

# ProSeq4: A user-friendly multiplatform program for preparation and analysis of large-scale DNA polymorphism datasets

Dmitry A. Filatov 

Department of Biology, University of Oxford, Oxford, UK

## Correspondence

Dmitry A. Filatov, Department of Biology, University of Oxford, South Parks Road, Oxford, UK.  
Email: [dmitry.filatov@biology.ox.ac.uk](mailto:dmitry.filatov@biology.ox.ac.uk)

## Funding information

Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/P009808/1

Handling Editor: Frederic Austerlitz

## Abstract

Preparation of DNA polymorphism datasets for analysis is an important step in evolutionary genetic and molecular ecology studies. Ever-growing dataset sizes make this step time consuming, but few convenient software tools are available to facilitate processing of large-scale datasets including thousands of sequence alignments. Here I report “processor of sequences v4” (proSeq4)—a user-friendly multiplatform software for preparation and evolutionary genetic analyses of genome- or transcriptome-scale sequence polymorphism datasets. The program has an easy-to-use graphic user interface and is designed to process and analyse many thousands of datasets. It supports over two dozen file formats, includes a flexible sequence editor and various tools for data visualization, quality control and most commonly used evolutionary genetic analyses, such as NJ-phylogeny reconstruction, DNA polymorphism analyses and coalescent simulations. Command line tools (e.g. *vcf2fasta*) are also provided for easier integration into bioinformatic pipelines. Apart of molecular ecology and evolution research, proSeq4 may be useful for teaching, e.g. for visual illustration of different shapes of phylogenies generated with coalescent simulations in different scenarios. ProSeq4 source code and binaries for Windows, MacOS and Ubuntu are available from <https://sourceforge.net/projects/proseq/>.

## KEYWORDS

coalescent simulations, data visualisation, DNA polymorphism, file conversion, population genetics, sequence alignment editing, software

## 1 | INTRODUCTION

The availability of genome and/or transcriptome sequence data from multiple individuals of a species provides a lot of power to researches in molecular ecology and evolutionary genetics fields. Many bioinformatic tools have already been developed for various evolutionary genetic analyses of DNA polymorphism data, such as scans for selection (DeGiorgio et al., 2016; Foll & Gaggiotti, 2008),

inference of demographic history (Liu & Fu, 2020; Schiffels & Wang, 2020), the extent of genetic exchange between populations or sub-species and reconstruction of most likely speciation scenarios for closely related species (Excoffier et al., 2021; Gronau et al., 2011; Gutenkunst et al., 2009). However, far fewer convenient open source tools are available for preparation of datasets for evolutionary genetic analyses, particularly so at the genomic or whole-transcriptome scale.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

Modern evolutionary genetic studies are often based on the analysis of DNA polymorphism and/or divergence at thousands of loci (e.g. (Wong & Filatov, 2023)). Datasets of this size can be challenging to prepare, check and correct. A typical workflow in preparation of such datasets includes sequence read mapping to a reference sequence (e.g. with bwa (Li & Durbin, 2010)), processing of resulting read alignments and single nucleotide polymorphism (SNP) calling (e.g. with SAMtools (Li et al., 2009)) and analysis of DNA polymorphism in the resulting VCF (Danecek et al., 2011) file(s). All these steps are done with command-line tools that are efficient in processing large amounts of data, but offer very limited options to visualize the datasets (e.g. SAMtools view command). Visualization of the datasets to check for any problems and the correction of the identified errors is an important step in workflows of experimental evolutionary genetic studies that is often omitted due to lack of convenient software tools with graphic user interface (GUI) suitable for processing and analysis of large-scale DNA polymorphism datasets. ProSeq4 was developed to fill this gap in the workflow.

A typical use of proSeq4 in high-throughput based DNA polymorphism studies is downstream to SNP-calling: it allows the user to load multisample VCF files (Danecek et al., 2011), convert them to sequence alignments (e.g. in FASTA format), visualize, check, correct, analyse the resulting sequence datasets and convert them to many file formats for downstream analyses in widely used evolutionary genetic programs. Due to its versatility, proSeq4 can also be used for many other purposes, e.g. to load, filter and call consensus for sequence reads from SAM files, as was recently done to reconstruct

the sequences for Y-linked genes (Filatov, 2024). ProSeq4 may also be useful for data processing in low-throughput sequencing studies, as it includes an editor for sequence chromatograms produced by 2nd generation capillary sequencing machines and it facilitates the assembly of these sequence reads into contigs.

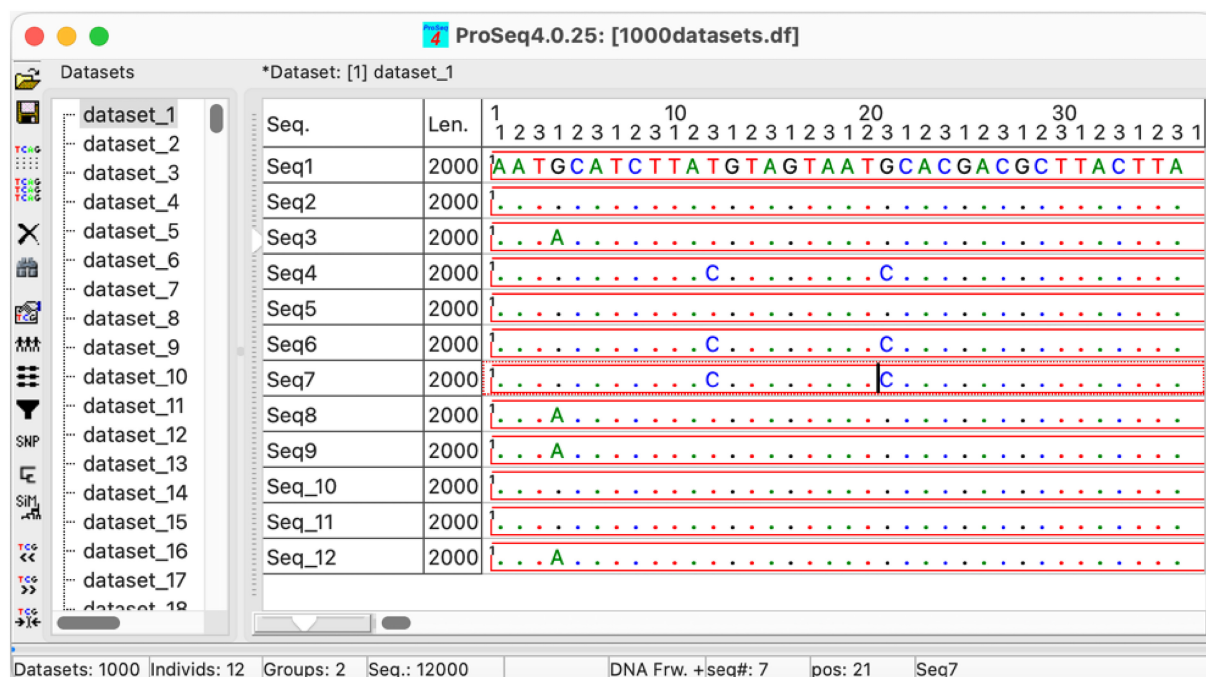
## 2 | MATERIALS AND METHODS

ProSeq4 is a 64-bit program written in Object Pascal. The older 32-bit versions (Filatov, 2002, 2009) were developed in Borland Delphi using Borland's VCF library that was Windows-based. ProSeq4 was rewritten in "Lazarus" – a free open source multiplatform Delphi-like programming environment (<https://www.lazarus-ide.org>), using LCL library that is available for many different platforms. ProSeq4 source code, manual, test data and binaries for Windows, MacOS and Ubuntu (v18.04+) are available from <https://sourceforge.net/projects/proseq/>.

## 3 | RESULTS AND DISCUSSION

### 3.1 | Data input/output

ProSeq4 supports data input and/or output for over two dozen different file formats, including widely used FASTA (Pearson & Lipman, 1988), FASTQ (Cock et al., 2010), SAM (Li et al., 2009) and VCF (Danecek et al., 2011). The data from multiple files, which can



**FIGURE 1** The main window of proSeq4 includes sequence editor, showing the currently active dataset, and the list of datasets in the project at the left. The dots in the sequence alignment denote the same nucleotide as in the first sequence. The sequences can be shown in this 'dotted mode' to highlight variable sites, or the 'standard' mode with all nucleotides shown. The numbers above the sequences show the position in the alignment and the first, second and third codon positions in the coding region. The red rectangles around the sequences show coding regions assigned to these sequences.

be in different file formats, are imported into a multilocus proSeq4 project that can be viewed and edited in the main window of the program (Figure 1).

ProSeq4 was tested with up to 40 thousand datasets (roughly corresponding to the number of genes in an eukaryotic genome) including up to a thousand individuals. Whole genome datasets from multiple individuals can only be handled for relatively small genomes (<150Mb). For larger genomes the analysis can be done separately for different chromosomal scaffolds. It is also possible to use command line tools included in proSeq4 distribution package to work with large genome-scale datasets. For example, vcf2fasta program converts VCF files into fasta format (one fasta file per contig or scaffold), while vcfSWpol command line program conducts sliding window DNA polymorphism analysis directly on a multisample VCF file. The command line tools in ProSeq4 package are designed for convenient integration into bioinformatic pipelines. For example, DNA polymorphism analysis in a gzipped VCF file could be done as follows: `zcat vcfFile.vcf.gz | ./vcfSWpol -i [other options]`.

A “dataset” in proSeq4 is an alignment of sequences from multiple individuals for a locus or a genomic contig. The project handling multiple datasets is implemented as an internal relational database that allows the user to (optionally) assign sequences to individuals and individuals to populations. Entering this information manually can be tedious for large projects and automatic linking of sequences to individuals, based on similarity of sequence and individual names, is implemented in proSeq4. Functional annotation can (optionally) be assigned to sequences in several ways, including import of annotation from GFF (Pfeifer et al., 2014), BED (Niu et al., 2022) or comma-delimited CSV files. The resulting multi-dataset project with functional annotation and population structure can be saved into a “data file” (\*.df) that is the binary file format native for proSeq4. Alternatively, the project can be saved as a set of text files, including the list of all files in the project, the population structure file (\*.s2i) and fasta files (\*.fa) along with their annotation (\*.gff). The advantage of the binary \*.df file is that it is substantially faster for data input and output operations compared to commonly used text-based file formats. The data from a proSeq4 project can also be exported into other program-specific file formats used in various popular evolutionary genetic programs: MEGA (Kumar et al., 2018), PAML (Yang, 2007), DnaSP (Rozas et al., 2017), Structure (Pritchard et al., 2000), Arlequin (Excoffier et al., 2007), fastPhase (Scheet & Stephens, 2006), BayeScan (Foll & Gaggiotti, 2008), SweepFinder (DeGiorgio et al., 2016), GPhoCS (Gronau et al., 2011), bpp (Yang, 2015), MSMC2 (Schiffels & Wang, 2020) and dadi (Gutenkunst et al., 2009). As such, proSeq4 can be used as a convenient converter of file formats for downstream analyses in commonly used evolutionary genetic programs. This proSeq4 functionality is comparable to that of a specialized file conversion tool PGDSpider (Lischer & Excoffier, 2012).

In addition to various sequence file formats, proSeq4 can open text files with multiple phylogenies in the widely used ‘brackets’ format (e.g. for species A, B and C the phylogeny can be written as ((A,B)C);). The loaded phylogenies are visualized in tree viewer window that can show trees one-by-one, or all together in a densiTree

(Bouckaert, 2010) -like style (Figure 2). The phylogenies can be edited as text in the brackets format on the “Data” page in the tree viewer window, which changes the way the phylogenies are shown on the “Tree” page. Visualization of phylogenies in proSeq4 is more basic than in recently published specialized TreeViewer program (Bianchini & Sanchez-Baracaldo, 2024), but it is comparable to that in MEGA (Kumar et al., 2018), though the latter cannot show multiple trees at the same time. The key advantage of proSeq4 is its versatility, facilitating dataset preparation, visualisation and analysis.

### 3.2 | Preparation of sequence data for analysis

The main proSeq4 window includes a panel listing all datasets in the project and a sequence editor showing the currently active dataset (Figure 1). It also (optionally) includes an outline sequence diagram at the bottom that shows the entire length of the current sequence, the functional regions assigned to that sequence and the region of the sequence visible in the editor. By default, the sequence diagram is switched off (as on Figure 1), but it can be invoked by “View/Sequence diagram” menu. Zooming out in the sequence editor would also show diagrams for all sequences in the dataset (Figure 3), though the functional annotation would not be shown, unless the row heights in the sequence editor are sufficiently increased (with the vertical slider at the left) to fit the annotations.

The sequence editor allows the user to visualize and manipulate sequences in various ways, including manual sequence editing and aligning, assigning functional regions, sequence searching, reverse-complementing, changing case, reordering sequences in multiple datasets, sequence trimming, masking and so forth. The editor includes tools for sequence translation, searching for open reading frames and site filtering.

ProSeq4 includes a legacy tool for visualization and correction of sequence chromatograms generated by capillary DNA sequencers. Sequence chromatograms can be loaded into proSeq4 in two file formats: \*.ab1 and \*.scf. The chromatogram can be visualized in a separate window (Figure 3), which allows the user to correct the sequence inferred from the chromatogram. The sequences with corrected chromatograms can be assembled into a contig and consensus for the contig can be called in the main proSeq4 editor window (Figure 3). This proSeq4 functionality provides a free alternative to such commercial software as Sequencher (GeneCodes Corp.).

ProSeq4 includes several tools for quality control of the data in the project. In particular, it can create summary reports for the datasets, sequences, coding regions and missing data in the project. This allows the user to identify errors with assignment of coding regions, such as the presence of premature stop codons and the regions with possible misalignment. For example, the “Gaps and missing data report” helps to identify and remove individuals with too much missing data in the assigned sequences. The site filtering tool allows the user to automatically filter alignment positions in all datasets in the project. For example, it is possible to exclude first and second codon positions as well as the sites with indels or missing data.

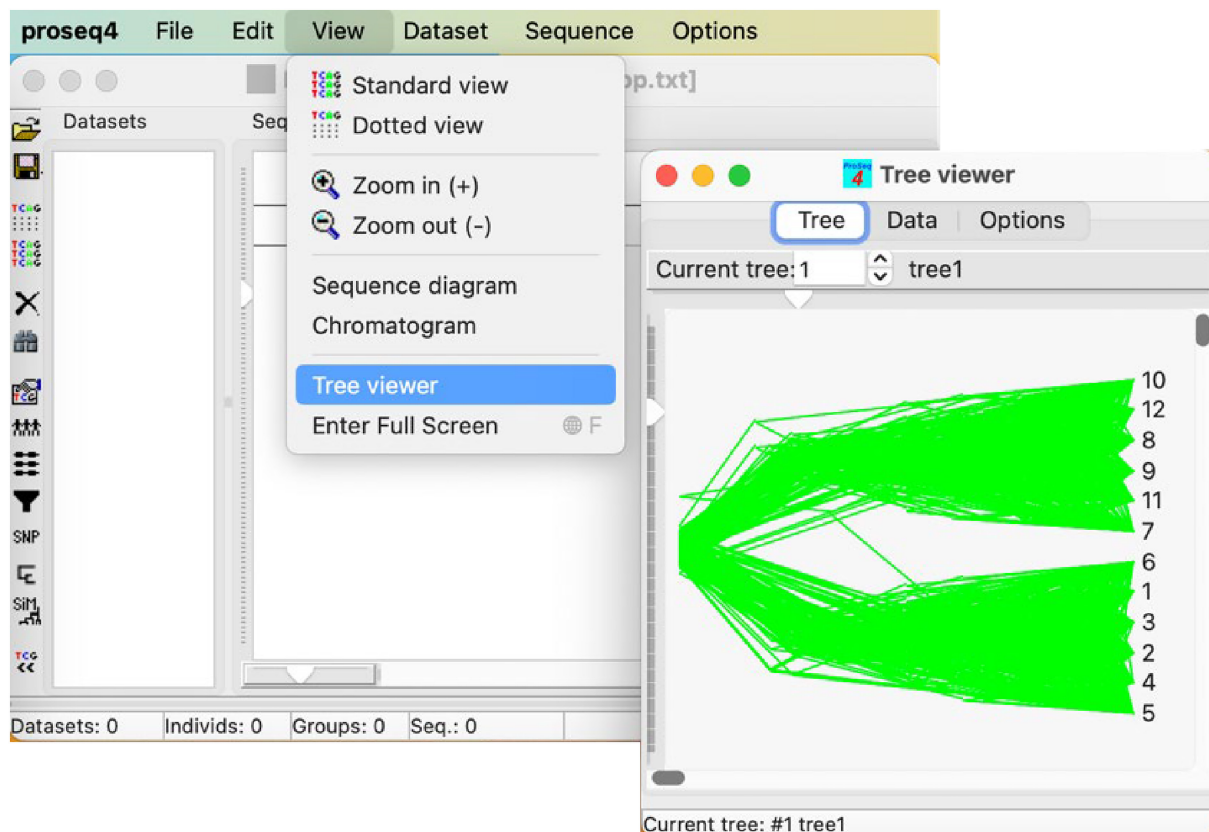


FIGURE 2 Tree viewer in proSeq4 can visualize multiple phylogenies together. The figure shows the phylogenies from file `simTrees_2pop.txt` included in the sample data in proSeq4 distribution. These phylogenies were simulated with coalescent simulations tool in proSeq4. The "current tree" in the multiTree mode determines the order of nodes shown at the right.

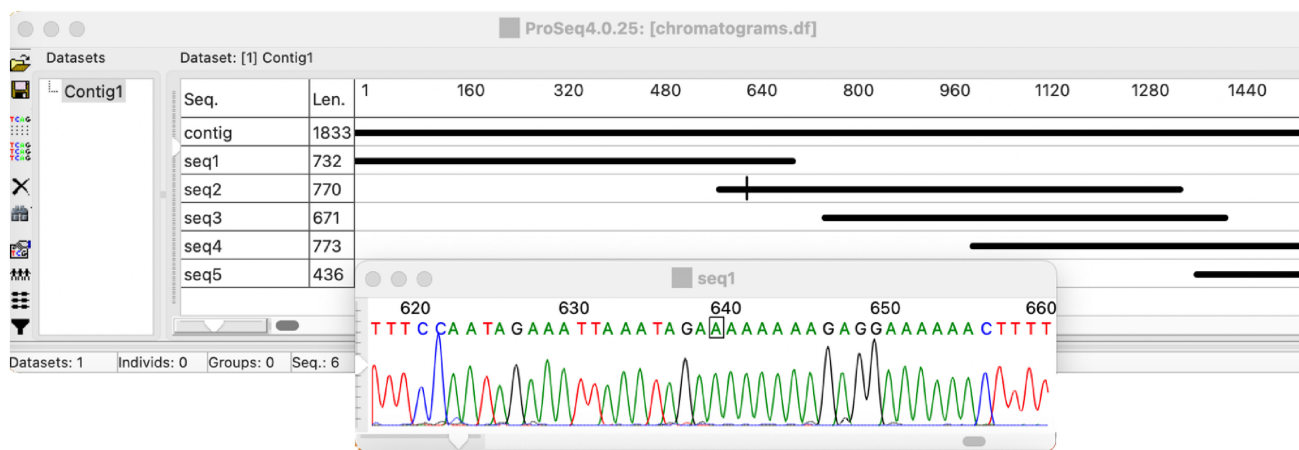


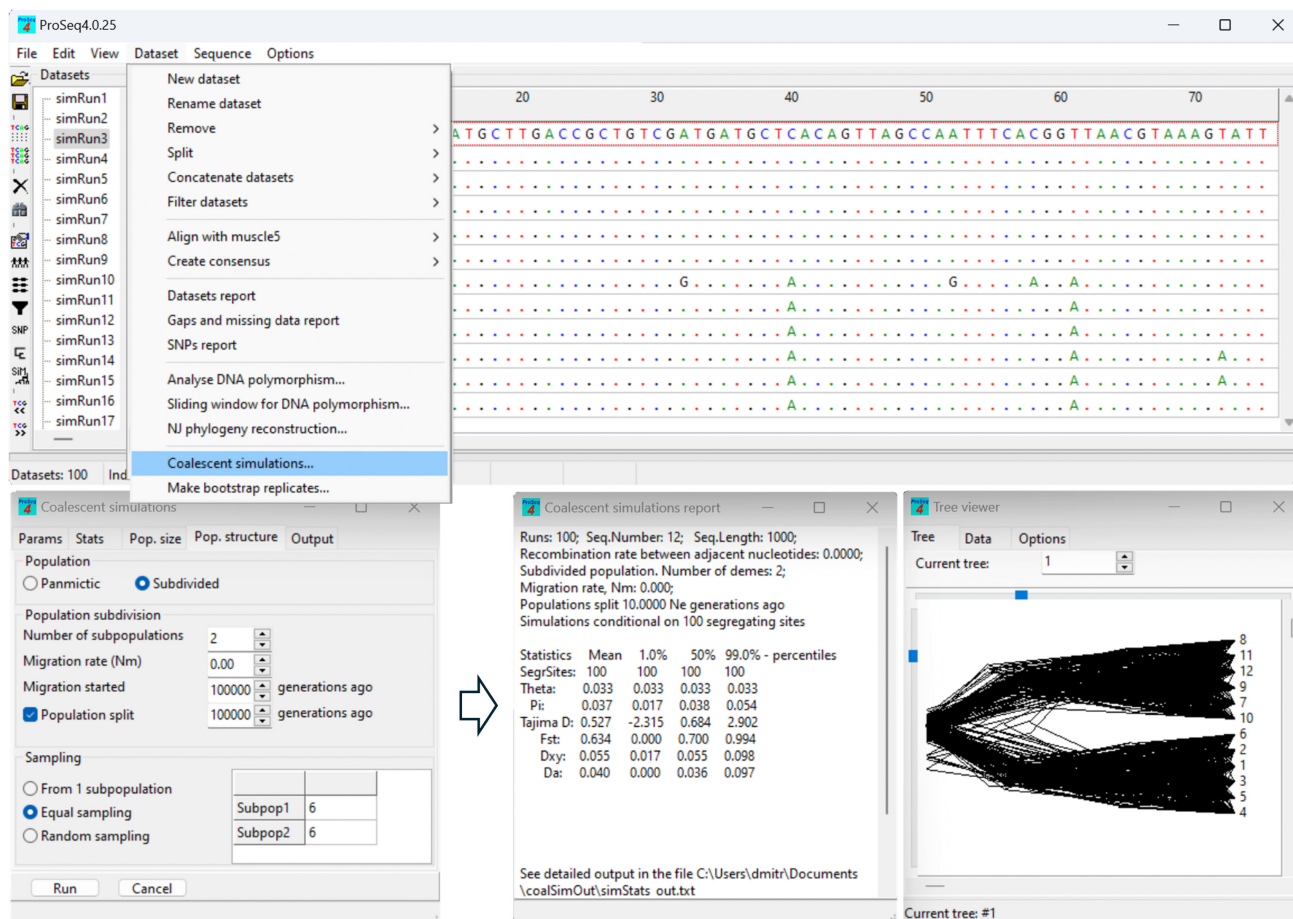
FIGURE 3 Sequence chromatogram correction and contig assembly.

### 3.3 | Phylogenetic analyses

Phylogeny reconstructions in proSeq4 are limited to Neighbour-Joining (Saitou & Nei, 1987) with the phylogenetic distance represented as the number of nucleotide differences, Jukes-Cantor (Jukes & Cantor, 1969) or Kimura's 2 parameter (Kimura, 1980) distance. In one go the phylogenies can be reconstructed for all datasets in the project and the resulting phylogenies are visualized in the tree

viewer (Figure 2). This allows the user to visually assess the extent of incongruence between the phylogenies reconstructed for different datasets (e.g. for different genes in the genome) and to identify potentially problematic datasets, e.g. containing highly divergent sequences that could indicate paralogy or problems with alignment. Although proSeq4 functionality for phylogenetic reconstructions is quite basic, it helps with checking and visually exploring the data before more detailed downstream analyses in specialized phylogenetic





**FIGURE 4** Coalescent simulations tool in proSeq4 can be used to generate critical values for the statistics and simulate gene trees and datasets under various demographic scenarios.

programs such as MEGA (Kumar et al., 2018), PhyML (Guindon & Gascuel, 2003) and PAML (Yang, 2007).

### 3.4 | Population genetic analyses

ProSeq4 implements many of the widely used DNA polymorphism summary statistics, such as average nucleotide diversity  $\pi$  (Nei, 1987), Watterson's  $\theta$  (Watterson, 1975), Tajima's D (Tajima, 1989), Kelly's  $Z_{ns}$  (Kelly, 1997) etc, which can be calculated for the current, or for all datasets in the project. These statistics can be calculated in a sliding window of given size to analyse the distribution of level and patterns of DNA polymorphism along the sequence length. If coding region(s) were assigned to a dataset, the statistics are calculated separately for silent and non-silent sites. An alternative way to analyse polymorphism at different types of sites is to run the analysis on pre-filtered datasets created with "Dataset/Filter datasets/Filter sites in all datasets" menu. If the project includes population structure (assigned with "Edit/Edit groups and individs" menu) with at least two populations (or groups of individuals), the program also calculates population subdivision statistics, such as the numbers of shared and fixed sites,  $F_{st}$  and  $K_{st}$  (Hudson et al., 1992), and their significance is tested with permutation (that is, randomly swapping

sequences between populations). The population genetic analyses implemented in proSeq4 are roughly similar to that in recent versions of DNAsp after an upgrade enabling that program to analyse multiple datasets (Rozas et al., 2017). However, unlike DNAsp, proSeq4 is designed not only to analyse but also to facilitate preparation of multigenic datasets for analyses.

A convenient way to obtain bespoke critical values for statistics in the particular dataset is to run coalescent simulations with the same size and level of polymorphism as in the experimental dataset (Hudson, 1992). For this purpose, proSeq4 includes coalescent simulations tool that can run simulations under various demographic scenarios, including population size change and population subdivision. The output of coalescent simulations includes the critical values for the statistics of interest as well as (optionally) the simulated datasets in the form of sequence alignments (Figure 4). These simulated datasets can be saved in any of the supported file formats for downstream analyses in other programs. Furthermore, it is possible to save and/or visualize the genealogies (Figure 4) generated as part of the coalescent simulations process (Hudson, 1992). This flexibility and versatility of proSeq4 is unmatched by other programs for coalescent simulations (e.g. Hudson's ms program (Hudson, 1992) or coalescent simulations tool in DNAsp (Rozas et al., 2017)).

### 3.5 | The use of proSeq4 for teaching

Students often find evolutionary genetic concepts counterintuitive. Dataset visualization facilitates understanding and illustrates the analyses that are done during evolutionary genetic practicals. Coalescent theory, with its upside down phylogenies and time running backwards (Hudson, 1992), is one of the most difficult topics for students to understand. Visualization of simulated phylogenies and datasets can help students to better grasp the concepts and the use of coalescent theory in evolutionary genetics. To this end, proSeq4 includes a tool (Figure 4) to run coalescent simulations and to simulate datasets under various scenarios. The simulated phylogenies can be visualized in tree viewer tool in proSeq4 (Figure 2). The simulated DNA polymorphism datasets (Figure 4) can be viewed in proSeq4 sequence editor, and these datasets can be used to calculate a statistic of user's choice (e.g. Tajima's D (Tajima, 1989)), which generates the null distribution of that statistic expected under the simulation scenario. This serves as a visual illustration for students how coalescent simulations can be used to generate a bespoke null distribution for the specific dataset, which is a common use of coalescent theory in experimental evolutionary genetics.

## 4 | CONCLUSIONS

ProSeq4 is a versatile program for dataset preparation and most common evolutionary genetic analyses. It can be useful at various steps in the workflows of molecular ecology and evolution studies involving compilation, quality control, visualization and analysis of DNA polymorphism datasets. The program facilitates handling and efficient analysis of large-scale high-throughput datasets but also includes tools for editing of data generated by legacy 2nd generation sequencing machines. Large number of file formats supported by proSeq4 makes it useful as a convenient and powerful file conversion tool. Given the versatility of this user friendly GUI program, it is likely to be useful for many users in Ecology and Evolution fields.

### ACKNOWLEDGEMENTS

The author acknowledges support from BBSRC grant BB/P009808/1.

### CONFLICT OF INTEREST STATEMENT

The author declared no conflict of interest.

### DATA AVAILABILITY STATEMENT

ProSeq4 source code and binaries for Windows, Mac and Ubuntu are available from <https://sourceforge.net/projects/proseq4/>.

### ORCID

Dmitry A. Filatov  <https://orcid.org/0000-0001-8077-5452>

## REFERENCES

- Bianchini, G., & Sanchez-Baracaldo, P. (2024). TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecology and Evolution*, 14(2), e10873. <https://doi.org/10.1002/ece3.10873>
- Bouckaert, R. R. (2010). DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10), 1372–1373. <https://doi.org/10.1093/bioinformatics/btq110>
- Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), 1767–1771. <https://doi.org/10.1093/nar/gkp1137>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: Increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12), 1895–1897. <https://doi.org/10.1093/bioinformatics/btw051>
- Excoffier, L., Laval, G., & Schneider, S. (2007). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1, 47–50.
- Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, 37, 4882–4885. <https://doi.org/10.1093/bioinformatics/btab468>
- Filatov, D. A. (2002). PROSEQ: A software for preparation and evolutionary analysis of DNA sequence data sets. *Molecular Ecology Notes*, 2(4), 621–624. <https://doi.org/10.1046/j.1471-8286.2002.00313.x>
- Filatov, D. A. (2009). Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics*, 25(23), 3189–3190. <https://doi.org/10.1093/bioinformatics/btp572>
- Filatov, D. A. (2024). Evolution of a plant sex chromosome driven by expanding pericentromeric recombination suppression. *Scientific Reports*, 14, 1373. <https://doi.org/10.1038/s41598-024-51153-0>
- Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, 180(2), 977–993. <https://doi.org/10.1534/genetics.108.092221>
- Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10), 1031–1034. <https://doi.org/10.1038/ng.937>
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696–704. <https://doi.org/10.1080/10635150390235520>
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10), e1000695. <https://doi.org/10.1371/journal.pgen.1000695>
- Hudson, R. R. (1992). Generating samples under a Wright-fisher neutral model of genetic variation. *Bioinformatics*, 18, 337–338.
- Hudson, R. R., Boos, D. D., & Kaplan, N. L. (1992). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, 9(1), 138–151. <https://doi.org/10.1093/oxfordjournals.molbev.a040703>
- Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism* (pp. 21–132). Academic Press.
- Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, 146(3), 1197–1206.

- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111–120.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547–1549. <https://doi.org/10.1093/molbev/msy096>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5), 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lischer, H. E., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2), 298–299. <https://doi.org/10.1093/bioinformatics/btr642>
- Liu, X., & Fu, Y. X. (2020). Stairway plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, 21(1), 280. <https://doi.org/10.1186/s13059-020-02196-9>
- Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.
- Niu, J., Denisko, D., & Hoffman, M. M. (2022). *The Browser Extensible Data (BED) format*. <https://samtools.github.io/hts-specs/BEDv1.pdf>
- Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8), 2444–2448. <https://doi.org/10.1073/pnas.85.8.2444>
- Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, 31(7), 1929–1936. <https://doi.org/10.1093/molbev/msu136>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, 34(12), 3299–3302. <https://doi.org/10.1093/molbev/msx248>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
- Scheet, P., & 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(4), 629–644. <https://doi.org/10.1086/502802>
- Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: The multiple sequentially Markovian coalescent. *Methods in Molecular Biology*, 2090, 147–166. [https://doi.org/10.1007/978-1-0716-0199-0\\_7](https://doi.org/10.1007/978-1-0716-0199-0_7)
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256–276. [https://doi.org/10.1016/0040-5809\(75\)90020-9](https://doi.org/10.1016/0040-5809(75)90020-9)
- Wong, E. L. Y., & Filatov, D. A. (2023). Pericentromeric recombination suppression and the 'large X effect' in plants. *Scientific Reports*, 13(1), 21682. <https://doi.org/10.1038/s41598-023-48870-3>
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yang, Z. H. (2015). A tutorial of BPP for species tree estimation and species delimitation. *Current Zoology*, 61(5), 854–865.

**How to cite this article:** Filatov, D. A. (2024). ProSeq4: A user-friendly multiplatform program for preparation and analysis of large-scale DNA polymorphism datasets. *Molecular Ecology Resources*, 24, e13962. <https://doi.org/10.1111/1755-0998.13962>