

Draft

## **PROSEQ V3 MANUAL**

Dmitry A. Filatov

Department of Plant Sciences, University of Oxford, South Parks Road,

Oxford, OX1 3RB, UK

e-mail: [dmitry.filatov@plants.ox.ac.uk](mailto:dmitry.filatov@plants.ox.ac.uk)

**Availability:** The program is available free of charge from

<http://dps.plants.ox.ac.uk/sequencing/proseq.htm>

### **CONTENTS**

WHY PROSEQ? .....	3
WHAT IS IT FOR? .....	3
WHAT'S NEW IN PROSEQ3? .....	3
INPUT/OUTPUT OPTIONS .....	4
SEQUENCE EDITING FUNCTIONS .....	5
Chromatogram and contig editing .....	6
Sequence alignment, editing and polymorphism verification .....	7
Functional region assignment.....	8
HANDLING MULTIGENIC DATA WITH RELATIONAL DATABASE IN PROSEQ....	9
DNA POLYMORPHISM ANALYSIS .....	10
ACKNOWLEDGEMENTS .....	11
REFERENCES .....	11

## **WHY PROSEQ?**

"ProSeq" stands for "Processor of sequences" and that's what it is – a program to handle ("process") DNA sequence data.

## **WHAT IS IT FOR?**

ProSeq was developed to facilitate routine work involved in preparation of DNA sequence polymorphism datasets, from the very first steps of checking sequence chromatograms to the final steps of checking the dataset before the analysis and preparing input files for many DNA polymorphism analysis programs. The DNA polymorphism analysis functions were added to ProSeq mainly to run the data via preliminary analyses, which often helps to identify and fix problems with the dataset, such as misalignment, the presence of paralogs etc. It also includes miscellaneous handy tools for dataset preparation, manipulation and analysis. These tools were added because we needed them in our day-to-day lab work and it was easier to add them to ProSeq than to write standalone small programs. For example, I added a sequence plate editor in order to simplify preparation of labbook reports of what was sequenced in what well of the sequencing plate. Thus, as such, ProSeq is mainly an advanced sequence editor with some (fairly limited) analysis functionality and it is advisable to use other programs for final analyses of clean datasets.

## **WHAT'S NEW IN PROSEQ3?**

ProSeq was originally developed as a sequence editor with some DNA polymorphism analysis capability (Filatov 2002). The previous version could handle only single gene datasets and was available only to Windows users. The new version was developed in Delphi7 with cross-platform CLX library, which makes it possible to compile the same code for Windows and Linux. The new version of ProSeq includes an internal relational database that links sequences to individuals and individuals to populations, simplifying handling and analysis of datasets containing multiple genes. The editor window was been completely redesigned to make it more flexible and convenient: in ProSeq3 there are three viewing modes, allowing the user to see the sequences, polymorphisms in the alignment and the functional regions assigned to the sequence. Using these modes the user can scroll along the sequence, zoom in to see a region of the sequence or zoom out to visualize the entire sequence with annotation shown in a graphical form. Many new input/output options have been added.

## INPUT/OUTPUT OPTIONS

The main (“native”) file format for ProSeq3 is “data file” or “database file”, \*.df. This is a binary file format and the details of its internal structure are available on request from the author. The format supports multiple datasets (alignments) of a single data project to be saved into a single file. The relational links between the sequences, individuals and populations are preserved if the contents of the database (data project) is saved in a \*.df file. Each sequence in the dataset can have multiple features (e.g. coding/non-coding regions), which are preserved if the data is saved in the \*.df format.

ProSeq3 also supports many other file formats (Table 1). It can create input files for such popular DNA polymorphism analysis programs as DNAsp (Librado and Rozas 2009), MEGA (Tamura, Dudley et al. 2007), PAML (Yang 2007), Arlequin (Excoffier, Laval et al. 2005) and Structure (Pritchard, Stephens et al. 2000).

In those cases when specific file format does not support multiple datasets, ProSeq3 automatically outputs different genes into separate files.

**Table 1. File formats supported in ProSeq3.**

Format	extention	Input	Output	Comment
ProSeq3	*.df	y	y	Native for ProSeq3; incompatible with ProSeq2
ProSeq2 data file	*.psf	y	y	Old proseq format that supports only a single dataset per file
Nexus	*.nex	y	y	
Phylip	*.phy	n	y	
Mega	*.meg	y	y	
Fasta	*.fasta	y	y	
PIR/NBRF	*.pir	y	y	
text files	*.txt	y	y	Sequence from text file; the name of file used as seq.name
GenBank	*.gb	y	y	Only a subset of annotation from GenBank files implemented
EMBL	*.emb	y	y	
ABI chromatograms	*.ab1	y	n	Chromatogram files from ABI sequencers
SCF chromatograms	*.scf	y	y	“Sequence chromatogram files”
Staden EXP files	*.exp	y	n	EXP files used in Staden package (similar to EMBL)
Staden-assemblies	*.contig	y	n	Contig assemblies generated by Staden+phred+phrap
Comma delimited	*.csv	y	y	Handy for exporting files into Excell
Hudson's ms	*.*	y	y	Files in format recognised by R.Hudson's ms
Clustal	*.aln	y	y	
Phrap ace	*.ace	y	n	
phred phd	*.phd	y	n	
Extended multifasta	*.xmfa	n	y	
HKA input file	*.hka	n	y	Input file for J.Hey's HKA program
IM input file	*.im	n	y	Input file for J.Hey and R.Nielsen's IM (or IMa) program
Arlequin project	*.arp	n	y	Input file for Arlequin program
Structure input file	*.txt	n	y	Input file for Structure program
Blast XML output	*.xml	y	n	Allows opening BLAST output in xml format (blast option -m7)

## SEQUENCE EDITING FUNCTIONS

The main project window includes alignment editor, sequence viewer, database navigator (data tree) and the database (link) editor (Fig 1).

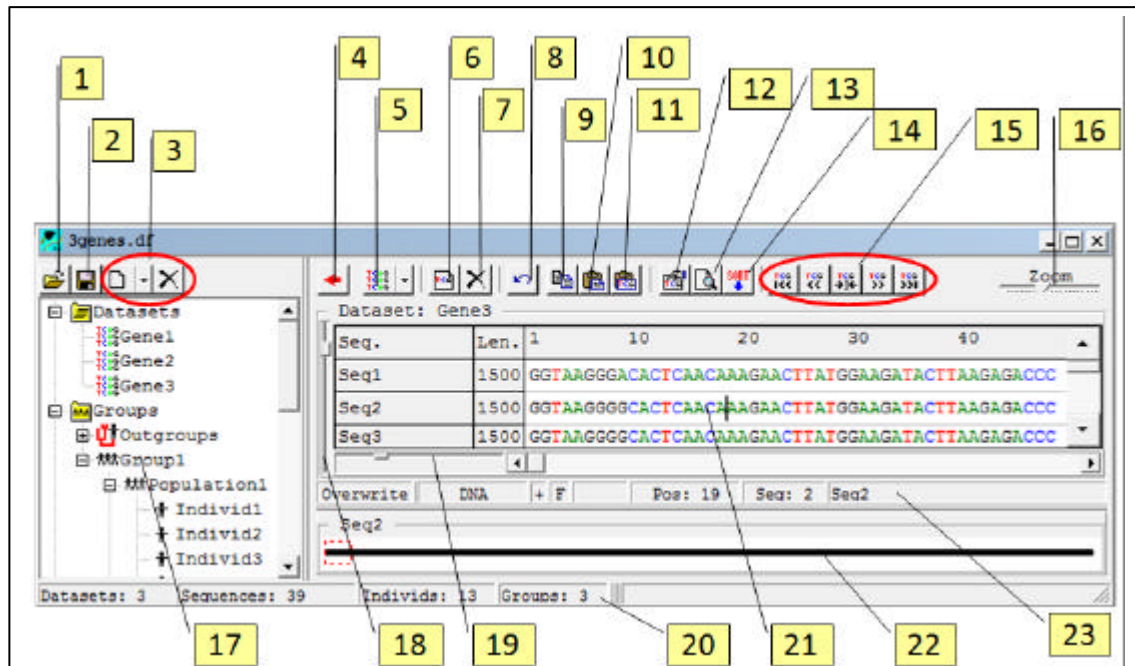


Figure 1 Project window.

- 1) File open button
- 2) Save project button
- 3) Add/delete items (project, dataset [gene], individual, group, sequence) to/from the project
- 4) Switch button to move between editor and database tabs
- 5) Switch button to choose the view mode of the sequence editor
- 6) Add new sequence button
- 7) Remove selected sequence button
- 8) Undo the last edit action
- 9) Copy selected sequence or part of it (if a region of the sequence is selected)
- 10) Paste button.
- 11) Paste as new sequence button.
- 12) Sequence properties
- 13) Find a region in the sequence
- 14) Sort sequences by name.
- 15) Move buttons. Move a sequence or several selected sequences left, right, to the very beginning or the end of the alignment, or to cursor position. If a region of a sequence is selected, the region will be moved within that sequence (e.g. handy to move indels when adjusting alignment manually).
- 16) Zoom slider to change the viewing scale of the sequences in the editor
- 17) Database navigator. Clicking on an item (gene, sequence, individual) opens the corresponding sequence or database editor.
- 18) Sequence height adjustment slider
- 19) Seq name width adjustment slider
- 20) Project status bar. Shows the number of genes, sequences, individuals and groups in the project.
- 21) Sequence editor
- 22) Outline sequence viewer. Shows which part of the sequence is shown in the editor window.
- 23) Dataset status bar. Shows the editor mode (insert/overwrite), type of sequence, sequence orientation (+/- strand and Forv/Rev) and the position of cursor in the alignment

Typically, the preparation of DNA polymorphism datasets includes the following steps, all of which are supported by ProSeq3 functionality.

**Chromatogram and contig editing:** The collection of DNA polymorphism datasets usually starts with PCR amplification and sequencing of genes (or non-coding regions) of interest from several individuals of the same or several closely related species. At this stage the processing of the data includes visual inspection of DNA sequence chromatograms to correct base-calling and sequencing errors. Chromatogram quality checking is followed by assembly of individual sequence reads into longer contigs. ProSeq3 allows the users to open chromatogram files in popular \*.ab1 and \*.scf formats, visualise chromatograms, edit the sequence and (semi-) manually to assemble sequence contigs (Fig 2).

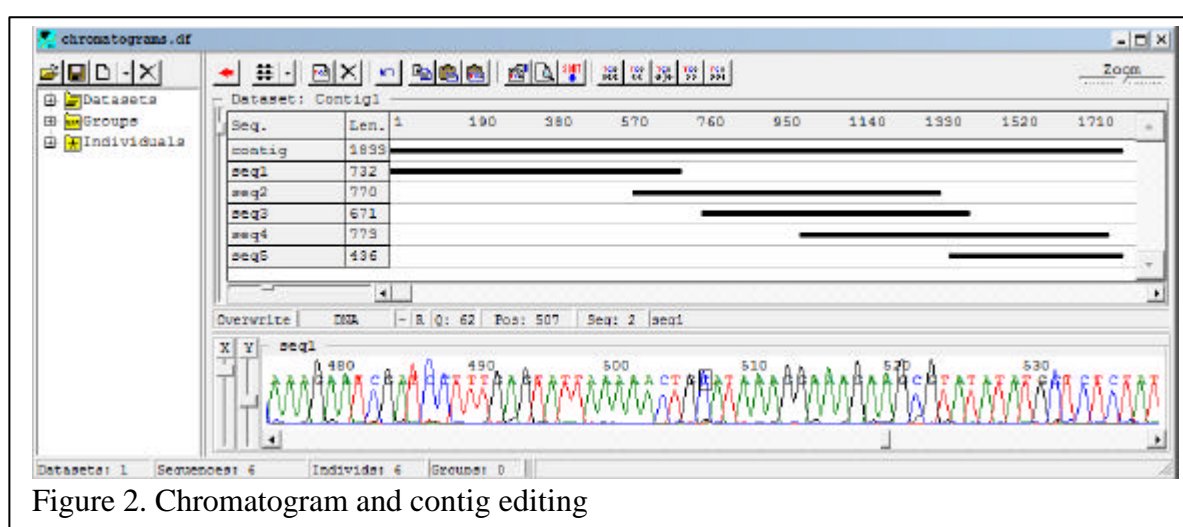


Figure 2. Chromatogram and contig editing

Integration with popular phred and phrap programs (Ewing and Green 1998; de la Bastide and McCombie 2007) makes it possible to automatically assess chromatogram quality and assemble contigs. Raw sequences with or without associated chromatogram and base quality information can be further edited in ProSeq3 to obtain finished sequences. Note that phred and phrap are not included with ProSeq3 due to licensing limitations. These programs should be obtained directly from the authors and binaries (phred.exe and phrap.exe) placed into proseq folder (the same folder where the proseq.exe [or ProSeq3.exe] is located). Once phred and phrap executables are in proseq folder and environmental variables for phred and phrap are set as described in the manual for these programs, ProSeq3 will run them automatically in the background when it needs to assess base quality or assemble contigs. Alternatively, users can run phred and phrap separately and open the output of these programs (\*.phd and \*.ace) in ProSeq3. Another option available for automatic contig assembly is to use Staden package with phred/phrap integrated into it (see

Staden help for details <http://staden.sourceforge.net/>) and open resulting \*.contig files in ProSeq3.

**Sequence alignment, editing and polymorphism verification:** The next step in dataset preparation is typically the alignment of finished contigs from multiple individuals. ProSeq3 includes several options facilitating this step. Firstly, it includes Clustal (Higgins, Thompson et al. 1996) based global multiple sequence aligner. Secondly, it includes an implementation of a pairwise alignment algorithm based on expansion of matching regions in two sequences. Thirdly, ProSeq3 can import gaps from an existing external alignment. The latter option is particularly useful if one needs to use an external alignment program that does not support sequence annotation available in ProSeq3. In that case "passing" sequences through such external program results in loss of annotation. To avoid such loss the user can save the dataset in the format suitable for external aligner, align sequences and then import alignment information (position and length of gaps) back into the annotated dataset in ProSeq3.

Following automated alignment, it is usually necessary to check, correct and trim the alignment manually, and check sequence differences between individual sequences, which is easily done in the sequence editor included in ProSeq3 (Fig 3). The editor is fairly flexible and the options and functionality available to the user make it quite a powerful tool. The sequence (or a selected part of it) can be edited, reversed, complemented, translated and shifted relative to the rest of the alignment. The editor includes three viewing modes, allowing the user to see the sequence, polymorphisms in the alignment and the functional regions assigned to the sequence. Using these modes the user can scroll along the sequence,

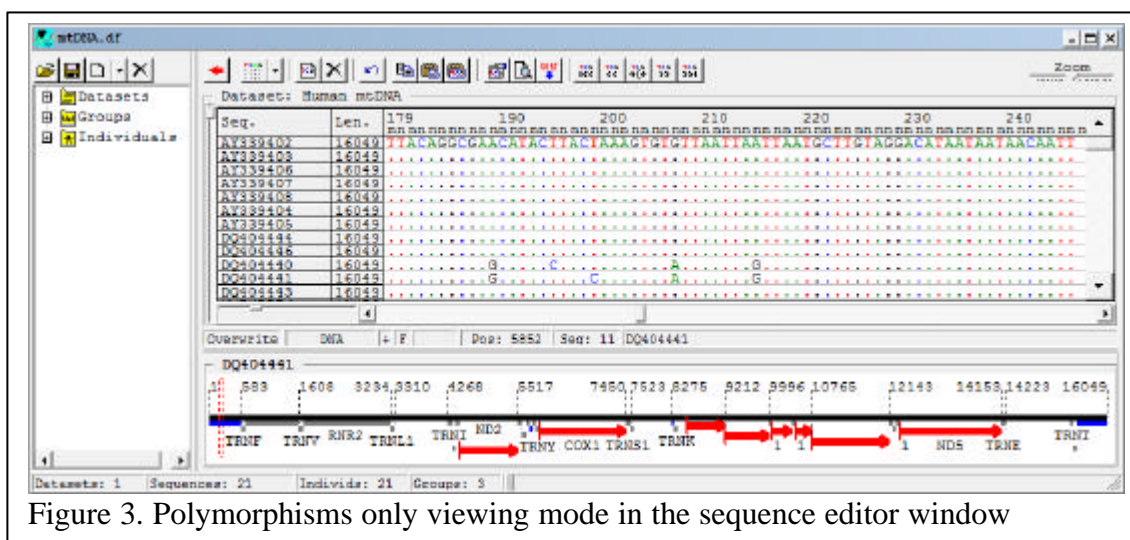


Figure 3. Polymorphisms only viewing mode in the sequence editor window

zoom in to see a region of the sequence or zoom out to visualize the entire sequence with annotation shown in a graphical form (Figs 1-3).



**Functional region assignment:** ProSeq3 supports functional annotation of individual sequences in the dataset (Figs 4, 5). Manual assignment of functional regions can be quite tedious. For example, in DNAsp one has to type in the positions of coding and non-coding regions in the dataset, which is an error-prone and slow process. ProSeq3 facilitates the input of functional annotation with several handy functions, such as selection and assignment of a functional region in the editor window, and the ability to copy assigned regions from another sequence in the dataset. All functional annotations are preserved if the dataset is saved in the data file (\*.df) "native" for ProSeq3. Annotation can also be preserved in several other file formats, such as Nexus (Maddison, Swofford et al. 1997) supported by ProSeq3.

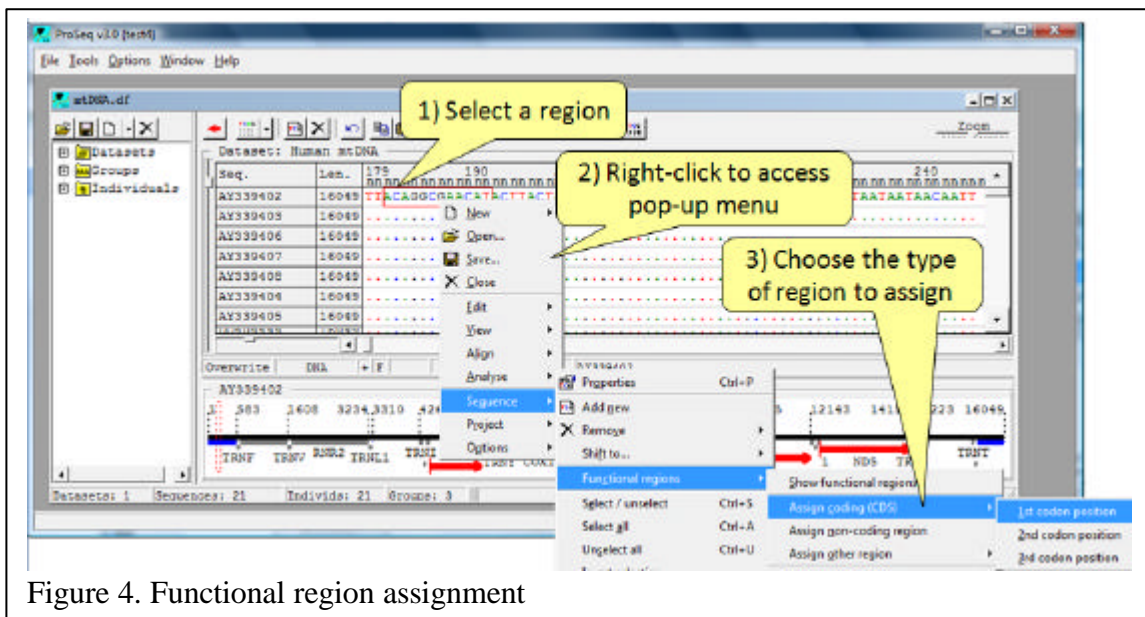


Figure 4. Functional region assignment

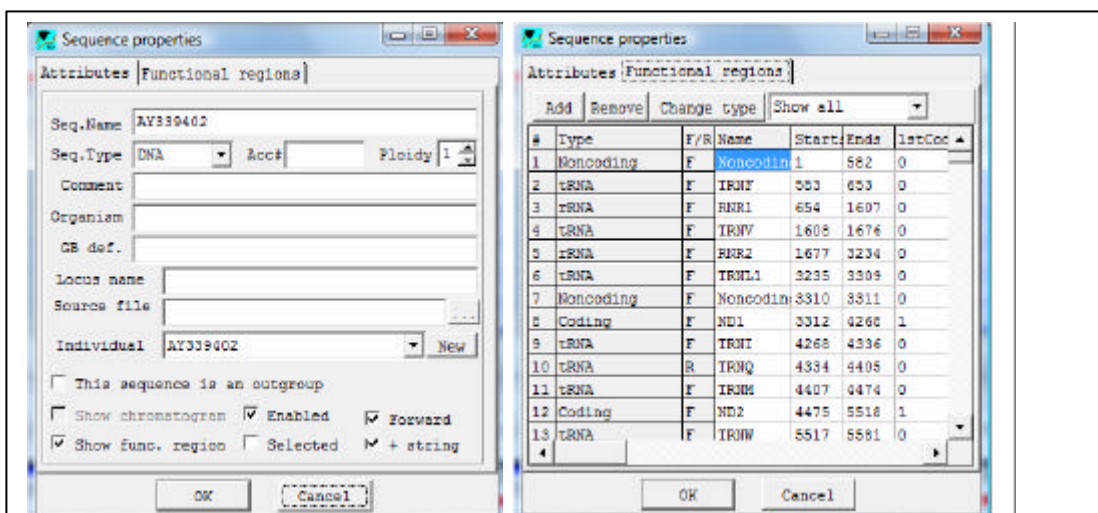


Figure 5. Sequence properties dialog (to invoke press ctrl-p)

### HANDLING MULTIGENIC DATA WITH A RELATIONAL DATABASE IN PROSEQ3

Tracking what sequence in what dataset comes from what individual becomes problematic when the number of sequenced genes is large. Even a trivial task of compiling a table that lists what genes are sequenced from what individuals (e.g. to identify what data is missing) may take several hours to complete when done manually for a project including multiple genes. ProSeq3 resolves this problem storing all the data in the project in an internal relational database where the sequences are "linked" to individuals and individuals can be combined into groups (populations). This data structure makes it trivial to manipulate multiple datasets in the project; e.g. exclusion of one individual from analysis can be done with a single mouse click, which results in automatic exclusion of all sequences linked to that individual. Grouping sequences into populations is also done at the level of individuals: if an individual is assigned to the particular population, all the sequences across multiple

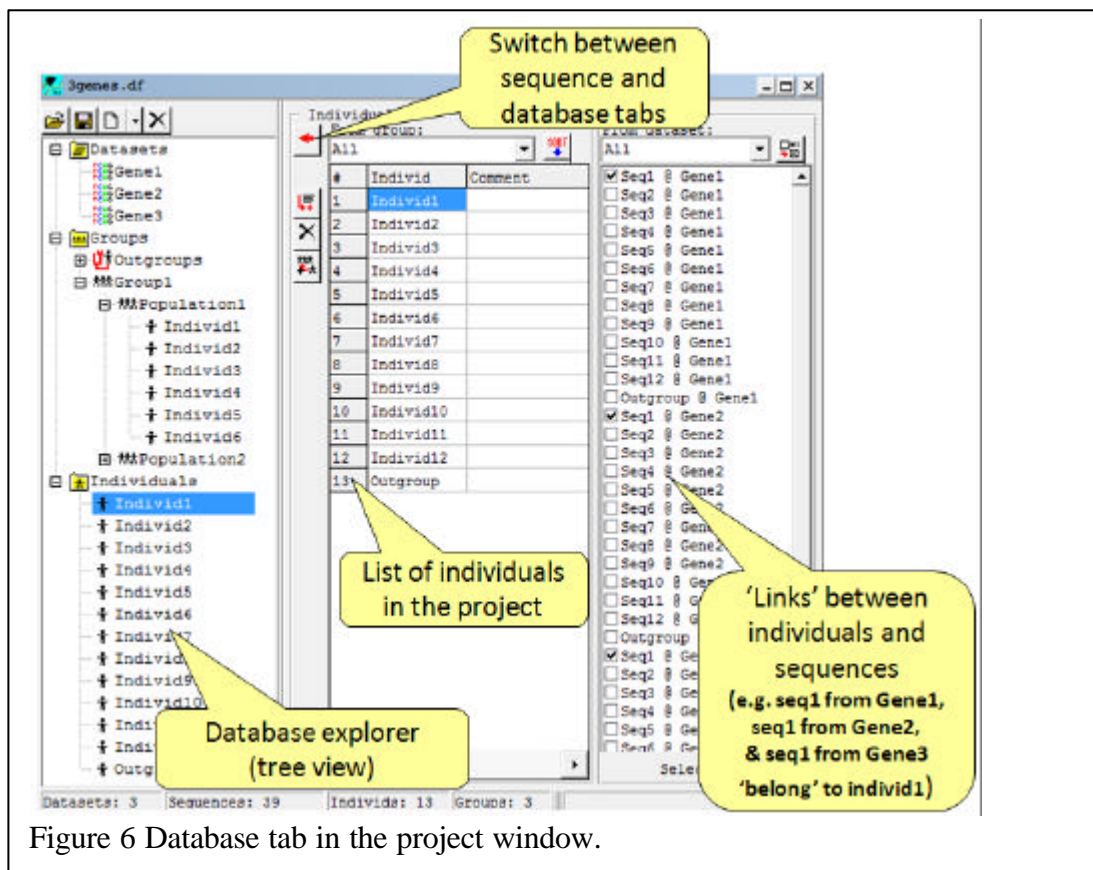


Figure 6 Database tab in the project window.

datasets in the project that are "linked" to that individual are automatically assigned to that population. This also applies to outgroup sequences/individuals: if an individual is marked as an outgroup, all sequences linked to that individual will be treated as outgroups in all the analyses. The assignment of sequences to individuals and individuals to groups can be done by simple drag and drop approach: dragging and dropping an individual to a group assigns that individual (and all linked sequences) to the group. The links between the sequences and



individuals can be visualised in a report that makes the task of tracking how many genes were sequenced from which individuals fairly trivial. Relational information of the database is preserved if the project is saved in the "native" (\*.df) ProSeq3 file format.

## DNA POLYMORPHISM ANALYSES

Once the alignments for several genes are complete and ready for analysis, they are usually analysed one by one using such programs as MEGA (Tamura, Dudley et al. 2007) or DNAsp (Librado and Rozas 2009). This process is relatively quick when there are only few genes, but it becomes prohibitively time-consuming with larger numbers of genes. ProSeq3 solves this problem allowing the user to run all the datasets in the project through the particular analysis. Several most commonly used population genetic analyses are implemented in ProSeq3: visualisation and analysis of single nucleotide polymorphisms (e.g. site frequency spectra for different types of sites), common statistics for DNA polymorphism ( $P_i$ , theta (Nei and Kumar 2000)), frequency spectrum-based neutrality tests, such as Tajima's D (Tajima 1989), linkage disequilibrium-based neutrality tests (ZnS (Kelly 1997) and B&Q (Wall 1999)) and various analyses of population subdivision/divergence. The distribution of DNA polymorphism or neutrality statistics along the length of a gene can be visualised with a sliding window option.

Although ProSeq3 was developed for population genetic analyses, it also includes a tool

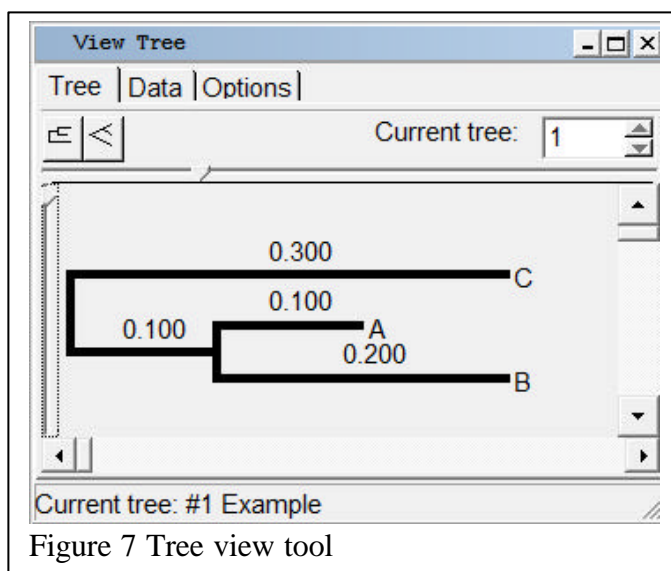


Figure 7 Tree view tool

for basic phylogenetic analysis (Fig 7) that can construct and visualise neighbor-joining trees (Nei and Kumar 2000). A combination of a sequence editor and tree visualisation tool in one program is particularly handy at the stage of preliminary evaluation and checking of the datasets, as oddities in the data, such as misalignment or sequencing errors make a sequence appear more

diverged, which is easily identifiable from the inspection of a gene tree and can be quickly fixed within the same program.

Other analysis options include the tool for creating bootstrap replicates of a dataset, and a tool for coalescent simulations (Hein, Schierup et al. 2005) that is capable of running simulations with or without intragenic recombination for panmictic or subdivided populations. Both these tools can generate pseudo-datasets that can be handled and analysed by ProSeq3 as a normal multi-dataset project, i.e. all the analyses mentioned above can be run for the pseudo-datasets generated by bootstrapping or coalescent simulations.

## ACKNOWLEDGEMENTS

I would like to thank the members of my lab for testing the program and several suggested improvements. This work was funded by the Natural Environment Research Council UK.

## References

- de la Bastide, M. and W. R. McCombie (2007). "Assembling genomic DNA sequences with PHRAP." Curr Protoc Bioinformatics **Chapter 11**: Unit11 4.
- Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.
- Excoffier, L., G. Laval, et al. (2005). "Arlequin (version 3.0): An integrated software package for population genetics data analysis." Evol Bioinform Online **1**: 47-50.
- Filatov, D. A. (2002). "PROSEQ: A software for preparation and evolutionary analysis of DNA sequence data sets." Mol. Ecol. Notes **2**: 621-624.
- Hein, J., M. H. Schierup, et al. (2005). Gene genealogies, Variation and Evolution. Oxford, Oxford University Press.
- Higgins, D. G., J. D. Thompson, et al. (1996). "Using CLUSTAL for multiple sequence alignments." Methods Enzymol **266**: 383-402.
- Kelly, J. K. (1997). "A test of neutrality based on interlocus associations." Genetics **146**(3): 1197-206.
- Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-2.
- Maddison, D. R., D. L. Swofford, et al. (1997). "NEXUS: an extensible file format for systematic information." Syst Biol **46**(4): 590-621.
- Nei, M. and S. Kumar (2000). Molecular Evolution and Phylogenetics. New York, Oxford University Press.
- Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-59.
- Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-95.
- Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-9.
- Wall, J. D. (1999). "Recombination and the power of statistical tests of neutrality." Genetical Research **74**(1): 65-79.
- Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-91.