the Fst (or Hst, Nst) over all loci.

The estimates of Nm (from the Fst -Hst or Nst values) are based on the island model of population structure (Wright 1951):

Haploids (Mitochondrial, Bacterial, Virus): Nm = (1 - Fst)/2Fst
Diploids (autosome): Nm = (1 - Fst)/4Fst
Diploids (X-chromosome): Nm = (1 - Fst)/3Fst
Diploids (Y-chromosome): Nm = (1 - Fst)/Fst
Triploids (autosome): Nm = (1 - Fst)/6Fst
Tetraploids (autosome): Nm = (1 - Fst)/8Fst

2.3 Genetic differentiation among all populations (for each region/MSA), and gene flow [*.VCF.GFlow.out]

- The Hs, Ht and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a, 3b and 2, respectively); see also note 1 below. [for phased data only].
- The Fw and Fb values. The Fw and Fb are computed per sequence (not per site). Be careful, in the Multi-MSA Data File Analysis (All Positions) module these statistics are computed on a per-site basis. These statistics are same than the Vw and Vb (Lynch and Crease 1990), but on per-site basis and (obviously) without applying the Jukes and Cantor correction. The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3). Fst = Fb / (Fw + Fb)

DnaSP estimates the gene flow levels (across all loci/MSA) as in point 2.2 above.

2.4 Pairwise genetic distances (differentation or Fst-related values) among populations [*.VCF.PW.out]

- The Dxy, Da, Hst and Fst values, for any population pair (represented as a semi-matrix).

Output

The results are saved on different text files with tab-separated values. These files are ready to be read by any spreadsheet application (such as Excel).

*.VCF.out -The results of the DNA polymorphism analysis across all loci
*.VCF.PopName1.out -The results of the DNA polymorphism analysis across all loci, for the PopName1 population
*.VCF.Hetz.out -The results of the heterozygosity values within each individual
*.VCF.Btw.out -The results of the DNA Divergence between populations and the Gene Flow estimates
*.VCF.GFlow.out -The results of the Genetic Differentiation among populations (for each region/MSA) and the Gene Flow estimates
*.VCF.PW.out -The pairwise genetic distances among populations

Note

DnaSP computes the Hs, Ht and Hst statistics (Hudson et al., 1992a) using the weighting factors recommended in page 144 (Hudson et al. 1992a); that is using the n-2 correction (only for cases where all populations have samples sizes greater than 2).

More information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Linkage Disequilibrium Tajima's Test, etc.

Computational issues/limitations

For sample sizes higher than 4000, DnaSP computes the Fu and Li's D*, Fu and Li's D* and Achaz Y* using a bootstrapping (approximate) algorithm instead of the analytical equations given in Achaz (2008 and 2009).
The ZnS statistic is not computed if the number of net segregating sites (NetSegSites) in a particular MSA/region is higher than 1000.

Abbreviations:

n.a., not available.
n.s., not significant.
n.d., not determined.

Haplotype Frequency Data File Analysis (*.arp)

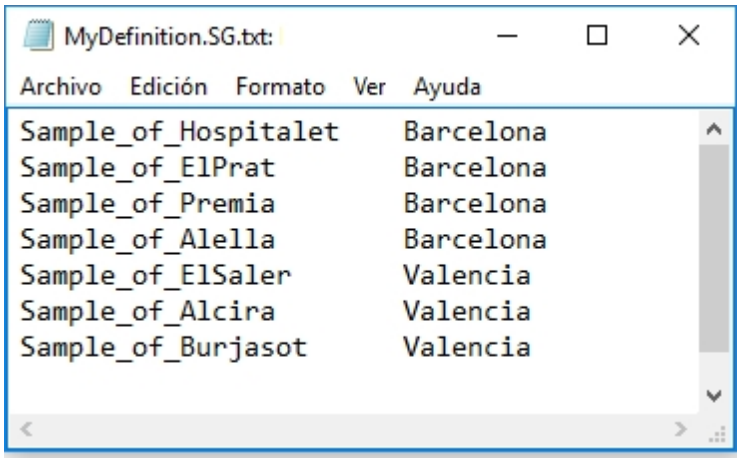
Haplotype Frequency File Analysis (*.arp)

References: Excoffier and Lischer 2010

This module allows the user to read and analyze an Arlequin *.arp (Arlequin project) data file with DNA sequence information (haplotype) and their frequency, for a single locus (genomic region). Moreover, the DNA sequence data must be aligned; that is, it represents a single MSA. This module is especially useful when the user has a huge number of DNA sequences (hundreds or thousands), but with a low level of genetic variation (low levels of haplotype diversity). In this case it makes sense to store the DNA sequence information by the haplotype sequence and their frequency. The input data file must have the sections indicated in the Arlequin File Format (see also the Arlequin manual: http://cmpg.unibe.ch/software/arlequin35/Arlequin35.html).

Population Assignment File (Definition of hierarchical groups of individuals or populations; *.SG.txt)

In addition of the information provided in the Structure section, the user can also define the hierarchical groups of samples by a separated file (*.SG.txt); in the same way as the one used in the RADseq Data File Analysis (see also the Multi-MSA Formats). If provided (is optional) DnaSP will be able to perform analyses separately within or between groups (e.g., between populations). This feature is equivalent to the Define Sequence Sets command (used for the standard 1-locus analyses). In case of using both systems to define groups (using the Structure section, and the *.SG.txt file), DnaSP will prioritize *.SG.txt file. Therefore, the user can perform a number of group assignations (different hierarchical categories) by simply updating the *.SG.txt file, without having to modify the original Arlequin data file (*.arp).
The structure of *.SG.txt text file is the same as that used by STACKS (a population map file; Catchen et al., 2011). This text file has two columns separated by a tab (or blanc space). The first column includes information of the Sample (sample/individual name) and the second column for the Group (population; an upper hierarchical category) name.



Analysis

DnaSP can conduct two types of analysis: DNA Polymorphism within populations and DNA Divergence among populations. These analyses are performed separately for each region/MSA.

1. DNA Polymorphism (within sample or within group)

1.1 GC content [*.DnaSP.out]

- G+Ctot, G+C content in the complete genomic region.

1.2 Haplotype/Nucleotide Diversity [*.DnaSP.out]

- The number of segregating sites, S
- The total number of mutations, Eta
- The number of haplotypes, Hap (Nei 1987, p. 259).
- Haplotype (gene) diversity (Hd), and its sampling variance(VarHd) (Nei 1987).
- Nucleotide diversity, Pi (π) (Nei 1987).
- The average number of nucleotide differences, k (aka ThetaK) (Tajima 1983).
- Watterson theta per site (ThetaWattNuc) from Eta (η) or from S (Watterson 1975; Nei 1987).
- Watterson theta per gene -sequence (ThetaWatt) from Eta (η) or from S (Watterson 1975; Nei 1987).

1.3 Neutrality tests [*.DnaSP.out]

- Tajima's D (Tajima 1989), and its statistical significance.
- Fu and Li's D* (Fu and Li 1993; computed for biallelic positions), and its statistical significance.
- Fu and Li's F* (Fu and Li 1993, Achaz 2009; computed for biallelic positions), and its statistical significance.
- Achaz's Y* (Achaz 2008, equation 21; computed for biallelic positions).
- Ramos-Onsins and Rozas R2 (Ramos-Onsins and Rozas 2002).

1.4 DNA Variability within groups -if the *.SG.txt file is provided [*.DnaSP.out]

In addition of the results of items 1.1, 1.2, 1.3, DnaSP also computes two different estimates of the net DNA divergence levels (net number of nucleotide substitution per site) among samples from a given group.

- Da_Unweigthed; computed as: $Da\_u = \pi_t - \left(\sum_{i=1}^{m} \pi_i\right)/m$

where π_t is the nucleotide diversity among a given group, π_i is the nucleotide diversity in sample i, and m is the number of samples in the group

- Da_Weigthed; computed as: $Da\_w = \pi_t - \left(\left(\sum_{i=1}^{m} \pi_i n_i\right) / \left(\sum_{i=1}^{m} n_i\right)\right)$

where n_i is the sample size (number of sequences) in sample i

2. DNA Divergence between samples and groups [*.DnaSP.Btw.out]

2.1 For a given pairwise sample comparison

- The sample size in sample 1 (S_size1) and in sample 2 (S_size2).
- The total number of net sites (total number of positions excluding missing data and alignment gaps).
- The total number of segregating sites in population 1 (S1), in population 2 (S2), or in the total sample (populations 1 plus population 2) (ST).
- The total number of mutations in the total sample (EtaT).
- The total number of fixed differences (mutations) between populations (Fix), segregating only in population 1 or population 2 (M1 and M2, respectively), or shared between populations (MSh).
- The number of haplotypes in population 1, population 2 and in the total sample (H1, H2 and HT, respectively).
- The haplotype diversity in population 1, population 2 and in the total sample (Hd1, Hd2 and HdT, respectively).
- The average number of nucleotide differences in population 1, population 2 and in the total sample (k1, k2 and kT, respectively).
- The average number of nucleotide differences between population 1 and population 2 (kxy).
- The nucleotide diversity in population 1, population 2 and in the total sample (Pi1, Pi2 and PiT, respectively).
- The average number of nucleotide substitutions per site between population 1 and population 2 (Dxy).
- The net number of nucleotide substitutions per site between population 1 and population 2 (Da).
- The Hs and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a; eq. 2); see also note 1 below.
- The Ks and Kst nucleotide-based statistics (Hudson et al., 1992a, eq. 10; eq. 9).
- The Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 36).
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).

2.2 Genetic differentiation among all groups [*.DnaSP.Btw.out]

- The Hs, Ht and Hst haplotype-based statistics (Hudson et al., 1992a, eq. 3a, 3b and 2, respectively); see also note 1 below.
- The Hw, Hb and Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).
- The Vw, Vb and Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 3, 15 and 36, respectively).

2.3 Gene Flow among all groups [*.DnaSP.Btw.out]

- The Hst haplotype-based statistic among populations (Hudson et al., 1992a, eq. 2); see also note 1 below.
- The Nm(Hst) parameter estimated from Hst.
- The Fst nucleotide-based statistics (Hudson et al., 1992b, eq. 3).
- The Nm(Fst) parameter estimated from Fst.
- The Nst nucleotide-based statistics (Lynch and Crease 1990, eq. 36).
- The Nm(Nst) parameter estimated from Nst.

DnaSP estimates the gene flow levels (Nm values) from the Hst, Fst or Nst values, assuming the island model of population structure (Wright 1951).

Haploids (Mitochondrial, Bacterial, Virus): Nm = (1 - Fst)/2Fst

Diploids (autosome): Nm = (1 - Fst)/4Fst
Diploids (X-chromosome): Nm = (1 - Fst)/3Fst
Diploids (Y-chromosome): Nm = (1 - Fst)/Fst
Triploids (autosome): Nm = (1 - Fst)/6Fst
Tetraploids (autosome): Nm = (1 - Fst)/8Fst

Output

The results are saved on different text files with tab-separated values. These files can be read by any spreadsheet application (such as Excel).
*.DnaSP.out -The results of the DNA polymorphism analysis, separately for samples and group of samples
*.DnaSP.Btw.out -The results of the DNA Divergence between populations and Gene Flow estimates

Note

DnaSP computes the Hs, Ht and Hst statistics (Hudson et al., 1992a) using the weighting factors recommended in page 144 (Hudson et al. 1992a); that is using the n-2 correction (only for cases where all populations have samples sizes greater than 2).

Computational issues/limitations

For sample sizes higher than 4000, DnaSP computes the Fu and Li's D*, Fu and Li's D* and Achaz Y* using a bootstrapping (approximate) algorithm instead of the analytical equations given in Achaz (2008 and 2009).

More information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Linkage Disequilibrium Tajima's Test, etc.

Abbreviations:

n.d., not determined.
n.a., not available.
n.s., not significant.

Open Unphase/Genotype Data File

Open Unphase/Genotype Data Files

References: Stephens et al. 2001 Stephens and Donnelly 2003 Scheet and Stephens 2006 Wang and Xu 2003
See Also: FASTA File Format Unfold a FASTA File with Ambiguity Codes Convert a FASTA File with Ambiguity Codes to Ns

DnaSP can not read directly unphased data (genotype data from diploid individuals). This option, however, allows the user to reconstruct the phases. The unphased data files should be in the standard FASTA format (see FASTA), but including the IUPAC nucleotide ambiguity codes to represent heterozygous sites.

Suppose a data set containing 5 diploid individuals (therefore a total of 10 sequences) with 16 positions each.

```
          *    *    *
Ind1  TRCAAGACCGGAGGCG
Ind2  .A.C..-.-........
Ind3  .A..M.......S...
```

```
Ind4  .A---.......C...
Ind5  .G..C..........
```

For instance, as the second site of Ind1 is heterozygous (R = Purine; A and G), Ind1 includes the following two sequences:

```
Ind1-1    TACAAGACCGGAGGCG
Ind1-2    .G..............
```

As there is not heterozygous site in Ind2, then the two composing sequences are:

```
Ind2-1    TACCAG--CGGAGGCG
Ind2-2    ......-........
```

This DnaSP module allows reconstructing the 10 sequences from the 5 individuals. DnaSP might handle and use the reconstructed data set (10 sequences of 16 nucleotides each) for further analysis.

Haplotype Reconstruction

DnaSP can reconstruct the haplotype phases from unphase data. This haplotype reconstruction is conducted using the algorithms provided by PHASE (Stephens et al. 2001; Stephens and Donnelly 2003), fastPHASE (Scheet and Stephens 2006) and HAPAR (Wang and Xu 2003).

PHASE 2.1 uses a coalescent-based Bayesian method to infer the haplotypes. It can also be used to estimate the recombination rate along the sequences.

fastPHASE 1.1 modifies the PHASE algorithm taking into account the patterns of linkage disequilibrium and its gradual decline with physical distance.

HAPAR uses a pure parsimony approach to estimate the haplotypes; the optimal solution is that which requires less haplotypes to resolve the genotypes. For positions not completely resolved, the user can choose between to replace unresolved positions as "N", or to assign the nucleotide variants randomly.

Note:

fastPHASE and HAPAR can handle only diallelic polymorphic positions. Nevertheless, polymorphic positions segregating for three or more variants can be resolved with PHASE.

Very important:

See the PHASE, fastPHASE or HAPAR documentation for more information and details.

Temporal Results

You can found the temporal results produced by PHASE, fastPHASE or HAPAR in the folders:

```
Users/YourUser/AppData/Roaming/DnaSPhase
Users/YourUser/AppData/Roaming/DnaSfPhase
Users/YourUser/AppData/Roaming/DnaSPHapar
```

IUPAC nucleotide ambiguity codes

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	

```
W          A or T
S          C or G
Y          C or T
K          G or T
V        A or C or G
H        A or C or T
D        A or G or T
B        C or G or T
X      G or A or T or C
N      G or A or T or C
```

Unfold a FASTA File (Diploid Individuals) with Ambiguity Codes to...

Unfold a FASTA File (Diploid Individuals) with Ambiguity Codes

See Also: FASTA File Format Open Unphase/Genotype Data Convert a FASTA File with Ambiguity Codes to Ns

DnaSP can not read and interpret data files including IUPAC nucleotide ambiguity codes (other than 'N' to indicate missing data). If your data file contain such symbols you can either:

- Use this module to randomly unfold IUPAC nucleotide ambiguity codes. Using this option you are considering that ambiguity codes represent heterozygous positions (genotype data). DnaSP will randomly assign the two variants of any heterozygous (biallelic) position to any of the two chromosomes. The converted file (FASTA format) could be read and interpret by DnaSP.
- Use the Open Unphase/Genotype Data command to perform the haplotype reconstruction (statistically sound; computationally demanding). Using this option you are considering that ambiguity codes represent heterozygous positions (genotype data). The converted file could be read and interpret by DnaSP.
- Use this module to convert IUPAC nucleotide ambiguity codes to 'N'. Using this option you are considering that ambiguity codes represent sequencing errors. The converted file (FASTA format) could be read and interpret by DnaSP.

Suppose a data set containing 5 diploid individuals (therefore a total of 10 sequences) with 16 positions each.

```
          *    *    *
Ind1  TRCAAGACCGGAGGCG
Ind2  .A.C..--........
Ind3  .A..M.......S...
Ind4  .A---.......C...
Ind5  .G..C...........
```

The converted file may have the following structure (R, M and S are randomly unfolded into the two chromosomes _0 and _1):

```
          *    *    *
Ind1_0  TGCAAGACCGGAGGCG
Ind1_1  .A..............
Ind2_0  .A.C..--........
Ind2_1  .A.C..--........
Ind3_0  .A.........C...
Ind3_1  .A..C...........
Ind4_0  .A---.......C...
Ind4_1  .A---.......C...
```

```
Ind5_0  .G..C...-------
Ind5_1  .G..C...-------
```

IUPAC nucleotide ambiguity codes

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	
W	A or T	
S	C or G	
Y	C or T	
K	G or T	
V	A or C or G	
H	A or C or T	
D	A or G or T	
B	C or G or T	
X	G or A or T or C	
N	G or A or T or C	

Convert a FASTA File with Ambiguity Codes to 'Ns'

Convert a FASTA File with Ambiguity Codes to Ns

See Also: FASTA File Format Open Unphase/Genotype Data Unfold a FASTA File with Ambiguity Codes

DnaSP can not read and interpret data files including IUPAC nucleotide ambiguity codes (other than 'N' to indicate missing data). If your data file contain such symbols you can either:

- Use this module to convert IUPAC nucleotide ambiguity codes to 'N'. Using this option you are considering that ambiguity codes represent sequencing errors. The converted file (FASTA format) could be read and interpret by DnaSP.
- Use the Open Unphase/Genotype Data command to perform the haplotype reconstruction (statistically sound; computationally demanding). Using this option you are considering that ambiguity codes represent heterozygous positions (genotype data). The converted file could be read and interpret by DnaSP.
- Use the Unfold a FASTA File with Ambiguity Codes command to randomly unfold IUPAC nucleotide ambiguity codes. Using this option you are considering that ambiguity codes represent heterozygous positions (genotype data). The converted file (FASTA format) could be read and interpret by DnaSP.

Suppose a data set containing 5 diploid individuals (therefore a total of 10 sequences) with 16 positions each.

```
          *    *    *
Ind1  TRCAAGACCGGAGGCG
Ind2  .A.C..-.........
Ind3  .A..M.......S...
Ind4  .A---.......C...
```

```
Ind5  .G..C....-------
```

The converted file will have the following structure:

```
          *    *    *
Ind1  TNCAAGACCGGAGGCG
Ind2  .A.C..--........
Ind3  .A..N.......N...
Ind4  .A---.......C...
Ind5  .G..C....-------
```

IUPAC nucleotide ambiguity codes

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	
W	A or T	
S	C or G	
Y	C or T	
K	G or T	
V	A or C or G	
H	A or C or T	
D	A or G or T	
B	C or G or T	
X	G or A or T or C	
N	G or A or T or C	

Output

Output

See Also: Graph Window

The output is displayed in three kinds of windows: text, table or grid (the output data are laid out in rows and columns like in a spreadsheet) and graphic (scatter graph and line chart).

All commands produce an output text window; moreover, some of them also produce a grid (table) window. Data in the grid can be used to create a graph (Graphs command in the Display menu). The data generated from DnaSP can be saved as an ASCII text file. The grid output data file can be easily used by other applications, such as spreadsheets, statistical or graphics applications, by simply removing the header.

Display Menu

Display Menu

This menu has four commands:

Graphs.

This command opens the Graphs Window where graphs from results given in the grid (table) can be displayed.

Data Info.

This command displays a summary of the data file:

The number of sequences; the number of sites; the data file format; the Genetic Code assigned; the organism's genomic type (diploid / haploid); the chromosome type where the nucleotide region is located (autosomal / X-chromosome, etc.).

View Data.

This command displays a window with the sequence data of the active data file. In this window you can get information about:

Coding and noncoding regions.

The status of a selected site (monomorphic, polymorphic, informative, synonymous, nonsynonymous, etc.).

View Data Options.

You can use this command to specify some options about displaying the nucleotide sequences:

To indicate by the dot symbol a nucleotide with identical nucleotide variant to the one in the first sequence.

To show polymorphic sites in Lower Case.

Graphs Window

Graphs Window

This window displays graphs from results given in a grid (table). The are the following commands:

Select Graph

Use this command to select the kind of graph. There are the following:

DNA Polymorphism command. Graph: Line chart
X axis: Nucleotide position; Y axis: Pi(p)
X axis: Nucleotide position; Y axis: Theta (per site)
X axis: Nucleotide position; Y axis: S

DNA Divergence between Populations. command. Graph: Line chart
X axis: Nucleotide position; Y axis: Pi(1) (pop 1)
X axis: Nucleotide position; Y axis: Pi(2) (pop 2)
X axis: Nucleotide position; Y axis: Dxy
X axis: Nucleotide position; Y axis: Da
X axis: Nucleotide position; Y axis: Pi(1) and Pi(2)
X axis: Nucleotide position; Y axis: Dxy and Da
X axis: Nucleotide position; Y axis: Pi(1), Pi(2) and Dxy
X axis: Nucleotide position; Y axis: Pi(1), Pi(2) and Da

Polymorphism and Divergence command. Graph: Line chart
X axis: Nucleotide position; Y axis: Pi(p) and K

Gene Conversion command. Graph: Line chart
X axis: Nucleotide distance; Y axis: Psi(y)

Linkage Disequilibrium command. Graph: Scatter graph
X axis: Nucleotide distance; Y axis: D
X axis: Nucleotide distance; Y axis: |D|
X axis: Nucleotide distance; Y axis: D'
X axis: Nucleotide distance; Y axis: |D'|
X axis: Nucleotide distance; Y axis: R
X axis: Nucleotide distance; Y axis: R^2

Population Size Change command.
Pairwise Number of Differences. Graph: Line chart
X axis: Pairwise differences; Y axis: Frequency
Segregating sites. Graph: Line and Bar chart:
X axis: Number of nucleotide variants in a site; Y axis: Frequency
X axis: Sample size; Y axis: Segregating sites

Fu and Li's tests command. Graph: Line chart
X axis: Nucleotide distance; Y axis: D*
X axis: Nucleotide distance; Y axis: F*
X axis: Nucleotide distance; Y axis: D
X axis: Nucleotide distance; Y axis: F

Tajima's test command. Graph: Line chart
X axis: Nucleotide distance; Y axis: D

Print Graph (Black/White)
Use this command to print in black and white the contents of the window (the graph) at the default printer.

Print Graph (color)
Use this command to print the graph in color at the default printer.

Save Graph (*.bmp)
Use this command to save the graph in a file (bmp format).

Copy Graph (clipboard)
Use this command to copy the graph to the clipboard (i.e. you can paste it to other applications).

Show Significant
This command displays the significant values in Linkage Disequilibrium analysis.

Display in Black/White
Use this command to display the graph in black and white.

Display Default Color
Use this command to display the graph in the default colors.

Colors
You can use this command to change the default colors of the graph.

UCSC Browser

UCSC Browser

***References*:** Kent et al. 2002a Kent 2002b

DnaSP allows you visualizing DNA sequence data and sliding window results, integrated with available genome annotations using the UCSC browser (Kent et al. 2002a). To display the genome annotations DnaSP requires that the information of genomic position of the data (chromosome and physical position) was defined.

DnaSP allows searching available genomes in UCSC. To define the genomic position of your data, choose the appropriate genome, and specify the chromosome and the physical position of the reference sequence (the first one). If you don't know this information you can obtain it:

- Performing a BLAT (Blast-Like Alignment Tool) search (Kent 2002b) against the appropriate UCSC genome.
- Searching the appropriate UCSC genome by key words, and import the output-information to DnaSP.

Genomic position assignations can be stored in NEXUS data files; for that, use the save/export or update commands.

Data Menu

Data Menu

Format

Use this command to indicate if the data file contains sequences of:

DNA or RNA

the chromosomal (genomic) type where the region is located:

Autosome
X chromosome
Y chromosome
Z chromosome
W chromosome
prokaryotic
mitochondrial
chloroplast

or the organism's genomic state:

Diploid or Haploid

Gaps in Sliding Window.

This command is used to exclude/include sites with alignment gaps in the length of the windows (Sliding Window method).

Gaps in Sequence Sets

Use this command to choose how to treat alignment gaps in sequence sets.

Segregating Sites/Mutations

Use this command to select between the number of segregating sites or the total number of mutations in computing some parameters of the Fu and Li's (and other) Tests, Fu and Li's (and other) Tests with an Outgroup, and Tajima's Test.

Assign Coding Regions

Use this command to assign noncoding and coding protein regions to a particular data file.

Assign Genetic Code

Use this command to assign the Genetic Code used for translation. There are 9 pre-defined Genetic codes:
Nuclear universal, the standard code (Table 1)
Mitochondrial of mammals (Table 2)
Mitochondrial of invertebrate, including Drosophila (Table 5)
Mitochondrial of Yeast (Table 3)
Mitochondrial of Mold, Protozoan and Coelenterate (Table 4)
Mitochondrial of Echinoderm (Table 9)
Mitochondrial of Flatworm (Table 14)
Nuclear of Ciliate, Dasycladacean and Hexamita (Table 6)
Nuclear of some Candida species (Table 12)
In parenthesis is indicated the GenBank translation table number. More information on the Genetic Codes used by GenBank in:
https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes

Assign Preferred / Unpreferred Codons Table

Define Sequence Sets

Filter / Remove Positions

Include / Exclude Sequences

Note:
This information will be stored if you save/export (or update) the data file as a NEXUS file format.

Gaps in Sliding Window

 Sliding Window

The sliding window method allows you to calculate some measures or parameters (for example the nucleotide diversity) across a DNA region. In this method a window (segment of DNA) is moved along the sequences in steps. The parameter is calculated in each window, and the value is assigned to the nucleotide at the midpoint of the window. Both the window length and the step size default values can be changed by the user. DnaSP allows you to perform sliding window analyses in non-overlapping windows; for that analysis you must assign the same values to both the window length and the step size.

The output of the sliding window analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph the parameter (Y axis) is plotted against the nucleotide position (X axis).

Gaps in Sliding Window
Sites with alignment gaps are not considered in the length of the windows (i.e. all windows have the same number of net nucleotides).
Windows with a fixed number of net nucleotides. All windows will have the same number of net nucleotides (i.e. the number of nucleotides excluding sites with alignment gaps). In the same way, the step size will also have the same number of net nucleotides.
Sites with alignment gaps are considered.
Windows with a fixed number of total nucleotides. All windows will have exactly the same number of nucleotides. For example, if we choose a window length of 50 nucleotides, and in a particular window the DNA region contains 4 sites with gaps the analysis will be performed in only 46 sites. Likewise, the step size will also have the same total number of nucleotides.

Assign Coding Regions

 Assign Coding Regions

This command allows you to assign noncoding and coding protein regions to the data file. This information might be needed for several analyses. The meaning of a specific codon will depend on the Genetic Code assigned. There is no maximum number of coding protein regions (exons).

You can assign a specific region as a noncoding region, or as a coding region. In the later case, you have to indicate what is the codon position of the first site selected (First, second, third); from this site DnaSP

will assign codons to the remainder sites following the reading frame.

Example: Assume that you have a data file including DNA sequences 34 nucleotides long, and you would like to indicate (assign):
exon 1, from site 6 to site 16,
exon 2, from site 24 to site 30,
noncoding, the rest of the sites.
Assuming a Universal Nuclear Genetic Code:

```
        10        20         30
         *         *          *
ATCTCTTATCGTCGATTTGTTGTTTGTATTTAAT
     LeuSerSerIl       eCysIle
```

You have to do two codon assignments:
i) In the dialog box, indicate as selected region: 6 - 16
Set the codon position of the first site as: 1 (First position)
ii) In the dialog box, indicate as selected region: 24 - 30
Set the codon position of the first site as: 3 (Third position)

You can see the current assignation using the View Data command. You will see the following:

```
NNNNNLeuSerSerIlNNNNNNNeCysIleNNNN
ATCTCTTATCGTCGATTTGTTGTTTGTATTTAAT
```

N, noncoding.

Examples:

DnaSP assigns codons in the following way (assuming the Universal Nuclear Genetic Code). Examples 3 and 4, show how DnaSP assigns codons in case of misassignations or alignment gaps.

Example#1

```
NNNNNLeuSerSerIleNNNNNNCysIleNNNN
ATCTCTTATCGTCGATTTGTTGTTGTATTTAAT
```

Example#2

```
NNNNNLeuSerSerIlNNNNNNNeCysIleNNNN
ATCTCTTATCGTCGATTTGTTGTTTGTATTTAAT
```

Example#3

```
     123123112312       3123123
NNNNNLeuSerS#erIlNNNNNNNeCysIleNNNN
ATCTCTTATCGTGCGATTTGTTGTTTGTATTTAAT
```

(#, wrong assignation)

Example#4

```
     12312312312       3123123
NNNNNLeu???SerIlNNNNNNNeCysIleNNNN
ATCTCTTA---TCGATTTGTTGTTTGTATTTAAT
     Leu   SerIl       eCysIle
```

(?, alignment gaps in nucleotide sequence)

Note:

The amino acid assignation corresponds to the first nucleotide sequence.
This information will be stored if you save/export the data file as a NEXUS file format.

Assign Preferred / Unpreferred Codons Table

Table for Preferred and Unpreferred Codons

See Also: Preferred and Unpreferred Synonymous Substitutions analysis
References: Akashi 1995 Akashi and Schaeffer 1997 Duret and Mouchiroud 1999 Kanaya et al 1999

Use this command to assign the specific table of preferred/unpreferred synonymous codons for the Preferred and Unpreferred Synonymous Substitutions analysis. There are 8 predefined tables. Nevertheless, the user can define its own table. That information could be included in the NEXUS data file; for that use the save/export or update commands.

Create New Table:
This button allows the user to define a new preference synonymous codons table. The table will be linked to a particular Genetic code.

Codes:
P, preferred codon.
?, unknown preference.
none, unpreferred.

Define Domain Sets

Define Domain Sets

This command allows you to define domain sets. A domain set is a partial fragment of the multiple alignment that could represent, for example, an exon, a gene, an intrón, etc. That definition allows DnaSP conducting analyses on specific functional regions using MultiDomain Analysis command.

Domain sets assignations can be stored in NEXUS data files; for that use the save/export or update commands.

Example:
In the data file DmelOsRegions.nex (included in DnaSP package) there are defined two genes (OS-E, OS-F) with three and four exons, respectively. Each gene would correspond to a domain, and each exon to a subdomain. Specifically, the three OS-E_gene (2334..2870) subdomains are: subdomain_1 (2334-2402), subdomain_2 (2468-2542) and subdomain_3 (2598-2870).

Note:
Subdomains does not have name nor can be contiguous.

Remove Positions

Filter / Remove Positions

See Also: Input Data Files Output

This command allows the user to remove some positions. DnaSP module generates a NEXUS Data File including information about the polymorphic sites.

Selected Positions:

DnaSP can select the following sorts of positions:
Coding and Noncoding positions;
First, Second and Third codon positions;
Zero, Two and Four-Fold Degenerate positions;

Example (using the nuclear universal genetic code):

How DnaSP select the X-fold degenerate positions

```
 3  6  9
 *  *  *
ATA TTA ACT
ATA TTA GAT
ATA TTA -CT
```

Positions 1, 2, 5, 7 and 8, are zero-fold degenerate positions.
Position 3, is a three-fold degenerate position.
Position 4 and 6, are two-fold degenerate positions.
Position 9 could be i) four-fold degenerate (codon ACT) or ii) two fold-degenerate (codon GAT). DnaSP will no include that position neither for two-fold degenerate positions nor for four-fold degenerate positions.
Codons with missing information or alignment gaps are not considered.

Positions with Alignment Gaps option:

Excluded: These sites are removed.
Included: These sites are included.
Included if there is a polymorphism: These sites are included if there is a polymorphism.

Positions option:

Remove Non-Selected Positions: Non-Selected positions will be definitively removed from the active data.
Generate a NEXUS File with selected: Selected positions will be included in a NEXUS data file. The active data file will maintain all the positions.

Define Sequence Sets

Define Sequence Sets

This command allows you to define sequence sets (groups of sequences). A sequence set is a group of related sequences that could represent, for example, a population, a species of an outgroup. That allows conducting analyses on a specific group of sequences.

Sequence sets assignations can be stored in NEXUS data files; for that use the save/export or update commands.

Include / Exclude Sequences

Include / Exclude Sequences

DnaSP allows you the analysis in a subset of sequences of the original data file. This command allows you to include (or exclude) sequences from the analysis. All analyses will be performed with the information of only the included sequences. Consequently, if you use the Save/Export Data As command, the saved/exported data file will not contain excluded sequences.
Note: DnaSP also allows you the analysis in a subset of sequences by using the Define Sequence Sets command.

Options

There are two options that deal with alignment gaps. Suppose the following original data file:

```
             10         20         30
              *          *          *
Seq1 ATCTCTTAGGGTCGATTTGTTGTTTGTATTTAAT
Seq2 AT-TCTTATTTTCGA-TTGTTGTTTGTATTTAAT
Seq3 ATCGCTTA---TCGATTTGT----TGTATTTAAT
Seq4 ATCTCTTA---TCGATTTGTTGTTTGTATTTAAT
Seq5 ATCTCTTA---TCGATTTGTTGTTTGTATTTAAT
```

DnaSP will not use any site with alignment gaps or missing data. Thus, if you are using the complete data file, DnaSP will not use sites 3, 9, 10, 11, 16, 21, 22, 23, 24 for further analysis.

If you exclude 2 sequences (for example, Seq2 and Seq4) from the previous original data file, the active data will be composed of:

```
             10         20         30
              *          *          *
Seq1 ATCTCTTAGGGTCGATTTGTTGTTTGTATTTAAT
Seq3 ATCGCTTA---TCGATTTGT----TGTATTTAAT
Seq5 ATCTCTTA---TCGATTTGTTGTTTGTATTTAAT
```

With the Sites with alignment gaps are excluded if they are present in the active subset option (default option), DnaSP will not use information of sites 9, 10, 11, 21, 22, 23, 24.

With the option Sites with alignment gaps in the original data file are excluded in all subsets DnaSP will not use information of sites 3, 9, 10, 11, 16, 21, 22, 23, 24 (i.e., all sites with alignment gaps in the original data file). This option is appropriate to analyze exactly the same sites in different subsets of sequences.

Note

Both options generate the same estimates of the nucleotide distance (see the Linkage Disequilibrium command).

Analysis Menu

Polymorphic Sites

Polymorphic Sites

See Also: Input Data Files Output

This command displays some general information about the polymorphisms on the data file: the number of sites with alignment gaps (or missing data), the number of monomorphic sites, the number of polymorphic sites segregating for two, three, or four nucleotides. DnaSP also indicates the total number of parsimony-informative sites (sites that have a minimum of two nucleotides that are present at least twice), and non-informative sites (singleton sites).

This command also displays information about the genetic code used for these data, and the regions that are protein coding and noncoding (if this information was included in the NEXUS file, or has been defined using the Assign Coding Regions command in the Coding Region Menu). In this case, for the coding region, DnaSP also displays the number of synonymous and nonsynonymous (replacement) substitutions (see how DnaSP estimates the number of Synonymous and Nonsynonymous changes in a codon)

Estimating Synonymous & Nonsynonymous Changes

Number of Synonymous and Nonsynonymous Changes

How DnaSP estimates Synonymous and Nonsynonymous changes in a codon:
In general DnaSP uses a conservative criterion to decide if a particular change in a nucleotide site is synonymous or nonsynonymous (replacement); see the following examples. Nevertheless, the user should check the complex cases (those triplets of sites segregating for several codons; i.e. in highly variable regions).

Example using the Nuclear Universal Genetic Code

```
  3   6   9  12  15  18  21  24  27
  *   *   *   *   *   *   *   *   *
AGT TCT ATT CCC AAT ATA AGT UAU UAU
AGC TCT ATT CCC AGG TTA AGT UAU UAU
AGA TCT CTG CAG ACT TTG AGA CUG CUG
AGG TCT CTG CAG ACT ATG AGA CUG CUG
```

Codon (1,2,3):
3 mutations in site#3: 1 replacement, 2 synonymous.

Codon (4,5,6):
Monomorphic.

Codon (7,8,9):
Site#7 is replacement; Site#9 is synonymous.
When there are two possible evolutionary paths:
Path#1: ATT (Ile) -> CTT (Leu) -> CTG (Leu) Site#7 Replacement; Site#9 Synonymous
Path#2: ATT (Ile) -> ATG (Met) -> CTG (Leu) Site#7 Replacement; Site#9 Replacement

DnaSP will choose path#1, the path that requires the minor number of replacements.

Codon (13,14,15):
Site#14 (2 replacements); Site#15 is Synonymous.
Here there are four possible paths:
Path#1: ACT (Thr) -> AAT (Asp) -> AGT (Ser) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#2: ACT (Thr) -> AAT (Asp) -> AAG (Lys) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#3: AAT (Asn) -> ACT (Thr) -> AGT (Ser) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#4: AAT (Asn) -> ACT (Thr) -> ACG (Thr) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Synonymous).
DnaSP will choose path#4, the path that requires the minor number of replacements.

Codon (16,17,18):
Site#16 (1 replacement); Site#18 (1 synonymous).
Here there is a circular path:
ATA (Ile) -> TTA (Leu)
¡ !
ATG (Met) <- TTG (Leu)
Let us suppose that the number of mutations were only two (one in site 16, and another in site 18), DnaSP must assume one recombination event, the recombination event that requires the lower number of replacement substitutions:
| TTG (Leu)
TTA (Leu) ->| recomb: ATG (Met)
| ATA (Ile)
Note: This kind of codons will be analyzed only for Nuclear Genetic Codes.

Codon (19,20,21):
1 replacement (site#21).

Codon (22,23,24):
There are 3 changes among codons. So that there are 6 putative evolutionary paths (in this particular example there are only 4 because we exclude paths that go through stop codons). DnaSP will choose randomly between:
2 replacements (Site#22 and Site 23), and 1 synonymous (Site#24) and
2 replacements (Site#23 and Site 24), and 1 synonymous (Site#22).

Codons not analyzed:
DnaSP does not estimate synonymous and replacement changes in some complex cases (ambiguous/complex codons; those sites segregating for several codons; i.e. in highly variable regions). The user should do manually.
DnaSP does not estimate synonymous and replacement changes in codons with alignment gaps.

NOTE: Estimates of the number of synonymous and nonsynonymous substitutions might be different than the number of the synonymous and nonsynonymous differences (see the Synonymous and Nonsynonymous Substitutions module).

DNA Polymorphism

DNA Polymorphism

See Also: Coalescent Simulations Graphs Window Input Data Files Output
References: Hutter et al. 2006 Jukes and Cantor 1969 Lynch and Crease 1990 Nei 1987 Nei and Miller 1990 Tajima 1983 Tajima 1989 Tajima 1993 Tajima 1996 Watterson 1975

This command computes several measures of the extent of DNA polymorphism and their variances.

Alignment gaps and missing data:

Sites with alignment gaps (or missing data) are not used (these sites are completely excluded).

Analysis:

DnaSP computes the following measures:

- Haplotype (gene) diversity and its sampling variance (Nei 1987, equations 8.4 and 8.12 but replacing 2n by n). The standard deviation (or standard error) is the square root of the variance.
- Nucleotide diversity, Pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987, equations 10.5 or 10.6; see also Nei and Miller 1990), and its sampling variance (Nei 1987, equation 10.7). The standard deviation (or standard error) is the square root of the variance.
- Nucleotide diversity (Jukes and Cantor), Pi (JC), the average number of nucleotide substitutions per site between two sequences (Lynch and Crease 1990, equations 1-2). Unlike the previous estimates (Nei 1987, equations 10.5 or 10.6), this one has been obtained using the Jukes and Cantor (1969) correction. The correction has been performed in each pairwise comparison; the Pi (π) estimates were obtained as the average of the values for all comparisons. Note that DnaSP does not use the simplification indicated in Nei and Miller 1990 (equation 25); i. e. to perform the Jukes and Cantor (1969) correction directly on Pi (π) (Nei 1987, equations 10.5). Nevertheless, for low levels of polymorphism both methods give similar estimates.
- Theta (per site) from Eta (η) or from S, i.e. the Watterson estimator (Watterson 1975, equation 1.4a, but on base pair basis; Nei 1987, equation 10.3). Theta (θ) = $4N\mu$ for an autosomal gene of a diploid organism (N and μ are the effective population size and the mutation rate per nucleotide site per generation, respectively), Eta (η) is the total number of mutations, and S is the number of segregating (polymorphic) sites. The variance of this estimator depends on the recombination between sites. The variances for no recombination and for free recombination are estimated from equations 4 and 8 of Tajima 1993, respectively. These variances are computed on a per nucleotide site basis:

  Variance(per nucleotide site) = Variance(per DNA sequence) / m*m

where m is the total number of nucleotides studied. The standard deviation (or standard error) is the square root of the variance.
Note: for no recombination, estimates of the variance of theta can be different from those obtained from equation 10.2 of Nei 1987 (see Tajima 1989 equations 33 and 34, Tajima 1993 equations 4 and 8).

- Finite Sites Model (four possible nucleotides per site). The total number of mutations Eta (η) (Fu and Li 1993) also referred as the minimum number of mutations (Tajima 1996). Estimates of theta (θ) per site. $\theta = 4N\mu$ for an autosomal gene of a diploid organism (N and m are the effective population size and the mutation rate per nucleotide site per generation, respectively).
  Theta (θ) per site from Pi(π) (Tajima 1996, equation 9)
  Theta (θ) per site from S (Tajima 1996, equation 10)
  Theta (θ) per site from Eta (η) (Tajima 1996, equation 16)
- The average number of nucleotide differences, k (Tajima 1983, equation A3).

Stochastic variance of k (no recombination), Vst(k) (Tajima 1993, equation 14).
Sampling variance of k (no recombination), Vs(k) (Tajima 1993, equation 15).
Total variance of k (no recombination), V(k) (Tajima 1993, equation 13).
Stochastic variance of k (free recombination), Vst(k) (Tajima 1993, equation 17).
Sampling variance of k (free recombination), Vs(k) (Tajima 1993, equation 18).
Total variance of k (free recombination), V(k) (Tajima 1993, equation 16).

- Theta (per DNA sequence) from S (Watterson estimator). Theta (θ) = 4Nu for an autosomal gene of a diploid organism (N and u are the effective population size and the mutation rate per DNA sequence per generation, respectively) (Tajima 1993, equation 3).

Variance of θ (no recombination) (Tajima 1993, equation 4).
Variance of θ (free recombination) (Tajima 1993, equation 8).

Note: Tajima (1993) uses M to indicate θ (per DNA sequence), and v to indicate the mutation rate per DNA sequence per generation.

Effective Population size

Sliding window option:

This option allows you to calculate the nucleotide diversity, theta (per site), and S (the number of segregating sites), by the Sliding Window method.
The output of the sliding window analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph the nucleotide diversity, theta or S (Y axis) is plotted against the nucleotide position (X axis).

Nucleotide diversity (gaps/missing data) option:

- Pairwise comparisons: The average number of nucleotide differences, k (Tajima 1983, equation A3), and nucleotide diversity Pi (π) (Nei 1987, equations 10.5 or 10.6) are calculated by the Pairwise-Deletion option (DnaSP will not compute their variances). Using this option, only those gaps/missing present in a particular pairwise comparison are ignored. Pairwise sequence comparisons with 0 sites (after excluding the gaps) are also ignored.
- Individual Sites (column by column): The average number of nucleotide differences k, and nucleotide diversity π, are calculated as described in Hutter et al. 2006 (equation I and II). The same criteria is applied to obtain θ.

Statistical significance by the coalescent:

DnaSP can provide the confidence intervals of the number of haplotypes, the haplotype diversity and the nucleotide diversity by computer simulations using the coalescent algorithm (see Coalescent Simulations).

Note:

n.a. , not applicable. When the proportion of differences is equal or higher than 0.75, the Jukes and Cantor correction can not be computed.

Effective Population Size

Effective Population Size

The mutation parameter θ (theta) is defined as $4N\mu$ for autosomal loci of diploid organisms, where N is the effective population size (diploid individuals) and m is the neutral mutation rate (per gene or per base pair) per generation.
Assuming equal population sizes of males and females, the parameter θ is $3N\mu$ for X-linked (or Z-linked) loci of diploid organisms. In the same way, the parameter θ is Nm for Y-linked (or W-linked) loci of diploid organisms. In both cases, N is the effective population size considering both males and females (diploid individuals). For Y-linked loci the parameter θ would be $2Nm\mu$, where Nm is the male effective population

size. For mitochondrial DNA (or haploid individuals) θ is $2N\mu$, where N is the effective population size of females.

Likewise, the recombination parameter C (or R) is 4Nc for autosomal loci of diploid organisms, where N is the effective population size and c is the is the recombination rate per generation. C = 3Nc and C = Nc for X-linked and Y-linked loci, respectively.

InDel (Insertion-Deletion) Polymorphism

InDel (Insertion-Deletion) Polymorphism

See Also: DNA Polymorphism

This module allows estimating several measures of the level of Insertion/Deletion (InDel) polymorphism (DIPs). In particular, DnaSP will infer the number of InDel events from the data.
Let me suppose the following example data file (13 sequences with 18 positions each).

```
          *    *    *
Seq1   AAAAAAGGGGGGGGGGG
Seq2   ..................
Seq3   ...C..--..........
Seq4   ..................
Seq5   ..---.............
Seq6   .........-------..
Seq7   ..---....-------..
Seq8   .........-------..
Seq9   .C............---..
Seq10  .........-------..
Seq11  .........-------..
Seq12  ..................
Seq13  ..................
```

In this data file we can identify 4 InDel events:
Event#1 (Seq5 and Seq7); InDel length = 3 nucleotides.
Event#2 (Seq3); InDel length = 2 nucleotides.
Event#3 (Seq6, Seq7, Seq8, Seq10 and Seq11); InDel length = 7 nucleotides.
Event#4 (Seq9); InDel length = 3 nucleotides.

Option#1: Diallelic

Only InDel diallelic states (gap event/not gap) will be considered. That is, positions 10-16 will be excluded from the analysis since InDel event#3 and event#4 overlap at positions 14-16.

Output
- Total number of InDel events analysed: 2 (event#1 and event#2)
- Average InDel length per event: 2.5 (the average length of event#1 and event#2)
- Average deletion length: 2.667 (2 sequences with 3 nucleotides deleted plus 1 sequence with 2 deleted nucleotides, divided by 3 -the number of analysed sequences with gaps-).

DnaSP also computes:
- The number of InDel Haplotypes: 3
- InDel Haplotype Diversity: 0.410

- InDel Diversity, k(i): 0.436 (this is the analogue of k, the average number of nuc. differences)
- InDel Diversity per site, Pi(i): 0.03963 (this is the analogue of Pi, the nucleotide diversity). Pi(i) is computed as k(i)/m, where m is the net number of positions analysed, 11 (18 minus the 7 positions with overlapping InDels)
- Theta (per sequence) from the number of InDel events: 0.644
- Tajima's D: -0.9092

Additionally, DnaSP allows generating a NEXUS file with ONLY InDel events information. The data file will be recoded as:

```
Seq1  AA
Seq2  ..
Seq3  .G
Seq4  ..
Seq5  G.
Seq6  ..
Seq7  G.
Seq8  ..
Seq9  ..
Seq10 ..
Seq11 ..
Seq12 ..
Seq13 ..
```

where, A and G represents the two InDel states (no InDel/InDel).

Option#2: Triallelic

Only Diallelic and Triallelic InDel states will be considered. In the example, all positions will be used.

- Total number of InDel events analysed: 4
- Average InDel length per event: 3.75
- Average deletion length: 5.111
- Number of InDel haplotypes: 6
- …

DnaSP will generate the following recoded NEXUS file:

```
Seq1  AAAA
Seq2  ....
Seq3  .G..
Seq4  ....
Seq5  G...
Seq6  ..G.
Seq7  G.G.
Seq8  ..G.
Seq9  ...G
Seq10 ..G.
Seq11 ..G.
Seq12 ....
Seq13 ....
```

Option#3: Tetrallelic

Only Diallelic, Triallelic and Tetrallelic InDel states will be considered.

Option#4: Multiallelic

All InDel events will be considered.

Option#5: "As Is"

DnaSP will no infer events from InDel information. DnaSP will generate the following recoded NEXUS file:

```
          *    *
Seq1   AAAAAAAAAAA
Seq2   ...........
Seq3   ...GG......
Seq4   ...........
Seq5   GGG........
Seq6   .....GGGGGGG
Seq7   GGG..GGGGGGG
Seq8   .....GGGGGGG
Seq9   .........GGG
Seq10  .....GGGGGGG
Seq11  .....GGGGGGG
Seq12  ...........
Seq13  ...........
```

Note:

Throughout this module, nucleotide substitution polymorphism is not considered; either in non-InDel sites (such as the nucleotide polymorphism at site#2 in the example data file), or in InDel positions (such as the nucleotide polymorphism at site#4).

DNA Divergence Between Populations

DNA Divergence Between Populations

See Also: Gene Flow and Genetic Differentiation Graphs Window Input Data Files Output
References: Hey 1991 Jukes and Cantor 1969 Nei 1987 Tajima 1983 Wakeley and Hey 1997

This command computes some measures of the extent of DNA divergence between populations taking into account the effect of the DNA polymorphism.

Data Files:

For the present analysis, at least two sets of sequences (one for each population) must be defined (see: Data | Define Sequence Sets command).

Alignment gaps and missing data:

Sites containing alignment gaps (or sites with missing data) in any population are not used (these sites are completely excluded).

Analysis:

The program estimates the following measures:

For each individual population:

- The average number of nucleotide differences (Tajima 1983, equation A3).

- The nucleotide diversity, Pi (Nei 1987 equation 10.5).
- Nucleotide diversity with Jukes and Cantor, Pi (JC) (Nei 1987, equations 10.19 and 5.3; Lynch and Crease 1990, equations 1-2).

Variance of Pi(JC) (Nei 1987, equation 10.7). The standard deviation (or standard error) is the square root of the variance.
These estimates may be different from those obtained by the DNA Polymorphism command. This is because in the present analysis all sites with alignment gaps in population 1 or in population 2 are not considered. That is, the total number of analyzed sites considered in this command can be equal or lower than those taken into account in the DNA polymorphism command.

For the total data:

- The average number of nucleotide differences (Tajima 1983, equation A3).
- The nucleotide diversity, Pi(total) (Nei 1987, equation 10.5).

Between populations:

- #The number of fixed differences between populations, nucleotide sites at which all of the sequences in one population are different from all of the sequences in the second population (Hey 1991).
- #Mutations that are polymorphic in population 1, but monomorphic in population 2.
- #Mutations that are polymorphic in population 2, but monomorphic in population 1.
- #The total number of shared mutations.
- The average number of nucleotide differences between populations.
- The average number of nucleotide substitutions per site between populations, Dxy (Nei 1987, equation 10.20).
- Dxy with Jukes and Cantor (Nei 1987, equation 10.20 using the Jukes and Cantor correction).
- The number of net nucleotide substitutions per site between populations, Da (Nei 1987, equation 10.21).
- Da with Jukes and Cantor (Nei 1987, equation 10.21 using the Jukes and Cantor correction).

Variance of Dxy(JC) (Nei 1987, equation 10.24). The standard deviation (or standard error) is the square root of the variance.
Variance of Da(JC) (Nei 1987, equation 10.23). The standard deviation (or standard error) is the square root of the variance.
#, This information can be used to estimate the 4 parameters (thetaA, theta1, theta2 and tau) that describes the isolation model (see Wakeley and Hey 1997, equations 1-3).

Sliding window option:

This option computes the nucleotide diversity for populations 1 and 2, Dxy, and Da by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph the nucleotide diversity, Dxy, or Da (Y axis) can be plotted against the nucleotide position (X axis).

DNA Divergence among Populations:

You can perform some analyses of the DNA divergence among populations by using the Gene Flow and Genetic Differentiation command.

Conserved DNA Regions

Conserved DNA Regions

See Also: DNA Polymorphism
References: Vingron et al., 2009

This command identifies conserved DNA regions along the data set, and might be useful for phylogenetic footprinting-based analyses.

The "Minimum Window Length" (MWL) and "Conservation Threshold" (CT) parameters are, respectively, the minimum length and the minimum conservation value required to identify conserved regions. Here, the conservation (C) is measured as the proportion of conserved sites in the alignment region.

Alignment gaps and missing data:

Positions of the alignment containing gaps or missing data in more than half of the sequences are not considered.

1. Dynamic Defined Parameters:

DnaSP will estimate MWL and CT parameters from your data according to the current levels of nucleotide variation.
DnaSP will calculate C, the average conservation of the data, from:

- Observed number of polymorphic/variable sites (S) in the data.
- The number of polymorphic sites (S) estimated from nucleotide diversity (assuming mutation-drift equilibrium). It is estimated from:

  $S = k \cdot a1;$

  where k is the average number of nucleotide differences (Tajima 1983, eq. A3),
  $a1 = \sum(1 / i)$ from i=1 to n-1,
  and n is the sample size.

Therefore the proportion of invariable/monomorphic sites (C) is,

$C = 1 - (S / L);$

where L is the net number of analyzed positions.

From C estimates, DnaSP will fix the CT parameter to C+0.1. DnaSP will estimate the MWL as the minimum length that allows a conserved region to be statistically significant at $\alpha = 0.05$.

2. User Defined Parameters:

The user can define both the MWL and the CT parameters.

P-value:

The p-value is computed assuming that the number of variable positions in the alignment region follows an hypergeometric distribution.

Example of How DnaSP estimates MWL and CT by the Dynamic Option:

Let's suppose the following example:

Number of sequences, n: 10
Net number of analyzed positions, L: 1000
Number of segregating sites, S: 200
Average number of nucleotide differences, k: 80

Using the dynamic parameter estimation given the nucleotide diversity, the estimate of S (from k) will be:

$S = k \cdot a1;$
$S = 80 \cdot 2.829;\ S = 226.32$

DnaSP will fix S to 226.

From S (observed or estimated), it is straightforward to compute C. Using the observed S value (S=200):
C = 1 - (S / L) = 0.8

DnaSP will fix the Conservation Threshold (CT) to:
CT = C + 0.1 = 0.9

With CT fixed to 0.9, DnaSP will estimate MWL at p-value < 0.05.

In the example,

MWL = 30 and CT = 0.9 (S=3); p-value = 0.1190
...
MWL = 49 and CT = 0.9 (S=5); p-value = 0.0507
MWL = 50 and CT = 0.9 (S=5); p-value = 0.0440
...
MWL = 60 and CT = 0.9 (S=6); p-value = 0.0272

Therefore, the Minimum Window Length (MWL) will be set to 50.

Output:

For each conserved DNA region, DnaSP reports:

- C, Conservation Index (proportion of conserved columns).
- H, Homozigosity (1-Heterozigosity).
- P-value (under the hypergeometric distribution).
- The conserved DNA sequence using IUPAC ambiguity codes to represent variable nucleotides.

Polymorphism and Divergence

Polymorphism and Divergence

See Also: Graphs Window Input Data Files Output
References: Jukes and Cantor 1969 Lynch and Crease 1990 Nei 1987 Nei and Gojobori 1986 Nei and Miller 1990 Watterson 1975

This command computes some measures of the extent of DNA polymorphism and divergence in synonymous, nonsynonymous, silent and in all sites.

Data Files:

For the present analysis, at least one set of sequences must be defined (see: Data | Define Sequence Sets command).
Analysis using one sequence set. The sequence set must include intraspecific data information. DnaSP will estimate some measures of the extent of DNA polymorphism.
Analysis using two sequence sets. One Sequence Set must contain the intraspecific data, while the other must contain sequences (one or more) from a different species (or from a different population). DnaSP will estimate some measures of the extent of DNA polymorphism and of divergence.

Alignment gaps and missing data:

Sites (or codons) with alignment gaps or missing data in any data file are not used, i.e. these sites (or codons) are completely excluded.

Output:

The program estimates the following measures:

From the intraspecific data set:

- Nucleotide diversity Pi (π) (Nei 1987 equation 10.5).

Nucleotide diversity with Jukes and Cantor correction, Pi (JC) (Lynch and Crease 1990, equations 1-2).

- Theta (per site) from Eta (η), the total number of mutations (Watterson 1975, equation 1.4a, but on base pair basis; Nei 1987, equation 10.3). Theta values will not be reported in some cases where codons might differ by multiple changes (this feature will indicated by n.a.).

From both data sets:

- Nucleotide divergence, (average proportion of nucleotide differences between populations or species), K (or Dxy) (Nei 1987, equation 10.20).
- K(JC), average number of nucleotide substitutions per site between populations or between species with Jukes and Cantor correction (Between populations, Dxy; Nei 1987, equation 10.20), (Between species, K; Nei 1987, equation 5.3, but computing as the average of all comparisons between sequences of data set 1 and 2).

Estimation of nucleotide diversity and divergence separately for synonymous and nonsynonymous sites is performed using Nei and Gojobori (1986), equations 1-3.

Implementation:

Estimation of nucleotide diversity (and of divergence) by the Jukes and Cantor (1969) correction is performed using the simplification indicated in Nei and Miller 1990 (equation 25). That is, the correction of Pi (and of K) is performed directly on the uncorrected value, and not in each pairwise comparison of two sequences. Nevertheless, for low levels of polymorphism (and of divergence) both methods give similar estimates. For high polymorphism and divergence levels the use of the DNA Polymorphism and Synonymous and Nonsynonymous Substitutions commands might be desirable.

The total number of synonymous and nonsynonymous sites for a set of sequences is estimated as the average of the number of synonymous and nonsynonymous sites of all sequences; these values are used for all sequences. Note than in the Synonymous and Nonsynonymous Substitutions command, the total number of synonymous and nonsynonymous sites is performed in every pairwise comparison. So that, nucleotide diversity estimates (in synonymous, nonsynonymous, and silent sites) based on the present and on the Synonymous and Nonsynonymous Substitutions command could be slightly different.

Sites Considered:

Silent (synonymous sites and noncoding positions): Only silent (both synonymous sites and noncoding positions) are used.

Noncoding Positions: Only noncoding positions are used.

Only synonymous sites: Only synonymous sites are used (substitutions in the coding region that cause no amino acid changes). This option works only if the data file contains sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).

Only Nonsynonymous sites: Only nonsynonymous sites are used (substitutions in the coding region that cause amino acid changes). This option works only if the data file contains sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).

Pi(a)/Pi(s) and Ka/Ks ratios: DnaSP will compute $_{W}$ ratios ($_{W}$ = Ka/Ks; also known as $_{W}$ = dN/dS) for the intraspecific and interspecific (if available) data sets. This option works only if the data file contains sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).

All sites: All sites are used (excluding substitutions in sites with gaps or missing data).

Note: See how DnaSP estimates the number of Synonymous and Nonsynonymous changes in a codon.

Sliding window option:

This option computes the nucleotide diversity (intraspecific data file) and divergence (between both data files) by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph Pi, the nucleotide diversity, and K, divergence, (Y axis) can be plotted against the nucleotide position (X axis).

Abbreviations:
n.a. not available.

Polymorphism and Divergence in Functional Regions

Polymorphism and Divergence in Functional Regions

See Also: Input Data Files Output
References: Jukes and Cantor 1969 Lynch and Crease 1990 Nei 1987 Nei and Gojobori 1986 Nei and Miller 1990 Watterson 1975

This command computes some measures of the extent of DNA polymorphism and divergence in synonymous, nonsynonymous, silent and in all sites. Unlike the Polymorphism and Divergence command, this command provides estimates of nucleotide diversity, divergence, and the number of mutations in functional regions; i.e. separately for noncoding regions, exons, introns, etc.

Data Files:
For the present analysis, at least one set of sequences must be defined (see: Data | Define Sequence Sets command).
Analysis using one sequence set. The sequence set must include intraspecific data information. DnaSP will estimate some measures of the extent of DNA polymorphism.
Analysis using two sequence sets. One Sequence Set must contain the intraspecific data, while the other must contain sequences (one or more) from a different species (or from a different population). DnaSP will estimate some measures of the extent of DNA polymorphism and of divergence.

Alignment gaps and missing data:
Sites (or codons) with alignment gaps or missing data in any data file are not used, i.e. these sites (or codons) are completely excluded.

Output:
The program estimates the following measures:
From the intraspecific data file:

- Nucleotide diversity Pi (π) (Nei 1987 equation 10.5).

Nucleotide diversity with Jukes and Cantor correction, Pi (JC) (Lynch and Crease 1990, equations 1-2).

- Theta (per site) from Eta (η), the total number of mutations (Watterson 1975, equation 1.4a, but on base pair basis; Nei 1987, equation 10.3). Theta values will not be reported in some cases where codons might differ by multiple changes (this feature will indicated by n.a.).

From both data files:

- Nucleotide divergence, (average proportion of nucleotide differences between populations or species), K (or Dxy) (Nei 1987, equation 10.20).
- K(JC), average number of nucleotide substitutions per site between populations or between species with Jukes and Cantor correction (Between populations, Dxy; Nei 1987, equation 10.20), (Between species, K; Nei 1987, equation 5.3, but computing as the average of all comparisons between sequences of data file 1 and 2).

Estimation of nucleotide diversity and divergence separately for synonymous and nonsynonymous sites is performed using Nei and Gojobori (1986), equations 1-3.

Implementation:
Estimation of nucleotide diversity (and of divergence) by the Jukes and Cantor (1969) correction is

performed using the simplification indicated in Nei and Miller 1990 (equation 25). That is, the correction of Pi (and of K) is performed directly on the uncorrected value, and not in each pairwise comparison of two sequences. Nevertheless, for low levels of polymorphism (and of divergence) both methods give similar estimates. For high polymorphism and divergence levels the use of the DNA Polymorphism and Synonymous and Nonsynonymous Substitutions commands might be desirable.
The total number of synonymous and nonsynonymous sites for a set of sequences is estimated as the average of the number of synonymous and nonsynonymous sites of all sequences; these values are used for all sequences. Note than in the Synonymous and Nonsynonymous Substitutions command, the total number of synonymous and nonsynonymous sites is performed in every pairwise comparison. So that, nucleotide diversity estimates (in synonymous, nonsynonymous, and silent sites) based on the present and on the Synonymous and Nonsynonymous Substitutions command could be slightly different.

Sites Considered:
Silent (Synonymous and Noncoding): The analysis is limited to both synonymous sites and noncoding positions.
Only Synonymous Sites: The analysis is restricted to synonymous sites.
Only Nonsynonymous Sites: The analysis is restricted to nonsynonymous sites.
All (total) sites: All sites will be used (excluding those sites with gaps or missing data).
The synonymous and nonsynonymous sites (and changes) will be computed if the data file contains sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).
Note: See how DnaSP estimates the number of Synonymous and Nonsynonymous changes in a codon.

Sites:
Silent (Synonymous and Noncoding): Indicates both synonymous sites (in coding region) and noncoding positions.
Synonymous Sites: Indicates sites in the coding region where all mutations result in synonymous substitutions (no amino acid changes).
Nonsynonymous Sites: Indicates sites in the coding region where all mutations cause amino acid changes. The analysis is restricted to nonsynonymous sites.

Abbreviations:
Tot, Total, analysis in total (all) sites.
Sil, analysis in silent (synonymous and noncoding) sites.
Syn, analysis in synonymous (coding region only) sites.
NoSyn, analysis in nonsynonymous sites.
SilSites, the total number of silent sites.
NSynSites, the total number of nonsynonymous sites.
SilMut, the total number of silent mutations (intraspecific data file).
NSynMut, the total number of nonsynonymous mutations (intraspecific data file).
n.a. not available.

Synonymous and Nonsynonymous Substitutions

Synonymous and Nonsynonymous Substitutions

See Also: Assign Coding Regions Input Data Files Output
References: Jukes and Cantor 1969 Lynch and Crease 1990 Nei 1987 Nei and Gojobori 1986 Nei and Miller 1990 Osawa et al. 1992 Watterson 1975

This command estimates Ka (the number of nonsynonymous substitutions per nonsynonymous site; also

denoted as d_N), and Ks (the number of synonymous substitutions per synonymous site; also denoted as d_S) for any pair of sequences (Nei and Gojobori 1986, equations 1-3); it also computes several measures of the extent of DNA polymorphism in protein coding regions, noncoding regions, or in regions with both protein coding and noncoding regions (i.e. regions with both exons and introns, or exons and flanking regions). One interesting feature of DnaSP is that both coding and noncoding protein regions can be included in the data file; DnaSP can thus estimate the nucleotide diversity for synonymous, nonsynonymous and silent (both synonymous and noncoding positions) sites. Four pre-defined genetic codes can be used: the universal nuclear code, and the mitochondrial code of Drosophila, mammals and yeast.

Alignment gaps and missing data:

Sites (or codons) with alignment gaps or missing data are not used, i.e. these sites (or codons) are completely excluded.

Implementation:

DnaSP can compute the nucleotide diversity in synonymous, nonsynonymous, and silent sites. The total number of synonymous and nonsynonymous sites is computed as Nei and Gojobori 1986. By silent sites we refer both to the synonymous sites and the noncoding positions. Synonymous sites are those sites in a codon where nucleotide changes result in synonymous substitutions. For computing synonymous and nonsynonymous sites, DnaSP will exclude all pathways that go through stop codons.
No stop codons should be found in the middle of coding regions, however, if DnaSP finds stop codons (in the middle of coding regions) they will be considered as if they would code for a new amino acid (the amino acid 21; for example Selenocysteine, Secys (Osawa et al. 1992).
DnaSP computes the synonymous and nonsynonymous differences between a pair of sequences as Nei and Gojobori 1986. When there are two or three nucleotide differences between the two codons compared, two or six putative pathways exit. DnaSP considers all pathways with equal probability, but it excludes those pathways that go through stop codons. Obviously, all nucleotide differences in noncoding positions are considered silent. Silent differences will include, therefore, both the synonymous differences (in coding regions) and all differences in noncoding positions.

Silent Substitutions Considered:

Substitutions in Coding Regions: Only synonymous substitutions (coding region) will be considered.
In Coding and Noncoding Regions: All silent substitutions will be considered (synonymous substitutions and changes in noncoding positions). If the data file does not contain assigned coding regions all sites will be considered as noncoding positions; i.e. all substitutions will be considered as silent.

Analysis:

- The average number of nucleotide differences per site between two sequences, or nucleotide diversity, Pi (π) (Nei 1987, equations 10.5 or 10.6).
- The average number of nucleotide substitutions per site between two sequences or nucleotide diversity, Pi (π), using the Jukes and Cantor (1969) correction (Lynch and Crease 1990, equations 1-2). The correction has been performed in each pairwise comparison of two sequences (Nei and Gojobori 1986, equations 1-3); the Pi (π) estimates was obtained as the average of the values of all comparisons (of Ks and Ka values); (see also the DNA Polymorphism command).

Note that DnaSP has not used the simplification indicated in Nei and Miller 1990 (equation 25); i. e. to perform the Jukes and Cantor (1969) correction directly on Pi (π) (Nei 1987, equations 10.5). Nevertheless, for low levels of polymorphism similar estimates are given by both methods.

- Theta values (per site) from Eta (η), i.e. the Watterson estimator (Watterson 1975, equation 1.4a, but on base pair basis; Nei 1987, equation 10.3). See how DnaSP estimates Synonymous and Nonsynonymous changes in a codon. Note that the number of mutations might be different than the number of Synonymous and Nonsynonymous differences obtained in each pairwise comparison (see below). Theta values will not be reported in some cases where codons might differ by multiple changes.

The DnaSP output shows also the following:

For each sequence:
- The total number of Synonymous (SS) and Nonsynonymous (NSS) sites.

For each pair of sequences:
- The total number of Synonymous, Nonsynonymous and silent sites,
- The total number of Synonymous, Nonsynonymous and silent differences,
- The estimates of Ka (the number of nonsynonymous substitutions per nonsynonymous site),

and Ks (the number of synonymous -or silent- substitutions per synonymous -or silent- site) (Nei and Gojobori 1986, equations 1-3).

Assign codons:
To assign noncoding and coding protein regions in a particular DNA sequence you should use the Assign Coding Regions command.

Genetic Code:
To compute synonymous and nonsynonymous substitutions DnaSP will use the defined Assign Genetic Code assigned (the default is the Nuclear Universal).

Notes and abbreviations:
n.a. , not applicable. When the proportion of differences is equal or higher than 0.75, the Jukes and Cantor correction can not be computed.
Seq 1 and Seq 2, the two sequences compared.
SynDif, the total number of synonymous differences.
SynPos, the total number of synonymous sites.
SilentDif, the total number of silent differences.
SilentPos, the total number of silent sites.
Ks, the number of synonymous (or silent) substitutions per synonymous (or silent) site.
NSynDif, the total number of nonsynonymous differences.
NSynPos, the total number of nonsynonymous sites.
Ka, the number of nonsynonymous substitutions per nonsynonymous site.

Codon Usage Bias

Codon Usage Bias

See Also: Input Data Files Output Window Menu
References: Morton 1993 Sharp et al. 1986 Shields et al. 1988 Wright 1990

This command computes some measures of the extent of the nonrandom usage of synonymous codons.

Data Files:
The present analysis requires only one data file. This command works only if the coding regions and the genetic code have been previously defined (more help in Assign Coding Regions and Assign Genetic Code).

Codon Bias Measures

RSCU, Relative Synonymous Codon Usage (Sharp et al. 1986)
For a given DNA sequence DnaSP shows the RSCU value at each codon (Codon Usage Table). The RSCU value of a codon is the observed frequency of that codon in the gene divided by that expected under the assumption of equal usage of synonymous codons. A RSCU value of 1 indicates that the frequency of that

codon is the expected for an equal codon usage; values less than 1 (or more than 1) indicates that the codons are used less often (or more often) than the expected.
Codon Usage Table
DnaSP shows, for a given codon, the observed frequency and its RSCU value (in parenthesis). For a given DNA sequence, the Codon Usage Table also shows the "Scaled" chi square value.

ENC, Effective Number of Codons (Wright 1990)
That measure quantifies the "effective" number of codons that are used in a gene. For the nuclear universal genetic code, the value of ENC ranges from 20 (only one codon is used for each amino acid; i.e., the codon bias is maximum) to 61 (all synonymous codons for each amino acid are equally used; i.e., no codon bias).

CBI, Codon Bias Index (Morton 1993)
CBI is a measure of the deviation from the equal use of synonymous codons. CBI values range from 0 (uniform use of synonymous codons) to 1 (maximum codon bias).

SChi2, Scaled Chi Square (Shields et al. 1988)
The "scaled" 2 (chi square) is a measure based on the chi square statistics; i.e., based on the difference between the observed number of codons and those expected from equal usage of codons. The sum of the chi square values is divided by the total number of codons in the gene excluding those codons coding for a unique amino acid; i.e. all codons excluding the Trp and Met codons (nuclear universal genetic code).
DnaSP can compute the "scaled" chi square with Yates' correction, and also assuming a given G+C content (by default the G+C content is 50%).

G+C content
G+Cn, G+C content at noncoding positions.
G+C2, G+C content at second coding positions.
G+C3s, G+C content at (synonymous) third coding positions; i.e. the G+C content in the third codon positions excluding the Trp and Met codons (nuclear universal genetic code) (Wright 1990).
G+Cc, G+C content at coding positions.
G+C, G+C content in the genomic (whole) region.

Preferred and Unpreferred Synonymous Substitutions

Preferred and Unpreferred Synonymous Substitutions

See Also: Codon Preference Table Input Data Files Output
References: Akashi 1995 Akashi 1999

This command determines the polarity status (ancestral -> derived) of the polymorphic (or fixed) substitutions, and it also estimates the number of preferred and unpreferred substitutions.

Data Files:
For the present analysis, at least two sets of sequences (one with the intraspecific data, and other with the outgroup sequences) must be defined (see: Define Sequence Sets command).

Alignment gaps and missing data:
Sites (or codons) with alignment gaps or missing data in any group of sequences (sequence sets) are not used, i.e. these sites (or codons) are completely excluded.

Analyze:

One Species with an Outgroup: Analysis of the polarity status (ancestral -> derived) of the polymorphic substitutions. The outgroup allows inferring that information.
One Species with two Outgroups: Analysis of the polarity status (ancestral -> derived) of the polymorphic substitutions (intraspecific Data) and also of the fixed differences (between the MRCA of the intraspecific data and the common ancestor of the close outgroup). The distant outgroup allows inferring that information.

Pref / Unpref Tables:

Use this command to assign the specific codon preference table to the data.
Options:
Non Coding Positions: This option allows analyzing the polarity of changes in noncoding positions.

Statistical significance:

DnaSP conducts the Mann-Whitney test to determine if the frequency distribution of preferred and unpreferred substitutions are significantly different. DnaSP can carry out the fdMWU test (that uses information of only polymorphism data; Akashi 1999), or the fddMWU test (that uses information of both polymorphic substitutions and fixed differences; Akashi 1999).

Ambiguous information:

In some cases, the polarity of some substitutions could not be unambiguously determined (see below). There are several sources of ambiguity (ancestral polymorphism; multiple substitutions; alignment gaps/missing data, etc). In that cases, DnaSP will list the ambiguous sites (or codons).

How DnaSP polarizes the nucleotide changes and assigns the preferred and unpreferred status (coding region):

DnaSP uses a conservative (parsimony) criterion to infer the ancestral nucleotide state: only unambiguous cases are used for the analysis (see the following examples). Once the polarity has been established, DnaSP will use the codon preference table to assign codons (or changes) as preferred or unpreferred.
Some examples using the Nuclear Universal Genetic Code with the D. melanogaster (Akashi 1995) codon preference table:

Intraspecific Data

```
  3   6   9  12  15  18  21  24  27  30  33
  *   *   *   *   *   *   *   *   *   *   *
CTT AAC CTT CTA AAT TTA CCC CTT CTT GGT GGT
CTT AAC CTT CTA AAT TTA CCA CTA CTA AGT GGT
CTA AAC CTA CTT AAC TTN CCT CTA CTT GGA GGT
```

Close Outgroup

```
CTT AAT GTC GTT A-T TTT CCT CTT CTT GGA AGT
```

Distant Outgroup

```
CTT AAT GTT GTT AAT TTT CCT CTG CTA GGA TGT
```

One Species with an Outgroup (Close Outgroup)
Codon (1,2,3). CTT -> CTA. Polymorphic synonymous change: U -> U
Codon (4,5,6). AAC. Monomorphic codon.
Codon (7,8,9). CTT <-> CTA. Ambiguous Change.
Codon (10,11,12). CTT -> CTA. Polymorphic synonymous change: U -> U
Codon (13,14,15). AAT -> AAC. Polymorphic synonymous change: U -> P
Codon (16,17,18). Not analyzed: Codon with missing data.
Codon (19,20,21). Not analyzed: Multiple substitutions.
Codon (22,23,24). CTT -> CTA. Polymorphic synonymous change: U -> U
Codon (25,26,27). CTT -> CTA. Polymorphic synonymous change: U -> U

Codon (28,29,30). Not analyzed: Multiple substitutions.
Codon (31,32,33). GGT. Monomorphic codon.

One Species with two Outgroups
Codon (1,2,3). CTT -> CTA. Polymorphic synonymous change: U -> U
Codon (4,5,6). AAT -> AAC. Fixed synonymous change: U -> P
Codon (7,8,9). There are two changes:
    CTT -> CTA. Polymorphic synonymous change: U -> U
    GTT -> CTT. Fixed nonsynonymous change: Val -> Leu
Codon (10,11,12). There are two changes:
    CTT -> CTA. Polymorphic synonymous change: U -> U
    GTT -> CTT. Fixed nonsynonymous change: Val -> Leu
Codon (13,14,15). Not analyzed: Codon with alignment gaps.
Codon (19,20,21). Not analyzed: Multiple substitutions.
Codon (22,23,24). CTT -> CTA. Polymorphic synonymous change: U -> U
Codon (25,26,27). CTT <-> CTA. Ambiguous polymorphic change (ancestral polymorphism ?).
Codon (28,29,30). Not analyzed: Multiple substitutions.
Codon (31,32,33). Not analyzed: Ambiguous fixed change.

Abbreviations:
MRCA, Most recent common ancestor
U -> U, Unpreferred to unpreferred change
U -> P, Unpreferred to preferred change
P -> U, Preferred to Unpreferred change
P -> P, Preferred to Preferred change
Syn, Synonymous change
NonSyn, Nonsynonymous change

Gene Conversion

Gene Conversion

See Also: Graphs Window Input Data Files Output
References: Betrán et al. 1997 Rozas and Aguadé 1994

DnaSP incorporates the algorithm developed by Betrán et al. (1997) to detect gene conversion tracts from two differentiated populations (referred to as subpopulations). These subpopulations could be, for example, two different chromosomal gene arrangements (Rozas and Aguadé 1994), or two sets of paralogous sequences.

Data Files:
For the present analysis, at least two sets of sequences (one for each population) must be defined (see: Define Sequence Sets command).

Minimum number of sequences in each set:
One sequence set must contain at least three sequences, and the other a minimum of five.

Alignment gaps and missing data:
Sites containing alignment gaps (or sites with missing data) in any population are not used (these sites are completely excluded).

Implementation:
DnaSP estimates the observed tract length in nucleotides as:

L = TR - TL +1 - G

where TL (left) and TR (right) are the site positions of the outermost informative nucleotide sites of a congruent tract, and G is the number of alignment gaps (if any) between TL and TR in the particular sequence where the gene conversion tract is detected (see Betrán et al. 1997 equation A1).
DnaSP also estimates the parameter ψ (Betrán et al. 1997, equation A4), which measures the probability per site of detecting a conversion event between two subpopulations. From this information it is possible to estimate the true number and length of the gene conversion tracts.

Sliding window option:
This option computes the parameter y by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph the parameter y (Y axis) can be plotted against the nucleotide position (X axis).

Gene Flow and Genetic Differentiation

Gene Flow and Genetic Differentiation

See Also: Define Sequence Sets (Define Populations) Input Data Files Output
References: Hudson et al. 1992a Hudson et al. 1992b Hudson 2000 Lynch and Crease 1990 Nei 1973 Nei 1982 Nei 1987 Tajima 1983 Wright 1951

This command computes some measures of the extent of DNA divergence among populations, and from these measures it computes the average level of gene flow. Additionally, DnaSP allows testing for population subdivision.

Data Files:
For the present analysis, at least two sets of sequences (one set for each population) must be previously defined (see: Define Sequence Sets command).

Missing data:
Sites containing missing data in any population are not used (these sites are completely excluded).

Include / Exclude Populations (set of sequences):
Use this command to include or exclude a particular population from the analysis. In any case, populations with an unique included sequence will not be used.

Sites with Alignment Gaps option:
1. Excluded: Sites with gaps (in any population) will be completely excluded from the analysis.
2. Considered (Gap as the fifth state): Gaps will be used. They will be considered as a different nucleotide variant.
3. Excluded only in pairwise comparisons: Using this option, gaps will be ignored only if they are present in a particular pairwise comparison. Note, this option does not work for estimating haplotype-based statistics; in that case, DnaSP will considered the gap as a fifth state.

Genetic Diversity Analysis:
The program estimates the following measures:

For each individual population:

- The number of haplotypes, h.
- The haplotype diversity, Hd (Nei 1987, equation 8.4).
- The average number of nucleotide differences, K (Tajima 1983, equation A3).
- The nucleotide diversity, Pi (π, Nei 1987, equation 10.5).
- Nucleotide diversity with the Jukes and Cantor correction, Pi(JC) (Lynch and Crease 1990, equations 1-2).

Present estimates may differ from those obtained by the DNA Polymorphism command. This is because in the present analysis all sites with alignment gaps (in any population) are excluded (if you are using the excluded option in the sites with alignment gaps). That is, the total number of analyzed sites considered in this command can be equal or lower than those taken into account in the DNA polymorphism command.

For the total data:

- The average number of nucleotide differences, k (Tajima 1983, equation A3).
- The nucleotide diversity, Pi (π, Nei 1987 equation 10.5).
- The average number of nucleotide substitutions per site between populations, Dxy (Nei 1987, equation 10.20).
- The number of net nucleotide substitutions per site between populations, Da (Nei 1987, equation 10.21).

Genetic Differentiation Analysis:

DnaSP conducts the following analyses:

Haplotype-based statistics:
Hs (Hudson et al. 1992a, eq. 3a); Hst (Hudson et al. 1992a, eq. 2).

Nucleotide Sequence-based statistics:
Ks (Hudson et al. 1992a, eq. 10); Kst (Hudson et al. 1992a, eq. 9).
Ks* and Kst* (Hudson et al. 1992a, eq. 11).
Z (Hudson et al. 1992a).
Z* (Hudson et al. 1992a).
Snn (Hudson 2000).

Statistical tests:
Chi-square test (haplotype data) (Nei 1987; Hudson et al. 1992a, eq. 1).
PM, Permutation (randomization) test (Hudson et al. 1992a).

Population Size Weighting factor (see Hudson et al. 1992a, p. 144):
DnaSP computes the statistics using the weighting factors recommended in Hudson et al. (1992a); i.e, using the n-2 correction.

Export Genetic Distances:

Use this command to export genetic distances into MEGA or PHYLIP format files. These files will allow performing subsequent phylogenetic analyses using the the MEGA or PHYLIP softwares

Precision Value
Number of decimal included in the distance data files.

Note:
DnaSP can not read MEGA / PHYLIP data files with genetic distance information. These files can be read by the MEGA or PHYLIP softwares which allows performing some phylogenetic analysis.
Any Word Processor could also be used to read/edit MEGA or PHYLIP files (they are just text files).

MEGA (Molecular Evolutionary Genetics Analysis) Software
The MEGA software is distributed by free from:http://www.megasoftware.net/

Gene Flow Analysis:

The gene flow estimates are computed using information about the organism's genomic type (haploid, diploid) indicated in the Data Menu. DnaSP computes the following measures:

From haplotype data information

- Nei 1973: Gst (Nei 1973, equation 9) and Nm. DnaSP calculates Gst as equations 5 and 6 in Hudson et al. (1992a).

From nucleotide sequence data information:

- Nei 1982: DeltaST (δst), equation 4; GammaST (γst), equation 5; and Nm.
  Note: DnaSP calculates PiS (πs), the average of the Pi (π) for over populations, using Nei (1982) equation 2; i.e. making use of the relative size of any population.
- Lynch and Crease 1990: Nst (equation 36) and Nm.
  Nst estimator is almost the same as Fst (Hudson et al. 1992b). The difference is that Nst uses the Jukes and Cantor (1969) correction.
- Hudson et al. 1992b: Fst (equation 3) and Nm (equation 4).

Wright (1951)

The estimates of Nm are based on the island model of population structure:
Haploids: Fst, γst, Nst = 1 / (1 + 2Nm)
Diploids (autosome): Fst, γst, Nst = 1 / (1 + 4Nm)
Diploids (X-chromosome): Fst, γst, Nst = 1 / (1 + 3Nm)
Diploids (Y-chromosome): Fst, γst, Nst = 1 / (1 + Nm)

Effective Population size

Note:

n.a. , not applicable. When the proportion of differences is equal or higher than 0.75, the Jukes and Cantor correction can not be computed.

Tips:

This module might run slowly than that in the previous DnaSP version. You might consider to use the old DnaSP version (you can execute both versions in your computer) www.ub.edu/dnasp/indexDnaSPv5, or use the new Multi-MSA Data File Analysis (All Sites) utilizing a *.SG.txt file.

Linkage Disequilibrium

Linkage Disequilibrium

See Also: Coalescent Simulations Graphs Window Input Data Files Output
References: Hill and Robertson 1968 Kelly 1997 Langley et al. 1974 Lewontin 1964 Lewontin and Kojima 1960 Rozas et al. 2001 Sokal and Rohlf 1981 Wall 1999 Weir 1996

This command calculates the degree of linkage disequilibrium (LD), or nonrandom association between nucleotide variants at different polymorphic sites. Sites containing alignment gaps, or polymorphic sites segregating for three or four nucleotides, are completely excluded from the analysis. The analysis can be

performed with all polymorphic sites in the data, or only with parsimony-informative sites (sites that segregate for only two nucleotides that are present at least twice).

Linkage disequilibrium between nucleotide variants:

The degree of LD is estimated by the following parameters:
D (Lewontin and Kojima 1960),
D' (Lewontin 1964),
R and R^2 (Hill and Robertson 1968).
DnaSP considers as coupling gametes those with the most or the least common variants (Langley et al. 1974).

Linkage disequilibrium for the whole data:

For the whole data DnaSP computes:
ZnS statistic (Kelly 1997, equation 3). ZnS is the average of R^2 (Hill and Robertson 1968) over all pairwise comparisons.
Za and ZZ statistics (Rozas et al. 2001). Za is the average of R^2 (Hill and Robertson 1968) over all pairwise comparisons between adjacent polymorphic sites; ZZ = Za - ZnS. ZZ statistic could be used for detecting intragenic recombination (see Recombination).
DnaSP can compute the Confidence Intervals of ZnS, Za, ZZ by coalescent-based simulations (see: Coalescent Simulations).

Association among nucleotide variants:

DnaSP also computes the B and Q statistics (Wall 1999).

Statistical significance of LD:

Both the two-tailed Fisher's exact test, and the chi-square test are computed to determine whether the associations between polymorphic sites are, or are not, significant (see Sokal and Rohlf 1981).
(*, P < 0.05; **, P < 0.01; ***, P < 0.001).
DnaSP also performs the Bonferroni correction for multiple tests (see Weir 1996). The Bonferroni procedure tries to avoid spurious rejections of the null hypothesis in multiple tests (assuming that all tests are independent). For an overall a' (a' is the probability that at least one test causes the rejection of a true null hypothesis), α (α is the probability that an individual test causes the rejection of a true null hypothesis; i.e., type I error of an individual test) is obtained from:

$$\alpha = 1 - (1 - \alpha')^{(1/L)}$$

where L is the number of tests performed. DnaSP obtain the probability associated with a particular chi-square value (with 1 degree of freedom) by the trapezoidal method of numeric integration. Significant disequilibrium by the Bonferroni procedure for an $\alpha' = 0.05$ is indicated by the letter B.
NOTE: The Bonferroni correction applied to non-independent tests (as in the LD tests) would be highly conservative.

Statistical significance by the coalescent:

DnaSP can also provide the confidence intervals of B, Q, ZnS, Za and ZZ statistics by computer simulations using the coalescent algorithm (see Coalescent Simulations).

LINKAGE DISEQUILIBRIUM AND PHYSICAL DISTANCE

DnaSP estimates the relationship of linkage disequilibrium with physical distance by the regression analysis (Sokal and Rohlf 1981).
DnaSP estimates the linear regression equation: Y = a + bX,
where Y is the LD value, and X is the nucleotide distance (measured in kilobases; kb).
The regression equation is performed for |D| (absolute value of D), |D'| (absolute value of D') and R^2 values.
For |D'| values, DnaSP gives two regression equations:
i) for all |D'| values (blue line in the graph -default colour-);

ii) for all |D'| values excluding values of |D'| = 1 (+1 and -1); (black line in the graph -default colour-).

Statistical significance:

The statistical significance of the regression coefficient could be conducted by the Student's t-test with n-2 degrees of freedom (n is the total number of values -pairwise comparisons-) (this test is not included in DnaSP). But, be careful! This test requires independent sample values, and certainly, it is not the case for LD.
Alternative. You could determine the confidence intervals of the ZZ test statistic by coalescent-based simulations (see: Coalescent Simulations).
Another alternatives (not included in present version of DnaSP): you might test the decay of LD with physical distance by the randomization (permutation) test (i.e., by random permutation of the polymorphic sites).

Nucleotide distance:

The nucleotide distance (Dist in the output), i. e. the distance in nucleotides between a given pair of polymorphic sites, is calculated as the average number of nucleotides that separate two particular polymorphic sites.
For example, the nucleotide distance between polymorphic sites 1 and 18 (marked with asterisks) in the following four sequences is 13:

```
      *                  *
seq_1 ATATACGGGGTTA---TTAGA
seq_2 CGATAC--GG-TA---TAACA
seq_3 AGATACGG-GATA---TAATA
seq_4 ATAAACGGGGATA---GTAGT
```

Output:

The output of the analysis is given in a grid (table). The columns Site1 and Site2 refer to the polymorphic sites analyzed (compared); Dist to the nucleotide distance between them; Fisher to the probability obtained by Fisher's exact test; and Chi-sq to the value of X^2. The results are also presented graphically (by a scatter graph). In the graph D, D', R, R^2 can be plotted against the nucleotide distance (X axis).

Recombination

Recombination

See Also: Coalescent Simulations Input Data Files Output
References: Hudson 1987 Hudson and Kaplan 1985 Rozas et al. 2001

This command computes some estimates of the Recombination parameter R = 4Nr (for autosomal loci of diploid organisms), where N is the population size and r is the recombination rate per sequence (per gene). In the literature, the recombination parameter is also indicated as C = 4Nc.
For the present analysis sites containing alignment gaps (or missing data) in the data files are not used (these sites are completely excluded). The program estimates the following measures:

Recombination parameter R = 4Nr (Hudson 1987)

The estimator is based on the variance of the average number of nucleotide differences between pairs of sequences, S2k (Hudson 1987, equation 1).
The estimator R is obtained after solving equation 4 (Hudson 1987). The solution of the function g(C,n) of equation 4 is obtained numerically (see the Appendix in Hudson 1987).
The output DnaSP shows the estimate of R (=4Nr), per gene (r, is the recombination rate per generation between the most distant sites) (Hudson 1987, from equation 4). DnaSP also calculates the estimate of R

between adjacent sites:

R (between adjacent sites) = R (per gene) / D

where D, is the average nucleotide distance (in base pairs) of the analyzed region (the average nucleotide distance after removing the alignment gaps; i.e., nucleotide distance); (see Nucleotide Distance in Linkage Disequilibrium command). Note that the average length is equal to the average nucleotide distance + 1.

The minimum number of recombination events RM (Hudson and Kaplan 1985)

The parameter indicates the minimum number of recombination events in the history of the sample (note that RM underestimates the total number of recombination events). RM is obtained using the four-gamete test (see Figure 1 and Appendix 2 in Hudson and Kaplan 1985). From RM it is possible to estimate R by computer simulations Coalescent Simulations.

The output shown by DnaSP is:

The RM value.
The list of all the pairs of sites with the four-gametic types.
The list of all RM pairs of sites where it is possible to assign at least one recombination event.

Note: for the present analysis sites segregating for three or four nucleotides are completely excluded from the analysis.

ZZ test statistic (Rozas et al. 2001)

This test statistic could be useful in detecting intragenic recombination (see Linkage Disequilibrium).

Statistical significance by the coalescent:

DnaSP can provide the confidence intervals of the RM statistic by computer simulations using the coalescent algorithm (see Computer Simulations).

Effective Population size

Other methods

The recombination parameter can also be estimated by the method described in Hey and Wakeley 1997 this method, however, is not included in the DnaSP software. That method is implemented in the SITES computer program, distributed by Jody Hey.
Jody Hey Web Page:
https://bio.cst.temple.edu/~hey/

Population Size Changes

Population Size Changes

See Also: Coalescent Simulations Graphs Window Input Data Files Output
***References*:** Harpending 1994 Ramos-Onsins and Rozas 2002 Rogers 1995 Rogers and Harpending 1992 Rogers et al. 1996 Slatkin and Hudson 1991 Tajima 1989a Tajima 1989b Watterson 1975
Abstracts: Ramos-Onsins and Rozas 2002

This command analyzes the frequency spectrum (for segregating sites) and the pairwise number of differences. DnaSP performs these analyses for constant size, and for growing size populations.

Alignment gaps and missing data:

Sites containing alignment gaps (or sites with missing data) in the data file are not used (these sites are

completely excluded).

1. Pairwise Number of Differences:

1.1 Constant Population Size

DnaSP shows (in tabular and graphic form) the distribution of the observed pairwise nucleotide site differences (also called mismatch distribution), and the expected values (at equilibrium for no recombination) in a stable population, i.e. population with constant population size (Watterson 1975; Slatkin and Hudson 1991, equation 1; Rogers and Harpending 1992, equation 3).

1.2 Population Growth-Decline

DnaSP shows (in tabular and graphic form) the distribution of the observed pairwise nucleotide site differences (also called mismatch distribution), and the expected values (for no recombination) in growing and declining populations (Rogers and Harpending 1992, equation 4). The model is based on three parameters: Theta Initial (theta before the population Growth or Decline), Theta Final (theta after the population Growth or Decline), and τ (Tau) is the date of the Growth or Decline measured in units of mutational time (Tau = 2ut; t is the time in generations, and u is the mutation rate per sequence and per generation) (Rogers and Harpending 1992). By letting Theta Final as infinite it is possible to estimate Theta Initial and Tau (2ut) from the data (Rogers 1995). DnaSP gives these estimates that can be used to obtain the expected values.

DnaSP also estimates the raggedness statistic, r (Harpending 1994, equation 1). This statistic quantifies the smoothness of the observed pairwise differences distribution. DnaSP can provide the confidence intervals of this statistic by computer simulations using the coalescent algorithm (see: Computer Simulations).
Nevertheless, the raggedness statistic has low statistical power for detecting population expansion. Therefore, it is better to use more powerful statistics as the Fu's Fs (see: Fu and Li's (and other) Tests) and the Ramos-Onsins and Rozas's R2; DnaSP can also provide (by computer simulations using the coalescent; see: Coalescent Simulations) the confidence intervals of the Fs and R2 statistics.

C.V., Coefficient of variation (see: Rogers and Harpending 1992, p. 554)
MAE, Mean Absolute Error (see: Rogers et al. 1996, p. 896).

2. Segregating Sites:

2.1 Constant Population Size

DnaSP shows (in tabular and graphic form) the distribution of the observed frequency spectrum (distribution of the allelic frequency in a site) (see Tajima 1989a, figure 6), and the expected values in a stable population, i.e. population with constant population size (Tajima 1989a, equation 50).

2.2 Population Growth-Decline

DnaSP shows (in tabular and graphic form) the distribution, at different times and for several sample sizes, of Sn(t), the expected number of segregating sites among n DNA sequences in generation t, and Sn(t) / a1 (at equilibrium this value is equal to theta) after a population growth or decline (Tajima 1989b, equation 9). The time is measured in N generations units, where N is the effective population size.

$a1 = \sum (1 / i)$ from i=1 to n-1
(n is the sample size, i.e., the number of nucleotide sequences)

Effective Population size

Fu and Li's (and other) Tests

Fu and Li's (and other) Tests

See Also: Coalescent Simulations Graphs Window Input Data Files Output Polymorphic/Variable Sites File
References: Achaz 2008 Achaz 2009 Ewens 1972 Fu and Li 1993 Fu 1995 Fu 1997 Kimura 1983 Simonsen et al. 1995 Strobeck 1987 Tajima 1983

This command calculates the statistical tests D* and F* proposed by Fu and Li (1993) for testing the hypothesis that all mutations are selectively neutral (Kimura 1983). In this command DnaSP also computes the Fu's Fs and the Strobeck's S statistics. These tests require data only on molecular polymorphism.

Alignment gaps and missing data:

Sites containing alignment gaps (or sites with missing data) are not used (these sites are completely excluded).

Minimum number of sequences in data files:

The data file must contain at least four sequences.

Analysis:

D* and F* tests are based on the neutral model prediction that estimates of η / a1, (n -1)η_s / n, and of k, are unbiased estimates of θ,
where,
η, is the total number of mutations
a1 = Σ (1 / i) from i=1 to n-1
n, the number of nucleotide sequences
η_s, is the total number of singletons (mutations appearing only once among the sequences).
k , is the average number of nucleotide differences between pairs of sequences (Tajima 1983, equation A3). (Note that Fu and Li use Pn to indicate k).
$\theta = 4N\mu$ (for diploid-autosomal; N and μ are the effective population size, and the mutation rate per DNA sequence per generation, respectively).

Fu and Li D* and F* test statistics

The D* test statistic is based on the differences between η_s, the number of singletons (sites segregating at frequency of 1/n or (n-1)/n); i.e. sites with nucleotide variants appearing only once at a particular site), and η, the total number of mutations (Fu and Li 1993, p. 700 bottom).
The F* test statistic is based on the differences between η_s, the number of singletons, and k, the average number of nucleotide differences between pairs of sequences (Fu and Li 1993, p. 702; see also Simonsen et al. 1995, equation 10; Achaz 2009).

Fu and Li D* and F* test statistics (DnaSP v5 and later versions)

In version 5.10 (and earlier), these tests were computed in a way slightly different than in version 6.
In version 6 (and later) the Fu and Li D* statistic is computed exactly as in version 5, but using only biallelic positions.
In version 6 (and later) the Fu and Li F* statistic is computed using the generic variances described in Achaz 2009 and using only biallelic positions. Therefore, the results can differ a little between versions.

Achaz Y* test statistic

The Y* test statistic is also based on the differences between two different estimates of θ, from the number of nonsingleton segregating sites and from k (Achaz 2008; equation 21). For the analysis only biallelic positions are used.

Fu's Fs statistic

The Fs test statistic (Fu 1997, equation 1) is based on the haplotype (gene) frequency distribution conditional the value of θ (Ewens 1972, equations 19-21).

Strobeck's S statistic

The Strobeck's S test statistic (Strobeck 1987; see also Fu 1997) is also based on the haplotype (gene) frequency distribution conditional the value of θ (Ewens 1972, equations 19-21). The S statistic gives the probability of obtaining a sample with equal or less number of haplotypes than the observed. DnaSP also provides the probability of obtaining a sample with a number of haplotypes equal to the observed. See also the Discrete Distributions command in the Tools Menu.

Total number of mutations vs. number of segregating sites:

The D* and F* test statistics can also be computed using S, the number of segregating sites instead of η, the total number of mutations (Simonsen et al. 1995, equations 9-10). Under the infinite site model (with two different nucleotides per site) both D* and F* values should be the same (S and η have the same value). However, if there are sites segregating for more than two nucleotides, values of S will be lower than those of η.

Effective Population size

Statistical significance:

DnaSP uses the critical values obtained by Fu and Li (1993) (two tailed test, Tables 2 and 4) to determine the statistical significance of D* and F* test statistics. Note that these values were obtained by computer simulations considering that the true value of θ falls into the interval [2, 20]; so that the critical values are not applicable when the true value of θ is not in that interval.
DnaSP will not determine the critical values for sample sizes larger than 300. For sample sizes 100-300 DnaSP uses the same critical values than for n=100; the reason is that the critical values increases (or decreases) with ln(n), so that when n is large the curve of critical values becomes flat (Fu, personal communication).
(n.d., not determined; #, P < 0.10; *, P < 0.05; **, P < 0.02).

Statistical significance by the coalescent:

DnaSP can also provide the confidence intervals of the Fu and Li's D* and F*, the Fu's Fs and the Achaz's Y* by computer simulations using the coalescent algorithm (see: Coalescent Simulations).

Sliding window option:

This option computes both D* and F* values, and their statistical significance, by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph D* and F* values (Y axis) can be plotted against the nucleotide position (X axis).

Fu and Li's (and other) Tests with an Outgroup

Fu and Li's (and other) Tests with an Outgroup

See Also: Coalescent Simulations Graphs Window Input Data Files Output Polymorphic/Variable Sites File
References: Achaz 2008 Achaz 2009 Fay and Wu 2000 Fu and Li 1993 Fu 1995 Kimura 1983 Simonsen et al. 1995 Tajima 1983 Zeng et al 2006

This command calculates the statistical tests D and F proposed by Fu and Li (1993) for testing the

hypothesis that all mutations are selectively neutral (Kimura 1983). These tests require data of the intraspecific variation (polymorphism) and data from an outgroup (one or more sequences from a related species).

Data Files:

For the present analysis, at least two sets of sequences (one with the intraspecific data, and other with the outgroup sequences) must be defined (see: Data | Define Sequence Sets command).

Minimum number of sequences in data files:

The intraspecific data file must contain at least four sequences.
The outgroup can contain more than one sequence, but the analysis will be performed in the first sequence one. Nevertheless, if there are more than one sequence in the outgroup, sites with alignment gaps (or with missing data) in any of the outgroup sequences will not be used (see below).

Alignment gaps and missing data:

Sites containing alignment gaps (or sites with missing data) in any data file are not used (these sites are completely excluded).

Ambiguous information:

In some cases, the polarity of some substitutions could not be unambiguously determined; for example:

Intraspecific Data

```
              10
               *
seq1 CTTAACCTTC
seq2 CATTATTTAC
seq3 CTATATTCCC
seq4 A-AAACCTAC
```

Outgroup

```
Seq5 CT-AAGGGAC
Seq6 CTA-AGCTAC
```

Site 1. The “A” in seq 4 is an external mutation (and derived substitution)
Site 2. Not used (alignment gaps in intraspecific data)
Site 3. Not used (alignment gaps in the used outgroup)
Site 4. Not used (alignment gaps in one sequence of the outgroup data file)
Sites 6-8. Not used (ambiguous information in outgroup)
Site 9. The “T” and the “C” on the intraspecific data file are singleton and also external mutations.

Ambiguous positions will not be used; DnaSP will list them.

Analysis:

These tests are based on the neutral model prediction that estimates of η / a1, η_e, and of k, are unbiased estimates of θ,
where,
η, is the total number of mutations
a1 = Σ (1 / i) from i=1 to n-1
n, the number of nucleotide sequences
η_e, is the total number of mutations in external branches of the genealogy.
k , is the average number of nucleotide differences between pairs of sequences (Tajima 1983, equation A3). (Note that Fu and Li use Pn to indicate k).
$\theta = 4N\mu$ (for diploid-autosomal; N and μ are the effective population size, and the mutation rate per DNA sequence per generation, respectively).

Fu and Li D and F test statistics

The D test statistic is based on the differences between η_e, the total number of mutations in external branches of the genealogy, and η, the total number of mutations (Fu and Li 1993, equation 32).
The F test statistic is based on the differences between η_e, the total number of mutations in external branches of the genealogy, and k, the average number of nucleotide differences between pairs of sequences (Fu and Li 1993, p. 702, top).

Fu and Li D and F test statistics (DnaSP v5 and later versions)

In version 5.10 (and older), these tests were computed in a way slightly different than in version 6.
In version 6 (and later) the Fu and Li D statistic is computed exactly as in version 5, but using only biallelic positions.
In version 6 (and later) the Fu and Li F statistic is computed using the generic variances described in Achaz 2009 and using only biallelic positions. Therefore, the results can be a little different.

Fay and Wu H, and Normalized Fay and Wu Hn test statistics

The H test statistic (Fay and Wu 2000, equations 1-3) is based on the differences between two estimators of θ: θ_π (or k), the average number of nucleotide differences between pairs of sequences, and θ_H (Fay and Wu 2000, equation 3), an estimator based on the frequency of the derived variants.
The normalized H statistic (Hn) is the scaled version of the H statistic (Zeng et al. 2006; equation 11).
From version 6, DnaSP only provides the value of the Hn statistic, which it is computed using only biallelic positions.

Zeng et al. E test statistic

The Z test statistic is a normalized statistic contrasting the differences between low- and high-frequency variants of the frequency spectrum (Zeng et al. 2006; equation 13). For the analysis only biallelic positions are used.

Achaz Y test statistic

The Y test statistic is based on the differences between two different estimates of θ, from the number of nonsingleton segregating sites and from k (Achaz 2008; equation 21). For the analysis only biallelic positions are used.

The number of mutations in external branches

Assuming the infinite sites model DnaSP calculates the total number of mutations in the external branches of the genealogy as follows: at a given particular polymorphic site, the number of mutations in external branches is counted as the number of distinct singleton nucleotide variants (in the intraspecific data file) that are not shared with the outgroup (a singleton mutation is a nucleotide variant that appears only once among the sequences). The total number of mutations in external branches of the genealogy is then computed as the sum of the number of mutations in external branches of every polymorphic site.

Total number of mutations vs. number of segregating sites:

The D and F test statistics can also be computed using S, the number of segregating sites instead of η, the total number of mutations (see Simonsen et al. 1995). Under the infinite sites model (with two different nucleotides per site) both D and F values should be the same (S and η have the same value). However, if there are sites segregating for more than two nucleotides, values of S will be lower than those of η.

Effective Population size

Statistical significance:

DnaSP uses the critical values obtained by Fu and Li 1993 (two tailed test, Tables 2 and 4) to determine the statistical significance of D and F test statistics. Note that these values were obtained by computer simulations considering that the true value of θ falls into the interval [2, 20]; so that, the critical values are

not applicable when the true value of θ is not in that interval. DnaSP will not determine the critical values for sample sizes larger than 300. For sample sizes 100-300 DnaSP uses the same critical values than for n=100; the reason is that the critical values increases (or decreases) with ln(n), so that when n is large the curve of critical values becomes flat (Fu, personal communication).
(n.d., not determined; #, P < 0.10; *, P < 0.05; **, P < 0.02).

Statistical significance by the coalescent:
DnaSP can also provide the confidence intervals of the Fu and Li's D and F, the Fay and Wu's Hn, the Zeng's E and the Achaz's Y by computer simulations using the coalescent algorithm (see Coalescent Simulations).

Sliding window option:
This option computes both D and F values, and their statistical significance, by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph D and F values (Y axis) can be plotted against the nucleotide position (X axis).

HKA, Hudson, Kreitman and Aguadé's Test

Hudson, Kreitman and Aguadé's Test (HKA Test)

See Also: Input Data Files Output
References: Hudson et al. 1987 Kimura 1983 Nei 1987

This command performs the Hudson, Kreitman and Aguadé's (1987) test (HKA test). The test is based on the Neutral Theory of Molecular Evolution (Kimura 1983) prediction that regions of the genome that evolve at high rates will also present high levels of polymorphism within species. The test requires data from one interspecific comparison of at least two regions of the genome, and also data of the intraspecific polymorphism in the same regions of at least one species.

Data Files:
For the present analysis, at least two sets of sequences (one with the intraspecific data, and other with the outgroup sequences) must be defined (see: Define Sequence Sets command).

Minimum number of sequences in data files:
The intraspecific data file must contain at least two sequences, while the interspecific data file can contain one or more sequences.

Alignment gaps and missing data:
Sites containing alignment gaps (or sites with missing data) in any data file are not used (these sites are completely excluded).

Implementation:
The test is performed considering intraspecific data from only one species (Hudson et al. 1987, equation 6). If there is more than one sequence in the interspecific data file, the intraspecific polymorphism will be ignored; however this information will be considered in computing the interspecific divergence. The estimate of D, the between species divergence, is obtained as the average number of differences between DNA sequences from species 1 and 2; that is, D is estimated in the same way as the Dxy (Nei 1987, equation 10.20) but per sequence.

Regions (loci):

DnaSP performs the HKA test of only two regions. These regions could be any two non-overlapping segments of sites of the data file.

Chromosomal location:

DnaSP assumes that the two regions (loci) are located in the same chromosome (the two compared regions are from the same data file); i.e., both are in autosomal chromosomes or in sex chromosomes. Even though, the statistical significance of the HKA test will be the same in both cases, different estimates of the divergence time or theta are expected; so that, it is convenient to indicate the chromosome where the region is located. DnaSP has considered that the expectation of π is $4N\mu$ for autosomal, $3N\mu$ for X-linked genes, and $N\mu$ for Y-linked genes [we have slightly modified the equations of Begun and Aquadro (1991) for comparisons involving autosomal (or X-linked) with Y-linked genes].
You can compare regions located in autosomes with regions in sex chromosomes using the module HKA test. Direct Mode.

Substitutions Considered:

All substitutions: All substitutions are used (excluding substitutions in sites with gaps or missing data).
Silent substitutions: Only silent substitutions are used (synonymous substitutions and changes in noncoding positions). If the data file does not contain assigned coding regions all sites will be considered as noncoding positions; i.e. all substitutions will be considered as silent.
Synonymous substitutions: Only synonymous substitutions are used (substitutions in the coding region that not result in amino acid changes). This option works only if the data file contain sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).
Note: see how DnaSP estimates Synonymous and Nonsynonymous changes in a codon.

Effective Population Size

Output:

The present module displays the following output:

- Estimates of the Time of divergence (measured in 2N generations, where N is the effective population size),
- Estimates of theta (θ) per nucleotide in region (locus) 1,
- Estimates of theta (θ) per nucleotide in region (locus) 2,
- The X-square value, and the statistical significance.

Statistical significance:

The statistical significance is obtained assuming a χ-square distribution with one degree of freedom.
DnaSP obtains the probability associated with a particular chi-square value (with 1 degree of freedom) by the trapezoidal method of numerical integration.
(#, P < 0.10; *, P < 0.05; **, P < 0.01; ***, P < 0.001).

See also:

To compare autosomal and sex-linked regions, or to perform the HKA test with polymorphism data with different number of sequences in the two regions, or with different number of sites for the intraspecific and interspecific comparison the module HKA test -Direct Mode- should be used.

McDonald and Kreitman's Test

McDonald and Kreitman's Test

See Also: Input Data Files Output
References: Fay et al. 2001 Kimura 1983 McDonald and Kreitman 1991 Rand and Kann 1996

This command conducts the test of the neutral hypothesis (Kimura 1983) proposed by McDonald and Kreitman (1991). The test is based on a comparison of synonymous and nonsynonymous (replacement) variation within and between species. Under neutrality, the ratio of replacement to synonymous fixed substitutions (differences) between species should be the same as the ratio of replacement to synonymous polymorphisms within species.

Alignment gaps and missing data:

Codons containing alignment gaps (or codons with missing data) in any species are not used (these codons are completely excluded).

Data Files:

For the present analysis, at least two sets of sequences (one for each species) must be defined (see: Define Sequence Sets command).
DnaSP performs the McDonald and Kreitman test from sequence information included in data files. DnaSP calculates:

- Number of (polymorphic) synonymous substitutions within species,
- Number of (polymorphic) nonsynonymous (replacement) changes within species,
- Number of synonymous substitutions fixed between species,
- Number of nonsynonymous (replacement) differences fixed between species,
- and for these information computes the 2 x 2 contingence table.

A fixed nucleotide site between species is a site at which all sequences in one species contain nucleotide variants that are not in the second species.

Silent Substitutions Considered:

Substitutions in Coding Regions: Only synonymous substitutions (coding region) will be considered.
In Coding and Noncoding Regions: All silent substitutions will be considered (synonymous substitutions and changes in noncoding positions). If the data file does not contain assigned coding regions all sites will be considered as noncoding positions; i.e. all substitutions will be considered as silent.

How DnaSP estimates Synonymous and Nonsynonymous changes in a codon (MK test):

In general DnaSP uses a conservative criterion to decide if a particular change in a nucleotide site is synonymous or replacement (see the following examples). Nevertheless, the user should check the complex cases (those triplets of sites segregating for several codons; i.e. in highly variable regions).

Example using the Nuclear Universal Genetic Code

Species#1

```
  3   6   9  12  15  18  21  24  27
  *   *   *   *   *   *   *   *   *
AGT TCT ATT CCC AAT ATA AGT UAU UAU
AGC TCT ATT CCC AGG TTA AGT UAU UAU
AGA TCT CTG CAG ACT TTG AGA CUG CUG
AGG TCT CTG CAG ACT ATG AGA CUG CUG
```

Species#2

```
AGG CCT ATT CCC GGA TTT GGA CUG CUG
AGG CCT ATT CCC GGA TTT GGA CUG CUG
AGG CCT ATT CAC GGA TTT GGT CUG CUU
AGG CCT ATT CAC GGA TTT GGT CUG CUU
```

Codon (1,2,3):
species#1: 3 mutations in site#3: 1 replacement, 2 synonymous.

species#2: Monomorphic.
within species: 1 replacement and 2 synonymous (site#3).
fixed differences: 0

Codon (4,5,6):
species#1: Monomorphic.
species#2: Monomorphic.
within species: 0.
fixed differences: 1 replacement (site 4).

Codon (7,8,9):
species#1: Site#7 is replacement; Site#9 is synonymous.
If there are two possible paths:
Path#1: ATT (Ile) -> CTT (Leu) -> CTG (Leu) Site#7 Replacement; Site#9 Synonymous
Path#2: ATT (Ile) -> ATG (Met) -> CTG (Leu) Site#7 Replacement; Site#9 Replacement
DnaSP will choose path#1, the path that requires the minor number of replacements (however, see the next codon).
species#2: Monomorphic.
within species: 1 replacement, 1 synonymous.
fixed differences: 0.

Codon (10,11,12):
species#1: Site#11 is replacement; Site#12 is replacement.
Here there are also two possible paths:
Path#1: CCC (Pro) -> CCG (Pro) -> CAG (Gln) Site#11 Replacement; Site#12 Synonymous
Path#2: CCC (Pro) -> CAC (His) -> CAG (Gln) Site#11 Replacement; Site#12 Replacement
However, DnaSP will choose path#2. If there are two possible paths, and one of the non-extant codons (e.g. CAC in this case) is found in the other species, DnaSP assume that the true evolutionary path is the path with that codon (i.e. path#2 in the present example).
species#2: Site#11 is replacement.
within species: 2 replacements (site#11 and site#12)
fixed differences: 0

Codon (13,14,15):
species#1: Site#14 (2 replacements); Site#15 is Synonymous.
Here there are four possible paths:
Path#1: ACT (Thr) -> AAT (Asp) -> AGT (Ser) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#2: ACT (Thr) -> AAT (Asp) -> AAG (Lys) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#3: AAT (Asn) -> ACT (Thr) -> AGT (Ser) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Replacement).
Path#4: AAT (Asn) -> ACT (Thr) -> ACG (Thr) -> AGG (Arg) Site#14 (2 Replacements); Site#15 (1 Synonymous).
DnaSP will choose path#4, the path that requires the minor number of replacements.
species#2: Monomorphic.
within species: 2 replacements (site#14), and 1 synonymous (Site#15)
fixed differences: 2, Site#13 is replacement; Site#15 is synonymous.
For computing fixed differences, DnaSP will check all paths between codons of the two species, and it will choose the path with the minor number of changes. If there are several paths with the same number of differences, DnaSP will choose the path with the lower number of replacement changes.

Codon (16,17,18):

species#1: Site#16 (1 replacement); Site#18 (1 synonymous).
Here there is a circular path:
ATA (Ile) -> TTA (Leu)
¡ !
ATG (Met) <- TTG (Leu)
Let us suppose that the number of mutations were only two (one in site 16, and another in site 18), DnaSP must assume one recombination event, the recombination event that requires the lower number of replacement substitutions:
| TTG (Leu)
TTA (Leu) ->| recomb: ATG (Met)
| ATA (Ile)
species#2: Monomorphic.
within species: Site#16 (1 replacement); Site#18 (1 synonymous).
fixed differences: 1, Site#18 is replacement.
Note: This kind of codons will be analyzed only for Nuclear Genetic Codes.

Codon (19,20,21):
species#1: 1 replacement (site#21).
species#2: 1 synonymous (site#21).
within species: 1 replacement (site#21).
If there is discordance between replacement and synonymous changes within species (for the same nucleotide variants), DnaSP will choose the case with more replacement substitutions.
fixed differences: 1 replacement (site#19).

Codon (22,23,24):
species#1: There are 3 changes between codons. So that there are 6 putative evolutionary paths (in this particular example there are only 4 because we exclude paths that go through stop codons). DnaSP will choose one of following paths:
2 replacements (Site#22 and Site 23), and 1 synonymous (Site#24) and
2 replacements (Site#23 and Site 24), and 1 synonymous (Site#22).
(however, see also the next codon).
species#2: Monomorphic.
within species: the same than for species#1.
fixed differences: 0.

Codon (25,26,27):
The present example is similar to the previous codon (22,23,24) example. Here, however, there is variation in species#2. In this case DnaSP will check the codons in species#2 to decide the assignation of species#1.
species#1: 2 replacements (Site#22 and Site 23), and 1 synonymous (Site#24).
species#2: 1 synonymous (Site#24).
within species: 2 replacements (Site#22 and Site 23), and 1 synonymous (Site#24).
fixed differences: 0.

Output:
Codons not analyzed:
DnaSP does not estimate synonymous and replacement changes in some complex cases (ambiguous/complex codons; those sites segregating for several codons; i.e. in highly variable regions). The user should do manually.
DnaSP does not estimate synonymous and replacement changes in codons with alignment gaps.
Neutrality Index: Indicates the extent to which the levels of amino acid polymorphism depart from the expected in the neutral model (Rand and Kann, 1996).
Alfa value (α): Indicates the proportion of amino acid substitutions driven by positive selection (Fay et al.

2001).

Statistical significance:
Both the two-tailed Fisher's exact test, and G-test of independence are computed to determine whether the deviations on the ratio of replacement to synonymous (fixed substitutions between species vs. polymorphisms within species) are or not significant. DnaSP obtains the probability associated with the G value (with 1 degree of freedom) by the trapezoidal method of numerical integration.

Tajima's Test

Tajima's Test

See Also: Coalescent Simulations Graphs Window Input Data Files Output
References: Kimura 1983 Tajima 1983 Tajima 1989

This command calculates the D test statistic proposed by Tajima (1989), equation 38, for testing the hypothesis that all mutations are selectively neutral (Kimura 1983). The D test is based on the differences between the number of segregating sites and the average number of nucleotide differences.

Minimum number of sequences in data files:
The data file must contain at least four sequences.

Alignment gaps and missing data:
Sites containing alignment gaps (or sites with missing data) are not used (these sites are completely excluded).

Analysis:
Tajima's test is based on the neutral model prediction that estimates of S / a1, and of k, are unbiased estimates of θ,
where,
S, is the total number of segregating sites.
a1 = $\sum$ (1 / i) from i=1 to n-1
n, the number of nucleotide sequences
k, is the average number of nucleotide differences between pairs of sequences (Tajima 1983, equation A3).
$\theta = 4N\mu$ (for diploid-autosomal; N and μ are the effective population size, and the mutation rate per DNA sequence per generation, respectively).

Total number of mutations vs. number of segregating sites:
The D test statistic can also be computed using η, the total number of mutations (see Fu and Li's test), instead of S, the total number of segregating sites. Under the infinite sites model (with two different nucleotides per site) estimates of the D test statistic based on S and on η should be the same (S and η have the same value). However, if there are sites segregating for more than two nucleotides, values of S will be lower than those of η.
Effective Population size

Tajima's D on synonymous, nonsynonymous and silent changes:
If the Coding Region has been defined (Assign Coding Regions command in the Coding Region Menu), DnaSP will also compute:

- Tajima's D on synonymous changes, D(Syn).
- Tajima's D on nonsynonymous changes, D(NonSyn).

- Tajima's D on silent (synonymous and noncoding) changes, D(Sil).
- Tajima's D(NonSyn)/D(Syn) ratio.

Statistical significance:

The confidence limits of D (two tailed test) is obtained assuming that D follows the beta distribution (Tajima 1989, equation 47), i.e. the confidence limits given in Table 2 of Tajima (1989). Note that the critical values will not be determined for sample sizes larger than 1000.
(n.d., not determined; #, P < 0.10; *, P < 0.05; **, P < 0.01; ***, P < 0.001).

Statistical significance by the coalescent:

DnaSP can also provide the confidence intervals of the Tajima's D by computer simulations using the coalescent algorithm (see: Coalescent Simulations).

Sliding window option:

This option computes the D test statistic and the confidence limits of D by the Sliding Window method. The output of the analysis is given in a grid (table). The results can also be presented graphically (by a line chart). In the graph the D value, can be plotted against the nucleotide position (X axis).

Overview Menu

Polymorphism Data

Polymorphism Data

See Also: Input Data Files Output Window Menu
References: Achaz 2008 Achaz 2009 Ewens 1972 Fu and Li1993 Fu 1997 Nei 1987 Strobeck 1987 Tajima 1993 Tajima 1989 Watterson 1975

This command computes a number of measures of the extent of DNA polymorphism and also performs some common neutrality tests. Use this command to obtain a summary of the data analysis.

Data Files:

The present analysis requires only one data file.

Analysis:

DnaSP computes the following measures:

G+C Content

- G+Cn, G+C content at noncoding positions.
- G+Cc, G+C content at coding positions.
- G+Ctot, G+C content in the complete genomic region.

Haplotype/Nucleotide Diversity

- The number of haplotypes NHap, (Nei 1987, p. 259).
- Haplotype (gene) diversity and its sampling variance (Nei 1987, equations 8.4 and 8.12 but replacing 2n by n).
- Nucleotide diversity, Pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987, equations 10.5 or 10.6), and its sampling variance (Nei 1987, equation 10.7).
- The average number of nucleotide differences, k (Tajima 1983, equation A3).
- Theta (per gene or per site) from Eta (η) or from S, (Watterson 1975, equation 1.4a; Nei 1987, equation 10.3). Theta (θ) = $4N\mu$ for an autosomal gene of a diploid organism (N and μ are the effective population size and the mutation rate -per gene or per site- per generation, respectively), Eta (η) is the total number of mutations, and S is the number of segregating (polymorphic) sites.

Neutrality tests

- Tajima's D, (Tajima 1989, equation 38).
- Fu and Li's D*, (Fu and Li 1993; computed for biallelic positions).
- Fu and Li's F*, (Fu and Li 1993, Achaz 2009; computed for biallelic positions).
- Achaz Y*, (Achaz 2008, equation 21; computed for biallelic positions).
- Fu's Fs, (Fu 1997, equation 1).
- The Strobeck's S (Strobeck 1987; see also Fu 1997).
- DnaSP also provides the probability of obtaining a sample with a number of haplotypes equal to the number observed.

More Information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Fu and Li's (and other) Tests with an Outgroup Tajima's Test

Computational issues/limitations
The sampling variance of Pi is not computed for sample sizes higher than 500.

Abbreviations:
n.a., not available.
n.s., not significant.
n.d., not determined.

Polymorphism/Divergence Data

Polymorphism/Divergence Data

See Also: Input Data Files Output Window Menu
References: Achaz 2008 Achaz 2009 Ewens 1972 Fu and Li1993 Fu 1997 Nei 1987 Strobeck 1987 Tajima 1993 Tajima 1989 Watterson 1975 Zeng et al 2006

This command computes a number of measures of the extent of DNA polymorphism and also performs some common neutrality tests. Use this command to obtain a summary of the data analysis.

Data Files:
The present analysis requires only one data file.

Analysis:
DnaSP computes the following measures:

G+C Content
- G+Cn, G+C content at noncoding positions.
- G+Cc, G+C content at coding positions.
- G+C, G+C content in the genomic region.

Haplotype/Nucleotide Diversity
- The number of haplotypes NHap, (Nei 1987, p. 259).
- Haplotype (gene) diversity and its sampling variance (Nei 1987, equations 8.4 and 8.12 but replacing 2n by n).
- Nucleotide diversity, Pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987, equations 10.5 or 10.6), and its sampling variance (Nei 1987, equation 10.7).
- The average number of nucleotide differences, k (Tajima 1983, equation A3).
- Theta (per gene or per site) from Eta (η) or from S, (Watterson 1975, equation 1.4a; Nei 1987, equation 10.3). Theta (θ) = $4N\mu$ for an autosomal gene of a diploid organism (N and μ are the effective population size and the mutation rate -per gene or per site- per generation, respectively), Eta (η) is the total number of mutations, and S is the number of segregating (polymorphic) sites.

Neutrality tests
- Tajima's D, (Tajima 1989, equation 38).
- Fu and Li's D*, (Fu and Li 1993; computed for biallelic positions
- Fu and Li's F*, (Fu and Li 1993, Achaz 2009; computed for biallelic positions).
- Fu's Fs, (Fu 1997, equation 1).
- Achaz's Y*, (Achaz 2008, equation 21; computed for biallelic positions)
- The Strobeck's S (Strobeck 1987; see also Fu 1997).
- DnaSP also provides the probability of obtaining a sample with a number of haplotypes equal to the number observed.

Tests using information from an outgroup

- Fu and Li's D, (Fu and Li 1993; computed for biallelic positions
- Fu and Li's F, (Fu and Li 1993, Achaz 2009; computed for biallelic positions).
- Fay and Wu's Hn (normalized) (Fay and Wu 2000, Zeng et al. 2006; computed for biallelic positions).
- Achaz's Y, (Achaz 2008, equation 21; computed for biallelic positions)
- Zeng et al. E (Zeng et al. 2006, equation 13; computed for biallelic positions)

Divergence

- K(JC), average number of nucleotide substitutions per site between species in data set 1 and the first sequence in data set 2, with Jukes and Cantor correction (Nei 1987, equation 5.3).

Note: For computing Fu and Li's D, Fu and Li's F and Fay and Wu H, DnaSP will use only the first sequence of the outgroup data set (Population/Species #2) to polarize the mutations.

More Information in the specific modules: Codon Usage Bias DNA Polymorphism Fu and Li's (and other) Tests Fu and Li's (and other) Tests with an Outgroup Tajima's Test

Abbreviations:
n.a., not available.
n.s., not significant.
n.d., not determined. The sampling variance of Pi will not be not computed if the sample size is higher than 500.

MultiDomain Analysis

MultiDomain Analysis

***See also*:** Define Domain Sets

DnaSP allows analyzing DNA polymorphism data in specific functional regions (see Define Domain Sets), for example; exons, introns, etc. It can compute a number of measures of the extent of DNA polymorphism and can also perform some common neutrality tests.

Haplotype/Nucleotide Diversity

- The number of Segregating Sites, S
- The total number of mutations, Eta
- The number of haplotypes NHap, (Nei 1987, p. 259).
- Haplotype (gene) diversity and its sampling variance (Nei 1987).
- Nucleotide diversity, Pi (π), (Nei 1987), and its sampling variance (not implemeneted yet) (Nei 1987, equation 10.7).
- The average number of nucleotide differences, k (Tajima 1983).
- Theta (per gene or per site) from Eta (η) or from S, (Watterson 1975; Nei 1987).

Neutrality tests

- Tajima's D, (Tajima 1989), and its statistical significance.
- Fu and Li's D*, (Fu and Li 1993), and its statistical significance.
- Fu and Li's F*, (Fu and Li 1993), and its statistical significance.
- Fu's Fs, (Fu 1997).
- G+Cn, G+C content at noncoding positions.

- G+Cc, G+C content at coding positions.

Output

Results are presented in a grid (table). You can save these results on a text file which can be opened by any spreadsheet (such as Excel).

Example

The following results represent the output of the OS-E_gene domain analysis (Data File Example: DmelOsRegions.nex):

Population	Domain	Region	n	Sites	NetSites
All_Seqs	OS-E_gene	2334..2870	17	417	417
All_Seqs	OS-F_gene	6059..7091	17	405	405

The ".." notation indicates that not all positions inside the domain range (2334..2870) have been analyzed. For instance, in OS-E_gene domain, only positions belonging to the subdomains have been analyzed (2334-2402, 2468-2542 and 2598-2870).
The "-" symbol indicates that all positions within the range have been analyzed.

Abbreviations:

n.d., not determined (not implemented yet).
n.a., not available.

Generate Menu

Concatenated Data File

Concatenated Data File

See Also: Input Data Files

DnaSP allows you to create a concatenated data file (NEXUS format), that is, a big data file containing DNA sequence information from a number of single data files.

Assumptions

All files must have the same number of sequences and in the same order.
DnaSP generates the concatenated data file by consecutive adding single data files to the right.

Individual Data Files Option

Real length: For a single data file, DnaSP will use the DNA sequence information selected in the Region to Analyse box.
Fixed length: All data files will contribute with a fixed (X nucs) number of sites. If the current (single) data file has less than X sites, DnaSP will complement with missing information. On the contrary, if the current data file has more than X sites, DnaSP will use only the firsts X sites.

Notes

Any codon assignation present in single data files will be saved on the concatenated file. The concatenated file will also save the population set information present in only the first single data file.

Shuttle to: DNA Slider

Shuttle to: DNA Slider

See Also: Input Data Files Output
References: Kimura 1983 McDonald 1996 McDonald 1998

The neutral theory of molecular evolution predicts that the levels of polymorphism will be correlated with levels of divergence between species (Kimura 1983; see also HKA test). McDonald (1996, 1998) has proposed some tests to detect heterogeneity in the polymorphism to divergence ratio across a region of DNA. These tests are based on the distribution of polymorphic sites and fixed differences across a DNA region. DnaSP searches for polymorphic sites and fixed differences and can generate a Data File that can be read by the DNA Slider program (McDonald 1998). The DNA Slider program will perform the tests described in McDonald (1996, 1998).

Data Files:

For the present analysis, at least two sets of sequences (one with the intraspecific data, and other with the outgroup sequences) must be defined (see: Data | Define Sequence Sets command).

Alignment gaps and missing data:

Sites containing alignment gaps (or sites with missing data) in any population are not used (these sites are

completely excluded).

Implementation:
If there are more than one sequence in the interspecific data file, DnaSP will assign one substitution as a fixed difference if (in a particular site) all nucleotide variants from file 1, differ from those of file 2.
Sites with three of four nucleotide variants are treated as if they were at adjacent sites, and polymorphism - fixed differences are put in the order that maximizes the number of runs (see McDonald 1996).

Substitutions Considered:
All substitutions: All substitutions are used (excluding substitutions in sites with gaps or missing data).
Silent substitutions: Only silent substitutions are used (synonymous substitutions and changes in noncoding positions). If the data file does not contain assigned coding regions all sites will be considered as noncoding positions; i.e. all substitutions will be considered as silent.
Synonymous substitutions: Only synonymous substitutions are used (substitutions in the coding region that not result in amino acid changes). This option works only if the data file contains sequences with assigned coding regions (more help in Assign Coding Regions and Assign Genetic Code).
Note: See how DnaSP estimates the number of Synonymous and Nonsynonymous changes in a codon.

Note:
DnaSP does not perform the tests described in McDonald (1996, 1998); but it can create the data file with the relevant information for the test. This data file can be read by the DNA Slider program.

DNA Slider program:
It is a Macintosh program that performs the heterogeneity tests described in McDonald (1996, 1998). You can download the program from the John McDonald Web Page:
http://udel.edu/~mcdonald
http://udel.edu/~mcdonald/aboutdnaslider.html

Filtered Positions Data File

Filter / Remove Positions

See Also: Input Data Files Output

This command allows the user to remove some positions. DnaSP module generates a NEXUS Data File including information about the polymorphic sites.

Selected Positions:
DnaSP can select the following sorts of positions:
Coding and Noncoding positions;
First, Second and Third codon positions;
Zero, Two and Four-Fold Degenerate positions;

Example (using the nuclear universal genetic code):
How DnaSP select the X-fold degenerate positions

```
  3   6   9
  *   *   *
ATA TTA ACT
ATA TTA GAT
ATA TTA -CT
```

Positions 1, 2, 5, 7 and 8, are zero-fold degenerate positions.
Position 3, is a three-fold degenerate position.
Position 4 and 6, are two-fold degenerate positions.
Position 9 could be either a i) four-fold degenerate (codon ACT), or ii) two fold-degenerate (codon GAT). DnaSP will no include that position neither for two-fold degenerate positions nor for four-fold degenerate positions.
Codons with missing information or alignment gaps are not considered.

Positions with Alignment Gaps option:
Excluded: These sites are removed.
Included: These sites are included.
Included if there is a polymorphism: These sites are included if there is a polymorphism.

Positions option:
Remove Non-Selected Positions: Non-Selected positions will be definitively removed from the active data.
Generate a NEXUS File with selected: Selected positions will be included in a NEXUS data file. The active data file will maintain all the positions.

Polymorphic Sites File

Polymorphic/Variable Sites Data File

See Also: Input Data Files Output

This module generates a NEXUS Data File including polymorphic sites information.

Sites with Alignment Gaps option:
Excluded: These sites are removed.
Included: These sites are included in the file.
Included if there is a polymorphism: These sites are included if there is a polymorphism.

Substitutions Considered:
All Substitutions: All polymorphic sites will be included.
Silent (Synonymous -coding region- and non coding region): Only silent (i.e. noncoding positions plus synonymous sites in the coding region) polymorphic sites will be included.
Only Synonymous: Only synonymous polymorphic sites (at the coding region) will be included.
Only Nonsynonymous: Only nonsynonymous polymorphic sites will be included.

Tips:
You can use this module to conduct the Fu and Li's (Fu and Li 1993) (or other tests) using only synonymous (or nonsynonymous) changes:

1. You should generate a data file with only synonymous (or nonsynonymous) substitutions.
   Generate->Polymorphism/Variable Sites Data File->only synonymous (or nonsynonymous substitutions).
2. Conduct the Fu and Li test on these new data files.

Haplotype Data File

Haplotype Data File

See Also: Input Data Files Output
References: Bandelt et al. 1999 Hudson et al. 1992 Schneider et al. 2000

This module generates Data Files with information on haplotype data. Results can be saved on a NEXUS or Roehl Data Files.

Sites with Alignment Gaps option:
Not considered: These sites are ignored (complete deletion).
Considered: Gaps are considered just like another nucleotide variant (fifth state).
Only gaps are considered: Only gaps information is considered to built haplotypes.

Invariable Sites option:
Removed: Invariable (monomorphic) sites will not be included in the output file.
Included: Invariable (monomorphic) sites will be included in the output file.

Generate option:
NEXUS Data File: Haplotype information will be stored on a NEXUS data file. Later, this file could be opened by DnaSP and might be exported in another data file format.
Arlequin Project File: DnaSP will create an Arlequin project file (*.arp) with haplotype information. This file format is the format accepted by the Arlequin software.
Roehl Data File: Haplotype information will be stored on a Roehl (Röhl) Data File (multistate data). This file format is the format accepted by the Network software. That program allows reconstructing intraspecific phylogenies (network analysis).

Arlequin software (Schneider et al. 2000)
Arlequin is a software for population genetics analysis, and it is distributed from:
http://cmpg.unibe.ch/software/arlequin35/Arlequin35.html

Network software (Bandelt et al. 1999)
The Phylogenetic Network Analysis software was written by Arne Röhl, and it is distributed for free from:
http://fluxus-engineering.com/sharenet.htm

Translate to Protein Data File

Translate to Protein Data File

See Also: Input Data Files

In this module DnaSP will translate the nucleotide sequence into an amino acid sequence, and generates a NEXUS Data File with that information. This command works only if the coding regions and the genetic code have been previously defined (more help in Assign Coding Regions and Assign Genetic Code).

Data File:
The present analysis requires only one data file.

Note:
DnaSP can not read NEXUS data files with Protein information. You can read that information with

MacClade or with any Word Processor.

Reverse Complement Data File

Reverse Complement Data File

See Also: Input Data Files

This module generates a NEXUS Data File including the sequence data in the reverse complement direction. This option would be interesting for the analysis of synonymous and nonsynonymous substitutions in data files with coding regions transcribed in both directions. DnaSP can only analyze nucleotide variation in synonymous and nonsynonymous sites if the coding regions (in data file) are in the 5' -> 3 direction. If the coding regions are transcribed in the opposite direction:

1. you should generate the reverse complement data file;
2. define the coding regions;
3. perform the appropriate analysis.

Prepare Submission to EMBL/GenBank Databases

Prepare Submission for EMBL / GenBank Databases

See Also: Input Data Files

This command generates a text file with the relevant information for a submission of DNA sequence information to the nucleotide sequence database (EMBL / GenBank / DDBJ). This command is appropriated for researchers wishing to submit multiple related sequences (see the Bulk Submissions in the EMBL Nucleotide Sequence Database Information). If the coding regions have been previously defined (see Assign Coding Regions command), DnaSP will include information on the exonic/intronic regions.

More Information on the EMBL / GenBank Databases:
http://www.ebi.ac.uk/ena/submit

Tools Menu

Tools Menu

References: Ewens 1972 Fu 1997 Jukes and Cantor 1969 Sokal and Rohlf 1981 Strobeck 1987 Tajima 1989 Watterson 1975

This menu has the following commands:

Coalescent Simulations

Coalescent Simulations (DnaSP v5)

HKA Test. Direct Mode

Discrete Distributions

Use this command to calculate probabilities, the expected value and the variance of some distributions: Binomial, Hypergeometric and Poisson.
The Ewens option allows computing the Strobeck's S statistic (Strobeck 1987; see also Fu 1997), the Fu's Fs statistic (Fu 1997), and the probability (and the expected value) of obtaining a particular number of haplotypes (Ewens 1972, equations 19-21, 24). See also the Fu and Li's Tests command.

Tests of Independence: 2 x 2 table

This command allows testing independence in a 2 x 2 tables (contingency tables). DnaSP performs three types of Independence tests: Fisher's exact test, Chi-square test (standard, and using Yates' correction) and G test (standard, and using Williams' or Yates' corrections); (see Sokal and Rohlf 1981). The probability associated with a particular chi-square or G value (with 1 degree of freedom) is obtained by the trapezoidal method of numerical integration.

Evolutionary Calculator

This command displays a calculator that allows computing some commonly used molecular evolutionary parameters:

a1 = $\sum$ (1 / i) from i =1 to n-1
where, n is the number of nucleotide sequences (Watterson 1975; Tajima 1989, equation 3)

a2 = $\sum$ (1 / i^2) from i =1 to n-1
where, n is the number of nucleotide sequences (Watterson 1975; Tajima 1989, equation 4)

K = (-3/ 4) Ln [1- (4p/3)] i
s the Jukes and Cantor (1969) correction, where p is the proportion of different nucleotides between two sequences.

Coalescent Simulations (1-locus | 1-pop model)

Coalescent Simulations (1-locus, 1-population model)

See Also: DNA Polymorphism Linkage Disequilibrium Population Size Changes Recombination Fu and

Li's (and other) Tests Fu and Li's (and other) Tests with an Outgroup Tajima's Test

References: Achaz 2008 Achaz 2009 Depaulis and Veuille 1998 Fay and Wu 2000 Fu and Li 1993 Fu 1997 Harpending 1994 Hudson 1983 Hudson 1990 Hudson and Kaplan 1985 Kelly 1997 Nei 1987 Press 1992 Ramos-Onsins and Mitchell-Olds 2007 Ramos-Onsins and Rozas 2002 Rozas et al. 2001 Simonsen et al. 1995 Tajima 1989 Wall 1999 Watterson 1975 Zeng et al. 2006

The Coalescent process

In this module DnaSP generates (under the coalescent and for a one-locus model), the empirical distributions of some summary test-statistics. These distributions are generated assuming different demographic scenarios. These distributions are using to compute the confidence limits for a given interval. Both one-tailed and two-tailed tests are provided.

Under this module all computer simulations are conducted given a particular value of theta (simulations given theta). That is, mutations are Poisson distributed along the lineages using the poidev routine (Press et al. 1992), and using the ran1 routine (Press et al. 1992) as a source of random numbers (random numbers uniformly distributed within a specified range).

If the user want to compute coalescent simulations given the number of segregating sites, should use the old coalescent simulation routine (Coalescent Simulations -DnaSP v5). The coalescent process could be conducted assuming No recombination (DnaSP generates the genealogy of the alleles using a modification of the routine make_tree; Hudson 1990), or under Intermediate levels of recombination (the genealogies are generated as described in Hudson 1983, 1990). The simulations are computing using the mlcoalsim (Ramos-Onsins and Mitchell-Olds, 2007) routines.

In this module DnaSP does not generates genealogies assuming Free recombination; if the user is interested in that option should use the old coalescent simulation routine (Coalescent Simulations -DnaSP v5).

Theta value (per gene)

Usually the theta value (θ) is unknown, in this case that value can be estimated from the data (see: DNA Polymorphism and Tajima's Test modules). Theta (per gene) can be estimated from:

i) k, the average number of nucleotide differences

ii) S / a1, where S is the total number of segregating sites

$a1 = \Sigma\ (1 / i)$ from i=1 to n-1

and n, the number of nucleotide sequences

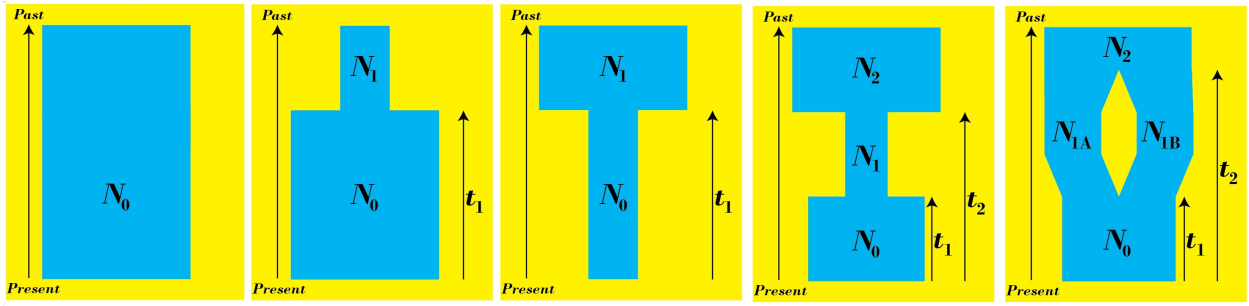
Recombination option

R, is the recombination parameter, R = 4Nr (for autosomal loci of diploid organisms), where N is the effective population size and r is the recombination rate per gene -sequence- (i.e., r is the recombination rate per generation between the most distant sites of the DNA sequence); see also: the Recombination module and Effective Population size information.

No Recombination (R = 0). It is assumed that there is no intragenic recombination (R = 0); e.g. mitochondrial DNA data.

Intermediate level of Recombination. This is the case for most nuclear genes. The user must indicate the per gene R value. The per site R value is automatically set from the number of sites.

Demographic models (Ho)



Standard Neutral Model | Population Growth | Population Decline | Population Bottleneck | Population Split/Admixture

DnaSP implements 5 different demographic models. Indeed, DnaSP estimates by coalescent simulations the empirical distribution (for each of summary statistic) under the corresponding null hypothesis (Ho). In all cases the alternative hypothesis (H1) states that the population data does not follow the specific Ho distribution.

1. Standard Neutral Model (SNM)
2. Population Growth
3. Population Decline
4: Population Bottleneck
5. Population Split and Admixture

Symbols and parameters:

t1, units of time before the present until the first demographic event. Units in 4No generations
t2, units of time before the present until the second demographic event (t2 >= t1). Units in 4No generations
No, current effective population size

1. Standard Neutral Model (SNM)

The coalescent simulations are computed assuming a large constant population size under the neutral infinite-sites model (Hudson 1990).

2. Population Growth

This model assumes that the population suffered a population growth, t1 units of time before the present. t1 is measured in 4No generations.

3. Population Decline

This model assumes that the population suffered a population decline, t1 units of time before the present. t1 is measured in 4No generations.

4: Population Bottleneck

This model assumes that the population suffered two demographic events. A population decline (t2 units of time before the present), and later a population growth (t1 units of time before the present). t1 and t2 are measured in 4No generations.

5. Population Split and Admixture

This model assumes that the population suffered two demographic events. A population split (t2 units of time before the present), generating two populations of N1A and N1B effective population sizes. Later (t1 units of time before the present) these two populations contacted and mixed (the admixture event). t1 and t2 are measured in 4No generations.

Summary statistics (on per gene basis)

DnaSP can generate the empirical distribution of the following summary statistics:

- Theta-K (θ) (Av. Number of Nucleotide Differences, k), or nucleotide diversity Pi (π) per gene (Nei 1987, equations 10.5 or 10.6; but on per gene basis).
- Theta-W (θ, from S), (Watterson 1975, equation 1.4a). This value is automatically set from the

observed number of segregating sites.

- Number of Segregating sites, S.
- Number of Haplotypes, h (Nei 1987, p. 259). See also Deapulis and Veuille 1998.
- Haplotype diversity, Hd (Nei 1987, equation 8.4 but replacing 2n by n). See also Depaulis and Veuille 1998, eq. 1. By careful; the H test defined in Depaulis and Veuille 1998 eq. 1, corresponds to: H = Hd * (n-1) / n
- Tajima's D, TD (Tajima 1989, equation 38).
- Fu and Li's D*, FLD* (Fu and Li 1993).
- Fu and Li's F*, FLF* (Fu and Li 1993, Achaz 2009).
- Achaz's Y*, AY* (Achaz 2008, equation 21).
- Fu's Fs, (Fu 1997, equation 1).
- Ramos-Onsins and Rozas's R_2 (Ramos-Onsins and Rozas 2002, equation 1).
- Raggedness, r (Harpending 1994, equation 1).
- Recombination, Rm, the minimum number of recombination events (Hudson and Kaplan 1985, Appendix 2).
- Linkage Disequilibrium, Kelly's ZnS (Kelly 1997, equation 3).
- Linkage Disequilibrium, Rozas's Za (Rozas et al. 2001; equation 2).
- Linkage Disequilibrium, Rozas's ZZ (Rozas et al. 2001; equation 1). ZZ = Za - ZnS.
- Wall's B (Wall 1999).
- Wall's Q (Wall 1999).
- Fu and Li's D, FLD (Fu and Li 1993).
- Fu and Li's F, FLF (Fu and Li 1993, Achaz 2009).
- Fay and Wu's FWHn, FWHn (Fay and Wu 2000, Zeng et al. 2006).
- Achaz Y, AY (Achaz 2008, equation 21).
- Zeng E, ZE (Zeng et al. 2006, equation 13).

Observed values

DnaSP captures the observed value of a particular statistic using information of the last analysis conducted in DnaSP. Nevertheless, the user can also (optionally) provide the observed value.
In the observed value is provided, DnaSP will also estimate the probability of obtaining values lower than the observed (one-tailed test). For example if the user indicates an observed value for the Tajima's D of TD(obs) = -1.73, and the output show that P(Sim <= Obs) = 0.01, means that the probability of obtaining Tajima's D values (under the corresponding demographic model) equal or lower than -1.73 (the observed) is 0.01.
Cells colored in yellow. Significant results (left tail); that is P(Sim <= Obs) < 0.05
Cells colored in orange. Cases where the P(Sim <= Obs) > 0.95. These cases represent extreme values (at the right tail), although not necessarily significant; indeed, the significant cases would be those with P(Sim <= Obs) >= 0.95. The user should check the values reported in the confidence interval (CI) columns to determine this feature.

Temporal Results

You can found the temporal results produced by mlcoalsim in the folder:
`Users/YourUser/AppData/Roaming/DnaSP`

Tips:

This new coalescent module does not allow to conduct the coalescent simulations using the "Free Recombination" option, and other minor things, such us computing the Fu and Li F* and F without using the Achaz (2009) equations. If you are interested in such options you can use the old coalescent simulation routine (Coalescent Simulations -DnaSP v5). This module also allows conducting coalescent simulations fixing the number of segregating sites.

Coalescent Simulations (n-loci | 1-pop model)

Coalescent Simulations (n-loci, 1-population model)

See Also: DNA Polymorphism Linkage Disequilibrium Population Size Changes Recombination Fu and Li's (and other) Tests Fu and Li's (and other) Tests with an Outgroup Tajima's Test
References: Achaz 2008 Achaz 2009 Depaulis and Veuille 1998 Fay and Wu 2000 Fu and Li 1993 Fu 1997 Harpending 1994 Hudson 1983 Hudson 1990 Hudson and Kaplan 1985 Kelly 1997 Nei 1987 Press 1992 Ramos-Onsins and Mitchell-Olds 2007 Ramos-Onsins and Rozas 2002 Rozas et al. 2001 Simonsen et al. 1995 Tajima 1989 Wall 1999 Watterson 1975 Zeng et al. 2006

The Coalescent process

In this module DnaSP generates (under the coalescent and for a n-loci model), the empirical distributions of some summary test-statistics. These distributions are generated assuming different demographic scenarios. These distributions are using to compute the confidence limits for a given interval. Both one-tailed and two-tailed tests are provided. DnaSP conducts the coalescent genealogies as in the Coalescent Simulations (1-locus; 1-Population model), using the mlcoalsim (Ramos-Onsins and Mitchell-Olds, 2007) routines.

Input data

There are two manners to input the data.

- Direct Input. The user must enter manually the relevant data for each loci: Theta per gene; sample size; the total number of sites analyzed; and the chromosomal status of each chromosomal region (autosomic, X-linked, Y-linked, mitochondrial).
- Input Data from Batch Mode. DnaSP can automatically read the relevant information from a *.MF.out file (an output file generated by DnaSP using the Multiple Data Files Analysis command (aka Batch Mode).

Chromosome parameter

DnaSP consider that the expectation of theta is $4N\mu$ for autosomal, $3N\mu$ for X-linked genes, and $N\mu$ for Y-linked (or mitochondrial) regions. See the Effective Population Sizes.

Observed values

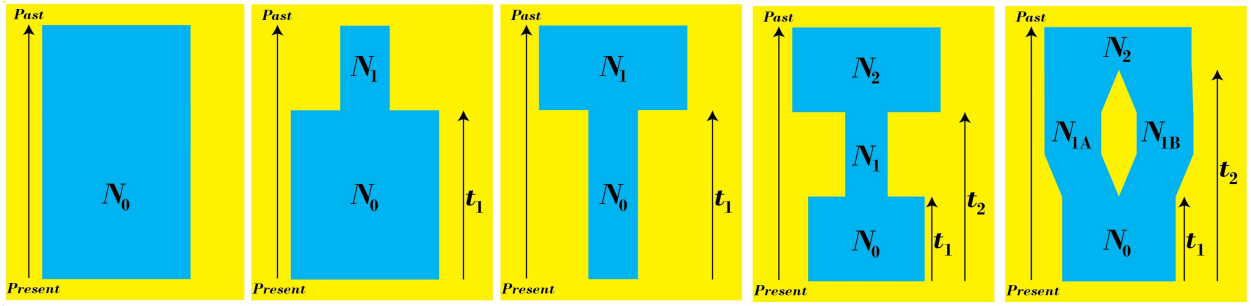
There are two manners to input the observed values (optional input).

- Direct Input. The user must enter the observed values for each statistic (the mean and variance across loci)
- Input Data from Batch Mode. DnaSP can automatically read the observed Mean and Variance values across loci from the *.MF.out (values of DNA polymorphism) or *.MFd.out (values of DNA polymorphism and Divergence) files.

Recombination values

Within this model, DnaSP assumes that there is free recombination between loci, but no intragenic (within locus) recombination.

Demographic models (Ho)



DnaSP implements 5 different demographic models. Indeed, DnaSP estimates by coalescent simulations the empirical distribution (for each of summary statistic) under the corresponding null hypothesis (Ho). In all cases the alternative hypothesis (H1) states that the population data does not follow the specific Ho distribution.

1. Standard Neutral Model (SNM)
2. Population Growth
3. Population Decline
4: Population Bottleneck
5. Population Split and Admixture

Symbols and parameters:

t1, units of time before the present until the first demographic event. Units in 4No generations
t2, units of time before the present until the second demographic event (t2 >= t1). Units in 4No generations
No, current effective population size

1. Standard Neutral Model (SNM)

The coalescent simulations are computed assuming a large constant population size under the neutral infinite-sites model (Hudson 1990).

2. Population Growth

This model assumes that the population suffered a population growth, t1 units of time before the present. t1 is measured in 4No generations.

3. Population Decline

This model assumes that the population suffered a population decline, t1 units of time before the present. t1 is measured in 4No generations.

4: Population Bottleneck

This model assumes that the population suffered two demographic events. A population decline (t2 units of time before the present), and later a population growth (t1 units of time before the present). t1 and t2 are measured in 4No generations.

5. Population Split and Admixture

This model assumes that the population suffered two demographic events. A population split (t2 units of time before the present), generating two populations of N1A and N1B effective population sizes. Later (t1 units of time before the present) these two populations contacted and mixed (the admixture event). t1 and t2 are measured in 4No generations.

Summary statistics (on per gene basis)

DnaSP can generate the empirical distribution of 12 statistics. Moreover for statistic DnaSP will determine the mean and the variance across loci. DnaSP will compute the variance using the Bessel's correction (i.e., dividing by n-1 instead of by n).

- Theta-K (θ) (Av. Number of Nucleotide Differences, k), or nucleotide diversity Pi (ρ) per gene (Nei

1987, equations 10.5 or 10.6; but on per gene basis).
- Haplotype diversity, Hd (Nei 1987, equation 8.4 but replacing 2n by n).
- Tajima's D, TD (Tajima 1989, equation 38).
- Fu and Li's F*, FLF* (Fu and Li 1993, Achaz 2009).
- Achaz's Y*, AY* (Achaz 2008, equation 21).
- Fu's Fs, (Fu 1997, equation 1).
- Ramos-Onsins and Rozas's R_2 (Ramos-Onsins and Rozas 2002, equation 1).
- Linkage Disequilibrium, Kelly's ZnS (Kelly 1997, equation 3).
- Fu and Li's F, FLF (Fu and Li 1993, Achaz 2009).
- Fay and Wu's FWHn, FWHn (Fay and Wu 2000, Zeng et al. 2006).
- Zeng E, ZE (Zeng et al. 2006, equation 13).
- Achaz Y, AY (Achaz 2008, equation 21).

Observed values

DnaSP can capture the observed values of the statistics (both the mean and the variance) using information of the *.MF.out and *.MFd.out files. If these values are provided, DnaSP will also estimate the probability of obtaining values lower than the observed (one-tailed test). For example. if the user indicates an observed value for the mean of the Tajima's D of TD(obs) = -1.73, and the output show that P(Sim <= Obs) = 0.01, means that the probability of obtaining mean values of the Tajima's D (under the corresponding demographic model) equal or lower than -1.73 (the observed) is 0.01.
Cells colored in yellow. Significant results (left tail); that is P(Sim <= Obs) < 0.05
Cells colored in orange. Cases where the P(Sim <= Obs) > 0.95. These cases represent extreme values (at the right tail), although not necessarily significant; indeed, the significant cases would be those with P(Sim <= Obs) >= 0.95. The user should check the values reported in the confidence interval (CI) columns to determine this feature.

Temporal Results

You can found the temporal results produced by mlcoalsim in the folder:
`Users/YourUser/AppData/Roaming/DnaSP`

Coalescent Simulations (DnaSP v5)

Coalescent Simulations (Method used in DnaSP v5 and earlier versions)

See Also: DNA Polymorphism Linkage Disequilibrium Population Size Changes Recombination Fu and Li's (and other) Tests Fu and Li's (and other) Tests with an Outgroup Tajima's Test
References: Depaulis and Veuille 1998 Fay and Wu 2000 Fu and Li 1993 Fu 1997 Harpending 1994 Hudson 1983 Hudson 1990 Hudson and Kaplan 1985 Kelly 1997 Nei 1987 Press 1992 Ramos-Onsins and Rozas 2002 Rozas et al. 2001 Simonsen et al. 1995 Tajima 1989 Wall 1999 Watterson 1975

This command provides the coalescent simulations module used in DnaSP version 5 (and earlier versions). In this module DnaSP generates the empirical distributions of some test-statistics using as as null hypothesis, the standard neutral model (SNM). From that distributions DnaSP can provide the confidence limits for a given interval. Both one-sided and two-sided tests can be conducted.

Statistics analysed (on per gene basis)

DnaSP can generate the empirical distribution of the following statistics:
- Haplotype diversity, Hd (Nei 1987, equation 8.4 but replacing 2n by n). (See also Depaulis and Veuille 1998, eq. 1). By careful; the H test defined in Depaulis and Veuille 1998 eq. 1, corresponds to: H = Hd *

(n-1) / n

- Number of haplotypes, h, (Nei 1987, p. 259). (see also Deapulis and Veuille 1998).
- Nucleotide diversity Pi (π) (Nei 1987, equations 10.5 or 10.6) but on per gene basis (that is, the average number of nucleotide differences).
- Theta (θ), (Watterson 1975, equation 1.4a).
- Linkage disequilibrium, ZnS statistic (Kelly 1997, equation 3).
- Linkage disequilibrium, Za statistic (Rozas et al. 2001; equation 2).
- Linkage disequilibrium, ZZ (Za - ZnS) statistic (Rozas et al. 2001; equation 1).
- Recombination, Rm, the minimum number of recombination events (Hudson and Kaplan 1985, Appendix 2).
- Tajima's D, (Tajima 1989, equation 38).
- Fu and Li's D*, (Fu and Li 1993, p. 700 bottom).
- Fu and Li's F*, (Fu and Li 1993, p. 702; see also Simonsen et al. 1995, equation 10).
- Fu and Li's D, (Fu and Li 1993, equation 32).
- Fu and Li's F, (Fu and Li 1993, p. 702, top).
- Fay and Wu's H, (Fay and Wu 2000, equations 1-3).
- Fu's Fs, (Fu 1997, equation 1).
- Raggedness, r (Harpending 1994, equation 1).
- Wall's B (Wall 1999)
- Wall's Q (Wall 1999)
- Ramos-Onsins and Rozas R_2 (Ramos-Onsins and Rozas 2002, equation 1).

The Coalescent process

The computer simulations are based on the coalescent process for a neutral infinite-sites model and assuming a large constant population size (Hudson 1990). DnaSP uses the ran1 routine (Press et al. 1992) as a source of uniform random deviates (i.e. random numbers uniformly distributed within a specified range).

No recombination. For no recombination DnaSP generates the genealogy of the alleles using a modification of the routine make_tree (Hudson 1990).

Intermediate level. For intermediate levels of recombination the genealogy is generated as described in Hudson (1983; 1990).

Free Recombination. For free recombination, DnaSP generates an independent genealogy for each segregating site. At each variable site, the number of sequences having one particular nucleotide variant (only two nucleotide variants per segregating site) is randomly obtained with probability proportional to their expected frequency (Tajima 1989, equation 50).

Simulations Given...

Theta (per gene). Mutations along the lineages are Poisson distributed using the poidev routine (Press et al. 1992).

Segregating Sites. The number of mutations (segregating sites) is fixed. Mutations are uniformly distributed (at random) along lineages.

Recombination option

No Recombination. It is assumed that there is no intragenic recombination (R = 0); e.g. mitochondrial DNA data.

Intermediate level (of Recombination). This is the case for most nuclear genes. You must indicate the value of the per gene recombination parameter (R).

Free Recombination. Maximum theoretical value of the Recombination parameter (R = ∞).

Recombination parameter, R

R, is the recombination parameter. R = 4Nr (for autosomal loci of diploid organisms), where N is the

effective population size and r is the recombination rate per gene -sequence- (i.e., r is the recombination rate per generation between the most distant sites of the DNA sequence); see also: the Recombination module and Effective Population size information.

Theta value (per gene)

Usually the theta value is unknown, in this case that value can be estimated from the data (see: DNA Polymorphism and Tajima's Test modules). Theta (per gene) can be estimated from:

i) k, the average number of nucleotide differences

ii) S / a1, where S is the total number of segregating sites

$a1 = \sum (1 / i)$ from i =1 to n-1

and n, the number of nucleotide sequences

Observed values

If the observed value is provided, DnaSP will estimate the probability of obtaining lower values than the ones observed.

For example, for the Tajima's D test statistic: P[D <= D(obs)] = 0.01

means that the probability of obtaining D values (under the neutral coalescent process) equal or lower than the observed is 0.01

Note:

You can perform computer simulations fixing the number of segregating sites. In this case the estimated values of theta (in different replicates) will be also fixed (because theta is estimated from the number of segregating sites).

Abbreviations:

(obs), observed value.

HKA test. Direct Mode

Hudson, Kreitman and Aguadé's Test (HKA Test). Direct Mode

See Also: Input Data Files Output

References: Begun and Aquadro 1991 Hudson et al. 1987 Kimura 1983

This command performs the Hudson, Kreitman and Aguadé's (1987) test (HKA test). The test is based on the Neutral Theory of Molecular Evolution (Kimura 1983) prediction that regions of the genome that evolve at high rates will also present high levels of polymorphism within species. The test requires data from an interspecific comparison of at least two regions of the genome, and also data of the intraspecific polymorphism in the same regions of at least one species.

In the present module DnaSP allows you to perform the HKA test when comparing autosomal and sex-linked regions (Begun and Aquadro 1991), or to perform the HKA test with polymorphism data with different number of sequences in the two regions, or with different number of sites for the intraspecific and interspecific comparisons. DnaSP has considered that the expectation of π is $4N\mu$ for autosomal, $3N\mu$ for X-linked genes, and $N\mu$ for Y-linked genes (so that, we have slightly modified the equations of Begun and Aquadro 1991 for comparisons involving autosomal (or X-linked) with Y-linked genes).

Effective Population size

Data:

The present module does not perform the HKA test from information of the DNA sequences included in the data file. Data from interspecific divergence and data on levels of intraspecific polymorphism must be entered in the dialog box. If you want that DnaSP obtains the information necessary to perform the HKA test directly from your sequences, you must use the HKA test module.

Output:

Estimates of the Time of divergence (measured in 2N generations, where N is the effective population size),
Estimates of theta (θ) per nucleotide in region (locus) 1,
Estimates of theta (θ) per nucleotide in region (locus) 2,
The X-square value, and the statistical significance.

Statistical significance:

The statistical significance is obtained assuming a χ-square distribution with one degree of freedom. DnaSP obtains the probability associated with a particular chi-square value (with 1 degree of freedom) by the trapezoidal method of numerical integration.
(# P < 0.10; * P < 0.05; ** P < 0.01; *** P, < 0.001).

Window & Help Menus

Window & Help Menus

This menu is provided with the following four commands:

Window Menu
Use this command to change the active window (windows with results, calculator, sequence data). The active window is the window that appears in the foreground.

Help Menu

Contents
This command provides information for using DnaSP (the commands open the present help file).

Search For Help on
This command displays Help's Search dialog box, where you can quickly find the information that you need by keywords.

DnaSP Bug Reports
This command displays DnaSP Bug Reports Web page.

Citation
This command displays a dialog box with the suggested citation for DnaSP.

DnaSP Home Page
This command displays DnaSP Web page.

About DnaSP
This command displays a dialog box with information about authors, and the DnaSP version number.

More Information & Distribution & Copyright

More Information

Distribution Policies & Copyright

For other uses, please get into contact with Julio Rozas:
E-mail: jrozas@ub.edu

Queries, comments and suggestions may be addressed via E-mail to Julio Rozas.

Availability

The program, the help file and some examples of the different data files are available from:
http://www.ub.es/dnasp

DnaSP updates and Bug reports will be advertised in the:
DnaSP Web in the Departament de Genètica, Universitat de Barcelona Web:
http://www.ub.es/dnasp

References

Citation

Citation and DnaSP papers

Abstracts: DnaSP v1 DnaSP v2 DnaSP v3 DnaSP v4 DnaSP v5

The suggested citation for the DnaSP version 6 is:

Julio Rozas[1], Albert Ferrer-Mata[1], Juan Carlos Sánchez-DelBarrio[1], Sara Guirao-Rico[2], Pablo Librado[1,3], Sebastián E. Ramos-Onsins[2] and Alejandro Sánchez-Gracia[1].
DnaSP v6: DNA Sequence Polymorphism Analysis of Large Datasets.
Mol. Biol. Evol. **34:** 3299-3302 (2017).

[1] Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona
[2] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB
[3] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

Authors

Authors

The DnaSP development started in 1992, and since then many people have contributed with the software

development:

List of authors in order of their first contribution:
Julio Rozas
Ricardo Rozas
Juan Carlos Sánchez-DelBarrio
Pablo Librado
Sara Guirao-Rico
Alejandro Sánchez-Gracia
Sebastian Ramos-Onsins
Albert Ferrer-Mata

Acknowledgements

Acknowledgements

Our thanks are due to the following people who made comments and suggestions, or tested the DnaSP program with their data. Particularly, we would like to thank those who are (or were) in the the Molecular Evolutionary Genetics group at the Departament de Genètica, Universitat de Barcelona:

M. Aguadé, D. Alvarez-Ponce, M. Alvarez-Presas (Ona), C. Arboleda, D. Balañà, A. Blanco-García, J. Braverman, J. L. Campos, S. Cirera, J. M. Comeron, D. De Lorenzo, T. Guebitz, S. Guirao-Rico, N. Khadem, S. O. Kolokotronis, H. Kuittinen, A. Llopart, J. M. Martín-Campos, A. Munté, A. Navarro-Sabaté, C. Nobrega, D. Orengo, M. Papaceit, J. Pérez, I. Pires, R. Pratdesaba, H. Quesada, U. Ramírez, S. E. Ramos-Onsins, C. Romero-Ibáñez, A. Sánchez-Gracia, C. Segarra, F. G. Vieira, A. G. Vilella.

Apart from the mentioned, special thanks are due to H. Akashi, A. Barbadilla, J. Bertranpetit, E. Betrán, C. H. Biermann, M. Blouin, F. Calafell, J. Castresana, F. González-Candelas, D. Govindaraju, R. R. Hudson, P. de Knijff, T. Mes, A. Navarro, D. Posada, C. Robin, A. P. Rooney, S. Schaeffer, W. Stephan, S. Wells and R. Zardoya for their comments, suggestions and help.

Finally, we also acknowledge D. R. Maddison for providing advice about the NEXUS file formats and for supplying us with precise instructions on this format.

This work was supported by the Dirección General de Investigación Científica y Técnica, The Ministerio de Educación y Ciencia and the Ministerio de Economía y Competitividad of Spain (grants PB91-0245, PB94-0923, PB97-0918, TXT98-1802, BMC2001-2906, BFU2004-02253, BFU2007-62927, BFU2010-15484, CGL2013-45211, CGL2016-75255).

References

References

ACHAZ, G. (2008). Testing for neutrality in samples with sequencing errors. Genetics 179: 1409-1424.

ACHAZ, G. (2009). Frequency spectrum neutrality tests: One for all and all for one. Genetics 183: 249-258.

AKASHI, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067-1076.

AKASHI, H. (1999). Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. Genetics 151: 221-238.

AKASHI, H. and W. SCHAEFFER. (1997). Natural selection and the frequency distributions of "silent" DNA polymorphisms in Drosophila. Genetics 146: 295-307.

BANDELT, H.-J., P. FORSTER and A. RÖHL, (1999). Median-Joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16: 37-48.

BEGUN, D. J. and C. F. AQUADRO. (1991). Molecular population genetics of the distal portion of the X chromosome in Drosophila: Evidence for genetic hitchhiking of the yellow-achaete region. Genetics 129: 1147-1158.

BETRÁN, E., J. ROZAS, A. NAVARRO and A. BARBADILLA. (1997). The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146: 89-99.

CATCHEN, J. M. et al. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. G3 (Bethesda) 1: 171–182.

DEPAULIS, F. and M. VEUILLE. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15: 1788-1790.

DANECEK et al. (2011). The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

DURET, L. and D. MOUCHIROUD. (1999). Expression pattern and, surprisingly,gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA 96: 4482-4487.

EATON, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30: 1844–1849.

EXCOFFIER, L. and H. E. L. LISCHER. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol Resources 10: 564-567.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles.Theor. Pop. Biol. 3: 87-112.

FAY, J. C. and C. I. WU. (2000). Hitchhiking under positive darwinian selection. Genetics 155: 1405-1413.

FAY, J., WYKCOFF, G. J. and WU, C. I. (2001). Positive and negative selection on the human genome. Genetics 158: 1227-1234.

FELSENSTEIN, J. (1993). Phylogeny Inference Package (PHYLIP). Version 3.5. University of Washington, Seattle.

FRÍAS-LÓPEZ, C., J. F. SÁNCHEZ-HERRERO, S. GUIRAO-RICO, E. MORA, M. A. ARNEDO, A. SÁNCHEZ-GRACIA and J. ROZAS. (2016). DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. Bioinformatics 32: 3753-3759.

FU, Y.-X. (1995). Statistical properties of segregating sites. Theor. Pop. Biol. 48: 172-197.

FU, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925.

FU, Y.-X. and W.-H. LI. (1993). Statistical tests of neutrality of mutations.Genetics 133: 693-709.

GILBERT, D. (1996). A biological sequence editor and analysis program. Indiana University.

HEY, J. (1991). The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics 128: 831-840.

HEY, J. and J. WAKELEY. (1997). A coalescent estimator of the population recombination rate. Genetics 145: 833-846.

HARPENDING, H. (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Human Biology 66: 591-600.

HILL, W. G. and A. ROBERTSON. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226-231.

HUDSON, R. R. (1983). Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. 23: 183-201.

HUDSON, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50: 245-250.

HUDSON, R. R. (1990). Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7: 1-44.

HUDSON, R. R. (2000). A new statistic for detecting genetic differentiation. Genetics 155: 2011-2014.

HUDSON, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

HUDSON, R. R. and N. L. KAPLAN. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.

HUDSON, R. R., M. KREITMAN and M. AGUADE. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

HUDSON, R. R., BOOS, D.D. and N. L. KAPLAN. (1992). A statistical test for detecting population subdivision. Mol. Biol. Evol. 9: 138-151.

HUDSON, R. R., M. SLATKIN and W. P. MADDISON. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583-589.

HUTTER, S., VILELLA, A. and ROZAS, J. (2006). Genome-wide DNA polymorphism analysis using VariScan. Bioinformatics 7: 409-419.

JUKES, T. H. and C. R. CANTOR. (1969). Evolution of protein molecules, pp.21-132. In H. N. Munro (ed.), Mammalian Protein Metabolism. Academic Press, New York.

KANAYA, S., Y. YAMADA, Y. KUDO, and T. IKEMURA. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and

species-specific diversity of codon usage based on multivariate analysis. Gene 238: 143-155.

KELLY, J. K. (1997). A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

KENT, W.J., SUGNET, C.W., FUREY, T.S., ROSKIN, K.M., PRINGLE, T.H., ZAHLER, A.M., HAUSSLER, D. (2002). The human genome browser at UCSC. Genome Research. 12: 996-1006.

KENT, W. J. (2002). BLAT- The Blast-like alignment tool. Genome Research. 12: 656-664.

KIMURA, M. (1983). The neutral theory of Molecular Evolution. Cambridge University Press, Cambridge, Massachusetts.

KREITMAN, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature 304: 412-417.

KUMAR, S., K. TAMURA and M. NEI. (1994). MEGA: Molecular Evolutionary GeneticsAnalysis software for microcomputers. Comput. Applic. Biosci. 10: 189-191.

LANGLEY, C. H., Y. N. TOBARI and K. KOJIMA. (1974). Linkage disequilibrium i natural populations of Drosophila melanogaster. Genetics 78: 921-936.

LEWONTIN, R. C. (1964). The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49-67.

LEWONTIN, R. C. and K. KOJIMA. (1960). The evolutionary dynamics of complex polymorphisms. Evolution 14: 458-472.

LIBRADO, p. and J. ROZAS. (2009). A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

LYNCH, M. and T. J. CREASE, (1990). The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. 7: 377-394.

McDONALD, J. H. (1996). Detecting Non-neutral heterogeneity across a region of DNA sequence in the ratio of Polymorphism to divergence. Mol. Biol. Evol. 13: 253-260.

McDONALD, J. H. (1998). Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. 15: 377-384.

MADDISON, W. P. and D. R. MADDISON. (1992). MacClade: Analysis of phylogeny and character evolution. Version 3. Sinauer Associates, Sunderland, Massachusetts.

MADDISON, W. P., D. L. SWOFFORD and D. R. MADDISON. (1997). NEXUS: an extensible file format for systematic information. System. Biol. 46: 590-621.

MAXAM, A. M. and W. GILBERT. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74: 560-564.

McDONALD, J. H. and M. KREITMAN. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

MORTON, B. R. (1993). Chloroplast DNA codon use: Evidence for selection at the psb A locus based on tRNA availability. J. Mol. Evol. 37: 273-280.

NEI, M. (1973). Analysis of gene diversity in subdivided populations. Proc.Natl. Acad. Sci. USA 70: 3321-3323.

NEI, M. (1982). Evolution of human races at the gene level, pp. 167-181. In B. Bonne-Tamir, T. Cohen, and R. M. Goodman (eds.), Human genetics, part A: The unfolding genome. Alan R. Liss, New York.

NEI, M. (1987). Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

NEI, M. and T. GOJOBORI. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418-426.

NEI, M. and J. C. MILLER. (1990). A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics 125: 873-879.

OSAWA, S., T. H. JUKES, K. WATANABE and A. MUTO. (1992). Recent evidence forEvolution of the genetic code. Microbiol. Rev. 56: 229-264.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING and B. P. FLANNERY. (1992).Numerical recipes in C. The art of Scientific Computing. Cambridge University Press, Cambridge.

RAMOS-ONSINS, S. E. and MITCHELL-OLDS, T. (2007). MIcoalsim: multilocus coalescent simulations. Evol. Bioinform. Online. 3: 41–44.

RAMOS-ONSINS, S. E. and J. ROZAS. (2002). Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. 19: 2092-2100.

RAND, D. M. and L. M. KANN. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol. Biol. Evol. 13: 735-748.

ROGERS, A. R. (1995). Genetic evidence for a pleistocene population explosion. Evolution 49: 608-615.

ROGERS, A. R. and H. HARPENDING. (1992). Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9: 552-569.

ROGERS, A. R., A. E. FRALEY, M. J. BAMSHAD, W. SCOTT WATKINS, and L. B. JORDE.(1996). Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. 13: 895-902.

ROZAS, J. and M. AGUADE. (1993). Transfer of genetic information in the rp49 region of Drosophila subobscura between different chromosomal gene arrangements. Proc. Natl. Acad. Sci. USA 90: 8083-8087.

ROZAS, J. and M. AGUADE. (1994). Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of Drosophila. Proc. Natl. Acad. Sci. USA 91: 11517-11521.

ROZAS, J. and R. ROZAS. (1995). DnaSP, DNA sequence polymorphism: an interactive program for estimating Population Genetics parameters from DNA sequence data. Comput. Applic. Biosci. 11: 621-625.

ROZAS, J. and R. ROZAS. (1997). DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. 13: 307-311.

ROZAS, J. and R. ROZAS. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15: 174-175.

ROZAS, J., M. GULLAUD, G. BLANDIN and M. AGUADÉ. (2001). DNA variation at the rp49 gene region of Drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147-1155.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

SAIKI, R. K., S. SCHARF, F. FALOONA, K. B. MULLIS, G. T. HORN, H. A. ERLICH and N. ARNHEIM. (1985). Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230: 1350-1354.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN,K. B. MULLIS and H. A. ERLICH. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.

SANGER, F., S. NICKLEN and A. R. COULSON. (1977): DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74: 5463-5467.

SCHAEFFER, S. W. and E. L. MILLER. (1993). Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of Drosophila pseudoobscura. Genetics 135: 541-552.

SCHNEIDER, S., ROESSLI, D., AND EXCOFFIER, L. (2000). Arlequin: A software for population genetics data analysis. Ver 2.001. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

SHARP, P. M., T. M. F. TUOHY and K. R. MOSURSKI. (1986). Codon usage in yeast:Cluster analysis clearly differentiates highly and lowly expressed genes.Nucleic Acids Res. 14: 5125-5143.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and F. WRIGHT. (1988). "Silent"sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons. Mol. Biol. Evol. 5: 704-716.

SIDMAN, K. E., D. G. GEORGE, W. C. BARKER and L. T. HUNT. (1988). The protein identification resource (PIR). Nucleic Acids Res. 16: 1869-1871.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

SLATKIN, M. and R. R. HUDSON. (1991). Pairwise comparisons of mitochondrialDNA sequences in stable and exponentially growing populations. Genetics 129: 555-562.

SOKAL, R. R. and F. J. ROHLF. (1981). Biometry. Second Edition. W. H. Freeman and Company. New York.

SCHEET, P. and STEPHENS, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics, 78: 629-644.

STEPHENS, M. and DONNELLY, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics, 73, 1162-1169.

STEPHENS, M., SMITH, N. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics, 68, 978-989.

STROBECK, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117: 149-153.

SWOFFORD, D. L. (1991). PAUP: phylogenetic analysis using parsimony, version3.0. Illinois Natural History Survey, Champaign.

TAJIMA, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.

TAJIMA, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

TAJIMA, F. (1989). The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

TAJIMA, F. (1993). Measurement of DNA polymorphism, pp. 37-59. In Takahata,N. and Clark, A. G. (eds), Mechanisms of Molecular Evolution, Sinauer Associates. Inc., Sunderland, Massachusetts.

TAJIMA, F. (1996). The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143: 1457-1465.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON. (1994). CLUSTAL W: improving the sensitivity of progressive sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

VINGRON, M., BRAZMA, A., COULSON, R., VAN HELDEN, J., MANKE, T., PALIN, K., SAND, O. and UKKONEN, E. (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. Genome Biology 10: 202-209.

WAKELEY, J. and J. HEY. (1997). Estimating ancestral population parameters.Genetics 145: 847-855.

WALL, J. D. (1999). Recombination and the power of statistical tests of neutrality. Genet Res 74: 65-69.

WANG, L. S. and XU,Y. (2003) Haplotype inference by maximum parsimony. Bioinformatics 19: 1773–1780.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. 7: 256-276.

WEIR, B. S. (1996). Genetic Data Analysis II. Sinauer Associates, Inc. Sunderland.

WRIGHT, S. (1951). The genetical structure of populations. Ann. Eugenics 15: 323-354.

WRIGHT, F. (1990). The "effective number of codons" used in a gene. Gene 87: 23-29.

ZENG, K., FU, Y., SHI, S. and WU, C. (2006). Statistical tests fir detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-1439