

Journal of Visualized Experiments

A Practical Guide to Phylogenetics for Non-Experts

--Manuscript Draft--

Manuscript Number:	JoVE50975R2
Full Title:	A Practical Guide to Phylogenetics for Non-Experts
Article Type:	Invited Methods Article - JoVE Produced Video
Keywords:	phylogenetics; multiple sequence alignments; phylogenetic tree; blast executables; basic local alignment search tool
Manuscript Classifications:	95.51.10: biology (general); 95.51.19: genetics (animal and plant); 95.51.9: biological evolution (terrestrial); 96.61.12: computer software
Corresponding Author:	Damien O'Halloran George Washington University Washington DC, DC UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author E-Mail:	damienoh@email.gwu.edu
Corresponding Author's Institution:	George Washington University
Corresponding Author's Secondary Institution:	
First Author:	Damien O'Halloran
First Author Secondary Information:	
Order of Authors Secondary Information:	
Abstract:	<p>Many researchers, across incredibly diverse foci, are applying phylogenetics to their research question(s). However, many researchers are new to this topic and so it presents inherent problems. Here we compile a practical introduction to phylogenetics for non-experts. We outline in a step-by-step manner, a pipeline for generating reliable phylogenies from gene sequence datasets. We begin with a user-guide for similarity search tools via online interfaces as well as local executables. Next, we explore programs for generating multiple sequence alignments followed by protocols for using software to determine best-fit models of evolution. We then outline protocols for reconstructing phylogenetic relationships via maximum likelihood and Bayesian criteria, and finally describe tools for visualizing phylogenetic trees. While this is not by any means an exhaustive description of phylogenetic approaches, it does provide the reader with practical starting information on key software applications commonly utilized by phylogeneticists. The vision for this article would be that it could serve as a practical training tool for researchers embarking on phylogenetic studies, and also serve as an educational resource that could be incorporated into a classroom or teaching-lab.</p>
Author Comments:	
Additional Information:	
Question	Response

A Practical Guide to Phylogenetics for Non-Experts

Author:

Damien M. O'Halloran^{1*}

Author: institution(s)/affiliation(s) for author:

Department of Biological Sciences and Institute for Neuroscience
George Washington University

Corresponding author:

*Damien M. O'Halloran

Keywords:

Phylogenetics, multiple sequence alignments, phylogenetic tree, blast executables, basic local alignment search tool, genomics

Short Abstract:

Here we describe a step-by-step pipeline for generating reliable phylogenies from nucleotide or amino acid sequence datasets. This guide aims to serve researchers or students new to phylogenetic analysis.

Long Abstract:

Many researchers, across incredibly diverse foci, are applying phylogenetics to their research question(s). However, many researchers are new to this topic and so it presents inherent problems. Here we compile a practical introduction to phylogenetics for non-experts. We outline in a step-by-step manner, a pipeline for generating reliable phylogenies from gene sequence datasets. We begin with a user-guide for similarity search tools via online interfaces as well as local executables. Next, we explore programs for generating multiple sequence alignments followed by protocols for using software to determine best-fit models of evolution. We then outline protocols for reconstructing phylogenetic relationships via maximum likelihood and Bayesian criteria, and finally describe tools for visualizing phylogenetic trees. While this is not by any

means an exhaustive description of phylogenetic approaches, it does provide the reader with practical starting information on key software applications commonly utilized by phylogeneticists. The vision for this article would be that it could serve as a practical training tool for researchers embarking on phylogenetic studies, and also serve as an educational resource that could be incorporated into a classroom or teaching-lab.

Introduction:

In order to understand how two (or more) species evolved, it is first necessary to obtain sequence or morphological data from each sample; these data represent quantities that we can use to measure their relationship through evolutionary space. Just like when measuring linear distance, having more data available (e.g. miles, inches, microns) will equate to a more accurate measurement. Ergo, the accuracy with which a researcher can deduce evolutionary distance is heavily influenced by the volume of informative data available to measure relationships. Furthermore, because different samples evolve at different rates and by different mechanisms, the method that we use to measure the relationship between two taxa also directly influences the accuracy of evolutionary measurements. Therefore, because evolutionary relationships are not directly observed but instead are extrapolated from sequence or morphological data, the problem of inferring evolutionary relationships becomes one of statistics. Phylogenetics is the branch of biology concerned with applying statistical models to patterns of evolution in order to optimally reconstruct the evolutionary history between taxa. This reconstruction between taxa is referred to as the taxa's *phylogeny*.

To help bridge the gap in expertise between molecular biologists and evolutionary biologists we describe here a step by step pipeline for inferring phylogenies from a set of sequences. Firstly, we detail the steps involved in database interrogation using the Basic Local Alignment Search Tool (BLAST¹) algorithm through the web based interface and also by using local executables; this is often the first step in obtaining a list of similar sequences to an unidentified query, although some researchers may also be interested in gathering data for a single group via web interfaces such as Phylota (<http://www.phylota.net/>). BLAST is an algorithm for comparing primary amino acid or nucleotide sequence data against a database of sequences to search for "hits" that resemble the query sequence. The BLAST program was designed by Stephen Altschul et al. at the National Institutes of Health (NIH¹). The BLAST server consists of a number of different programs, and here is a list of some of the most common BLAST programs:

i) *Nucleotide-nucleotide BLAST (blastn)*: This program requires a DNA sequence input and returns the most similar DNA sequences from the DNA database that the user specifies (e.g. for a specific organism).

ii) *Protein-protein BLAST (blastp)*: Here the user inputs a protein sequence and the program returns the most similar protein sequences from the protein database that the user specifies.

iii) *Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)*: The user input is a protein sequence which returns a set of closely related proteins, and from this dataset a conserved profile is generated. Next a new query is generated using only these conserved “motifs” which is used to interrogate a protein database and this returns a larger group of proteins from which a new set of conserved “motifs” are extracted and then used to interrogate a protein database until an even larger set of proteins are returned and another profile is generated and the process repeated. By including related proteins into the query in each step this program allows the user to identify sequences that are more divergent.

iv) *Nucleotide 6-frame translation-protein (blastx)*: Here the user provides a nucleotide sequence input which is converted into the six-frame conceptual translation products (i.e. both strands) against a protein sequence database.

v) *Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)*: This program takes a DNA nucleotide sequence input and translates the input into all six-frame conceptual translation products which it compares against the six-frame translations of a nucleotide sequence database.

vi) *Protein-nucleotide 6-frame translation (tblastn)*: This program uses a protein sequence input to compare against all six reading frames of a nucleotide sequence database.

Next, we describe commonly used programs for generating a Multiple Sequence Alignment (MSA) from a sequence dataset, and this is followed by a user guide to programs that determine the best-fit models of evolution for a sequence dataset. Phylogenetic reconstruction is a statistical problem, and because of this, phylogenetic methods need to incorporate a statistical framework. This statistical framework becomes an evolutionary model that incorporates sequence change within the dataset. This evolutionary model is comprised of a set of assumptions about the process of nucleotide or amino-acid substitutions, and the best-fit model for a particular dataset can be selected through statistical testing. The fit to the data of different models can be compared via likelihood ratio tests (LRTs) or information criteria to select the best-fit model within a set of possible ones. Two common information criteria are the Akaike information criterion (AIC)² and the Bayesian information criterion (BIC)³. Once an optimal alignment is generated, there are many different methods to create a phylogeny from the aligned data. There are numerous methods of inferring evolutionary relationships; broadly, they can be divided into two categories: distance-based methods and sequence-based methods. Distance-based methods compute pairwise distances from sequences, and then use these distances to obtain the tree. Sequence-based methods use the sequence alignment directly, and usually search the tree space using an optimality criterion. We outline two sequence-based methods for reconstructing phylogenetic relationships: these are PhyML⁴ which implements the maximum likelihood framework, and MrBayes⁵ which uses Bayesian Markov Chain Monte Carlo inference. Likelihood and Bayesian methods provide a statistical framework for phylogenetic reconstruction. By providing user information on commonly used tree-building tools, we introduce the reader to the necessary data required to infer phylogenetic relationships.

Protocol text:

1.) Basic Local Alignment Search Tool (BLAST): online interface

1.1) Click on this link to visit the BLAST¹ web server at the National Center for Biotechnology Information (NCBI). - <http://blast.ncbi.nlm.nih.gov/Blast.cgi> (Figure 1).

1.2) Input a FASTA formatted text sequence (see Figure 2 for example) into the query box.

1.3) Click the appropriate BLAST program and relevant database or individual species of interest to use in the search and then click "BLAST".

Note: FASTA formatted sequence begins with a description line indicated by a ">" sign. The description must follow immediately after the ">" sign, the sequence (i.e. nucleotides or amino acids) follow the description on the next line. The output from the BLAST search is viewed as HTML, plain text, XML, or hit tables (Text or csv) with the default set to HTML (Figure 3).

2.) Basic Local Alignment Search Tool (BLAST): local executables

2.1) Download the latest BLAST command-line BLAST executables from this link:

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/> -

2.2) For PC users: double-click the latest blast win32.exe file and accept the license agreement and click install.

Note: The default installation directory is C:\ncbi-blast-2.2.27+.

2.3) Configure the PC environment variable as follows:

2.3.1) Click the PC "start" button, and then right click "computer",

2.3.2) Click "Properties" and in the pop-up click on the "advanced" tab

2.3.3) Click the "Environment Variables button" and in the new pop-up click the "new" button under the "User variables for user" section

2.3.4) In the pop-up add the variable name "Path" and variable value "C:\ncbi-blast-2.2.27+\bin."

Note: the bin directory contains the executable (i.e. blastp etc.).

2.4) For Mac users: Open the Terminal application (to do this just open "Finder" and search "Terminal" and this will display the "terminal" icon). Into the terminal window type:

```
>ftp ftp.ncbi.nih.gov
```

Note: can also type the URL used above in the example for PC

2.5) To access the NCBI ftp site type “anonymous” for Name and Password, and then type:

```
>cd blast/executables/LATEST
```

2.6) List the executables by typing:

```
>ls
```

2.7) Get the latest version by typing the following (or whatever the latest version currently is):

```
>get ncbi-blast-2.2.7-macosx.tar.gz
```

2.8) Exit the NCBI ftp server site by typing “exit”.

2.9) Decompress the downloaded files by typing:

```
>tar -xzf ncbi-blast-2.2.7-macosx.tar.gz
```

2.10) Add the location of the binaries for the blast executable to your path so that the shell can search through this directory when looking for commands by typing:

```
>PATH=$PATH:new_folder_location
```

2.11) Check if this added the location to your path by typing:

```
>echo $PATH
```

2.12) Download a preformatted BLAST databases (which are updated daily) by clicking [here](http://ftp.ncbi.nlm.nih.gov/blast/db/):

[ftp://ftp.ncbi.nlm.nih.gov/blast/db/](http://ftp.ncbi.nlm.nih.gov/blast/db/)

2.13) Place the database into the “db” folder.

2.14) On a PC: open a MS-DOS prompt (to do this click “start” and type “cmd” in the search bar) and change the directory to the ncbi-blast folder by typing:

```
C:\Users>cd ..\ [moves up one folder]
```

```
C:\>cd ncbi-blast-2.2.27+
```

This will change the directory to:

```
C:\ncbi-blast-2.2.27+>
```

2.15) Create the database using the following “makedb” command:

```
>makedb -in db/briggsae.fasta -dbtype prot -out db/briggsae
```

Note: In the example below (**Figure 4**) the database is named “briggsae” and is comprised of one linkage group from the organism *Caenorhabditis briggsae*

2.16) Create a query protein sequence called “test” by inserting a FASTA formatted protein text sequence into the “db” folder.

2.17) Interrogate the database via a blastp search by typing the following command:

```
>blastp -query db/test.txt -db db/briggsae -out text.txt
```

2.18) On a Mac: download a database for local Blast searches by accessing the NCBI ftp website as per the instructions above (2.4) and then type:

```
>lcd ../databases/
```

2.19) Download the genome or sequence of interest by typing:

```
>get NC_[Accession #].fna
```

Note: “.fna” refers to the FASTA formatted nucleotide sequence and “.faa” refers to the FASTA formatted amino acid sequences.

2.20) Type “quit” to exit the ftp site.

2.21) Make the database by typing:

```
>makeblastdb -in db/mouse.faa -out mouse -dbtype prot
```

2.22) Insert a FAST formatted query sequence into the “bin” folder and interrogate the database with the following command:

```
> blastp -query “your query.fasta” -db “your database” -out results.txt
```

3.) Generating Multiple Sequence Alignments

3.1) Click on these links to access commonly used Multiple Sequence Alignment (MSA) programs:

ClustalW⁶ <http://www.clustal.org/>

Kalign⁷ <http://msa.sbc.su.se/cgi-bin/msa.cgi>

MAFFT^{8,9} <http://mafft.cbrc.jp/alignment/software/>

MUSCLE¹⁰ <http://www.drive5.com/muscle/>

T-Coffee¹¹ <http://www.tcoffee.org/Projects/tcoffee/>

PROBCONS¹² <http://toolkit.tuebingen.mpg.de/probcons>

3.2) Click on this link - <http://tcoffee.crg.cat/apps/tcoffee/do:regular> - and input FASTA formatted sequence data into the query box

Note: A sample output from T-Coffee can be seen in **Figure 5**, similar residues are color coded.

3.3) Download the Clustal MSA as a command line version (ClustalW) or a graphical version (ClustalX) by clicking this link: <http://www.clustal.org/clustal2/> - then click on the appropriate executable (i.e. win, linux, macosx).

3.4) Upload data as FASTA formatted sequence text and align (**Figure 6**)

4.) Determining best-fit models of evolution

4.1) Click here to download the ProtTest¹³ program:

<http://darwin.uvigo.es/our-software/>

4.2) Once ProtTest is downloaded, double-click on the ProtTest.jar file

4.3) Once ProtTest is launched, click on “select file” and load the sequence data (**Figure 7**).

4.4). Then click “start” and the program will begin (**Figure 8**).

Note: After completion of the run (**Figure 8**), the program will indicate the best model based on criteria e.g. “Best model according to AIC: WAG+I+G”

5.) Inferring sequence based phylogenies by maximum likelihood or Bayesian inference

5.1) Downloaded PhyML⁴ here:

<https://code.google.com/p/phyml/>

5.2) Launch the executable by double clicking the appropriate application (i.e. phym1 windows, phym1 linux etc.) and the interface window will pop up (**Figure 9**).

5.3) Load the input sequence as a PHYLIP formatted sequence by typing:

>“file name”.phy

Note: To convert between sequence formats, use the “Readseq” web program available at - <http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi>.

5.4) Launch the program by typing “Y”.

5.5) Download MrBayes⁵ here:

<http://mrbayes.sourceforge.net/download.php>

5.6) To start the program click on the executable file and read NEXUS formatted sequence data into the program by typing:

```
>execute "file name".nex
```

5.7) Set the evolutionary model.

5.8) Select the number of generations to run by typing:

```
>mcmcp ngen = 1000000 [this sets the number of generations to 1000000]
```

```
>sump burnin =10000 [this sets the burnin to 10000]
```

5.9) Save the branch lengths in the results file by typing:

```
>mcmcp savebrlens = yes
```

5.10) Run the analysis by typing:

```
>mcmc
```

5.11) Summarize the trees using the "sumt" command.

6.) Visualizing Phylogenies

6.1) View a list of tree viewer programs here:

<http://www.treedyn.org/overview/editors.html>

6.2) Download the TreeView¹⁴ program here:

<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

Representative Results:

Finding similarities to a query allows researchers to ascribe a potential identity to new sequences and also infer relationships between sequences. The file input type for BLAST¹ is FASTA formatted text sequence or GenBank accession number. FASTA formatted sequence begins with a description line indicated by a ">" sign (**Figure 2**). The description must follow immediately after the ">" sign, the sequence (i.e. nucleotides or amino acids) follow the description on the next line. When saving and editing sequence files, it is best to use a text editor such as "Notepad" on PC, or TextWrangler (<http://www.barebones.com/products/textwrangler/>) for Mac. The BLAST algorithm performs "local" alignments, which searches for short stretches of sequence

similarity. After the algorithm has looked up all possible "stretches" from the query sequence and maximally extended these sequences, it then assembles alignments for each query sequence pair. It is then important to understand how good these matches are, and so BLAST applies statistics to each hit which comprise an expect value (E) and a bit score. The E value gives an indication of the statistical significance for a match. The lower the E-value, the more significant the hit, for example a sequence alignment with an E-value of 0.05 means that the likelihood of this match occurring by chance alone is 5 in 100. The bit score uses a specific scoring matrix to provide an indication of how good the alignment is. The higher the bit score, the better the alignment. Similar to the online version of BLAST, there are a number of parameters that can be set via commands using the local BLAST executable. A comprehensive resource describing these commands can be found here - <http://www.ncbi.nlm.nih.gov/books/NBK1762/>. The output of the local search is a text file just like the output from the online BLAST interface (**Figure 4**).

A Multiple Sequence Alignment (MSA) is a sequence alignment of three or more primary sequences composed of amino acids, DNA, or RNA. ClustalW⁶ released in 1994, is one of the most popular MSA tools for biologists. A user friendly online interface that provides one-stop access to several popular MSA tools can be found at the EMBL-EBI server here - <http://www.ebi.ac.uk/Tools/msa>. The input for each program can be FASTA formatted sequence data (**see Figure 2**) although many different formats are also accepted, and numerous mirror sites for each can be found online. Numerous parameters like gap penalties and output formats can be easily chosen. A sample output from the MSA T-Coffee can be seen in **Figure 5**, where similar residues are color coded. In some cases, the MSA tool can also be downloaded and executed locally. Clustal can be downloaded as a command line version (ClustalW) or a graphical version (ClustalX) from this website - <http://www.clustal.org/clustal2/>. To download, just click on the appropriate executable (i.e. win, linux, macosx). For Windows the program executable will download and a pop-up menu will require the user to click "Run", and then installation will begin. The program is very intuitive, sequences can be loaded from a text file containing sequences formatted as NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF, and GDE. Sequences are aligned by clicking "do complete alignment" from the "alignment" menu. A sample alignment of six protein sequences aligned using ClustalX can be seen in **Figure 6**. Various parameters such as font size and color can be easily modified, and editing of sequences is done by clicking on the "Edit" menu. Manually refined alignments are often superior to fully automated methods and because of this, MSA tool development is a very active area of research. Some common alignment editors can be found at the following links: Se-Al - <http://tree.bio.ed.ac.uk/software/seal/>; BSEdit - <http://www.bsedit.org/>; JalView - <http://www.jalview.org/>; SeaView - <http://pbil.univ-lyon1.fr/software/seaview.html>.

For amino-acid alignments the program ProtTest¹³ is used to determine the selection of best-fit models of amino acid replacements within the data. ProtTest makes this selection by finding the model from the list of candidate models with the smallest Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) score or Decision Theory Criterion (DT). The latest version of ProtTest (ver. 3.2) includes 15 different rate

matrices that result in 120 different models. The user must have Java Runtime on their system to run ProtTest. Java Runtime is freely available here - <http://www.java.com/en/download/chrome.jsp>. Sequences are inputted as PHYLIP or NEXUS format. To convert between sequence formats, use the "Readseq" web program available at - <http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi>. Click on "select file" and load the sequence data. Then click "start" and the program will begin. To modify the number of models selected, you can click the "models" button. Once the program begins it will display a progress bar at the bottom and list the models as they are being analyzed (**Figure 8**). The developer's manual provides a comprehensive user guide to ProtTest (can be found in the "manual" folder when the program is downloaded) and also provides descriptions of the models and parameters. More background information can be found here - <https://code.google.com/p/prottest3/wiki/Background>. There is also an online web interface for ProtTest which functions just like the downloaded version except that it can only handle a limited number of sequences. This web interface can be accessed by clicking here - http://darwin.uvigo.es/software/prottest2_server.html. For nucleotide datasets the program jModelTest¹⁵ is used to examine the statistical selection of best-fit models of nucleotide substitutions by implementing the AIC, BIC, and DT criteria outlined above and also hierarchical and dynamical likelihood ratio tests (hLRT and dLRT). jModelTest is optimized for MacOSX. For the input, multiple formats are permitted. A clear step-by-step guide is available by the developers here - <http://computing.bio.cam.ac.uk/local/doc/jmodeltest.pdf>

PhyML is a program that estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences. PhyML will incorporate a large number of substitution models coupled to various options to search tree topology space (**Figure 10**). The program will save results into two text files. The first file will contain the ML tree in Newick format which can easily be viewed using a Tree viewer (see Step 6), and the other file will contain the statistics (filename, model, Log-likelihood scores etc.) of the analysis. All parameters are very easily set by following the Menu items. More detailed descriptions of each Menu option are explained in the PhyML manual available on the PhyML download page - <https://code.google.com/p/phyml/downloads/list>. MrBayes⁵ is a program that utilizes Bayesian MCMC inference across a number of evolutionary models to reconstruct phylogenetic relationships. The program behaves the same on all platforms and once downloaded the installer will install the executable. To start the program, simply click on the executable. There are numerous models that can be set and details of each model and their commands can be found here - <http://mr bayes.sourceforge.net/wiki/index.php/Tutorial>. Another help option is to type "help lset" – this will provide details on Model setting. For example "Prset aamodelpr=mixed" will permit mixed modeling or "prset aamodelpr=fixed(wag)" will set the amino acid model to the WAG model. An outgroup can be easily set by specifying the Taxon number "outgroup 30"; the program automatically lists the sequences/Taxa by number. If an outgroup is not specified the tree will be unrooted. Once the program is running (**Figure 11**) the progress can be viewed in specific intervals which can be set using the "printfreq=X" command. More details on when to stop the analysis (i.e. how many generations to run for) can be found in the user's manual. Clade values on a

cladogram are provided in the results alongside a phylogram which is also provided in Newick format that can easily be viewed using a tree viewer (see Step 6).

Once a phylogenetic tree is generated, the topology needs to be visualized. There are many online tools and downloadable applications used to visualize tree topologies. A partial list of popular programs can be viewed here - http://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software, and a more comprehensive list can be found here - <http://www.treedyn.org/overview/editors.html>. TreeView¹⁴ and TreeDyn¹⁶ are two popular choices. Both are very user friendly and easy to become familiar with the various options. TreeView runs on Mac and Windows, using almost identical interfaces. The input can be one of several formats including NEXUS, PHYLIP, Hennig86, MEGA, and ClustalW/X. TreeView (**Figure 12**) also includes a tree editor that allows the user to move branches, re-root trees, and rearrange the appearance of the tree.

Figures:

Figure 1: NCBI BLAST web-page.

The BLAST web server contains a suite of BLAST programs and is hosted by the National Center for Biotechnology Information (NCBI).

Figure 2: FASTA formatted sequence.

FASTA format begins with a description line indicated by a “>”. The description must follow immediately after the “>” sign, the sequence (i.e. nucleotides or amino acids) follow the description on the next line.

Figure 3: HTML output from a BLAST search.

The output from the BLAST search illustrates the areas of identity within the query sequence, and also provides bit-scores, expect values and pairwise alignments with each match.

Figure 4: A sample output from a local BLAST executable search.

The output of this search is a text file just like the output from the online BLAST interface, that include the expect value and bit score, as well as match description,

Figure 5: Output of a MSA using T-Coffee.

The output highlights similar sites and weights the match by color. Gaps are inserted as “-” signs and the residue or nucleotide position is preserved for each taxon.

Figure 6: A sample alignment using ClustalX. Similar matches are color coded and gaps are inserted as a “-” sign. The menu bar is seen in the top-left.

Figure 7: The ProtTest program interface.

Figure 8: The ProtTest console. ProtTest console while running an analysis. The progress bar indicates how many models have been completed, and the main window displays the log likelihood score for each model.

Figure 9: The PhyML interface.

Figure 10: The PhyML interface menu. Once sequences are loaded into PhyML the first menu appears, which can be navigated by typing the letter or symbol in the square bracket. Sub-menus can be reached by typing the “+” sign.

Figure 11: MrBayes Interface. When MrBayes is launched the progress can be viewed in specific intervals set using the “printfreq=X” command. Although the program cannot be stopped during a run, after the specified number of generations are computed the user will be asked if they want to run more generations.

Figure 12: The TreeView interface. In this Figure the TreeView window displays a sample tree of proteins from *Flybase* (<http://flybase.org/>). Files are imported by clicking the “open” option, and selecting an appropriate file type (e.g. Newick format).

Discussion:

Our hope for this article is that it will serve as a starting point to guide researchers or students that are new to phylogenetics. Genome sequencing projects have become less expensive over the last few years and as a consequence the user demand for this technology is increasing, and now the production of large sequence datasets is commonplace in small labs. These datasets often provide researchers with sets of genes that require a phylogenetic framework to begin to understand their function. Furthermore, because phylogenetics is finding a home in an ever increasing number of research labs, we also intend for this article to serve as an educational device for students interested broadly in biological research. By providing user information on the “why”, “how”, and “where” for commonly used tree-building tools, we provide a framework for the reader to begin to familiarize themselves with these applications and how they work. However, we advise the reader to play around with all the settings within each tool in an attempt to understand how the various parameters can influence their sequence data, and to ensure compatibility between platform and software in each case. The analysis outlined above was computed using a Dell Optiplex 990 with Intel core i7 processor and a MacBook laptop with an Intel Core 2 Duo processor, however, the speed of analysis and also the specific binaries (e.g. 32bit or 64bit) will depend on the user’s platform.

A challenge when compiling a user guide like this one for phylogenetics, is that the field of phylogenetics, and bioinformatics as a whole, is a rapidly expanding area of research that constantly releases new software aimed at providing better alignments, similarity predictions, or phylogenetic trees. To mitigate this problem, we tried to focus on programs that have been around for a number of years and are still popular on account of how well they work. That said, we want to point out that there are many other tools available to tackle the problems we have outlined in this article, and so encourage the reader to exploit this and incorporate multiple applications into their analyses.

Acknowledgments: We thank members of the O’Halloran lab for comments on the manuscript. We thank The George Washington University for Funding to D. O’Halloran.


Disclosures: We have nothing to disclose.

References:

1. Altschul SF, Carroll RJ, Lipman DJ. Weights for data related by a tree. *J Mol Biol.* 1989;207(4):647-653.
2. Akaike H. *IEEE Transactions on Automatic Control.* 1974;19(6):706-723.
3. Schwarz G. *The Annals of Statistics.* 1978;6(2):461-464.
4. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 2003;52(5):696-704.
5. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 2001;17(8):754-755.
6. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22(22):4673-4680.
7. Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics.* 2005;6:298. doi: 10.1186/1471-2105-6-298.
8. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005;33(2):511-518. doi: 10.1093/nar/gki198.
9. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 2002;30(14):3059-3066.
10. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792-1797. doi: 10.1093/nar/gkh340.
11. Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 2000;302(1):205-217. doi: 10.1006/jmbi.2000.4042.
12. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* 2005;15(2):330-340. doi: 10.1101/gr.2821705.
13. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics.* 2011;27(8):1164-1165. doi: 10.1093/bioinformatics/btr088; 10.1093/bioinformatics/btr088.

14. Page RD. TreeView: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci*. 1996;12(4):357-358.
15. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: More models, new heuristics and parallel computing. *Nat Methods*. 2012;9(8):772. doi: 10.1038/nmeth.2109; 10.1038/nmeth.2109.
16. Chevenet F, Brun C, Banuls AL, Jacq B, Christen R. TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*. 2006;7:439. doi: 10.1186/1471-2105-7-439.

*Figure 1
[Click here to download high resolution image](#)

 **BLAST®** *Basic Local Alignment Search Tool*

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

► **NCBI/ BLAST Home**

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

☐ [Human](#)

☐ [Mouse](#)

☐ [Rat](#)

☐ [Arabidopsis thaliana](#)

☐ [Oryza sativa](#)

☐ [Bos taurus](#)

☐ [Danio rerio](#)

☐ [Drosophila melanogaster](#)

☐ [Gallus gallus](#)

☐ [Pan troglodytes](#)

☐ [Microbes](#)

☐ [Apis mellifera](#)

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query

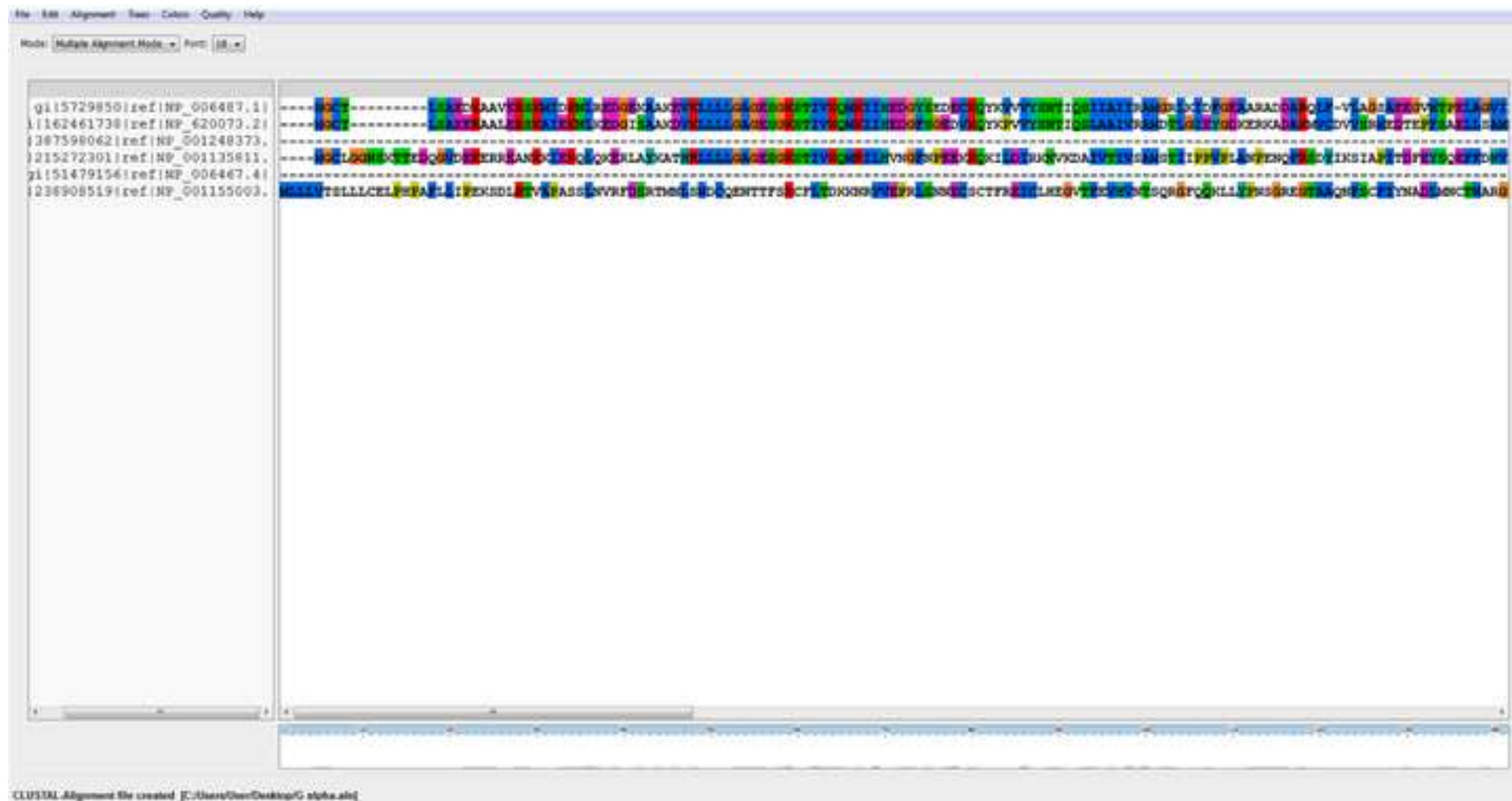
*Figure 2
[Click here to download high resolution image](#)

>gi|1164999|emb|X93022.1| A.thaliana mRNA for voltage gated potassium ion channel
ATAACAAATGGGTCAACAAGTTGCCATAAAGGTCCCAACAGAAGAAGAAAAAGAGAAGTTTCCCCGTGTCACCTTTGTCACAAGCTTCATCACTATTTATAACTACCTT
GCCAAGAAGAAAGCCTAAAAAGGTATAACAGTTTCTTTGTCTTTAGAGATCATCAAAAAGATGTCGATCTCTTGACTCGAAATTTCTTCGAAAGATTCTGCGTCGAGG
AATACAATATAGACACCATAAACAGAGTAGTTTCTCTCTGCCGATCTTCTACCATCTCTTGGAGCCAGGATCAACCAATCTACTAAGCTCCGCAAAACACATAATCTC
TCCTTTTAATCCACGTTACAGAGCGTGGGAGATGTGGCTAGTATTACTAGTTATTTACTCAGCTTGGATTTGCCCATTTCAATTTGCTTTTCATCACCTATAAAAAAGACG
CGATTTTCATCATCGACAACATTGTTAATGGCTTCTTCGCCATTGATATTATTCTCACCTTCTTCGTCGCTTATCTCGATAGCCACTCCTATCTTCTAGTTGACAGTCCT
AAGAAAATAGCAATAAGGTACCTTTTCGACGTGGTTCGCTTTTCGATGTTTGTTCACAGCACCATTTCAGCCACTAAGCCTCTTGTTAACTACAACGGAAGCGAACTA
GGATTCAGAATTCTTAGCATGCTCAGGTTATGGCGTCTCCGGCGAGTTAGCTCGCTATTTGCAAGGCTTGAGAAAAGATATCCGTTTCAACTATTTCTGGATACGTTGC
ACAAAACTCATTTCCGGTCACTTTGTTTCGCTATACATTGTGCTGGATGTTTCAACTACCTGATTGCAGATAGATATCCTAATCCAAGAAAGACATGGATTGGAGCTGTGT
ATCCAAATTTCAAAGAAGCAAGTCTATGGAATAGATATGTGACTGCTCTTTACTGGTCCATTACGACATTAACGACCACGGGATATGGAGATTTTCATGCTGAGAACCC
AAGAGAAATGCTTTTTGACATTTTCTTCATGATGTTCAACCTCGGTTTGACAGCTTACCTCATTGGAAATATGACCAACCTCGTCGTTTCATTGGACTAGCCGAACCAGA
ACCTTTAGGGATTCAGTGAGAGCTGCTTCAGAGTTTGCTTCAAGAAATCAACTCCACATGACATAGAAGATCAAATGTTATCACACATTTGCTTAAAGTTCAAAACAG
AGGGCTTGAAACAACAAGAGACCTTGAACAATCTGCCAAAAGCAATCCGGTCAAGCATTGCAAACATTTATTCTTCCCCATTGTTCAACAACATTTACCTCTTTCAAGG
AGTTTCTCGTAACTTCCTCTTTCAATTGGTTTCAGATATAGACGCTGAGTATTTCCACCAAAGAAGATATAATTCTACAAAACGAAGCTCCTACTGATCTTTACATTC
TGGTGTCAAGGAGCAGTGGACTTCACTGTCTACGTTGATGGACATGATCAGTTTCAAGGGAAAGCAGTAATTGGAGAAACATTTGGAGAGGTTGGAGTTTATACTATA
GACCACAACCATTCACAGTAAGAACAACCGAGCTGTCTCAAATACTACGGATAAGCAGAACATCGCTGATGAGTGCGATGCATGCTCATGCTGACGATGGACGAGTC
ATCATGAACAATCTTTTCATGAACTTAGAGGGCAACAGTCAATAGCAATAGATGATTGCAATACTAGTGGTCACGAAAACAGAGATTTCAAAGCATGGGATGGGAA
GAGTGGAGAGATTCAAGAAAAGATGGCTATGGTTTAGATGTTACGAATCCGACTTCCGACACTGCTCTAATGGATGCGATTACAAGGAAGATACTGAAATGGTTAAG
AAGATACTTAAGGAACAAAAGATAGAGAGAGCCAAAGAGGAAAAGATCTAGTAGTGAAAGCGCTGGAAGAAGTTACGCTAACGATTTCATCGAAAAAAGATCCATATTGC
AGCTCAAGCAACCAATCATCAAGCCATGCAAACGAGAAGAAAAGAGAGTTACCATCCACATGATGTCTGAGAGCAAGAACGGGAAGTTGATACTCGTACCATCATC
CATAGAAGAGCTTCTAAGACTTGCAAGTGAGAAGTTTGGAGGCTGCAACTTCACAAAGATCACCAATGCGGACAACGCTGAGATTGATGATTTAAATGTCATTTGGGA
TGGTGATCATTTGTATTTTTCATCAAATTGAGTTTGAAAACCTCGACTTCATTTATAGAGCATGTATATCTGCAGACAATGTATTTTACCCGGTTTCATAGAAAAGTCTAG
ATTATCAAAAAAAAAAAAAA

*Figure 3
[Click here to download high resolution image](#)



*Figure 6
[Click here to download high resolution image](#)



*Figure 7
[Click here to download high resolution image](#)

[About](#) [WWW](#) [Help](#)

ProtTest

(PAL & PhymI based)

Alignment

alignment

Select file

Tree

☒ BIONJ tree
☐ User tree

Select file

Program options

Optimization strategy...

Fast (opti...
▲▼

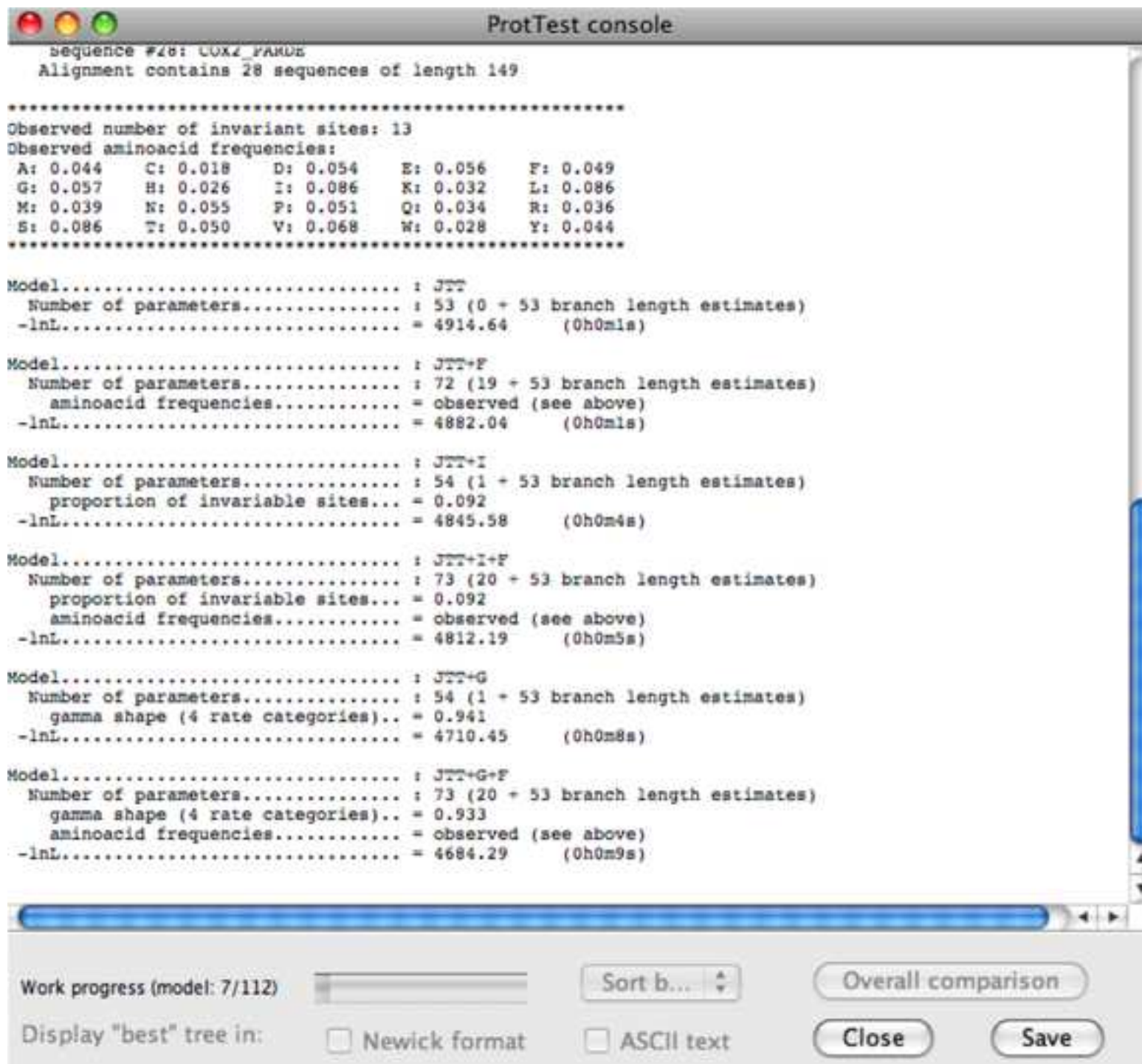
Set of candidate...

models

Exit

Stop

*Figure 8
[Click here to download high resolution image](#)



*Figure 10
Click here to download high resolution image

```

oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooo

          ---  PhyML 20110526  ---

  A simple, fast, and accurate algorithm to estimate large phylogenies by maxi
num likelihood
          Stephane Guindon & Olivier Gascuel

          http://www.atgc-montpellier.fr/phyml

          Copyright CNRS - Universite Montpellier II

oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooo

          .....
          Menu : Input Data
          .....

          [+] ..... Next sub-menu
          [-] ..... Previous sub-menu
          [Y] ..... Launch the analysis

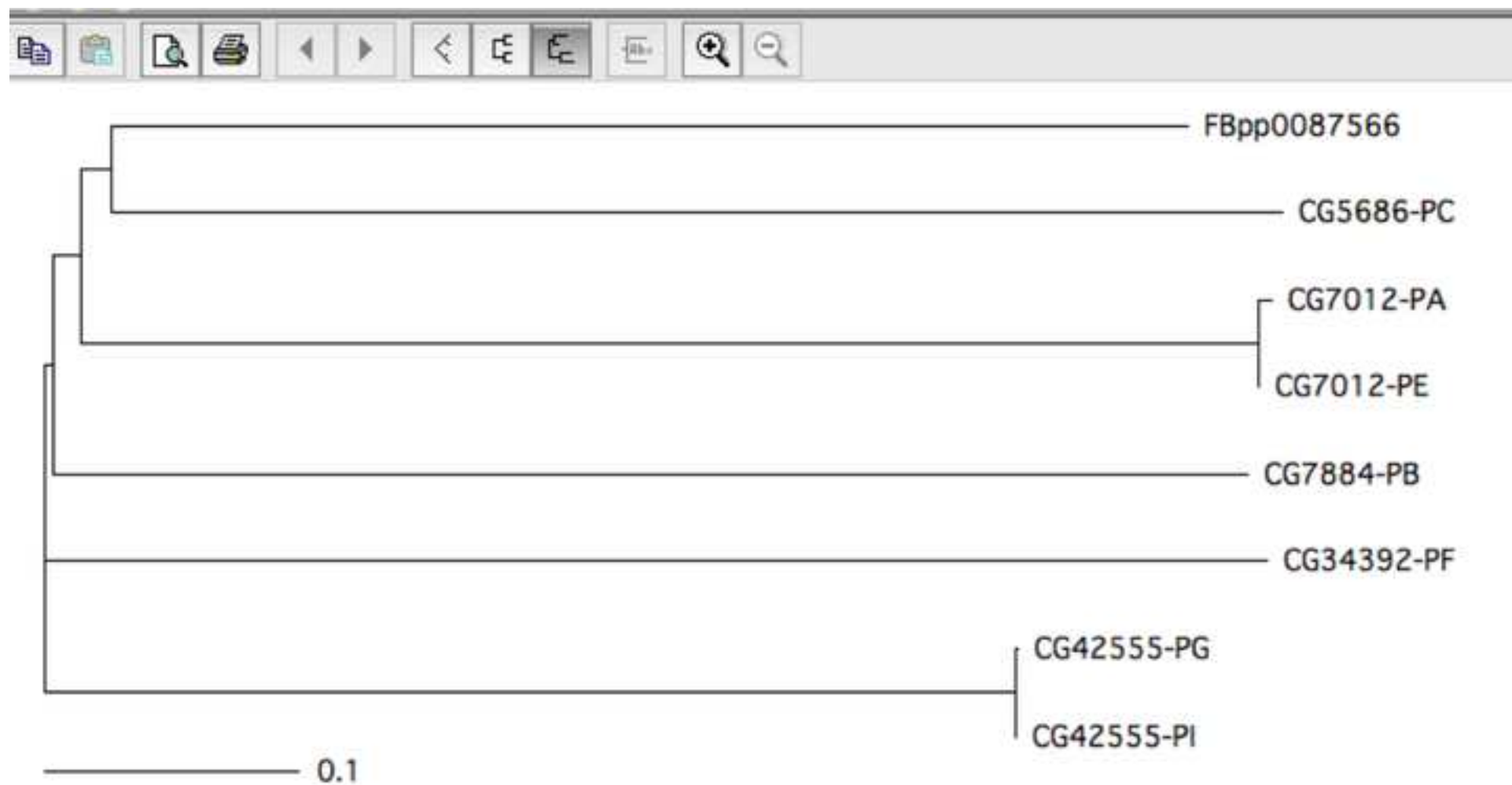
          [D] ..... Data type (DNA/AA/Generic)  DNA
          [I] ..... Input sequences interleaved (or sequential)  interlea
ved
          [M] ..... Analyze multiple data sets  no
          [R] ..... Run ID  none

. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to ch
ange)
```

*Figure 11
[Click here to download high resolution image](#)

Chain results:											
1	—	[−75717.735]	[−76930.858]	[−75343.419]	[−73657.639]	=	[−74183.283]	[−76998.215]	[−76262.816]	[−76855.122]	
100	—	[−69998.825]	[−71832.528]	[−67508.599]	[−68815.484]	=	[−68177.288]	[−69364.483]	[−70367.726]	[−72169.888]	11:06:36
200	—	[−65919.493]	[−68414.818]	[−61354.868]	[−63119.567]	=	[−65328.287]	[−65368.397]	[−63435.898]	[−66688.686]	9:43:13
300	—	[−61182.114]	[−61885.466]	[−57427.898]	[−58439.748]	=	[−56583.375]	[−62352.955]	[−59893.891]	[−60889.168]	7:15:23
400	—	[−57984.538]	[−59297.317]	[−55668.824]	[−56351.325]	=	[−53478.867]	[−60589.886]	[−55326.618]	[−56518.388]	9:01:27
500	—	[−55515.243]	[−56218.617]	[−54462.468]	[−54264.885]	=	[−58874.458]	[−56891.818]	[−52885.596]	[−54714.212]	8:53:04
600	—	[−53894.973]	[−53889.318]	[−51325.854]	[−53231.211]	=	[−49579.617]	[−53478.875]	[−51365.358]	[−52317.612]	8:47:27
700	—	[−51739.217]	[−52496.243]	[−50527.971]	[−51598.566]	=	[−47267.373]	[−52867.929]	[−50826.585]	[−50698.998]	8:43:26
800	—	[−51194.833]	[−51965.892]	[−49573.384]	[−50311.641]	=	[−45888.492]	[−49568.351]	[−49171.827]	[−48349.547]	8:48:25
900	—	[−50862.689]	[−50995.132]	[−48837.889]	[−49372.649]	=	[−45252.483]	[−47768.185]	[−48866.146]	[−47884.516]	8:38:03
1000	—	[−49821.648]	[−50456.766]	[−47498.438]	[−48643.142]	=	[−44937.872]	[−47229.114]	[−47419.437]	[−46325.581]	8:36:09
Average standard deviation of split frequencies: 0.152888											
1100	—	[−48882.447]	[−49351.552]	[−46781.842]	[−47848.316]	=	[−43588.198]	[−46584.988]	[−46238.196]	[−45626.848]	8:34:35
1200	—	[−48176.575]	[−48915.274]	[−46447.618]	[−46873.188]	=	[−43368.247]	[−45883.128]	[−45782.846]	[−44885.164]	8:33:16
1300	—	[−47492.264]	[−47759.387]	[−46338.284]	[−46672.572]	=	[−43167.618]	[−44778.898]	[−45657.189]	[−44443.255]	8:32:09
1400	—	[−47325.491]	[−46691.528]	[−46118.296]	[−46505.558]	=	[−42638.533]	[−44473.978]	[−44118.618]	[−44248.182]	8:31:11
1500	—	[−45978.852]	[−45185.432]	[−45988.788]	[−46171.498]	=	[−42111.872]	[−43328.786]	[−43784.489]	[−43687.278]	8:30:28
1600	—	[−45868.414]	[−44445.179]	[−44686.266]	[−45934.941]	=	[−41483.293]	[−43849.584]	[−43119.599]	[−43149.592]	8:29:36
1700	—	[−45652.885]	[−43871.979]	[−43272.189]	[−45887.557]	=	[−41311.925]	[−42766.894]	[−42969.599]	[−42334.771]	8:28:56
1800	—	[−45597.528]	[−43178.775]	[−42955.631]	[−44413.118]	=	[−41187.576]	[−42439.132]	[−42988.524]	[−41231.627]	8:28:28
1900	—	[−45488.295]	[−42375.584]	[−42588.191]	[−44275.225]	=	[−41832.814]	[−42385.514]	[−41813.315]	[−40958.454]	8:27:48
2000	—	[−45367.286]	[−41948.243]	[−41135.391]	[−44869.718]	=	[−40918.595]	[−42115.218]	[−41525.458]	[−40554.585]	8:27:19
Average standard deviation of split frequencies: 0.124478											
2100	—	[−45837.383]	[−41784.114]	[−40468.471]	[−43883.822]	=	[−40881.898]	[−41985.164]	[−41817.335]	[−39888.338]	8:34:47
2200	—	[−44185.536]	[−41641.194]	[−40321.336]	[−43788.859]	=	[−40881.882]	[−41673.651]	[−40834.737]	[−39738.959]	8:34:01
2300	—	[−43831.842]	[−41553.539]	[−40164.879]	[−42957.291]	=	[−40744.418]	[−41582.887]	[−40481.781]	[−39635.975]	8:33:18
2400	—	[−43186.677]	[−41486.281]	[−39998.914]	[−42625.819]	=	[−40729.564]	[−40589.852]	[−40185.512]	[−39559.235]	8:32:39
2500	—	[−42945.692]	[−41438.548]	[−39876.435]	[−42548.165]	=	[−40729.888]	[−40533.668]	[−39669.233]	[−39499.288]	8:32:03
2600	—	[−42618.431]	[−41419.752]	[−39829.879]	[−42448.741]	=	[−40788.928]	[−40458.771]	[−39586.123]	[−39425.235]	8:31:29
2700	—	[−42316.818]	[−41318.488]	[−39687.323]	[−42315.435]	=	[−40662.628]	[−40358.637]	[−39516.183]	[−39352.224]	8:30:57
2800	—	[−42121.683]	[−41143.832]	[−39588.473]	[−41844.783]	=	[−40688.261]	[−40292.248]	[−39387.818]	[−39894.868]	8:30:28
2900	—	[−42086.852]	[−40886.832]	[−39458.288]	[−41296.832]	=	[−40597.663]	[−40154.382]	[−39286.131]	[−39872.292]	8:30:08
3000	—	[−41976.784]	[−40787.868]	[−39297.829]	[−41223.388]	=	[−40553.856]	[−40139.982]	[−39199.654]	[−39825.681]	8:35:07
Average standard deviation of split frequencies: 0.145138											
3100	—	[−41979.114]	[−40343.392]	[−39286.951]	[−41879.886]	=	[−40554.382]	[−40135.515]	[−39166.487]	[−38943.924]	8:34:31
3200	—	[−41966.368]	[−40886.876]	[−39269.514]	[−40798.359]	=	[−40499.875]	[−39956.856]	[−39854.813]	[−38912.758]	8:33:58

*Figure 12
[Click here to download high resolution image](#)



Name of Reagent/ Equipment	Source
BLAST webpage	http://blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST executables	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Preformatted BLAST databases	ftp://ftp.ncbi.nlm.nih.gov/blast/db/
Clustal	http://www.clustal.org/
Kalign	http://msa.sbc.su.se/cgi-bin/msa.cgi
MAFFT	http://mafft.cbrc.jp/alignment/software/
MUSCLE	http://www.drive5.com/muscle/
T-Coffee	http://www.tcoffee.org/Projects/tcoffee/
PROBCONS	http://toolkit.tuebingen.mpg.de/probcons
Se-AI	http://tree.bio.ed.ac.uk/software/seal/
BSEdit	http://www.bsedit.org/
JalView	http://www.jalview.org/
SeaView	http://pbil.univ-lyon1.fr/software/seaview.html
ProtTest	https://code.google.com/p/prottest3/
Java Runtime	http://www.java.com/en/download/chrome.jsp
Readseq	http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi
jModelTest	https://code.google.com/p/jmodeltest2/
PhyML	https://code.google.com/p/phyml/
MrBayes	http://mrbayes.sourceforge.net/download.php
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
TreeDyn	http://www.treedyn.org/

This piece of the submission is being sent via mail.



REVIEWER #1:

Major Concerns:

1) Major concerns is that within the protocol some steps are taken/performed without an explanation. However they are then explained in the sample results section. I found this counter intuitive and could be improved by supplementing some of the protocols with text. I've suggested splitting up some of the protocol steps i.e ML and Bayesian reconstruction. Finally my biggest problem is that the author flips between PC and MAC software. I'm familiar with all softwares but found this a little tricky myself. I think one should be taken as the standard (MAC or PC) and then at the end of each mini protocol section alterations to the other OS should be mentioned. Also online and local analysis is confusing. Maybe these could be split up into discrete protocols also!

Response: *The protocol has to be strictly only the step-by-step guide for the video and thus the discussion section contains explanations. Regarding the MAC to PC switch: I believe it is important to design the protocol to facilitate both MAC and PC users so as maximize reader interest and applicability.*

Minor Concerns:

1) In keywords "basic local alignment tool" should be "basic local alignment search tool"

Response: *I have made this change throughout.*

2) In the Short abstract last line change "comparative analysis" to "phylogenetic analysis"

Response: *I have made this change.*

3) In the long abstract change "maximum likelihood and Bayesian probability" to "maximum likelihood and Bayesian criteria"

Response: *I have made this change.*

4) In the long abstract change "embarking on comparative studies" to "embarking on phylogenetics studies".

Response: *I have made this change.*

5) In the introduction having more available data does not necessarily lead to more accurate phylogenetic inferences, the data must also be phylogenetically informative. It is important to make this differentiation.

Response: *Changed to: "volume of informative data available to measure..."*

6) In the introduction: "because different samples evolve (change)" no need to have change in brackets after evolve

Response: *I have made this change.*

7) In the introduction change "The BLAST program was designed by Stephen Altschul, Warren Gish,



Webb Miller, Eugene Myers, and David J. Lipman at the National Institutes of Health (NIH1)." Change to "The BLAST program was designed by Stephen Altschul and colleagues at the National Institutes of Health (NIH1)."

Response: *I have made this change.*

8) In the introduction change "these are PhyML which uses Maximum Likelihood" to "these are PhyML which implements the maximum likelihood framework"

Response: *I have made this change.*

9) Protocol text change 'Basic Local Alignment Tool (BLAST)' to 'Basic Local Alignment Search Tool (BLAST)'

Response: *I have made this change.*

10) Protocol 1.2: explain briefly what a fasta formatted file looks like. I.e starts with a greater than sign. For clarity might be better if figure 2 did not have functional description as may confuse individuals. Basically sequence name and functionality of gene is not required. Also worth mentioning sequence itself with leading ">" is permitted as well.

Response: *Added the following to the "NOTE" after 1.3: "FASTA formatted sequence begins with a description line indicated by a ">" sign. The description must follow immediately after the ">" sign, the sequence (i.e. nucleotides or amino acids) follow the description on the next line"*

11) Protocol 1.3: change "and genome to use in the search" to and "relevant database or individual species of interest"

Response: *I have made this change.*

12) Protocol 2.4) Downloading for MAC users, why use ftp in this case and not for PC. To keep consistency I'd suggest downloading mac executables from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/> as well.

Response: *Added a "Note" stating - Note: can also type the URL used above in the example for PC*

13) Protocol 2.13) Place the database into the "db" folder. Should list the name of the DB in this case briggsae.fasta as is confusing in later steps.

Response: *Out of scope as I already have a Note clearly stating this.*

14) Protocol 2.14 does windows have "terminals" should it not be msdos prompt?

Response: *I have made this change.*

15) Protocol 2.16: need to specify that it's a protein sequence file

Response: *I now state that it is a "protein" sequence.*

16) Protocol 2.17: would be worth including an expectation "e" value and briefly explain it's significance as well as suggested cutoff (10⁻⁵)

Response: *This is discussed in detail in the representative results section.*



17) Protocols 3 for generating MSA needs to be improved. Seen as the author is doing BLAST searches locally and online it would be an idea to do MSA online and locally as well. The author should also include manual manipulation of alignments using for example Se-Al (for mac) or Jalview for PC. Maybe an additional step?

Response: *Out of scope. I already include details about alignment editors in the representative results section.*

18) For model inference the link is incomplete should be <http://darwin.uvigo.es/our-software/>
This package also has an online server so this should be highlighted also!

Response: *I have made this change.*

19) Protocol 5: need to describe what a phylip formatted sequence looks like.

Response: *Added the following "Note" to show the reader how to change between file type: Note: To convert between sequence formats, use the "Readseq" web program available at - <http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi>.*

20) For phylml and mrbayes there is no description how to set the evolutionary model, maybe will be clearer in video.

Response: *out of scope, discussed in representative results.*

21) Might be a better idea to split ML and Bayesian phylogeny reconstruction into 2 separate sections as is confusing together.

Response: *out of scope, as the goal here is to get the reader downloading the software and inputting files. The discussion section provides detailed user guides and resources.*

22) For Bayesian "ngen" and "sump burnin" need to be explained briefly

Response: *These are explained in the square brackets in each point*

REVIEWER #2:

1) Major comments:

It would be helpful to include an outline of the pipeline, or a schematic of the pipeline, near the beginning of the introduction so that readers know they are going.

Response: *Out of scope, we provide an overview in the introduction and provide step-by-step protocol.*

2) A short section on diagnosis of MCMC output may be required for the interpretation of the MrBayes results to guide the reader on how to evaluate whether they should run their analysis for more generations.

Response: *Out of scope, the goal of this paper is to introduce the reader as to which software to use to build reliable phylogenies, how to download them and input data into them. I provide detailed*



descriptions and resources in the discussion section on interpretation.

3) It is worth mentioning that if an outgroup is not specified a tree remains unrooted, and how that impacts how we interpret the evolutionary pattern of the phylogeny.

Response: *I have added the following: "If an outgroup is not specified the tree will be unrooted"*

Minor comments/edits:

1) To reduce wordiness, move the BLAST citation to the references rather than having it in the text.

Response: *Done: please refer to Minor Point #7 from Reviewer 1.*

2) Some researchers may be interested in gathering sequences from genbank for a single group via web interfaces such as phylota rather than by BLAST searches. This could be incorporated into the introduction and would broaden the audience for which this tutorial would be helpful.

Response: *Added the following to the Introduction: "...unidentified query, although some researchers may also be interested in gathering data for a single group via web interfaces such as Phylota (<http://www.phylota.net/>)"*

3) Given that Clustal is the alignment tool used in the tutorial, consider citing Wong et al. 2008 Science paper as a resource for more information.

Response: *This article is not specifically related to Clustal.*