DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets

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]

[1]Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[2]Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Bellaterra, Spain

[3]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

*Corresponding author: E-mail: jrozas@ub.edu.

Associate editor: Keith Crandall

Abstract

We present version 6 of the DNA Sequence Polymorphism (DnaSP) software, a new version of the popular tool for performing exhaustive population genetic analyses on multiple sequence alignments. This major upgrade incorporates novel functionalities to analyze large data sets, such as those generated by high-throughput sequencing technologies. Among other features, DnaSP 6 implements: 1) modules for reading and analyzing data from genomic partitioning methods, such as RADseq or hybrid enrichment approaches, 2) faster methods scalable for high-throughput sequencing data, and 3) summary statistics for the analysis of multi-locus population genetics data. Furthermore, DnaSP 6 includes novel modules to perform single- and multi-locus coalescent simulations under a wide range of demographic scenarios. The DnaSP 6 program, with extensive documentation, is freely available at http://www.ub.edu/dnasp.

Key words: **Population Genetics Software, RADseq Analysis, SNPs, VCF, Coalescent Simulations.**

Recent advances of high-throughput sequencing (HTS) technologies, including genomic partitioning methods, are generating massive high-quality DNA sequence and single-nucleotide polymorphism (SNP) data sets (Bleidorn 2016), facilitating the study of nonmodel organisms. Analyzing HTS data can provide new insights into the evolutionary forces shaping biodiversity (Ellegren 2014), which has multiple applications in animal and plant breeding, conservation genetics, biomedicine, forensics, and systematics.

DNA Sequence Polymorphism (DnaSP) is a bioinformatics tool designed for the comprehensive analysis of DNA sequence data variation, using a friendly Graphic User Interface (Rozas and Rozas 1995; Librado and Rozas 2009; Rozas 2009). The program allows the detailed characterization of the levels and patterns of DNA sequence variation at different time scales, using polymorphic variants (intraspecific data), divergence data (interspecific or interpopulation data), or a combination of both. Version 6 incorporates novel capabilities especially suitable for the analysis of thousands of DNA sequence regions in one-go, a feature increasingly demanded for RADseq-based studies, as well as in many disciplines, such as population genomics, molecular ecology, or clinical virology. Furthermore, DnaSP 6 also includes new functions to conduct coalescent simulations under a wide range of demographic scenarios.

Major Upgrades

Analyses of Multi-MSA Data Sets

Methods for genomic partitioning are cost-effective approaches for molecular population genetics, phylogeographic, and phylogenomic studies, especially in nonmodel organisms (Lemmon and Lemmon 2013; McCormack et al. 2013). These mainly include RADseq, and hybrid enrichment. The first one embraces the original Restriction-site Associated DNA sequencing (RADseq), as well as some methodological variants, such as double-digest RADseq or genotype-by-sequencing (Puritz et al. 2014; Andrews et al. 2016). The most popular hybrid enrichment techniques are anchored hybrid enrichment (Lemmon et al. 2012) and target capture of ultraconserved elements (UCE; Faircloth et al. 2012).

Raw sequencing reads from these approaches can be preprocessed and assembled using a number of available pipelines, such as PyRAD (Eaton 2014) and STACKS (Catchen et al. 2011) for RADseq, or PHYLUCE (Faircloth 2016) for UCE data. RADseq files typically store many thousands of short (from 100 to 300 bp) DNA sequences randomly distributed across the genome (RAD-loci), each encompassing one or a few SNPs, whereas hybrid enrichment generally collects data from a small number of longer loci. These programs generate curated data either in the form of a single multi-MSA file, or in the form of multiple files, each containing information of an individual marker (a single MSA).

Table 1. Performance Benchmark of DnaSP 6.

Data Files[a]	Data Information	DnaSP Module[b]	MB[c]	n[d]	MSA[e]	Total Pos[f]	Variable Pos[g]	PC/Windows[h]	PC/Windows[i]	Linux[j]	Macintosh[k]
*.fa	Phased—Genotype (Diploid)	Multi-MSA1	437	30	98,876	14,337,020	17,220	231	287	309	325
*.loci	Phased	Multi-MSA1	62	28	55,454	2,548,620	103,946	62	90	133	129
*.vcf	Phased—Genotype (Diploid)	Multi-MSA2	10	5,008	10	950	940	49	156	119	236
	Phased—Genotype (Diploid)	Multi-MSA2	10	5,008	1	968	963	124	174	143	187
	Phased—Genotype (Diploid)	Multi-MSA2	100	5,008	10	9,680	9,630	672	2,021	1,374	2,996
	Phased—Genotype (Diploid)	Multi-MSA2	200	5,008	1	19,394	19,322	1,095	1,318	1,724	1,666
	Phased—Genotype (Diploid)	Multi-MSA2	232	40	1,000	968,000	101,000	35	75	57	78
*.vcf	Unphased—Genotype (Diploid)	Multi-MSA2	56	82	12,065	22,290	20,834	12	15	30	29
*.vcf	Phased/Unphased—Genotype (Diploid)	Multi-MSA2	30	240	3,967	5,914	5,863	20	35	56	65
*.arp	Phased—Haplotype Data	HapFreq	0.4	316,976	1[l]	340	267	48	130	162	180

NOTE.—Computation time required to complete the DNA Polymorphism analysis referred in supplementary table S1, Supplementary Material online, using different data files and computer systems.
[a]Data files specification: *.fa (STACKS; Catchen et al. 2011); *.alleles and *.loci (PyRAD; Eaton 2014); *.vcf (Danecek et al. 2011).
[b]Data DnaSP modules as in supplementary tables S1 and S2, Supplementary Material online.
[c]Data file size measured in megabytes.
[d]Sample size, the number of chromosomes analyzed.
[e]Number of MSA included in the data file; i.e., the number of RAD loci or scaffolds.
[f]Total number of positions included in the data file (including monomorphic positions).
[h]Computation Time in seconds. PC-Windows computer with an Intel i7-6700 processor (3.4 GHz; 4 cores -8 threads), 32 GB RAM; Windows 10 (64 bits).
[g]Total number of polymorphic positions analyzed.
[i]Computation Time in seconds. PC-Windows computer with an Intel i7-6500 U processor (3.1 GHz; 2 cores -4 threads), 8 GB RAM; Windows 10 (64 bits).
[j]Computation Time in seconds. Linux (Mint 18.1 64 bits; 16 GB RAM), with an Intel i5-4690 processor (3.50 GHz; 4 cores -4 threads); VirtualBox 5.1.22 with Windows 8, 64 bits (8 GB RAM in the virtual machine).
[k]Computation Time in seconds. MacBook Pro (MacOS Sierra -10.12-; 8 GB RAM), with and Intel Core i5-5257 U processor (2.7 GHz; 2 cores; 4 threads); VirtualBox 5.1.22- Windows 8, 64 bits (4 GB RAM in the virtual machine).
[l]A single MSA with 116 samples.

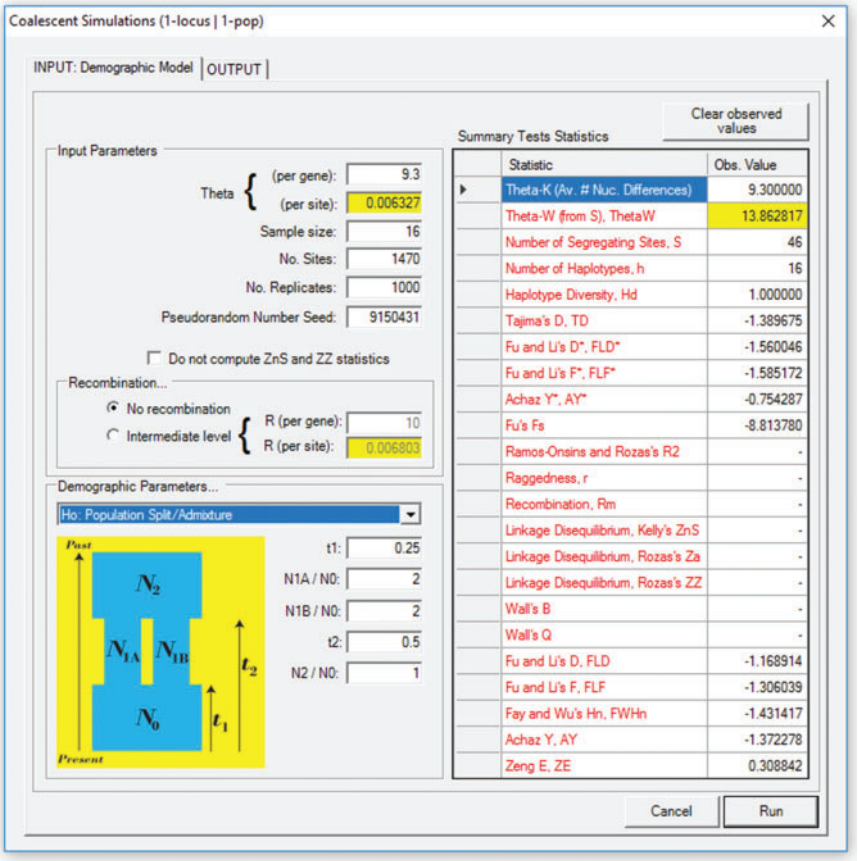

Fig. 1. Input for simulating a single locus under the coalescent.

Thanks to the new multithreading capabilities, DnaSP 6 can now efficiently process and analyze the massive data generated by PyRAD and STACKS programs, namely *.alleles, *.loci, and *.fa format files. The program can also handle data from other partitioning approaches (see Lemmon and Lemmon 2013 for a review), including data from low-coverage whole-genome sequencing projects generated, for example, by the DOMINO pipeline (Frías-López et al. 2016), or variation data stored in the popular Variant Calling Format (VCF) (Danecek et al. 2011). All these features allow an easy integration of DnaSP 6 with standard HTS pipelines.

Using multi-MSA files as the input data, DnaSP 6 will carry out most of the comprehensive population genetic analyses available for a single locus in DnaSP 5 (Librado and Rozas 2009). These typically involve the estimation of a large set of statistics that summarize the patterns and levels of DNA sequence variation both within and between populations, including estimates of linkage disequilibrium and gene flow (supplementary tables S1 and S2, Supplementary Material online; see also the DnaSP documentation). DnaSP 6 additionally estimates the observed individual heterozygosity from genotype data, a measure often used as proxy for inbreeding (Balloux et al. 2004). This new version also incorporates new neutrality tests, including the Zeng's E (Zeng et al. 2006), especially devised to pinpoint loci that recently underwent a positive selection event, and Achaz's Y* and Y (Achaz 2009), devised to mitigate the impact of HTS sequencing errors. Furthermore, DnaSP 6 can analyze full DNA sequence information or variable positions only (SNP data), phased or unphased SNP data, or genotype data with different ploidy levels.

Processing and Analysis of Haplotype-Frequency Data

Population-based studies of viruses are rapidly increasing our knowledge about the molecular mechanisms driving their evolution, revolutionizing molecular epidemiology and pathogenesis (Acevedo et al. 2014; Quiñones-Mateu et al. 2014). These studies usually involve millions of samples from small DNA regions. To handle such huge sample data sets, we extensively redefined DnaSP 5 variables, increasing their precision boundaries, and implemented efficient algorithms for multithreading calculations (table 1). DnaSP 6 is now capable of processing the commonly used Arlequin format, which stores frequency information of haplotype sequences (*.arp; Excoffier and Lischer 2010).

Multi-Locus Coalescent Simulations

The coalescent theory, which describes the statistical properties of gene genealogies, is a fundamental tool for understanding the evolutionary dynamics of natural populations (Hudson 1990). DnaSP 6 widely extends its capabilities to analyze DNA samples under the coalescent, by incorporating the algorithms described in Hudson (2002) and Ramos-Onsins and Mitchell-Olds (2007). The new coalescent modules automatically capture summary statistics calculated from the observed data, using either the single-locus mode or new batch routines for multilocus analyses (fig. 1). The current

version allows evaluating the likelihood of the summary statistics (by reporting their *P* values and confidence intervals), not only under the standard neutral model (already available in the version 5), but also under wide range of demographic scenarios, such as population growth (or decline), population bottleneck, and population split with admixture.

System and Benchmarking

To facilitate the analyses of large data sets, we have migrated DnaSP from Visual Basic 6 to VB.NET (Visual Studio 2015). This new Windows programming language supports multithreading computation, optimizes RAM memory usage, enables 64-bit variables and executables, and facilitates interoperability with the Internet. These features are primary requirements for user-friendly analyses of large data sets using personal computers or workstations.

We benchmarked DnaSP 6 performance using diverse data sets, file formats, and computer configurations (including Macintosh and Linux operating systems, using virtual machines) (table 1). We found that DnaSP 6 can efficiently manage large data files, storing >100,000 MSAs, >100,000 SNPs, or thousands of individuals (up to 500 MB in total). The software is able to conduct a complete DNA Polymorphism analysis (which computes 17 summary statistics and neutrality tests; supplementary table S1, Supplementary Material online) in a few seconds (or minutes, depending on the data file). For example, using an Intel Core-i7-6700 3.4-GHz processor and 32 GB of RAM (table 1), the analysis of a VCF data file of 30 MB (120 diploid individuals, $n = 240$; 3,967 scaffolds) takes 20 s, a VCF File of 232 MB ($n = 40$; 1,000 scaffolds; 101,000 SNPs) 35 s, and a multi-FASTA data file of 437 MB ($n = 30$; 98,876 MSAs; 17,220 SNPs) 231 s. Similar performance resulted using Arlequin file formats, completing the analysis of a data set with 316,976 sequences in 48 s. Therefore, the software is appropriated to analyze representative data files from diverse genome partitioning methods.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all beta testers, whose feedback helped us to significantly improve the software, and especially to Fernando González-Candelas and Jose Castresana for their valuable comments and suggestions. This work was supported by grants of the Ministerio de Economía y Competitividad, Spain (BFU2010-15484, CGL2013-45211, AGL2013-41834-R, AGL2016-78709-R, and CGL2016-75255), and by the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2009SGR-1287 and 2014SGR-1055). J.R. was partially supported by ICREA Academia (Generalitat de Catalunya), and S.G.-R. by a Beatriu de Pinós Postdoctoral Fellowship (AGAUR; 2014 BP-B 00027).

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 505(7485):686–690.

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183(1):249–258.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*. 17(2):81–92.

Balloux F, Amos W, Coulson T. 2004. Does heterozygosity estimate inbreeding in real populations. *Mol Ecol*. 13(10):3021–3031.

Bleidorn C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Syst Biodivers*. 14(1):1–8.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda)* 1(3):171–182.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. 2011. The Variant Call Format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Eaton DAR. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30(13):1844–1849.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol*. 29(1):51–63.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 10(3):564–567.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 61(5):717–726.

Faircloth BC. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32(5):786–788.

Frías-López C, Sánchez-Herrero JF, Guirao-Rico S, Mora E, Arnedo MA, Sánchez-Gracia A, Rozas J. 2016. DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. *Bioinformatics* 32(24):3753–3759.

Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxf Surv Evol Biol*. 7:1–45.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenetics. *Syst Biol*. 61(5):727–744.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst*. 44(1):99–121.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451–1452.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol*. 66(2):526–538.

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE. 2014. Demystifying the RAD fad. *Mol Ecol*. 23(24):5937–5942.

Quiñones-Mateu ME, Avila S, Reyes-Teran G, Martinez MA. 2014. Deep sequencing: becoming a critical tool in clinical virology. *J Clin Virol*. 61(1):9–19.

Ramos-Onsins SE, Mitchell-Olds T. 2007. mlcoalsim: multilocus coalescent simulations. *Evol Bioinformatics* 2:41–44.

Rozas J, Rozas R. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Comput Appl Biosci*. 11(6):621–625.

Rozas J. 2009. DNA sequence polymorphism analysis using DnaSP. In: Posada D, editor. Bioinformatics for DNA sequence analysis; methods in molecular biology series. Vol. 537. NJ: Humana Press. p. 337–350.

Zeng K, Fu Y, Shi S, Wu C. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174(3):1431–1439.