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

Ryan K Schott<sup>\*1,#</sup>, Daniel Gow<sup>\*2</sup>, Belinda SW Chang<sup>1,2</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, <sup>2</sup>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 publicationready 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: <u>https://github.com/ryankschott/BlastPhyMe</u>

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.

## **M**ETHODS

BlastPhyMe is a Microsoft Windows application coded in the C# programming language on the Microsoft .NET Framework platform, version 4.0 (<u>http://msdn.microsoft.com/en-us/library/zw4w595w(v=vs.100).aspx</u>). The database engine that BlastPhyMe interacts with for storing data is the Microsoft SQL Server 2014 Express LocalDB engine (<u>http://msdn.microsoft.com/en-ca/library/hh510202(v=sql.120).aspx</u>). For communicating with the NCBI BLASTN web service, BlastPhyMe uses the .NET Bio open source library (<u>https://github.com/dotnetbio/bio</u>). 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 (<u>http://www.ncbi.nlm.nih.gov/books/NBK25499/</u>). BlastPhyMe also communicates with NCBI's BLAST web service using the aforementioned .NET Bio libraries, which make use of the QBlast URL API (<u>http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html</u>).

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; <u>http://www.megasoftware.net/</u>) for visualization and editing. Multiple sequence alignments can be performed using PRANK (Löytynoja and Goldman 2008; <u>http://wasabiapp.org/software/prank/</u>) and MUSCLE (Edgar 2004; <u>http://www.drive5.com/muscle/</u>). Phylogenetic trees can be inferred using PhyML (Guidon et al. 2010; <u>http://www.atgc-montpellier.fr/phyml/binaries.php</u>). Resulting phylogenetic trees can be sent to TreeView (<u>https://code.google.com/archive/p/treeviewx/</u>) for visualization. Selection analyses can be performed using PAML (Yang 2007; <u>http://abacus.gene.ucl.ac.uk/software/paml.html</u>). Finally, sequences and PAML results tables can be exported to Microsoft Excel (<u>www.microsoftstore.com</u>).

## INSTALLATION

BlastPhyMe comes with a complete installation package available at:

<u>https://github.com/ryankschott/BlastPhyMe/releases</u>. 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.

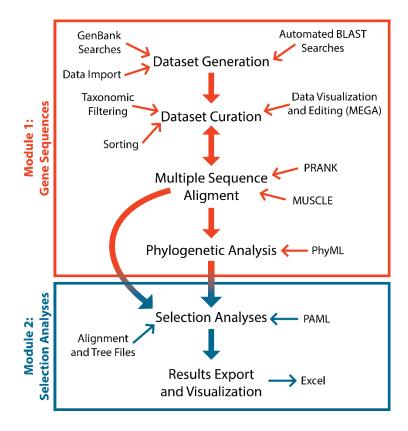
## FIGURES

## A. Module 1: Gene Sequences

	es 🤞 Selection A	nalyses (PAML)	earch GenBank					
Move to		Similar Sequences 鷸 Up Delete   📑 Export to 👻 S1 Aligned			from BLAST I	NCBI 🚰 Alig	n with 👻 📕 G	enerate Tree with PhyMI
Select all 0	of 0 records selected	ł					Showing 0 of 0 s	sequences Apply Filter
Gene	Organism	Definition	Length	GenBank	Source	Processed by BLASTN	Added At	Updated At

## 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.

🕝 BlastPhyMe - v1.5.0.5 Beta	- C:\BlastPhyMe Den	nonstration\BlastPhyMe Demonstration.mdf - De	mo Project				c	- O X
File Gene Sequences M	/lanage Help							
🖉 Gene Sequences 🛛 🛁 Sele	ection Analyses (PAML)							
🍸 🎦 New Dataset 📑 Open 🛛	Dataset 💥 Close Da	taset 🗧 Search GenBank						
🕴 🛋 Set Gene Name 🛭 😂 BLA	AST for Similar Sequer	nces 🍰 Update from GenBank 😪 Annotate fro	m BLAST N	CBI 🚰 Alig	n with 👻	Genera	ate Tree with Pl	nyML
🕴 🔂 Move to 🗎 Copy to	🗙 Delete 🛛 🗋 Ey	ort to + 🕞 Import from +						
All RH1 LWS SWS	1							
Select all 0 of 0 records	selected					Showing 0	of 0 sequences	Apply Filter
Gene Organism	n Definiti	on	Length	GenBank	Source	Process by BLASTI	Added At	Updated At
· · · · · · · · · · · · · · · · · · ·								
Search GenBank for Nucleotides								
	Nucleotide search:	LWS[Title]				^		
						-		
	Search history:	SWS1[Title] (1,173 results)						
	Sedicir filatory.	LWS[Title] (2,206 results)						
		RH1[Title] OR RH0[Title] (50,569 results)						
		Tilter by Taxonomy	<u>d</u>	Search	C <sup>4</sup> Cano			
								• .;;
L								

**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.

unic	2,206 nucleotide sequences.				
	Definition	Length	Accession	GenBank	In Proje
	Columba livia LWS opsin (lws) gene, complete cds	2,907	AH007800.2	http://www.ncbi	
	Epinephelus bruneus LWS mRNA for red opsin, complete cds	1,074		http://www.ncbi	
	Eubalaena glacialis long-wavelength sensitive cone opsin (LWS) pseudogene mRNA, partial sequen	627	KU363818.1	http://www.ncbi	
1	Thamnophis proximus long-wavelength-sensitive opsin (LWS) mRNA, complete cds	1,095	KU306727.1	http://www.ncbi	
	Gomphosus varius clone c11290_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete cds	1,394	KP881286.1	http://www.ncbi	
	Labroides dimidiatus clone c30919_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete c	1,718	KP881285.1	http://www.ncbi	
	Halichoeres chrysus clone c7660_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete cds	1,409	KP881284.1	http://www.ncbi	
	Cimhilabrus punctatus clone c33387_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete	1,343	KP881283.1	http://www.ncbi	
	Halichoeres omatissimus clone c6444_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complet	1,523	KP881282.1	http://www.ncbi	
	Thalassoma lunare clone c29104_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete cds	1,467	KP881280.1	http://www.ncbi	
	Coris gaimard clone c31276_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete cds	1,807	KP881278.1	http://www.ncbi	
	Halichoeres chloropterus clone c32345_g1_i1 long wavelength sensitive opsin (LWS) mRNA, compl	1,379	KP881275.1	http://www.ncbi	
	Choerodon fasciatus clone c18935_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete	1,480	KP881274.1	http://www.ncbi	
	Bodianus mesothorax clone c45355_g1_i1 long wavelength sensitive opsin (LWS) mRNA, complete	1,402	KP881273.1	http://www.ncbi	
	Pyropyga nigricans LW-sensitive opsin (LWS opsin) mRNA, partial cds	777	KR150964.1	http://www.ncbi	
	Pyractomena dispersa LW-sensitive opsin (LWS opsin) mRNA, complete cds	1,140	KR150963.1	http://www.ncbi	
	Photuris sp. 1 GJM-2015 LW-sensitive opsin (LWS opsin) mRNA, complete cds	1,143	KR150962.1	http://www.ncbi	
	Photuris sp. 2 GJM-2015 LW-sensitive opsin (LWS opsin) mRNA, complete cds	1,143	KR150961.1	http://www.ncbi	
	Photinus pyralis LW-sensitive opsin (LWS opsin) mRNA, complete cds	1,137	KR150960.1	http://www.ncbi	
	Photinus marginellus LW-sensitive opsin (LWS opsin) mRNA, complete cds	1,137	KR150959.1	http://www.ncbi	

**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.

Bos taurus opsi	n 1 (cone pigments), short-wave-se	ensitive (OPN1	SW), mRNA			ß
Details Aligner	d from Query Sequences					
Definition:	Bos taurus opsin 1 (cone pigments), s	hort-wave-sensi	tive (OPN1SW), mRNA			
Organism:	Bos taurus opsin 1 (cone pigments), short-wave-sensitive (OPN1SW), mRNA Bos taurus Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; I SWS1 Accession: NM_174567.1 Locus: NM_174567 27807026 http://www.ncbi.nlm.nih.gov/nucleotide/27807026 ArGAGGAAGAGTGTCACAGGAGGAGGTGTTCTTCGTGTCAAGAACATCTCCTTGGTGGGGCCGTGGGATGGAACCTCAGTACCACCTCGGCGCCT ACGAGTGTTATGGGGACGACGCATCTATCTTGGTCTATGGGTTGGGGGCCGTGGGATGGACCTCAGTACCACCTCGGCGCCT AccAGTGTTATGGGGACGACCACTATATTCTGGGTCAAGGAGCCTCCTGGGGGGCCGTGGGATGGACCTCAGTGACCACGTGGCCTCA AccAGTGTTTAGGGTACGTCGTCTGGCCGCCAGTGTGTGTGCGCGGACGTCCGGGGCGTGGAGGGCGGGGGCGGTGGACAGGCTGGGTGGCCACCTGGGCCGCCGACGTGGGCCGCAGGGCCGCAGGGGCCGCAGGGCGGGC	Taxonomy: Eukaryota		heria; Li		
Gene:	SWS1	Query Sequences         taurus opsin 1 (cone pigments), short-wave-sensitive (OPN1SW), mRNA         taurus       Taxonomy:       Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; L         S1       Accession:       NM_174567.1       Locus:       NM_174567         07026       http://www.ncbi.nlm.nih.gov/nucleotide/27807026       *         saacchachterCoAdsGedeGadeGadeTTCTTCTGTTCAAGAACATCTCCTTGETGEGGECCCTCAATGCCACCTCCACGCCACCTCGEGGEGCCACCTCCACTCCA				
GenBank ID:	27807026	http://www.	ncbi.nlm.nih.gov/nucleotide/27807	<u>7026</u>		
Sequence:	TCTGGGCCTTCCACCTCCAGGCAGT CAGAAAGTTGCGGCAGCCACTCAAC ACCAGTTGTTATGGGTACTTCGTCT TGGCCTTCTTGGCCTTCGAGCGCTA CACCTGGACCATCGGTATTGGTGT TGGTACACCGTGGGGCACCAAGTATT CCTACTCTCACCTGCTGGGGGCCCTT GGTGGTGGTCATGGTGGGATCCTTT TTACGGCTTGTCACCATTCCTGCCT	CTTCATGGGTT TATATTCTGGT TCGGCCGCCAT CATCATCATCAT TCCATCCCACC ACAGCGAGTAC CAGAGCTGTTC TGTCTCTGTTA TCTTCTCCAAC	TTGTCTTCTTTGTAGGGACGCCA CAACGTGTCTCTGGGGGGGCTTCA GTCTGTGCCCTGGAGGCCTTCCT GTAAACCCTTCGGCAGCCGATCCG CTTCTTTGGCTGGCTGGAGCGATCG TATACCTGGTTCCTCTTCATCTT CGGCTCAACAGCAGGAGTCAGCT ACACACCCTACGCTGCCCTGGCCA GAGTGCTTGTGTCTACAATCCCAT	CTCAATGCCACG TCTACTGCATCT GGGCTGTACAGC TTCGACCCCAAGC TCCCTGAGGGCC CTGCTACATTGT TCGACCCAGAAG TGTATATAGTCA CATCTACTGCTT	GTGCTGGTGGCCACACTGCGC TCTCTGTCTTCATCGTCTTCA CAGGTCTGGTGACAGGCTGGTC CCACGCCTGATGGTGGTCCCG TGCAGTGTCCTGTGGTCCCG GCCTCTCCCTCATCTGGTCT GCCTGACCGGAGGTGACCACC LACAACCGTAACCATGGGGTGG TCATGAATAAGCAGTTCCGAGC	TA TC AC GC AC CT AT AC TT
Length:	1,050 (CDS: 1,050)	]			<u>View Source Se</u>	quence
Notes:						*
Updated At:	2016-06-13 03:47 by User			Source:		se

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

anine and animes	Manage <u>H</u> elp										
ne Sequences 🦂	Selection Analyses (PAML)										
lew Dataset 🛐 O	pen Dataset 💥 Close Dataset 🛛 🗧	Search GenBank									
et Gene Name	BLAST for Similar Sequences	Update from GenBank	S Annotate from	BLAST NCBI 🚰 Align with 📕 Generate Tree w	ith PhyML	Aove to	Copy to	× Dele	te 🕞 Expo	ort to +	Import from
RH1 LWS			<b>S</b>								
	ecords selected							Chausing 0	-( 172		nge Filter Cle
select all got 3 h	ecorus selecteu							Showing 5	or 172 sequer	ices <u>unar</u>	
Gene	Organism	Definition	Alignments	100	p lastes	100		12	10000		
SWS1			Results Query Se	quences Request History Progress Messages							
SWS	Antilocapra americana Cavia porcellus	Intilocapra americar Ovia porcellus viole	Select alignme	nts to add to your dataset. To view additional d	etails for an align	ment, doub	le-click o	n a row or	right-click a	and select	"Details"
SW51	The second first theory		Select all	68 of 148 records selected			Show	ving 148 of	148 alignments	s Exclude	d by GenBank I
SW51	Lama glama Moschus moschifenus	Lana glama short w Moscaus moschifen	Definition			Sequence	Max	Query	Aligned	Identities	GenBank
WS1	Okapia johnstoni	Okapia ohnstoni she				Length	Score	Cover	Length	Match	
SWS1	and the second sec	Pecari tancu short v		talapoin blue cone pigment gene		977	1224	100 %	1,007	89 %	http://www
SWS1 SWS1	Pecari tajacu Rangifer tarandus tarandus	Rangfer talendus ta		: Chlorocebus sabaeus opsin 1 (cone pigments), short-wave		979	1224	100 %	1,007	89 %	http://www
SWS1 SWS1	Thylamys elegans	Thylamys elevans sh		es opsin 1 (cone pigments), short-wave-sensitive (OPN1SW		1,009	1222	100 %	1,043	88 %	http://www
SWS1 SWS1	Tragulus napu	Tragulus napu hort		Rousettus aegyptiacus opsin 1 (cone pigments), short-way		1,007	1220	100 %	1,008	89 %	http://www
3W31	Tragulus napu	Tragulus napu non		: Callithrix jacchus opsin 1 (cone pigments), short-wave-sens		981	1216	101 %	1.047	88 %	http://www
ASTN Options		8		Pan paniscus opsin 1 (cone pigments), short-wave-sensitiv		1.009	1216	100 %	1.043	88 %	http://www
earch Program				s opsin 1 short-wave-sensitive (OPN1SW) mRNA, complete		1,009	1211	100 %	1,043	88 %	http://www
atabase name: n	f			ricana opsin 1 (cone pigments), short-wave-sensitive (OPN1		1,006	1207	99 %	1,029	88 %	http://www
Optimize for:	Highly similar sequences (megablast)			s opsin 1 (cone pigments), short-wave-sensitive (OPN1SW)		1,009	1205	100 %	1.043	88 %	http://www
C	) More dissimilar sequences (discontin	uous megablast)		: Propithecus coquereli opsin 1 (cone pigments), short-wave		1,004	1205	100 %	1,041	88 %	http://www
C	) Somewhat similar sequences (blastn)		and the second second second	: Gorilla gorilla gorilla opsin 1 (cone pigments), short-wave-se		1,009	1205	100 %	1.043	88 %	http://www
loorithm Parameters				: Nomascus leucogenys opsin 1 (cone pigments), short-wav		1,009	1205	100 %	1,043	88 %	http://www
Max target sequence	s: 100 🚔			nstruct Homo sapiens clone IMAGE:100062210, MGC:1902		1,009	1203	100 %	1,036	88 %	http://www
Expect threshold:	10.00			: Marmota marmota marmota opsin 1 (cone pigments), short-		1,007	1201	100 %	1,036	88 % 88 %	http://www
Nord size:				: Ictidomys tridecemineatus opsin 1 (cone pigments), short-		1,007	1196	100 %	1,036	88 %	http://www
			the second se	: Microcebus murinus opsin 1 (cone pigments), short-wave-	ensitive (OPNTS	981	1196	98 %	1,035	88 %	http://www
Aatch/Mismatch sco	res: 12 -		Committee Constantion of Charles	hus blue cone pigment gene :: Equus asinus opsin 1 (cone pigments), short-wave-sensitiv	ODNICING D	981	1192	100 %	1,022	88 %	http://www.
iap costs:	Linear 👻			: Equus asinus opsin 1 (cone pigments), snort-wave-sensitiv : Mandrillus leucophaeus opsin 1 (cone pigments), short-war			1187	100 %	1.043	87 %	http://www
	Submit	Cancel		: Mandnilus leucophaeus opsin 1 (cone pigments), short-wa concinnus short-wave sensitive type 1 opsin mRNA, partial c		9/9 873	1187	84 %	1,043	8/ % 91 %	http://www
		Cancer		: Pongo abelii opsin 1 (cone pigments), short-wave-sensitive			1177	100 %	1.043	87%	http://www
				: rongo abelii opsin 1 (cone pigments), short-wave-sensitive : Cercocebus atys opsin 1 (cone pigments), short-wave-sensitive		977	1170	100 %	1.043	87 %	http://www
				: Macaca nemestrina opsin 1 (cone pigments), short-wave-seri		977	1170	100 %	1.043	87 %	http://www
			and the second se	icularis opsin 1 (cone pigments), short-wave-sensitive (OPN)		979	1170	100 %	1.043	87 %	http://www
				Papio anubis opsin 1 (cone pigments), short-wave-sensitive (of the		986	1165	100 %	1.043	87 %	http://www
				: Macaca mulatta opsin 1 (corre pigmenta), short-wave-sensativ	- (	977	1165	100 %	1.043	87 %	http://www
									.,	21.14	
			Add to:	SWS1 • Update from	m GenBank						C Close

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

e Gene Sequence	s Manage Help										
Gene Sequences 🖂	Selection Analyses (PAML)										
New Dataset 📋 🤇	Open Dataset 💥 Close Dataset 🛛 😪	Search GenBank									
Set Gene Name	BLAST for Similar Sequences and U	ndate from GenBank 😂 Annotate fro	m BLAST NO	BI 🚰 Align with 🔹 📕 Generate Tree wi	ith PhyMI	×	Delete	Export to	· Import from ·		
RH1 LWS		9									
	records selected	Configure PRANK				8			Showing 279 of 27	9 sequences Apply F	-
		Alignment Options						Processed	0.0000 2.000 2.0		
Gene	Organism	Trust insertions (-F)		Codon alignment (-codon)	Output guide tree (-showtree)		e	by BLASTN	Added At	Updated At	
SWS1	Abromis chloronotus	Keep gap characters (-ke	ep)	Standard translation (translate)	Output ancestral sequences (-show	inc)	nk Search		2016-06-13 03:10 PM	2014-12-24 12:00 AM	4
SWS1	Abromis humei			Mitochondrial translation (-mttranslate)	Output events per branch (showever	ents)	nk Search		2016-06-13 03:10 PM	2014-12-24 12:00 AM	A
SWS1	Abromis maculipennis	Iterations:	5 🌲				nk Search		2016-06-13 03:10 PM	2014-12-24 12:00 AM	A.
	Abudefduf sexfasciatus						nk Search		2016-06-13 03:10 PM	2012-10-23 12:00 AM	A,
SWS1	Acinonyx jubatus	Guide tree:					N (NCBI)		2016-06-13 03:38 PM	2016-06-13 03:47 PM	4
SWS1	Aluroedus buccoides	Job Options					nk Search		2016-06-13 03:10 PM	2012-07-31 12:00 AM	A.
SWS1	Aluroedus crassirostris	PRANK.exe	C:\prank\t	in \prank.exe			nk Search		2016-06-13 03:10 PM	2012-07-31 12:00 AM	A
SWS1	Aluropoda melanoleuca	Working Directory;	C:\BlastPh	vMe Demonstration			N (NCBI)		2016-06-13 03:38 PM	2016-06-13 03:47 PM	4
SWS1	Aluropoda melanoleuca	troning choosing:		,			N (NCBI)		2016-06-13 03:38 PM	2016-06-13 03:47 PM	A
SWS1	Amblycmis inormatus				Preserve Output	les	nk Search		2016-06-13 03:10 PM	2012-07-31 12:00 AM	A
SWS1	Amblycmis inormatus				🚳 Run 🤇 🥙 Cano		nk Search		2016-06-13 03:10 PM	2012-07-31 12:00 AM	ē.

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

e	Gene Sequences	Manage Help						
ier	ne Sequences 🛁	Selection Analyses (PANL)						
N	ew Dataset 👔 Op	en Dataset 💥 Close Dataset   😂 Search G	enBank					
	_		om GenBank 🗧 Annotate from BLAST NCBI 🚰 Align with 🕶 🔲 Gen	erate Tree with PhyML	ove to 📑 Copy to	X Delete 🗔 Expo	rt to + 👍 Import from +	
	-	WS1 SWS1 Aligned						
201		cords selected					She	wing 113 of 113 sequences Apply
	ou un our note	Sous selected						might sort is sequences Augur
	Gene	Organism	Definition	Leng	gth GenBank	Source	Processed by Added At BLASTN Added At	Updated At
	SWS1	Acinonyx jubatus	PREDICTED: Acironyx jubatus opsin 1 (cone pigments), short-wave-sensitive	Configure PhyML	-			
7	SWS1	Aluropoda melanoleuca	PREDICTED: Alumpoda melanoleuca opsin 1 (cone pigments), short-wave-se	Substitution Model Options				
	SWS1	Aluropoda melanoleuca	PREDICTED: Aluropoda melanoleuca opsin 1 (cone pigments), short-wave-se		GTR			
	SWS1	Antiocapra americana	Antilocapra americana short wavelength sensitive opsin 1 (SWS1) gene, partia	Substitution model:	UIK	•		
		Aotus nancymaae	PREDICTED: Actus nancymaae opsin 1 (cone pigments), short-wave-sensitive	Equilibrium frequencies:	Optimized	Empirical 🔘 Fixed		
		Balaenoptera acutorostrata scarrmoni	PREDICTED: Balaenoptera acutorostrata scammoni opsin 1 (cone pigments),					
		Bison bison	PREDICTED: Bison bison opsin 1 (cone pigments), short-wave-sensitive	Transition/Transversion ratio:	Estimated	hxed		
	SW51	Bos mutus	PREDICTED: Bos mutus opsin 1 (cone pigments), short-wave-sensitive IOPN	Proportion of invariable stes:	Estimated	Fixed 0.00000 -		
	SWS1	Bostaurus	Bostaurus opsin 1 (cone picments), short-wave-senstive (OPNISW), mRNA			L		
	SWS1	Bubalus bubalis	PREDICTED: Bubalus bubalis opsin 1 (cone pigments), short-wave-sensitive	# of substitution rate categories:	4			
	SWS1	Calithrix jacchus	PREDICTED. Callithrix jacchus opsin 1 (cone pigments), short-wave-sensitive	Gamma shape parameter:	Estimated	Fixed		
	SWS1	Camelus bactrianus	PREDICTED: Camelus bactrianus opsin 1 (cone pigments), short-wave-sensiti					
	SWS1	Camelus dromedarius	PREDICTED: Camelus dromedarius opsin 1 (cone pigments), short-wave-sens	Tree Searching Options				
	SWS1	Camelus ferus	PREDICTED: Camelua forua opain 1 (conc pigmenta), short wave aenaitive (0	Starting tree:	🖲 BIONJ 🔘 Fie	c		
	SWS1	Carie lupus familiarie	PREDICTED: Canie lupue familiarie opein 1 (cone pigmente), short-wave-senet	Type of tree improvement:	SPR & NN	-		
	SWS1	Capra hiroue	PREDICTED: Capra hircus opein 1 (cone pigments), short-wave-sensitive (OP					
	SWS1	Cavia porcellus	Cavia porcellus opsin 1 (cone pigments), short-wave-sensitive (color blindness	Number of random starting trees:	5 🚖			
	SWS1	Cavia porcellus	Cavia porcellus videt-sensitive visual pigment (SWS1) mRNA, complete cds	Optimize:	🗸 Topology 📝	Franch lengths 👿 Subst	turion rate parameters	
	SWS1	Ceratotherium simum simum	PREDICTED: Ceratotherium simum simum opsin 1 (cone pigments), short-wave					
m	SWS1	Cercocebus atys	PREDICTED: Cercocebus atys opsin 1 (cone pigments), short-wave-sensitive	Branch Support Options				
	SWS1	Chinchilla lanigera	PREDICTED: Chinchilla lanigera opsin 1 (cone pigments), short-wave-sensitive	Fast likelihood based method:	SH like branch	a aupporta ak 📼		
	SWS1	Chlorocebus sabaeus	PREDICTED: Chlorocebus sabaeus opsin 1 (cone pigments), short-wave-sens	Perform bootstrap:	1			
	SW51	Chrysochloris asiatica	PREDICTED: Chrysochloris asiatica opsin 1 (cone pigments), short-wave-sens					
	SWS1	Colobus angelensis paliatus	PREDICTED: Colcbus angolensis pallatus opsin 1 (cone pigments), short-way	Job Options				
	SWS1	Condylura cristata	PREDICTED: Condylura cristata opsin 1 (cone pigments), short-wave-sensitive	Sequence label format:	{Definition}			
	SWS1	Cricetulus griseus	PREDICTED: Cricetulus griseus opsin 1 (cone pigments), short-wave-sensitive	Create a copy of the tree file without support values:	V			
	SWS1	Cricetulus griseus	PREDICTED: Cricetulus griseus opsin 1 (cone pigments), short-wave-sensitive	PhyML exe	C:\PhyML-3.1\PhyM	JI -3.1 win 32 eve		
	SWS1	Didelphis aurta	Didelphis aurita short-wave sensitive type 1 opsin (SWS1) mRNA, complete co					
	SWS1	Dipodomys ordii	PREDICTED: Dpodomys ordii opsin 1 (cone pigments), short-wave-sensitive (	Working Directory:	C:\BlastPhyMe Den	nonstration		
	SWS1	Echinops telfairi	PREDICTED: Echnops telfairi opsin 1 (cone pigments), short-wave-sensitive (					Preserve Output Files
	SWS1	Bephantulus edwardii	PREDICTED: Elephantulus edwardii opsin 1 (cone pigments), short-wave-sen	Reset to Default Options	1 C			🔅 Run 🤇 Cancel
	SWS1	Eptesicus fuscus	PREDICTED: Eptesicus fuscus opsin 1 (cone pigments), short-wave-sensitive	Preset to Default Options	J			Cancel

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

hyMe - v1.5.0.5 Beta - C:\BlastPhyMe Demonstration\BlastPhyMe Demonstration.mdf - Demo Project						
Selection Analyses (PAML) Manage Help						
Sequences 🥰 Selection Analyses (PAML)						
w PAML Job 📄 PAML Job History 🕴 🛱 Move to 🔛 Copy to 🗙 Delete   📮 Export to • 🕞 In	mport from 👻	🎦 New Dataset 📋 Open Data	aset 💥 Close [	Dataset		
ct all 0 of 0 records selected					Showing 0	of 0 sequences Apply Filt
Job Title Tree Model	_	k/w Start InL k	Run At	_	Se	quences File
igure PAML 🛛	Edit Tree Configu	uration				×
figure models to run for each tree and sequence file:	Tree File:					
ee File Sequences File # Seq. Length Title	Sequences File:					
BastPhyMe Demonstration∖ C:∖BlastPhyMe Demonstrati 113 1908 Demo Project - SWS1	Title:					
	Show Additional (	Options				Extract from Control File
	Analyses:	Model	Site Models	Categories	Карра	Omega
		Random Sites (M0, M1a, M2a, M	0, 1, 2, 3, 7, 8	10	2.00 - 3.00 @ 1.00	0.00 - 2.00 @ 1.00
		Random Sites (M2a_rel)	22	10	2.00 - 3.00 @ 1.00	0.00 - 2.00 @ 1.00
		Random Sites (M8a)	8	10	2.00 - 3.00 @ 1.00	1.00 (fixed)
		Branch Model (Alt)	0	3	2.00 - 3.00 @ 1.00	1.00
		Branch Model (Null)	0	3	2.00 · 3.00 @ 1.00	1.00 (fixed)
Add 📑 Edit 🗈 Copy XRemove		Branch-Site Model (Alt)	2	3	2.00 - 3.00 @ 1.00	1.00
		Branch-Site Model (Null)	2	3	2.00 - 3.00 @ 1.00	1.00 (fixed)
figure job options:		Clade Model C (Alt) Clade Model D (Alt)	2	3	2.00 - 3.00 @ 1.00	1.00
leML.exe: C:\Users\Ryan\Sync\paml4.8\bin\codeml.exe		Clade Model D (Alt)			2.00 - 3.00 @ 1.00	1.00
rking Directory: C:\BlastPhyMe Demonstration		🔒 Add 📑 Edit	Copy	XRemove	1	
Processes: 6 - CPU Priority: Below Normal   Processes: Verserve Working Folders					, 🛃 Si	ave Cancel
Title: Demo Project - SWS1 Aligned						_
📖 Run 🤇 Cancel		Add Analysis Configu	ration		[	8
		Model: Rand	om Sites (MO, M1	a, M2a, M3, M7,	. M8) 🔻	
			<b>▼</b> 1 <b>▼</b> 2			
		Categories: 10	A.			
				0 🖨 Interva	: 1.00 🚔 🗔 Fixed	
					: 1.00 🚔 🔲 Fixed	
		anoga otak.	<u> </u>			
				🔒 Ad	dd Cancel	

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

XII FILE	HOME INSERT P.	AGE LAYOL	л	FORMULAS		ML Res	ults.xlsx - E REVIEW	Excel VIEW					? 🗹	] —	□ × Sign
Paste	X         Calibri • 11           □□ •         B         I         U         •         □	• A	A I			General \$ - 9 -00 -3.0	- F	Condition Format a Cell Style	s Table ▼ s ▼		E Insert Delete Format	-	₩ - #1 		
Clipb	oard 🕞 Font		E.	Alignment	Fa	Numbe	r Gi	5	ityles		Cells		Editing		
022		<i>fx</i>													
	А	В	С	D	E	F	G	Н	Ι	JK	L	м	N	0	
1	Tree	Model	np	InL	k			meters		Nu		df	Р		
	emo Project - SWS1 Aligned	M0	225	-20504.87	3.65		0.1	16094		n/a	1				
3 4		M1a	226	-20084.32	3.90		0.835	0.165		M	841.093	1	0.0000		
4 5		IVITA	220	-20084.32	3.90	p: w:	0.835	1.000		IVIC	841.093	1	0.0000		
5							0.054	1.000							
7		M2a	228	-20084.32	3.90	p:	0.835	0.008	0.157	M1	a 0.000	2	1.0000		
8						w:	0.094	1.000	1.000						
9															
0		M2a_rel	228	-19960.63	3.72	p:	0.676	0.067	0.257	M1	a 247.369	2	0.0000		
1						w:	0.050	1.000	0.332						
.2 .3		M3	220	-19959.97	3.72		0.661	0.255	0.085		1089.796		0.0000		
.5		IVID	229	-15555.57	5.72	p: w:	0.001	0.255	0.883	IVIC	1089.790	4	0.0000		
5							0.017	0.002	0.000						
.6		M7	226	-19983.08	3.74	p:	0.46718	q:	1.81560	n/a	1				
.7															
8		M8a	227	-19958.16	3.74	p:	0.685	q:	4.255	n/a	1				
.9						p1:	0.059	w:	1.000						[
20				40050.45			0.676				40.075	_	0.0000		
21 22		M8	228	-19958.15	3.75	p:	0.676	q: w:	4.124	M7 M8		-	0.0000		4-1
22						pr:	0.050	w:	1.020	IV/8	0.052	1	0.0000		-
	Sites Branch	Branch	Sito	CmC	CmD		+)	: [	4						Þ

**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. *Science* 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.