SelectionMap: a tool for detection and visualisation of natural selection patterns within codon alignments

I. Description

SelectionMap is a free Windows computer program designed to detect and graphically illustrate the types and degree of natural selection acting along gene alignment(s). In addition the program allows comparison of natural selection patterns between two or three alignments of homologous genes. The program can either take as input codon alignment(s) in FASTA format for detection of selection and plotting of the selection map, or it can take as input the output files generated by the selection detection methods FUBAR (Murrell et al. 2013) and/or MEME (Murrell et al. 2012).

The program detects selection using the FUBAR method implemented in HYPHY (Pond et al. 2005) and then produces a selection map in which every codon site is represented by a vertical bar with a colour and height respectively indicating the type and strength of selection acting at that codon. In general, when gene alignments of related species are loaded, the figure that is produced indicates sites that are evolving under: (1) negative selection favouring the same amino acid in the different species; (2) negative selection favouring different amino acids in the different species; (3) negative selection in only one of the species; (4) positive selection; and (5) episodic positive selection.

II. Download and installation

The program can be downloaded from
http://web.cbio.uct.ac.za/~brejnev/downloads/ComputationalTools/

1. Extract the **SelectionMap-v1.0.zip** file into a temporary folder.
2. In the temporary folder double-click on the file “SETUP.EXE and follow the instructions of the setup program. **NB.** Please use the default installation directory “**C:\ SelectionMap-v1.0**” for installation.
3. The installation directory will contain the following:
    - **SelectionMap-v1.0.exe**: the main program executable file.
    - **smap.ini, SelectionMap-v1.0.exe.manifest** and **ST6UNST.LOG**: program configuration files.
    - **HyPhy-CLI2.2.6.exe**: the HyPhy Command Line version 2.2.6 installation file.
    - **README_SMap-v1.0.pdf**: SelectionMap user guide.
    - **A bin directory**: which contains sequence alignment programs (muscle3.8.31_i86win32.exe; Edgar 2004) and programs producing phylogenetic

trees (FastTree.exe; Price et al. 2010 and PhyML3.0_win32.exe; Guindon et al. 2010).

- **A test_data directory**: which contains nine files to be used to test the program once installation is completed (three codon alignments in FASTA format, and three FUBAR and three MEME selection data output files in CSV format).

4. Inside the installation folder, run the **"HyPhy-CLI2.2.6.exe**" file to install HYPHY. After HyPhy is installed successfully then SelectionMap will be able to run.

III. Uninstalling the program

To uninstall the program go to the control panel then click on Programs and Features (for Window 7 and above) or Add/Remove programs (for Windows XP) , scroll down to SelectionMap-v1.0, click on it and select Uninstall/Remove program. Go to C:\ drive check whether the C:\SelectionMap-v1.0 directory was removed, if not remove it manually.

IV. Running the program

1. Starting the program

To plot a selection map you simply have to double-click on the **SelectionMap-v1.0.exe** file. When the program has launched you can either choose to load codon alignment(s) and detect selection with FUBAR before plotting the selection map, or load FUBAR and/or MEME CSV output file(s) and their corresponding codon alignment(s) and then plot the map. The Start-up screen (Fig 1.) displays these two options of loading the input files.

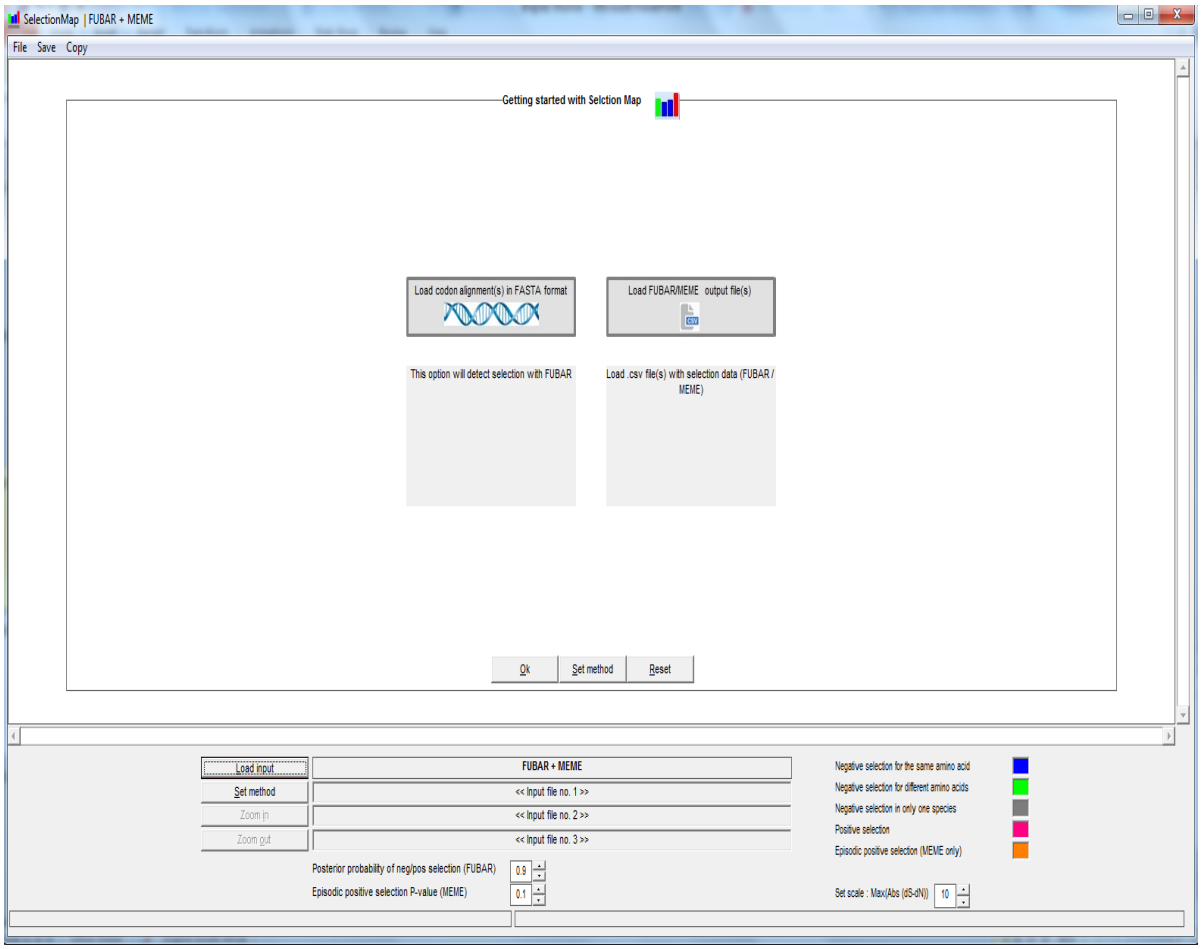

Fig 1. Start-up screen

2. Option 1: Loading the input codon alignment(s)

Once the user clicks the “Load codon alignment(s) in FASTA format” the “Loading codon alignment(s) screen” (Fig 2.) will appear. To load the input the user can either click on the frame in the middle or drag-and-drop the codon alignment file(s) to the frame. Once the alignments are loaded, the user can click on the command button to “detect selection”.

NB: only codon alignments can be loaded, and these should contain neither stop codon nor incomplete codons (i.e. codons containing one or two gaps).

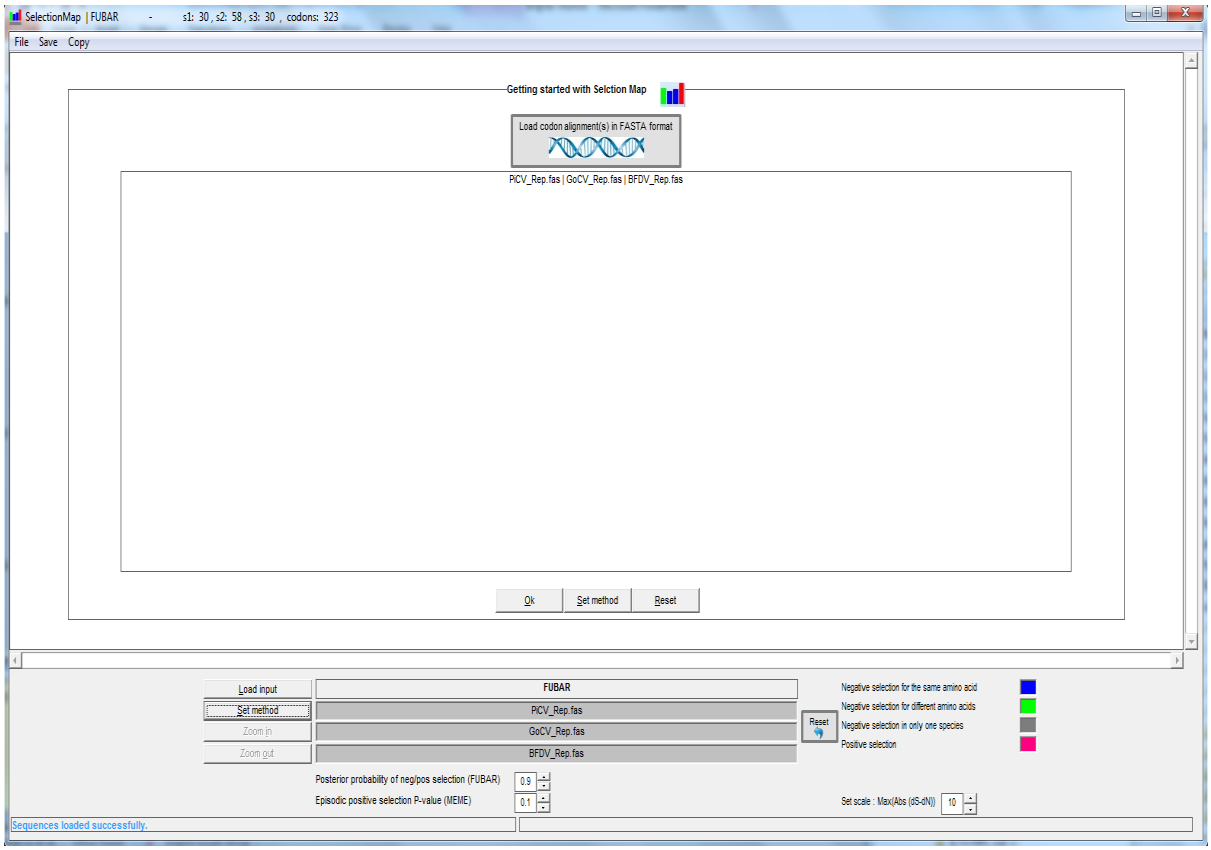

Fig 2. Loading codon alignment(s) screen

In cases where two or three codon alignments with different lengths (i.e. not aligned to each other) are loaded, the program will codon-align each input alignment individually, perform a profile alignment of the three alignments with one another, and then re-separate the alignments prior to selection analysis. This is to ensure that homologous codon sites are aligned with one another so as to enable inter-species comparisons in the figures that are generated by SelectionMap.

3. Option 2: Loading FUBAR/MEME output files and corresponding codon alignments

Once the user clicks the “Load FUBAR/MEME output files” button on the Start-up screen (Fig 1.), the program will allow FUBAR and/or MEME csv output files to be loaded. If the method of selection is set to “FUBAR”, the “FUBAR input screen” (Fig 3.) will appear to allow FUBAR CSV output file(s) and their corresponding codon alignment(s) to be loaded. If the method is set to “MEME”, the “MEME input screen” (Fig 4.) will be displayed to allow MEME CSV output file(s) to be loaded. If the method is set to “FUBAR+MEME”, the “FUBAR+MEME input screen” (Fig 5.) will appear to allow the loading of both FUBAR and MEME CSV output file(s). Use the “Set method” command button to switch between these selection detection options.

To load input files one can click on the appropriate cell and use the open dialog box to load the file(s), or the user can drag and drop file(s) on the appropriate cell.

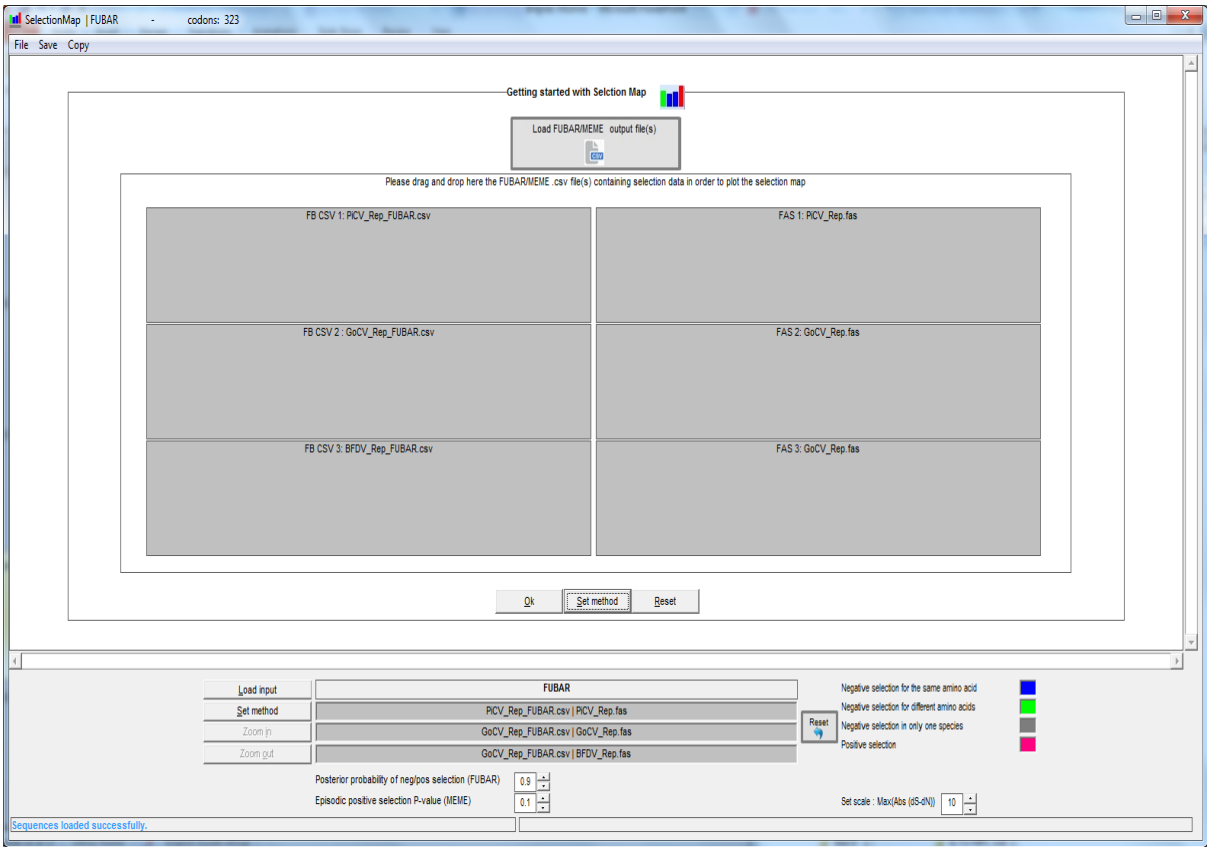

Fig 3. FUBAR input screen

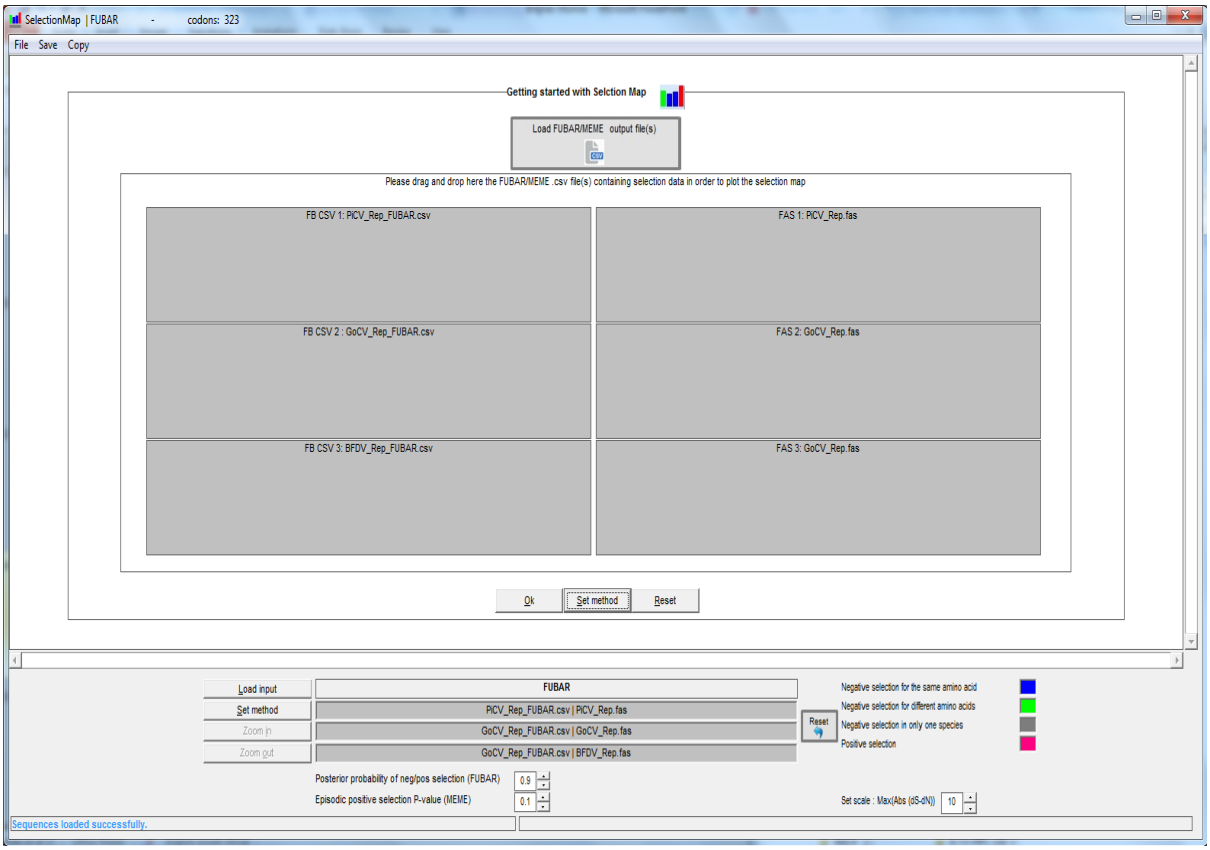

Fig 4. MEME input screen

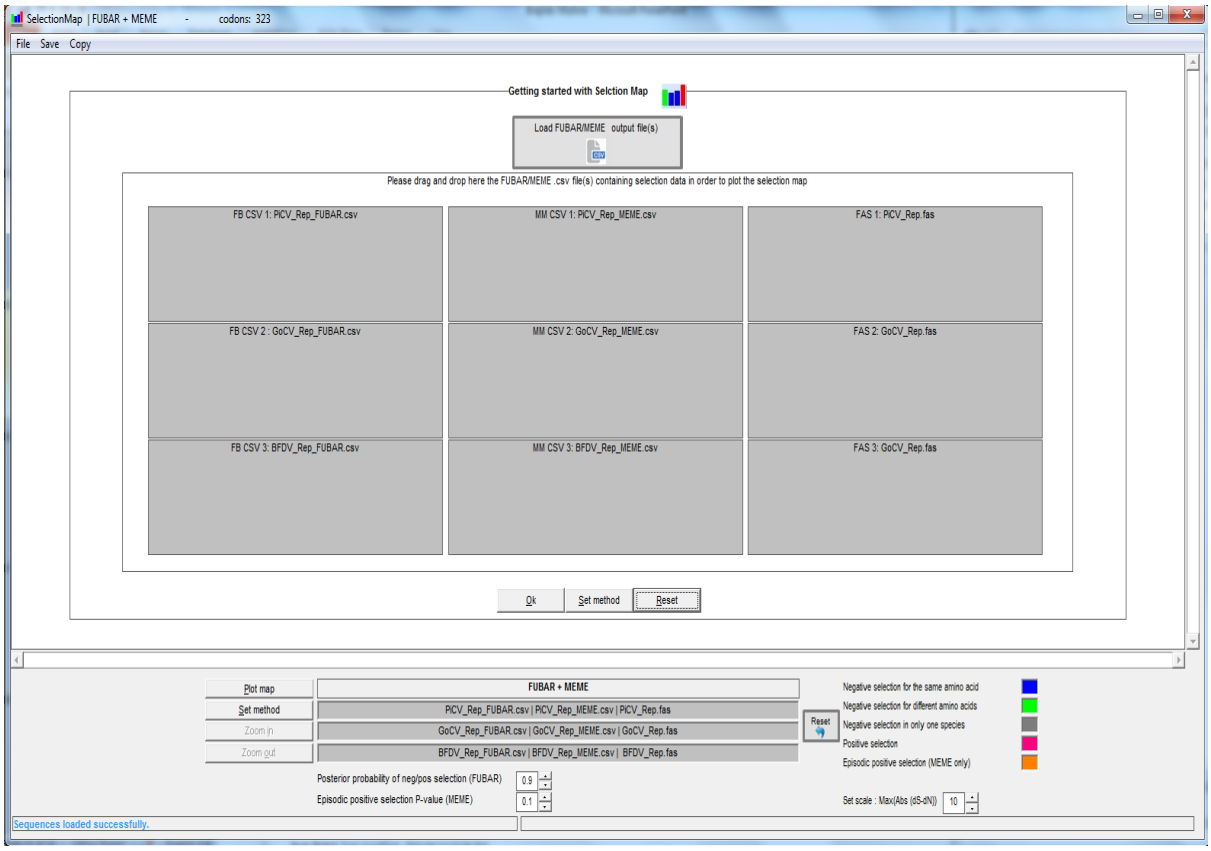

Fig 5. MEME input screen

Once the alignments are loaded, the “OK” command button must be pressed followed by the “Plot map” button, which will produce the selection map.

In cases where two or three selection analysis results are compared and the original alignments were not aligned to each other before detecting selection, the program will first perform a profile alignment of the input codon alignments and then will map the FUBAR/MEME selection data to their appropriate codon sites such that homologous codon sites are aligned to permit inter-species comparison.

V. Program features

1. Menus

a. The **File** menu gives the option load input files and exit the program

b. The **Save** menu contains several submenus allowing to save files in different formats:

i. **Save image as EMF:** allows selection maps to be saved in EMF format.

ii. **Save image as PNG:** allows selection maps to be saved in PNG format..

iii. **Save FUBAR result 1**: allows FUBAR selection results for input codon alignment 1 to be saved in CSV format.

iv. **Save FUBAR result 2**: allows FUBAR selection results for input codon alignment 2 to be saved in CSV format.

v. **Save FUBAR result 3**: allows FUBAR selection results for input codon alignment 3 to be saved in CSV format.

vi. **Re-save codon align 1:** allows input codon alignment 1 to be saved after it has been re-aligned to other input codon alignments.

vii. **Re-save codon align 2:** allows input codon alignment 2 to be saved after it has been re-aligned to other input codon alignments.

viii. **Re-save codon align 3:** allows input codon alignment 3 to be saved after it has been re-aligned to other input codon alignments.

ix. **Save translated align 1:** allows a translated version of the input codon alignment 1 to be saved.

x. **Save translated align 2:** allows a translated version of the input codon alignment 2 to be saved.

xi. **Save translated align 3:** allows a translated version of the input codon alignment 3 to be saved .

xii. **Save consensus AA sequence align 1:** allows the production and saving of a consensus amino-acid sequence for the input codon alignment 1.

xiii. **Save consensus AA sequence align 2:** allows the production and saving of a consensus amino-acid sequence for the input codon alignment 2.

xiv. **Save consensus AA sequence align 3:** allows the production and saving of a consensus amino-acid sequence for the input codon alignment 3.

c. The **Copy** menu allows the map to be copied and pasted in either EMF or PNG format, into another program such us MS PowerPoint or MS Word. An alternative option is to write-click on the image and select copy.

Fig 6. SelectionMap main interface

2. Command buttons

a. Plot map: also labelled “Load input” and “Detect selection” is a multi-purpose command button that allows the loading of input files, the running of selection detection analyses, and the plotting of selection maps following either the completion of selection analyses (Option 1) or the successful loading of selection data (Option 2).

b. Set method: allows the selection of the FUBAR, MEME and FUBAR+MEME methods.

c. Zoom in and **Zoom out:** these command buttons allow resizing of the map.

3. Posterior probability and P-value toggle buttons: these allow the setting of thresholds at which a codon sites are considered to be under negative/positive selection (for FUBAR method), or episodic positive selection (for the MEME method). The default values for the

posterior probability (FUBAR method) and P-value (MEME method) are 0.9 and 0.1 respectively.

4. **Set scale: Max(Abs(dN-dS)) toggle button:** allows setting the maximum value of the height of each selection map that is produced (y-axis).
5. **Reset:** this command button allows the program to be reset, removing all loaded input files and restoring all program variables to their start-up values.
6. **Colour boxes**: allow customisation of colour representations of different of types of selection.

Please cite (Stenzel et al. 2014)**:**

Stenzel, T. et al., 2014. Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of Beak and feather disease viruses. *The Journal of general virology*.

VI. References

dgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792–7.

Guindon, S. et al., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), pp.307–321.

Murrell, B. et al., 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*, 8(7), p.e1002764.

Murrell, B. et al., 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*, 30(5), pp.1196–205.

Pond, S.L.K., Frost, S.D.W. & Muse, S. V, 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5), pp.676–9.

Price, M.N., Dehal, P.S. & Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), p.e9490.

Stenzel, T. et al., 2014. Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of Beak and feather disease viruses. *The Journal of general virology*.

Authors: [1]Brejnev Muhire, [1]Penelope Hartnady, [2]Arvind Varsani and [1]Darren Martin

[1]Institute of Infectious Diseases and Molecular Medicine, Computational Biology Group, University of Cape Town, South Africa

[2]School of Life Sciences and Biodesign Institute, Arizona State University, Arizona State University, Tempe, AZ 85287-5401, USA

Bug reporting: mhrbre001@myuct.ac.za or mubrejnev@gmail.com