BlastPhyMe: A toolkit for rapid generation and analysis of protein-coding sequence datasets

Ryan K Schott[*1,#], Daniel Gow[*2], Belinda SW Chang[1,2]

[1]Department of Ecology and Evolutionary Biology, [2]Department of Cell and Systems Biology, University of Toronto, Ontario, Canada

#Current Address: Department of Vertebrate Zoology, National Museum of National History, Smithsonian Institution, Washington DC, USA

[*]Equal contribution

ABSTRACT

SUMMARY: We present BlastPhyMe (BLAST, Phylogenies, and Molecular Evolution) a new application to facilitate the fast and easy generation and analysis of protein-coding sequence datasets. The application uses a portable database framework to manage and organize sequences along with a graphical user interface (GUI) that makes the application extremely easy to use. BlastPhyMe utilizes several existing services and applications in a unique way that save researchers considerable time when building and analyzing protein-coding datasets. The application consists of two modules that can be used separately or together. The first module enables the assembly of coding sequence datasets. BLAST searches can be used to obtain all related sequences of interest from NCBI. Full GenBank records are saved within the database and coding sequences are automatically extracted. A feature of particular note is that sequences can be sorted based on NCBI taxonomic hierarchy before export for visualization using existing tools, such as fast. The application provides GUIs for automatic alignment of sequences with the popular tools MUSCLE and PRANK, as well as for reconstructing phylogenetic trees using PhyML. The second module incorporates selection analyses using codon-based likelihood methods. The alignments and phylogenetic trees generated with the dataset module, or those generated elsewhere, can be used to run the models implemented in the codeml PAML package. A GUI allows easy selection of models and parameters. Importantly, replicate analyses with different parameter starting values can be automatically performed in order to ensure selection of the best-fitting model. Multiple analyses can be run simultaneously based on the number of processor cores available, while additional analyses will be run iteratively until completed. Results are saved within the database and can be exported to publication-ready Excel tables, which further automatically compute the appropriate likelihood ratio test between models in order to determine statistical significance. Future updates will add additional options for phylogenetic reconstruction (eg, MrBayes) and selection analyses (eg, HYPHY). BlastPhyMe saves researches of all bioinformatics experience levels considerable time by automating the numerous tasks required for the generation and analysis of protein-coding sequence datasets using a straightforward graphical interface.

AVAILABILITY: Installation package and source code available from: https://github.com/ryankschott/BlastPhyMe

CONTACT: schottr@si.edu

INTRODUCTION

With continued advances in Next Generation Sequencing (NGS) technology and ongoing genome sequencing projects (e.g., Genome 10K project; Koepfli et al. 2015) the amount of publically available sequence data is growing exponentially. This data provides a valuable resource for biologists, but as the number of available sequences continues to grow the generation and analysis of comparative sequence datasets has become a daunting task, especially for researchers that lack bioinformatics and scripting experience. While databases such as GenBank provide access to the vast number of publically available sequences, searching for, finding, and extracting sequences from Genbank through a web browser is extremely time consuming. Additionally, many bioinformatics tools for the alignment and analysis of sequences are command-line based, which can be difficult to use, produce outputs that are difficult to understand, and are time consuming to manually parse. To address these difficulties we have developed **BlastPhyMe** (BLAST, Phylogenies, and Molecular Evolution) a new application that automatically gathers, organizes, and analyzes gene sequences within a portable database framework using an intuitive graphical user interface (GUI). BlastPhyMe consists of two modules that can be used separately or together: Gene Sequences and Selection Analyses (PAML). The Gene Sequences module enables the assembly of coding sequence datasets. This includes searches and automatic extraction of sequences from GenBank using direct and BLAST searches, import of user data, multiple sequence alignments, and phylogenetic analyses. The Selection Analyses (PAML) module incorporates selection analyses using codon-based likelihood methods. The alignments and phylogenetic trees generated with the Gene Sequences module, or those generated elsewhere, can be used to run the models implemented in the codeml PAML package using a series of GUIs. Results are saved within the database and can be exported to publication-ready Excel tables, which further automatically compute the appropriate statistical tests to evaluate model significance.

METHODS

BlastPhyMe is a Microsoft Windows application coded in the C# programming language on the Microsoft .NET Framework platform, version 4.0 (http://msdn.microsoft.com/en-us/library/zw4w595w(v=vs.100).aspx). The database engine that BlastPhyMe interacts with for storing data is the Microsoft SQL Server 2014 Express LocalDB engine (http://msdn.microsoft.com/en-ca/library/hh510202(v=sql.120).aspx). For communicating with the NCBI BLASTN web service, BlastPhyMe uses the .NET Bio open source library (https://github.com/dotnetbio/bio). BlastPhyMe has been designed to the standard of an n-Tier application. The application code is separated between three distinct layers: user interface, middle tier, and database.

The User Interface layer includes all code necessary to display visual interfaces for the user to interact with. User Interface code does not interact directly with the Database layer or third-party systems (e.g., NCBI) and is abstracted from both by objects within the Middle-Tier layer. The Middle-Tier layer includes all code necessary to transfer data between the User Interface and Database layer. Code within the Middle-Tier interprets the database architecture into objects that are exposed for the User Interface to interact with. The Middle-Tier is responsible for all direct communication with the database, and all interaction with third-party systems and products (e.g., .NET Bio) is contained within the Middle-Tier layer. The Database layer comprises all of the architecture necessary to store data for the BlastPhyMe application as well as code to manipulate that data within the database. The BlastPhyMe database exposes stored procedures to handle all data collection and modification processes performed by the Middle-Tier layer. BlastPhyMe is capable of exporting data from its database to a “data file”, distinguished by the “.bpmd” file extension. A BlastPhyMe data file is an XML document compressed via the GZip compression algorithm.

BlastPhyMe communicates with NCBI's GenBank nucleotide database to search for and download GenBank records. This communication is performed via HTTP requests of the E-utilities web services hosted by NCBI (http://www.ncbi.nlm.nih.gov/books/NBK25499/). BlastPhyMe also communicates with NCBI's BLAST web service using the aforementioned .NET Bio libraries, which make use of the QBlast URL API (http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html).

In addition to communicating with NCBI, BlatPhyMe utilizes several third party programs. These programs are accessed and run through the BlastPhyMe interface. Gene sequences and alignments can be exported to MEGA (Tamura et al. 2013; http://www.megasoftware.net/) for visualization and editing. Multiple sequence alignments can be performed using PRANK (Löytynoja and Goldman 2008; http://wasabiapp.org/software/prank/) and MUSCLE (Edgar 2004; http://www.drive5.com/muscle/). Phylogenetic trees can be inferred using PhyML (Guidon et al. 2010; http://www.atgc-montpellier.fr/phyml/binaries.php). Resulting phylogenetic trees can be sent to TreeView (https://code.google.com/archive/p/treeviewx/) for visualization. Selection analyses can be performed using PAML (Yang 2007; http://abacus.gene.ucl.ac.uk/software/paml.html). Finally, sequences and PAML results tables can be exported to Microsoft Excel (www.microsoftstore.com).

INSTALLATION

BlastPhyMe comes with a complete installation package available at: https://github.com/ryankschott/BlastPhyMe/releases. Prerequisites will be installed automatically. Third party programs need to be downloaded and installed separately. A complete installation guide is available.

FEATURES AND USE

Upon running BlastPhyMe for the first time the user will be asked to create a database. This database will store all of the sequences, analyses, and results that you generate, or import into, BlastPhyMe. The database is stored in a .mdf file. Each user will only need a single database and these can be shared between users. After creating a database file the user will be prompted to create a project. Projects are a way to organize sets of similar data and each database can have multiple projects.

BlastPhyMe consists of two distinct modules: (1) Gene Sequences and (2) Selection Analyses (PAML). These can be accessed via the tabs as shown in Figure 1. Each module can be used completely independently of the other, but they are also designed to offer a continuous workflow from dataset generation to selection analyses as shown in Figure 2.

Gene Sequences

The Gene Sequences module is organized into datasets. Each dataset can contain any number of sequences from one or more genes and species. To create a dataset click the new dataset button. Sequences can be added to the dataset using the 'Search GenBank' function. User data (e.g., FASTA file) can also be imported using the 'Import from' function. The 'Search GenBank' function opens a dialog box allowing search terms to be submitted to NCBI GenBank (Fig. 3). All search terms supported by the Genbank website are also supported when submitting with BlastPhyMe. Search results will appear in a separate window (Fig. 4). Sequences that the user wishes to download can be selected and then added to an existing or new dataset using the 'Add to' button. This will save the complete GenBank record for the selected sequences.

Once added to a dataset, sequences can be doubled clicked to open up a separate window with additional information about the sequence including a link to the GenBank page (Fig. 5). Sequences can

be ordered by clicking on any of the column headings. Sequences can be filtered using both text matching and taxonomic filtering using the 'Apply Filter' function. Sequences can be deleted and moved or copied to another dataset using the 'Move to', 'Copy to', and 'Delete' buttons. Sequences can be exported to different formats, including FASTA, or opened directly into MEGA for visualization and editing, using the 'Export to button'.

The dataset can be further expanded using the 'BLAST for Similar Sequences' function. This function will submit selected sequences for BLAST analysis against the full NCBI nucleotide database (Fig. 6). Resulting BLAST hits will be automatically combined, removing duplicates, and can be downloaded and added to a dataset (Fig. 6). This function allows the user to quickly expand a dataset to include all available sequences of a particular gene or set of genes.

Selected sequences from a dataset can be aligned using either PRANK or MUSCLE using the 'Align with' function. This will open a dialog box allowing parameters to be set (Fig. 7). It will be necessary to add the location of PRANK.exe and MUSCLE.exe on first use. Sequences submitted for PRANK codon alignment will be automatically trimmed to the last complete codon. Upon completion of the alignment a new window will open allowing the aligned sequences to be selected and added to a new (aligned) dataset. The output of the alignment will also be save to the selected working directory. The aligned dataset (or an alignment imported separately using the 'Import from' function) can then be submitted for phylogenetic analysis using the 'Generate Tree with PhyML' function. This function will open a dialog box to initiate phylogenetic inference using PhyML (Fig. 8). The location of PhyML.exe will need to be added at first use. Upon completion a dialog box will open with links that will open the resulting tree in TreeView for visualization. Treeview can also be used to label foreground branches/clades for the PAML branch, branch-site, and clade models implemented in the Selection Analyses (PAML) module (see PAML manual for details). Results will be automatically saved in the selected working directory.

A complete history of BLASTN, MUSCLE, PRANK, and PhyML jobs is stored and can be accessed from the Gene Sequences dropdown menu.

Selection Analyses (PAML)

The Selection Analyses (PAML) module is similarly organized into datasets except that instead of storing sequences and GenBank records they store the results of PAML analyses. PAML jobs are initiated using the 'New PAML Job' button. This will open a dialog box allowing PAML analyses to be set-up (Fig. 9). Upon first use the location of codeml.exe will need to be specified. The number of processes to use and the working (output) directory should also be specified.

When creating a new PAML job, clicking the 'Add' button will open a new dialog box (Fig. 9). Here a tree and sequence alignment file will need to be specified. If a phylogenetic analysis was performed using BlastPhyMe these can be found in the PhyML output folder that was specified. The 'Add' button in this dialog window allow models to be specified to run with the selected tree and alignment files. Models are specified using the drop-down menu, and for the site models multiple models can be specified simultaneously using the check boxes (Note: completion time is often faster when sites models are added individually). Starting values for kappa and omega can be set and specifying a range of values will automatically set-up replicated analyses with each combination of starting values in the range. Multiple sets of models can be run for each tree and alignment pair and multiple tree and alignment pairs can be run with each PAML Job. Once executed with the 'Run' button, BlastPhyMe will initiate codeml for the specified number of process and will run sequentially through each specified model until complete. A progress window will be displayed but this can be close and the job will continue to be run in the background.

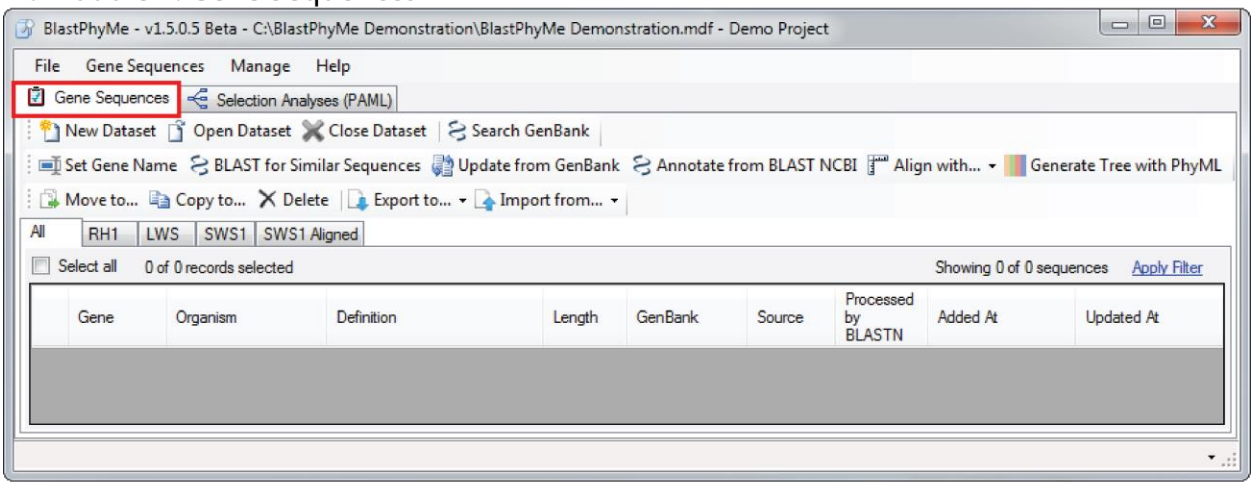
Upon completion of a PAML job a results dialog box will appear. Each line will display the best replicate for the model specified. These can be double clicked to see the results of all replicates. Select results can be added to a new or existing dataset. From within a dataset selected results can be exported to a preformatted Excel table using the 'Export to' function (Fig. 10). This produces a publication ready table that automatically computes the appropriate likelihood ratio tests to determine significance.

CONCLUSIONS

BlastPhyMe saves researchers of all bioinformatics experience levels considerable time when generating and analyzing sequence datasets. BlastPhyMe incorporate existing services and utilities in a novel portable database framework with intuitive user interfaces not offered by other bioinformatics programs BlastPhyMe is still under active development with planned updates including additional options for phylogenetic reconstruction (eg, MrBayes) and selection analyses (eg, HYPHY), as well as a new module to facilitate evolutionary medicine studies.

FIGURES

A. Module 1: Gene Sequences



B. Module 2: Selection Analyses (PAML)

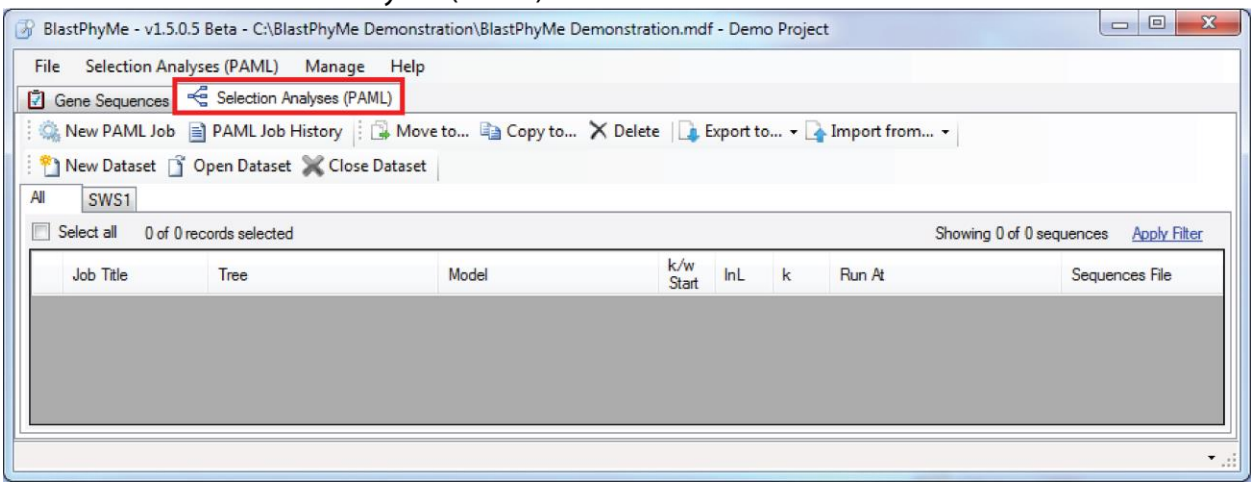


Figure 1. The two distinct modules of BlastPhyMe: Gene Sequences and Selection Analyses (PAML).

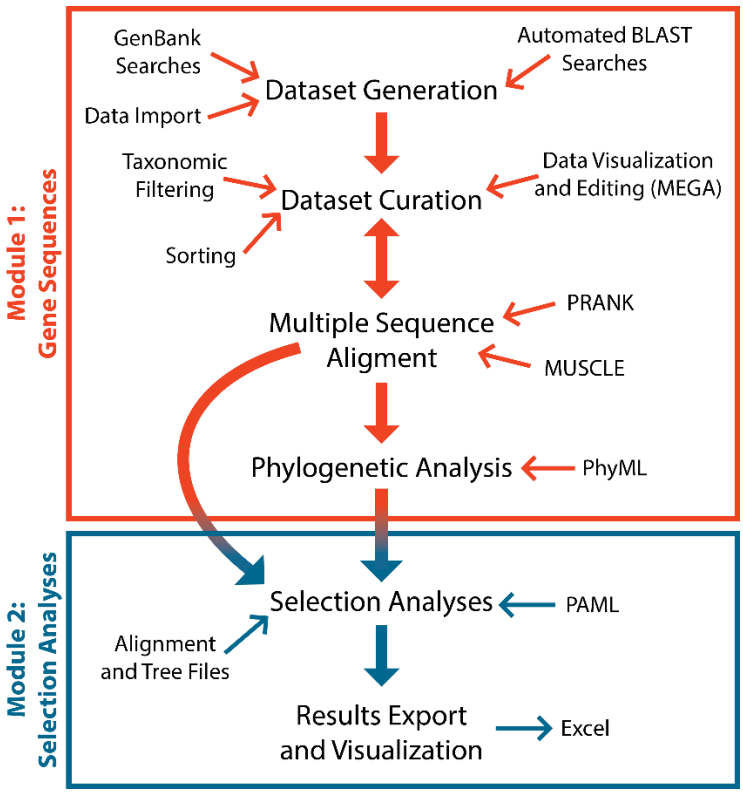


Figure 2. BlastPhyMe workflow. The two modules are distinct, but results from Module 1 provide the necessary input files for Module 2.

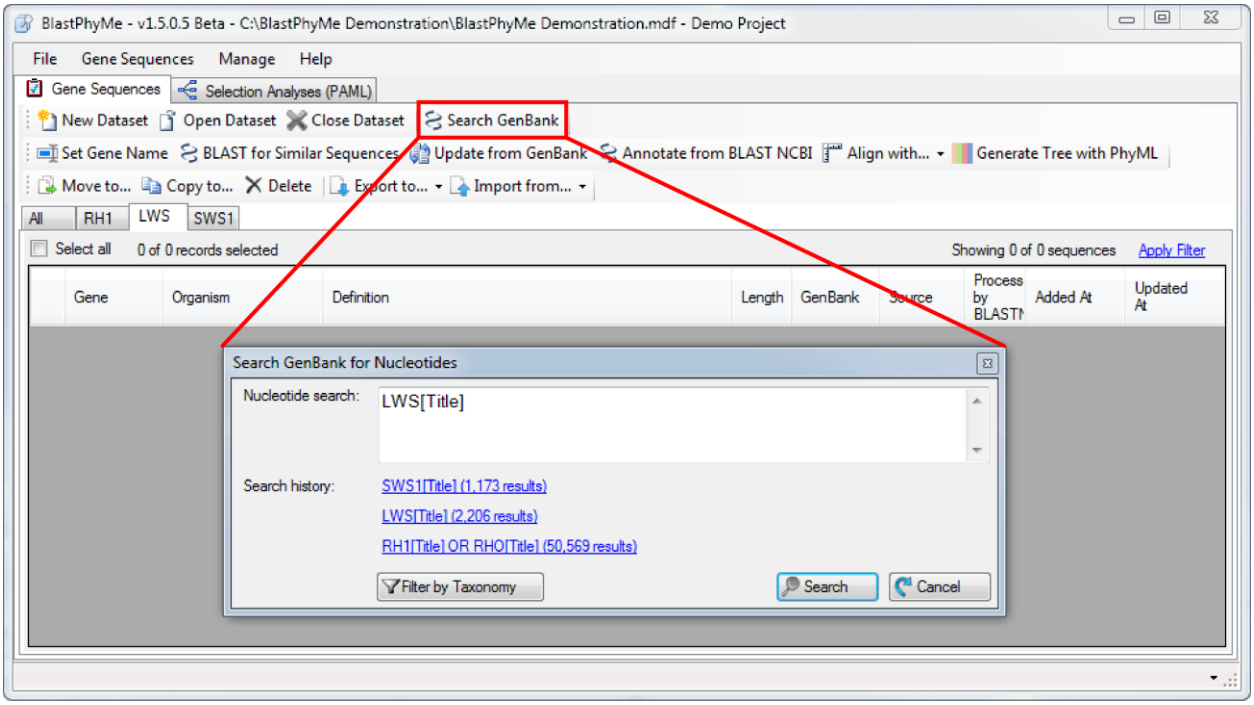


Figure 3. GenBank Search Function. This function directly searches the GenBank nucleotide database. All search terms used on the webpage can be used with BlastPhyMe.

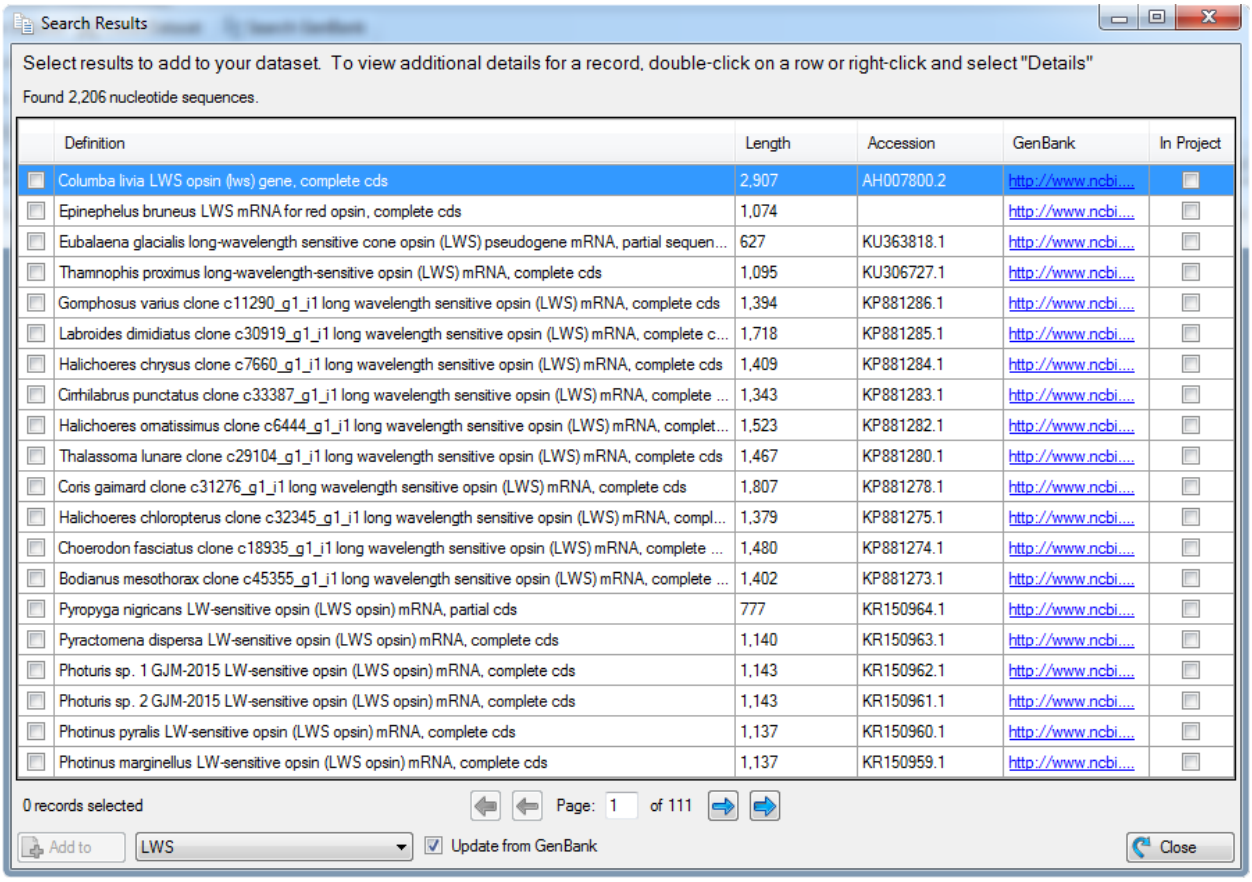


Figure 4. Genbank Search results window. Sequences can be selected using the check boxes and added to a new or existing dataset using the 'Add to' button.

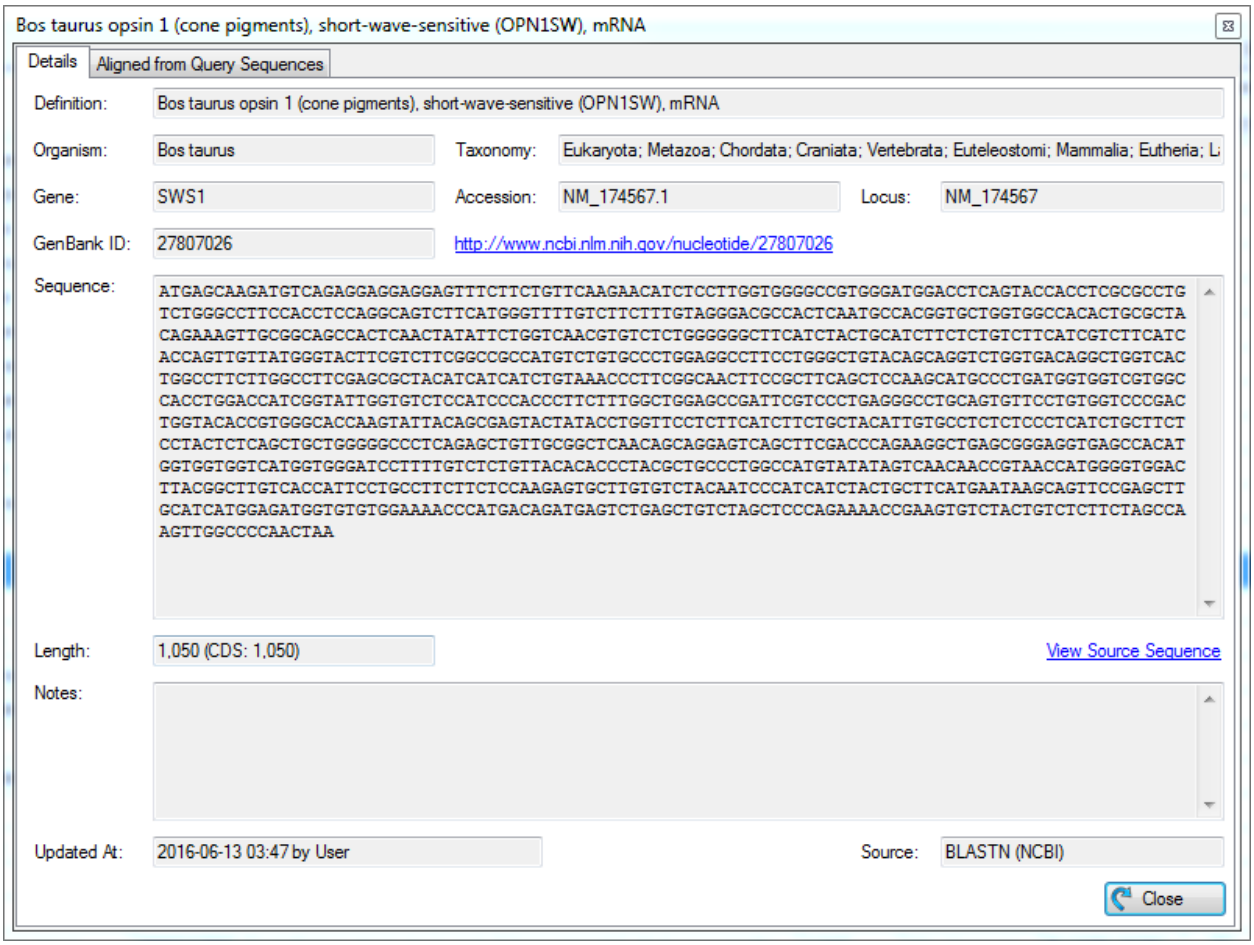


Figure 5. Additional data available for each sequence stored within the database accessible by double clicking on a sequence entry.

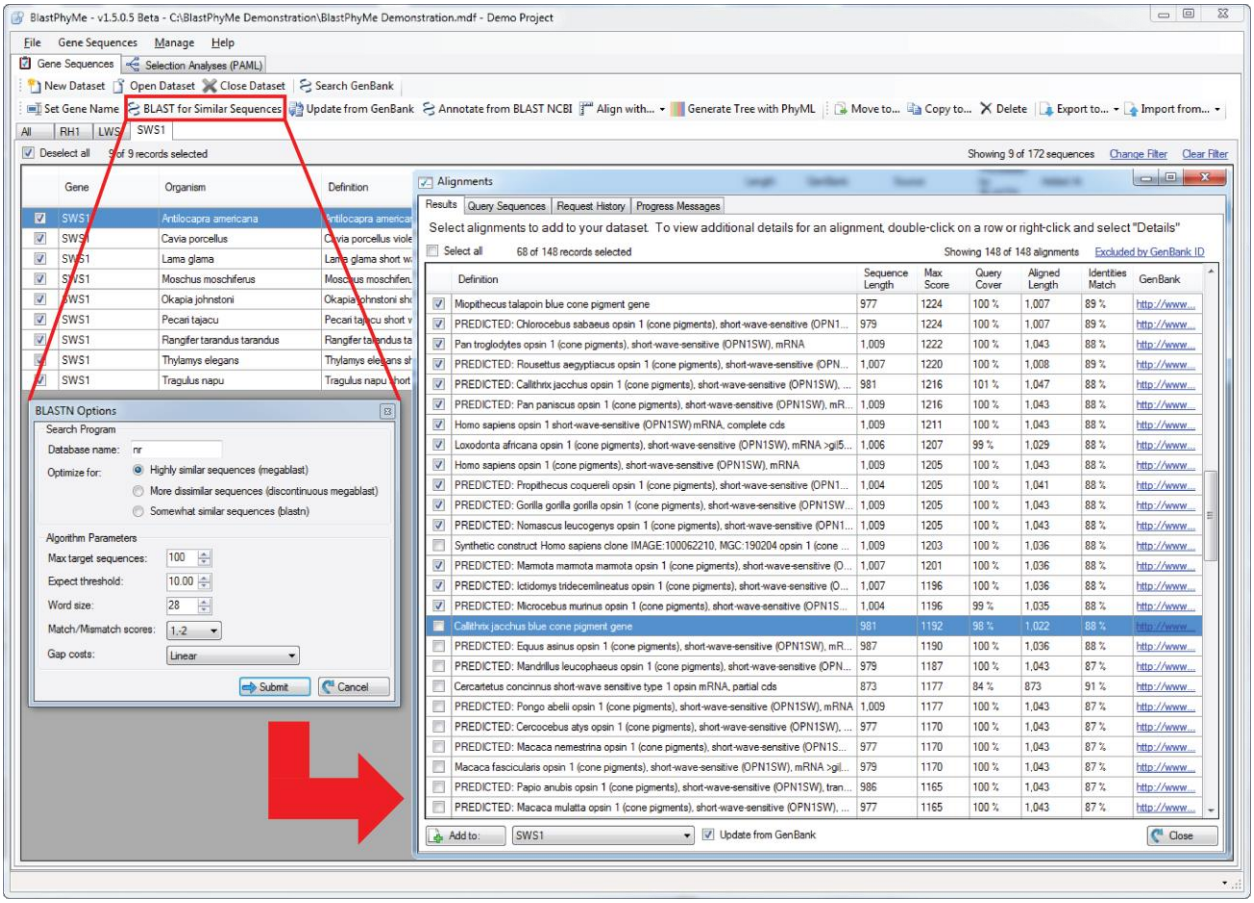


Figure 6. BLASTN search and results windows used to expand a dataset to include additional related sequences from the NCBI nucleotide database.

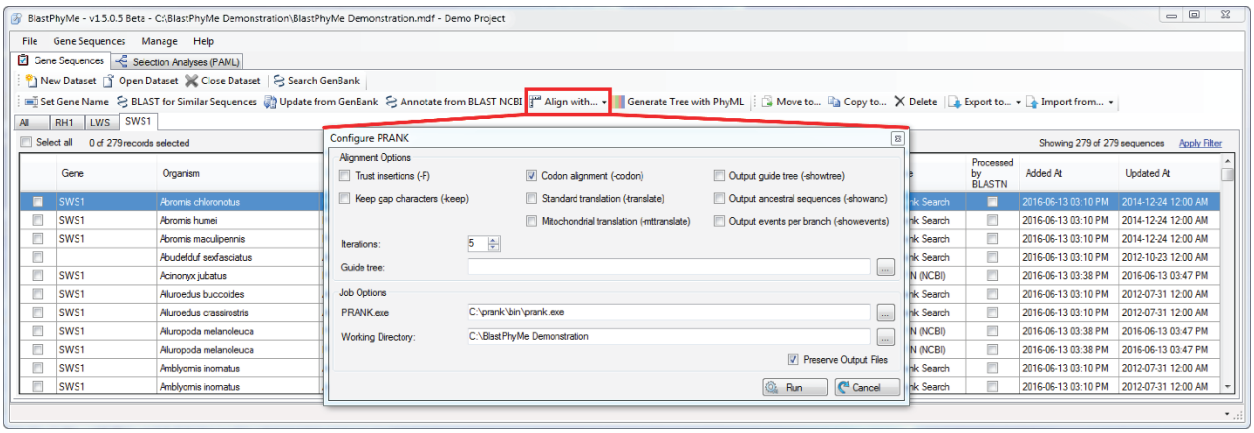


Figure 7. PRANK alignment window. A similar window is also available to set up alignments using MUSCLE.

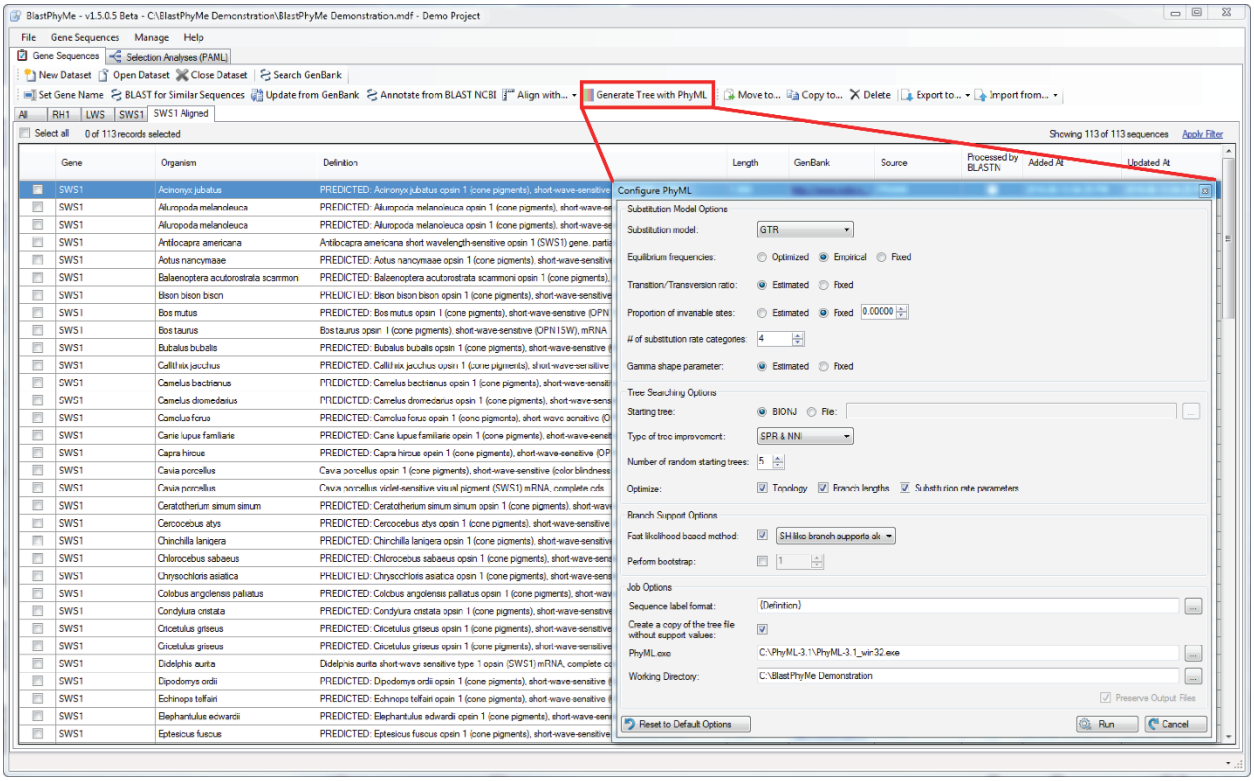


Figure 8. PhyML window to set up a phylogenetic analysis.

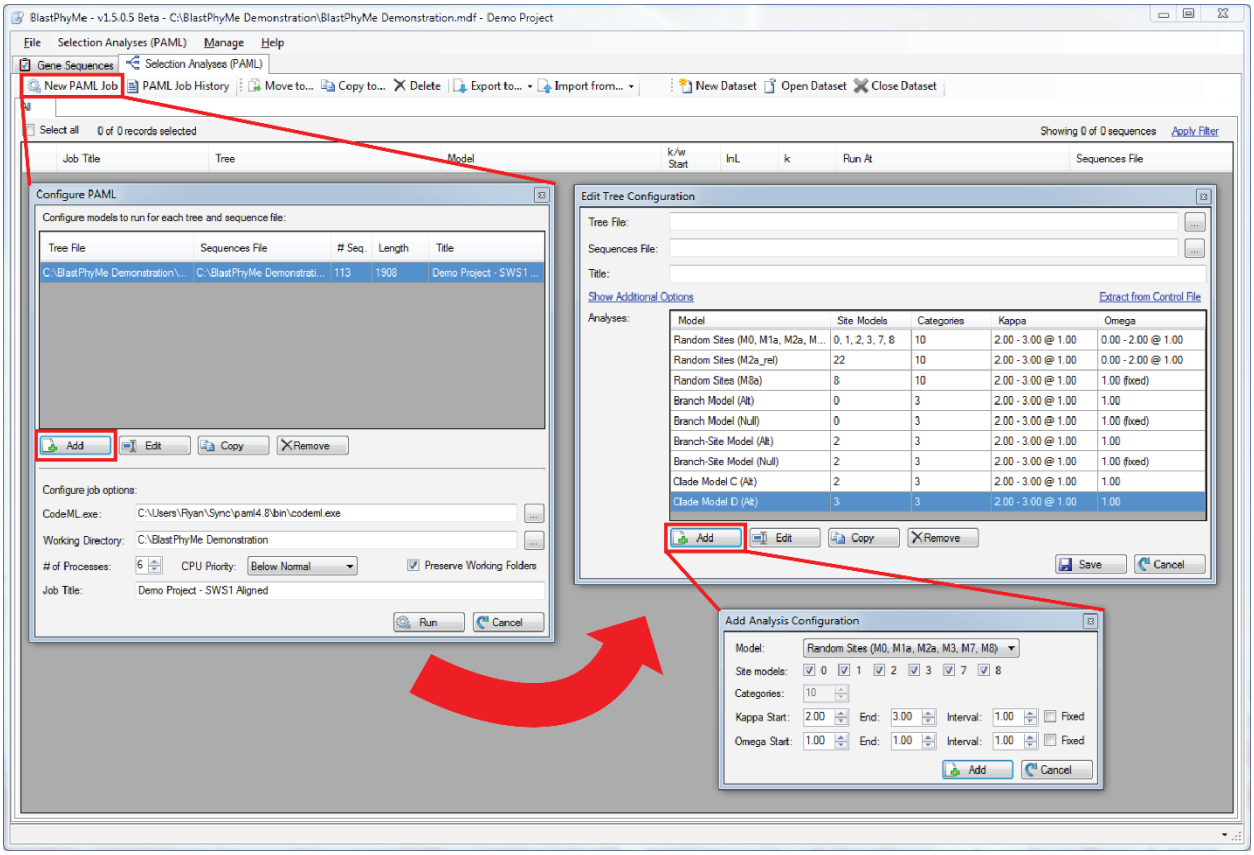
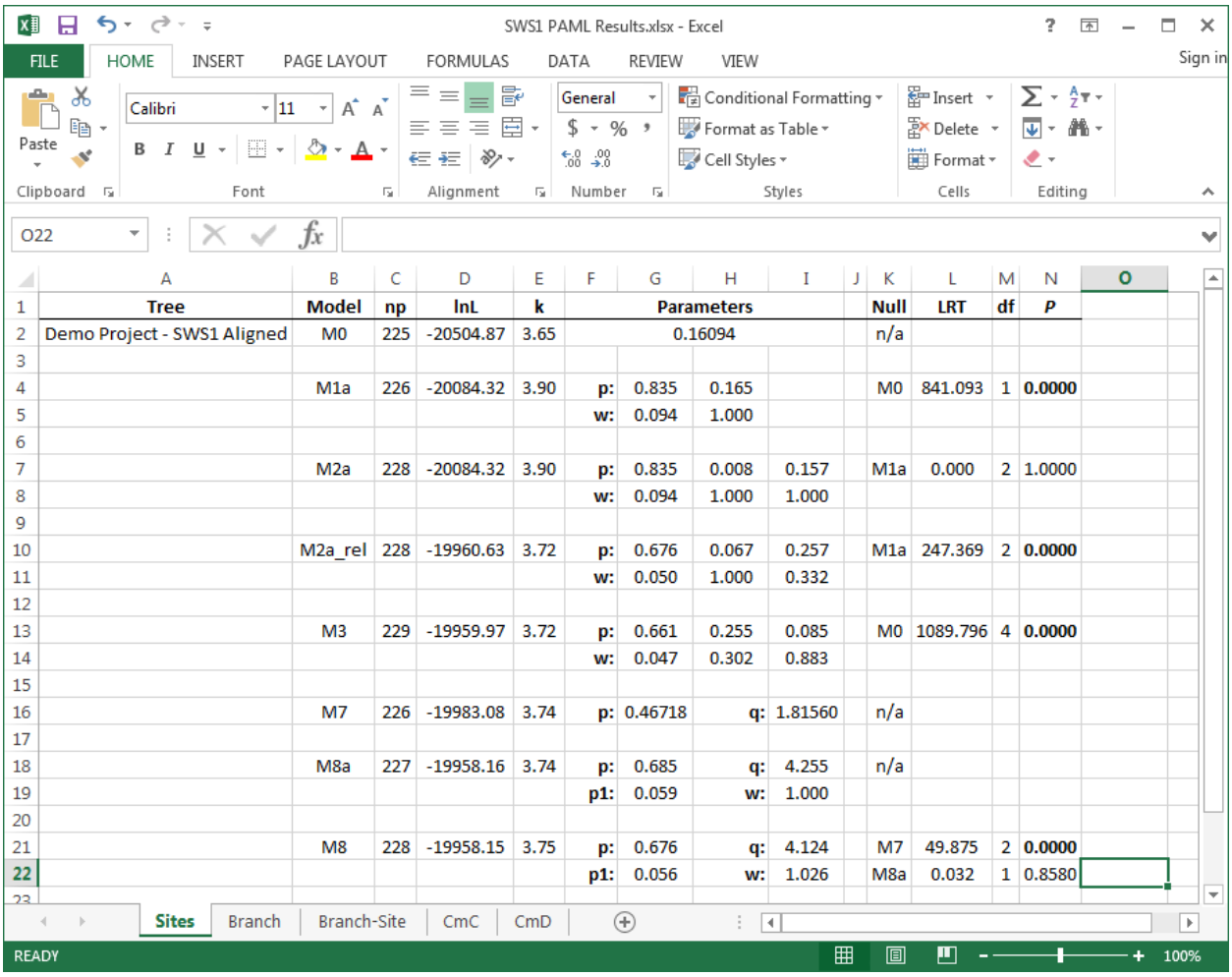


Figure 9. Interfaces used to set up and run PAML jobs.



	Tree	Model	np	lnL	k	Parameters				Null	LRT	df	P
2	Demo Project - SWS1 Aligned	M0	225	-20504.87	3.65			0.16094		n/a			
4		M1a	226	-20084.32	3.90	p:	0.835	0.165		M0	841.093	1	**0.0000**
5						w:	0.094	1.000					
7		M2a	228	-20084.32	3.90	p:	0.835	0.008	0.157	M1a	0.000	2	1.0000
8						w:	0.094	1.000	1.000				
10		M2a_rel	228	-19960.63	3.72	p:	0.676	0.067	0.257	M1a	247.369	2	**0.0000**
11						w:	0.050	1.000	0.332				
13		M3	229	-19959.97	3.72	p:	0.661	0.255	0.085	M0	1089.796	4	**0.0000**
14						w:	0.047	0.302	0.883				
16		M7	226	-19983.08	3.74	p:	0.46718	q:	1.81560	n/a			
18		M8a	227	-19958.16	3.74	p:	0.685	q:	4.255	n/a			
19						p1:	0.059	w:	1.000				
21		M8	228	-19958.15	3.75	p:	0.676	q:	4.124	M7	49.875	2	**0.0000**
22						p1:	0.056	w:	1.026	M8a	0.032	1	0.8580

Figure 10. Excel table exported from BlastPhyMe using the 'Export to' function. The table comes automatically organized and formatted with test statistics computed.

REFERENCES

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24:1586–1591.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum–likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307–321.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Research* 32:1792–1797.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. S*cience* 320:1632–1635.

Koepfli K-P, Paten B, O'Brien SJ, Genome KCS. 2015. The Genome 10K Project: A Way Forward. *Annual Review of Animal Biosciences* 3:57–111.