Chapter 17

DNA Sequence Polymorphism Analysis Using DnaSP

Julio Rozas

Abstract

The analysis of DNA sequence polymorphisms and SNPs (single nucleotide polymorphisms) can provide insights into the evolutionary forces acting on populations and species. Available population-genetic methods, and particularly those based on the coalescent theory, have become the primary framework to analyze such DNA polymorphism data. Here, I explain some essential analytical methods for interpreting DNA polymorphism data and also describe the basic functionalities of the DnaSP software. DnaSP is a multi-propose program that allows conducting exhaustive DNA polymorphism analysis using a graphical user-friendly interface.

Key words: Molecular population genetics software, DNA sequence polymorphisms, SNPs, Neutrality tests, Coalescent methods, DnaSP.

1. Introduction

The analysis of DNA sequence polymorphisms and SNPs (single nucleotide polymorphisms) can provide insights into the evolutionary significance of DNA polymorphisms, and into the selective and demographic factors acting on populations and species *(1–3)*. DNA polymorphism data, furthermore, are invaluable powerful tools (as molecular markers) in a wide range of disciplines such as biomedicine, animal and plant breeding, conservation genetics, epidemiology genetics, or forensics.

Modern high-throughput DNA sequencing and polymorphism detection methodologies are generating huge high-quality DNA sequence variation and SNPs data sets. This massive amount of data has stimulated the development of bioinformatics and analytical methods for handling, analyzing, and interpreting

David Posada (ed.), *Bioinformatics for DNA Sequence Analysis,* Methods in Molecular Biology 537

DOI 10.1007/978-1-59745-251-9_17

DNA polymorphism information. Current population genetics methods, and particularly those based on the coalescent theory, have become the primary framework to analyze DNA polymorphism data *(3, 4)*. Specifically, the comparative analysis of within-species DNA polymorphism and between-species variation is noticeably an effective approach to understand the evolutionary process and to obtain insights into the functional significance of genomic regions. In this context, the detection of both positive and negative selection is of major interest.

At present, there is a wealth of freely available computer programs for analyzing DNA polymorphism data (for a review *see* ref. *5*). Here, I explain the basic analytical methods implemented in DnaSP *(6)*; this software package is a multi-propose program that allows conducting exhaustive analysis using a graphical user-friendly interface. The main features of the software are as follows: (i) accommodates large data sets; (ii) computes many population genetic statistics describing the level and patterns of DNA polymorphism within and between populations; (iii) conducts computer simulation coalescent-based tests; (iv) generates graphical outputs rendering the information readily understandable.

2. Program Usage

2.1. Data Files

DnaSP accepts five input data file formats: FASTA, MEGA, NBRF/PIR, NEXUS, and PHYLIP *(6, 7)*. The data files should store a multiple DNA sequence alignment with polymorphism data (within-species variation), interspecific nucleotide variation (between-species variation), or any combination of both (*see* **Note 1**). Since all formats are in plain ASCII (text) files can, therefore, be viewed and edited in any plain-text editor. The software allows exporting and converting data files in the above mentioned formats, as well as in the format used by Arlequin *(8)* and NETWORK *(9)*.

2.2. Managing Data Information

Before conducting an analysis the data should be adequately prepared. For instance, to compute any statistic dealing with the number of synonymous substitutions, the user should specify the coding region positions, the genetic code, and the reading frame. DnaSP provides a user-friendly graphic interface (**Fig. 17.1**), where the user can specify these and many other features (codon preferences, genome type, ingroups and outgroups, chromosomal location, sites and sequence subsets, etc.). Certainly, not all analyses will require all data specifications; it is very convenient, however, to first define these features and save them on a NEXUS file format *(10)*. As the NEXUS format

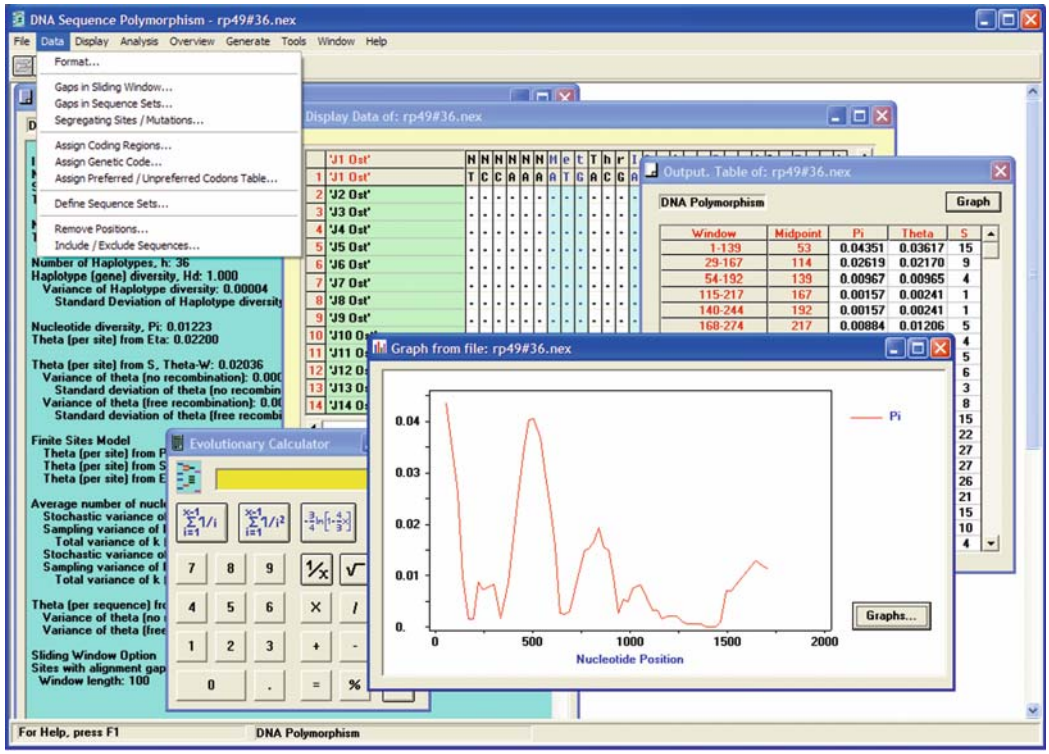


Fig. 17.1. DnaSP graphical user interface.

allows storing such information, the user no longer needs to define these data features again. Most importantly, data specifications written in DnaSP may be read by other population genetics and molecular evolution programs that make use of NEXUS files, such as MacClade *(11)*, PAUP*(12)*, and MEGA*(13)*; (*see* also ref. *5*) (*see* **Note 2**). DnaSP also includes a convenient DNA sequence browser, where the user can highlight alignment features such as polymorphic (variable) sites, singletons, parsimony-informative sites, invariant sites, synonymous and non-synonymous substitution sites, etc.

2.3. Analyses

The DnaSP software conducts exhaustive molecular population genetic analyses including those based on the coalescent theory *(4, 6)*. The software (i) measures the levels of DNA sequence polymorphism and SNPs (within and between populations), and divergence levels between species; (ii) estimates variation in synonymous and non-synonymous sites; (iii) analyzes the patterns of linkage disequilibrium, gene flow, and recombination; and (iv) computes the *P*-values of a number of neutrality tests using coalescent simulations. Furthermore, many of these analyses can be performed by the sliding window method, and

plotted. In this section, I describe some of the most representative and characteristic summary statistics and methods employed in the interpretation of DNA polymorphism data.

2.3.1. Summary Statistics

DnaSP computes most commonly used statistics quantifying the levels of DNA polymorphism *(14)*, such as the number of segregating sites, the average number of nucleotide differences, the number of haplotypes, and haplotype diversity, to analyze the distribution pattern of DNA variation, or to compare alternative evolutionary scenarios.

2.3.1.1. Site-by-Site Based Statistics

The number of segregating sites (S) is just the number of variable positions (i.e., polymorphic) in a sample of DNA sequences. Since the statistic does not utilize the information of the frequency of nucleotide variants, it is very sensitive to the sample size.

Nucleotide diversity (π), or the average number of nucleotide differences per site (i.e., the probability that two random sequences are different at a given site), is defined as

$$\pi = k/m \tag{1}$$

where m is the total number of nucleotide positions (including monomorphic positions, but excluding sites with alignment gaps) and k (often denoted as Π) is the mean number of nucleotide differences (*see* **Note 3**):

$$k = \frac{2}{n(n-1)} \sum_{i<j} d_{ij} \tag{2}$$

where n is the number of sequences (i.e., the sample size) and d_{ij} is the number of nucleotide differences between sequences i and j. The mean number of nucleotide differences can also be computed as

$$k = \sum_{i=1}^{m} h_i \tag{3}$$

where h_i is the heterozygosity at site i, which can be estimated as

$$h_i = \frac{n}{n-1} \left(1 - \sum_{j=1}^{4} x_{ij}^2 \right) \tag{4}$$

where x_{ij} is the relative frequency of nucleotide variant j ($j = 1, 2, 3$, and 4 correspond to A, C, G, and T) at site i. For large nucleotide diversity values ($\pi > 0.1$) it is convenient to apply the Jukes and Cantor multiple-substitution correction *(15)*. DnaSP allows computing these statistics primarily in the *Analysis|DNA Polymorphism* and *Analysis|DNA Divergence Between Populations* commands (*see* **Note 4**). DnaSP also provides a graphical representation of the site-frequency spectrum, i.e., the frequency distribution of segregating sites [*Analysis|Population Size Changes* command *Segregating Sites* option] (**Fig. 17.2**).

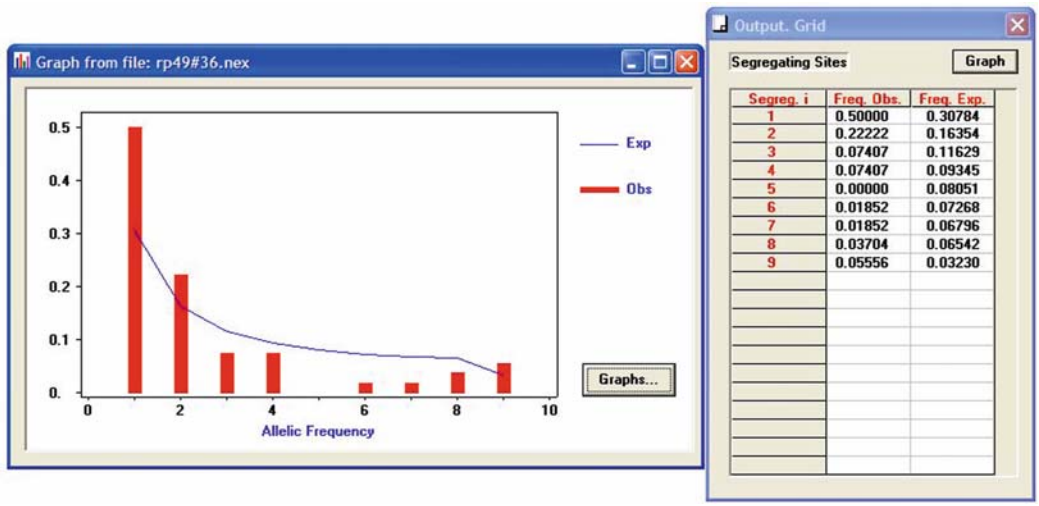


Fig. 17.2. Folded frequency spectrum of segregating sites.

2.3.1.2. Haplotype-Based Statistics

This category includes summary statistics that capture haplotype sequence information, for instance, haplotype diversity, linkage disequilibrium, and recombination statistics. Therefore, phased haplotype data is required. Hence, genotype (unphased) data – current routine SNP genotyping methods normally do not provide phase information – must be previously inferred to calculate these type of statistics; this step cannot be conducted with the current DnaSP version (*see* **Note 5**) and should be performed using other available algorithms and software *(16)*.

Haplotype diversity (*H*) – also known as gene/allele diversity or expected heterozygosity – is the probability that two random sequences are different, and is defined as

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{h} p_i^2\right) \tag{5}$$

where n is the number of sequences, h (often seen in the literature as K) the number of haplotypes (i.e., different DNA sequences), and p_i is the relative frequency of haplotype i. DnaSP calculates the haplotype diversity statistic in *Analysis|DNA Polymorphism*, *Analysis|Gene Flow and Genetic Differentiation*, and *Generate|Haplotype Data File* commands.

2.3.1.3. Mismatch Distribution-Based Statistics

This category includes descriptive statistics based on the mismatch distribution, namely the distribution of the number of differences between pairs of DNA sequences *(17,18)*. In spite of the popular nature and usefulness of these statistics in the analysis of past demographic events, they are very conservative, especially for recombining DNA regions *(19)*.

The raggedness *r* statistic *(20)* captures information of the shape of the mismatch distribution, which is affected by past population events. More specifically, *r* values differ between constant-size and growing populations. DnaSP provides a module to calculate *r* – and other commonly used statistics – [*Analysis*|*Population Size Changes* command], to estimate the relevant population parameters under a population growth scenario [*Model for Expected values* option] and to visualize the empirical and theoretical mismatch distributions (**Fig. 17.3**).

2.3.1.4. Neutrality-Based Statistical Tests

A neutrality test is a statistical method designed to test the neutral hypothesis, i.e., that all mutations are either neutral or strongly deleterious. There is a great variety of tests, which can be classified into different non-exclusive categories: (1) tests drawing information from synonymous non-synonymous substitutions (such as McDonald-Kreitman or d_N/d_S-based tests) *(21, 22)*; (2) tests that use only polymorphism (such as Tajima or Fu and Li's tests) *(23, 24)* or polymorphism and divergence data (such as Hudson, Kreitman, and Aguadé, or Fay and Wu tests) *(25, 26)* (*see* **Note 6**).

Here I will describe just one of the most popular DNA sequence-based statistical test, Tajima's *D* test *(23)*. Tajima's test employs polymorphism frequency spectrum data without taking into account synonymous and non-synonymous substitution information. Tajima's *D* statistic is defined as the standardized difference between two estimators of the population mutation rate parameter θ. For autosomal regions of diploid individuals $\theta = 4N\mu$, where N is the effective population size, and μ is the per-generation mutation rate. Specifically,

$$D = \frac{k - S/a_n}{\sqrt{Var(k - S/a_n)}} \tag{6}$$

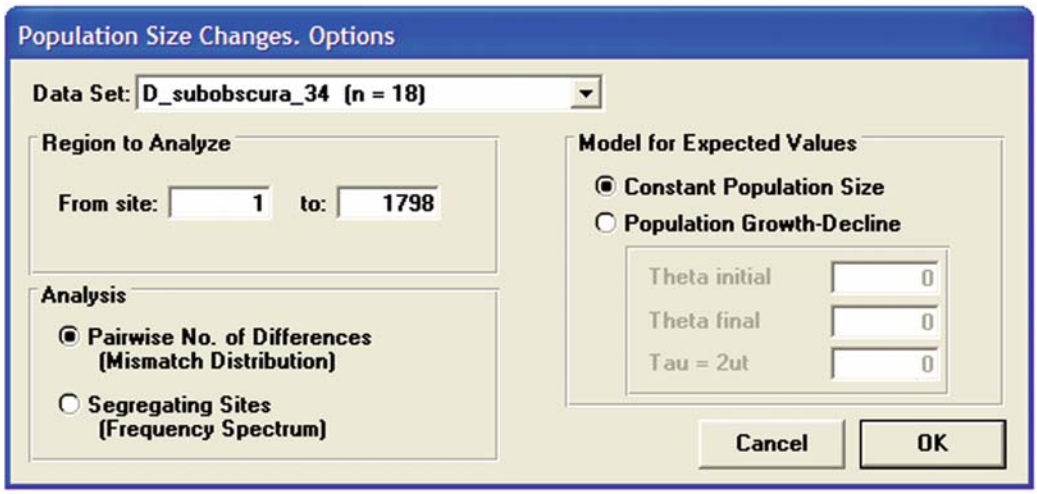
where

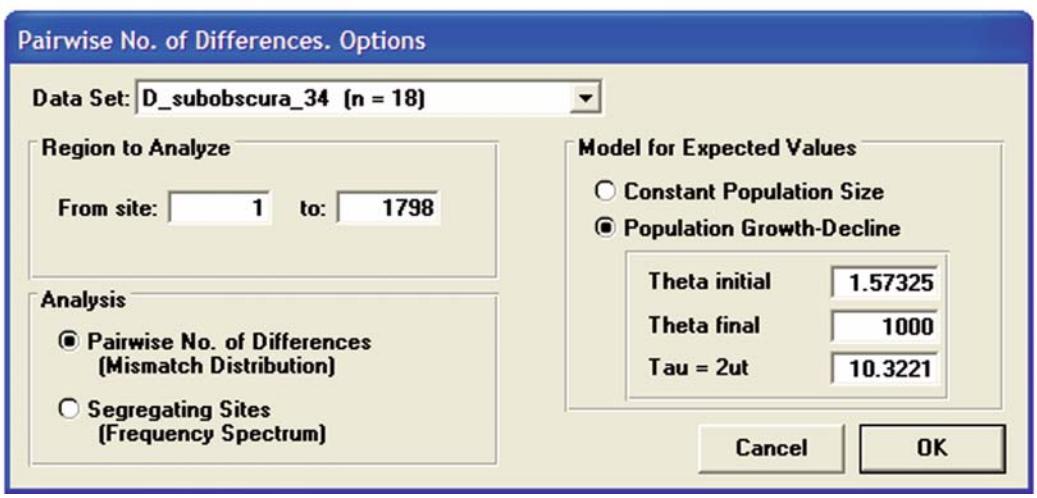
$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \tag{7}$$

Since under the standard neutral model the expected values of k and S are

$$E[k] = \theta \tag{8}$$
$$E[S]/a_n = \theta$$

the two estimates of θ (θ estimates provided by equations (8) and (9) are often symbolized as θ_π and θ_W, respectively) should provide roughly equal values and therefore D should be close to zero. To use this statistic as a test – particularly to contrast if a D value is significantly different from zero – it is necessary to know the





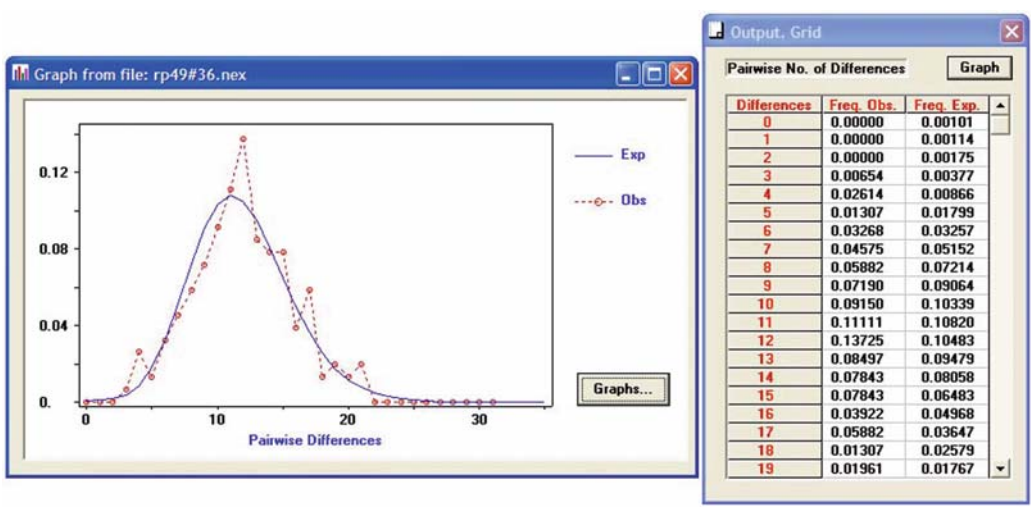


Fig. 17.3. Mismatch distribution plot. (**A**) Options for the first run. (**B**) Options for the second run. (**C**) Outcome results.

distribution of *D*. This distribution is usually obtained by computer simulations based on the coalescent process (see below). DnaSP provides specific modules to conduct these analyses [*Analysis*|*Tajima's Test* and *Tools*|*Coalescent Simulations* commands].

2.3.2. Sliding Window

The sliding window is a useful tool for locating genomic regions with atypical patterns of variation (*see* also ref. *27*); in DnaSP this method is implemented for a number of statistics [*Sliding Window*|-*Compute* option]. For instance, DnaSP can compute π values along a sliding window thus providing a graphical representation of the results (**Fig. 17.4**). Following equation (1), nucleotide diversity can also be calculated for synonymous and non-synonymous sites [*Analysis*|*Synonymous and NonSynonymous substitutions* command]; in addition, synonymous and non-synonymous variation within species (polymorphism) can be compared with between species (divergence) levels [*Analysis*|*Polymorphism and Divergence* command]. All these features are very useful for exploratory data analyses – such as detecting genomic regions with unusual patterns of variation – which, in turn, can provide insights into the functional constraints acting on genes or genomic regions.

2.3.3. Coalescent Simulations

The coalescent theory is a stochastic population-genetics model describing the statistical properties of gene trees *(4, 28)*. The coalescent provides very suitable methods for interpreting DNA polymorphism data – in fact it underlies the development of most neutrality tests– and for conducting efficient computer simulations. Indeed, the coalescent allows simulating samples under several different models (neutral, but also demographic or selective). These methods are essential for the detection of the signature of positive natural selection and for the distinction from the similar patterns generated by other (e.g., demographic) processes.

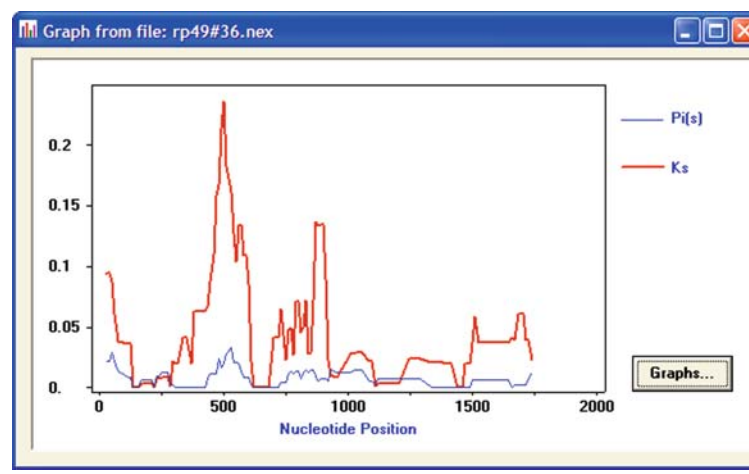


Fig. 17.4. Sliding-window plot of polymorphism and divergence levels.

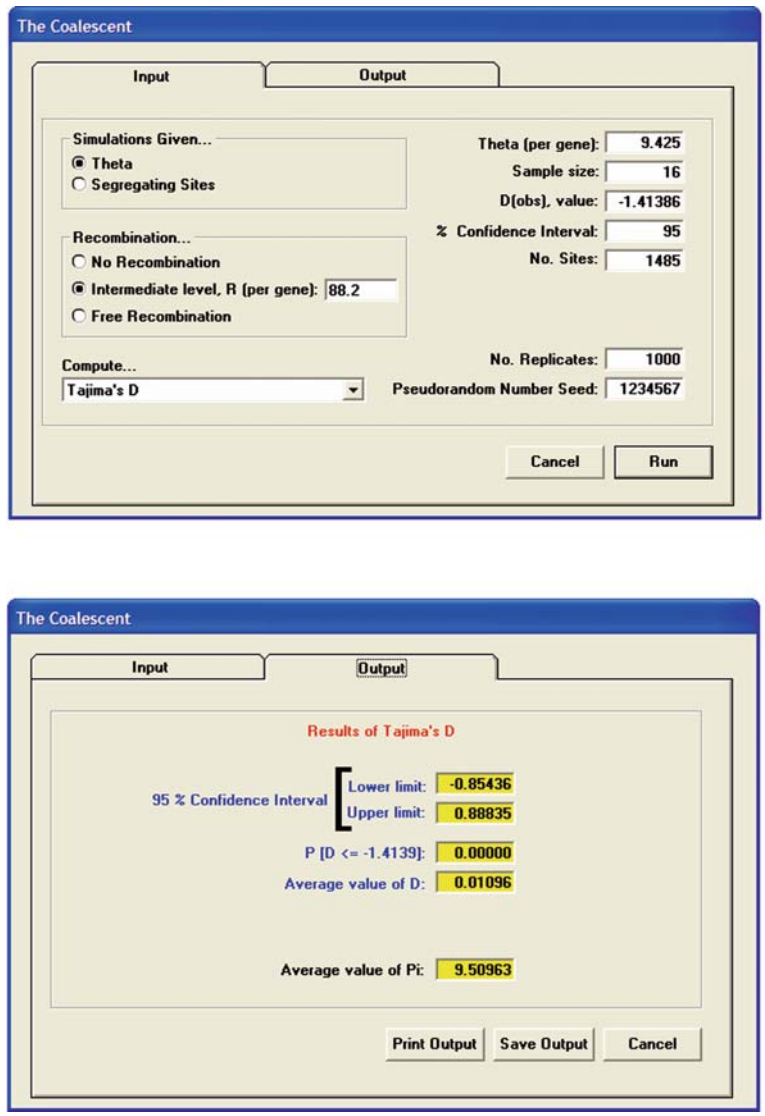


Fig. 17.5. Determining the confidence interval of Tajima's *D* by coalescent simulations. (**A**) Input Tab. (**B**) Output Tab.

DnaSP incorporates a specific module [*Tools*|*Coalescent Simulations*] for conducting coalescent-based computer simulations (**Fig. 17.5**). Gene trees are simulated under the null neutral coalescent model by applying different recombination rate values ($R = 4Nr$, where N is the effective population size, and r is the per-generation recombination rate between the most distant sites), and by fixing either the value of θ or the number of segregating sites S (*see* **Notes 7, 8**). Simulated samples are used to estimate the

empirical distribution of a wide variety of statistics, which, in turn, allow determining the confidence interval of the statistical tests, and conducting one-tailed or two-tailed tests (*see* **Notes 9, 10**).

3. Examples

The DnaSP software can be downloaded for free from http://www.ub.es/dnasp. The software package includes documentation (a windows help file) and some sample data sets. Here, I will use the rp49#36.nex file (data file distributed with the DnaSP software) as an example. The data set includes 36 multiple aligned DNA sequences of the ribosomal protein 49 gene region (~1.8 kb) from three species of Drosophila (ref. *29*). The data set contains four data subsets: D_subobscura_st (16 sequences of chromosomal arrangement O_{ST} from *D. subobscura*), D_subobscura_34 (18 sequences of chromosomal arrangement O_{3+4} from *D. subobscura*), D_madeirensis (one sequence from *D. madeirensis*), and D_guanche (one sequence from *D. guanche*).

3.1. Frequency Spectrum of Segregating Sites

Figure 17.2 shows a graph of the folded frequency spectrum of segregating sites [*Analysis*|*Population Size Changes* command *Segregating Sites* option]. The analysis was conducted using the D_subobscura_34 ($n = 18$) subset which contains $S = 54$ segregating sites. The *x*-axis of the graph represents the number of individuals carrying the least frequent nucleotide variant. The *y*-axis represents the proportion of segregating sites in the sample. For example, there are 12 polymorphic sites (12/54 = 0.2222) where the least frequent nucleotide variant is segregating in just two individuals. The expected values represent the values under the standard neutral model, i.e., a stationary constant-size population (ref. *23*; equation 50) (*see* **Note 11**).

3.2. Mismatch Distribution Analyses

The analysis was conducting using the D_subobscura_34 ($n = 18$) subset (**Fig. 17.3**). DnaSP generates the theoretical population growth distribution by a two run steps procedure [*Analysis*|*Population Size Changes* command]. On the first run (**Fig. 17.3A**) the software will estimate the relevant population growth parameters (*see* **Note 12**); in the second run (**Fig. 17.3B**) DnaSP will conduct the final analysis. The total number of pairs of sequences is $n(n-1)/2 = 153$ and the raggedness value $r = 0.0103$. The abscissa (**Fig. 17.3C**) gives the number of differences between pairs of DNA sequences (i.e., pairwise differences), and the ordinate gives the fraction of pairs that differ by that number. For instance, there are 14 pairs of sequences (14/153 = 0.0915; *y*-axis) differing by 10

differences (*x*-axis). The expected values are calculated assuming the sudden population growth model (ref. *18*; equation 4), and using the following parameter estimates: θ_0 = 1.573, θ_1 = 1,000, $\tau = 10.322$.

3.3. Sliding Window Analyses

Fig. 17.4 shows a sliding-window plot of the polymorphism (intraspecific data: *D_subobscura_st* subset) and divergence (interspecific data: *D_guanche* subset) [*Analysis*|*Polymorphism and Divergence* command]. The subset data file includes m = 1,485 sites (excluding alignment gaps). Polymorphism and divergence were expressed as π_S (silent nucleotide diversity per-site) and K_S (number of silent substitutions per silent site), respectively. Sites/Changes Considered option: *Silent* (*Synonymous & Noncoding*). Sliding window options: Window length = 50 bp; step size = 10 bp.

3.4. Coalescent Simulations

This example illustrate how to use DnaSP to determine the confidence interval of Tajima's *D* by computer coalescent simulations (**Fig. 17.5**) [*Analysis*|*Tajima's Test* and *Tools*|*Coalescent Simulations* commands]. The analysis was performed using the D_subobscura_st (n = 16) subset. This data includes m = 1,485 sites (excluding alignment gaps), being the observed Tajima's *D* value, D_{obs}= –1.41386. The computer simulations were conducted fixing the θ value (θ_π = 9.425), and considering recombination (R = 88.2); θ and R values are expressed on a per-sequence basis. The outcome (**Fig. 17.5B**) shows that the 95% confidence interval of Tajima's *D* lies between –0.85436 and 0.88835. Consequently, the observed *D* value is very unlikely under the standard neutral model (null hypothesis); indeed, the probability of obtaining values equal or lower than the observed $P[D \leq D_{obs}]$ is zero (*see* **Note 10**).

4. Notes

1. Although the DnaSP software has been designed to work with DNA sequence polymorphism data (including both monomorphic and polymorphic sites), it can also use data sets with only polymorphic positions (such as SNP haplotypes). In the latter case, however, not all analyses and methods will be applicable.
2. The NEXUS format is very convenient since it can store various types of information and can be read by many computer programs. This information, however, will be lost after exporting the data to simpler file formats.

3. Nucleotide diversity can be expressed on a per-site (π) or on a per-sequence (i.e., per-gene) basis (k). Many authors, however, use the same symbol (π) for both. The same is valid for θ, which is used either on a per-site or per-sequence basis.
4. DnaSP make most (but not all) estimates using the complete-deletion option. All multiple alignment columns with missing data or gaps are ignored.
5. DnaSP does not accept input files with genotype (diplotype) data. Such data should be previously converted to haploid phase information. Moreover, a number of analyses (i.e., haplotype-based analyses such as linkage disequilibrium or haplotype diversity) also require knowledge of the haplotype phase *(16)*.
6. Currently there is a high number of selective neutrality tests and test statistics. To determine which test should be used it is necessary to consider several factors such as the hypothesis that is being tested, the assumptions of the test, the study design, as well as the available information; in addition, the test should be chosen before observing the data. The choice of the test will depend on the specific reasonable alternative hypothesis; a statistical test performing well for detecting neutrality departures caused by hitchhiking events might be conservative against bottlenecks. Moreover, for a similar alternative scenario (e.g., population growth) may have different test statistics; in this case, the most powerful statistic should be chosen, i.e., the most powerful against a reasonable set of parameters of the alternative hypothesis (*see* ref. *19*).
7. The coalescent simulation module requires that θ and R values are expressed on a per-sequence (per-gene) basis.
8. Coalescent simulations can be conducted given a θ value (θ_π estimates) or by fixing the number of segregating sites. The former is recommended.
9. It should be stressed that a significant result (a significant departure from the null hypothesis) cannot be interpreted directly as evidence for positive selection; there are several putative alternative hypotheses to the single neutrality null hypothesis. For instance, the deviation may also be caused by demographic factors. Additional analyses, therefore, are required to determine the role of natural selection in shaping the patterns of nucleotide variation.
10. The specific *P*-value is obtained by comparing the observed (real data) test statistic value with the distribution of values (empirical distribution) obtained by computer simulations.

11. In the presence of an outgroup (unfolded frequency spectrum), the *x*-axis represents the number of individuals (i) carrying the derived nucleotide variant ($1 \leq i \leq n-1$). If there is no outgroup available, the frequency spectrum must be "folded" (folded frequency spectrum) to combine the indistinguishable categories (it is not possible to distinguish which nucleotide is ancestral and which is derived); in this case, the value $i = 1$ (*x*-axis) would indicate singleton configurations, i.e., where 1 individual (or $n - 1$ individuals) present a particular nucleotide variant, etc. In a folded frequency spectrum, therefore, $1 \leq i \leq n/2$.
12. DnaSP provides per-sequence estimates of the relevant population growth parameters: θ_0, θ_1, and τ; following Rogers *(30),* the estimates are obtained by fitting the empirical data to the theoretical expectations (method of moments).

Acknowledgments

We thank Sergios-Orestis Kolokotronis for his helpful comments and suggestions on the manuscript. This work was funded by grant BFU2007-62927 from the Dirección General de Investigación Científica y Técnica (Spain), and by grant 2005SGR00166 from Comissió Interdepartamental de Recerca i Innovació Tecnològica de Catalunya (Spain).

References

1. Przeworski, M., Hudson, R. R., and Di Rienzo, A. (2000) Adjusting the focus on human variation. *Trends Genet* **16,** 296–302.
2. Nordborg, M., and Innan, H. (2002) Molecular population genetics.*Curr Opin Plant Biol* **5,** 69–73.
3. Rosenberg, N. A., and Nordborg, M. (2002) Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat Rev Genet* **3,** 380–90.
4. Hudson, R. R. (1990) Gene genealogies and the coalescent process, in *Oxford Surveys in Evolutionary Biology* (Futuyma, D. J., and Antonovics, J. D., Eds.), Oxford University Press, New York. pp. 1–44.
5. Excoffier, L., and Heckel, G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* **7,** 745–58.
6. Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19,** 2496–7.
7. Rozas, J., and Rozas, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15,** 174–5.
8. Excoffier, L., Laval, G., and Schneider, S. (2005) Arlequin (version 3): An integrated software package for population genetics data analysis.*Evol Bioinf Online* **1,** 47–50.
9. Bandelt, H.-J., Forster, P., and Röhl, A. (1999) Median-Joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16,** 37–48.
10. Maddison, W. P., Swofford, D. L., and Maddison, D. R. (1997) NEXUS: an extendible file format for systematic information. *Syst Biol* **46,** 590–621.
11. Maddison, D. R., and Maddison, W. P. (2000) *MacClade 4*, version 4. Sinauer, Sunderland, MA.

12. Swofford, D. L. (1998) *PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods)*. Sinauer Associates, Sunderland, MA.
13. Kumar, S., Tamura, K., and Nei, M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinf* **5,** 150–63.
14. Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.
15. Jukes, T. H., and Cantor, C. R. (1969) Evolution of protein molecules, in *Mammalian Protein Metabolism* (Munro, H. N. Ed.), Academic Press, New York, pp. 21–132.
16. Stephens, M., and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73,** 1162–9.
17. Slatkin, M., and Hudson, R. R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129,** 555–62.
18. Rogers, A. R., and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9,** 552–69.
19. Ramos-Onsins, S. E., and Rozas, J. (2002) Statistical properties of new neutrality tests against population growth. *Mol Biol Evol* **19,** 2092–100.
20. Harpending, H. (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum Biol* **66,** 591–600.
21. Mcdonald, J. H., and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature* **351,** 652–4.
22. Yang, Z., and Bielawski, J. P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15,** 496–503.
23. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–95.
24. Fu, Y.-X., and Li, W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709.
25. Hudson, R. R., Kreitman M., and Aguadé, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–9.
26. Fay, J. C., and Wu, C. I. (2000) Hitchhiking under positive darwinian selection. *Genetics* **155,** 1405–13.
27. Hutter, S., Vilella, A, J., and Rozas, J. (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinf* **7,** 409.
28. Nordborg, M. (2001) Coalescent theory, in *Handbook of Statistical Genetics* (En Balding, D., Bishop, M., and Cannings, C., Eds.), John Wiley & Sons, Chichester, pp. 179–212.
29. Rozas, J., Segarra, C., Ribó, G., and Aguadé, M. (1999) Molecular population genetics of the *rp49*gene region in different chromosomal inversions of *Drosophila subobscura*. Genetics **151,** 189–202.
30. Rogers, A. R. (1995) Genetic evidence for a Pleistocene population explosion. *Evolution* **49,** 608–15.