- 1 Using the Free Program MEGA to Build Phylogenetic Trees from Molecular Data
- 2 (ABT-2015-0089) in the September issue of ABT (78,7), 2016

3 Lucas Newman

- 4 Amanda L. J. Duffus
- 5 Cathy Lee\*
- 6 Department of Biology, Gordon State College, Barnesville, GA
- 7 \* Corresponding Author: clee@gordonstate.edu
- 8 Abstract:
- 9 Building evolutionary trees can be an excellent way for students to see how different gene
- 10 sequences or organisms are related to one another. Molecular Evolutionary Genetics Analysis
- 11 (MEGA) software is a free package that lets anyone build evolutionary trees in a user-friendly
- 12 setup. There are several different options to choose from when building trees from molecular
- 13 data in MEGA, but the most commonly used are Neighbor-Joining and Maximum Likelihood,
- 14 both of which give good estimates on the relationship between different molecular sequences.
- 15 In this article, we describe how to collect data from GenBank, insert it into a text editor, import
- 16 that data into MEGA, and create phylogenetic trees from the collected data.
- 17
- 18 Key Words: MEGA, evolutionary trees, molecular data, Neighbor-Joining, Maximum-likelihood
   19
- 20 Phylogenetics is the study of the evolutionary relatedness between different groups of
- 21 organisms (Nei and Kumar 2000). These groups can be on small scales (e.g., mammals) or large
  - 1

scales (e.g., different domains of life). The results of phylogenetic analyses are usually
 presented in the form of evolutionary trees, where different branches represent different gene
 sequences or species used to build them. The branching pattern of the tree illustrates how the
 sequences or species are related.

26 Here we present how to build evolutionary trees using the MEGA software package (newest 27 version 7; www.megasoftware.net). It contains two of the most commonly used methods to 28 infer evolutionary relationships among species by using their gene sequences. In this paper, we 29 will introduce students to these two methods: Neighbor-Joining (NJ) and Maximum-Likelihood 30 (ML) method. According to MEGA authors, it is frequently used in educational setting in 31 advanced classes (pers. comm. Sudhir Kumar; Ryan et al. 2013). However, many K-12 32 instructors are not familiar with it, and its potential to introduce the concepts of evolutionary 33 biology to students in a hands-on, discovery-based pedagogy using the gene sequences. There 34 are multiple online resources that provide such gene sequences for a multitude of species (e.g., 35 GenBank), which is available from National Center for Biotechnology Information (NCBI) (Hall 36 2013). Both DNA and protein sequences are available, and there are several informative 37 tutorials provided on the NCBI website on how to use these. There is literally unlimited 38 sequence data from thousands of genes from animals, plants, protists, bacteria, and viruses 39 available through GenBank.

40 **Project Goal**:

To build an evolutionary tree using the rcbL gene sequence, which is commonly used to study
the evolutionary relationships between plants (see Newmaster et al. 2006). Sequence data
pertaining to the rcbL genes from many plant species are available through GenBank and we

- 44 will gather our data set from this resource. The rcbL gene sequence data will then be imported
- 45 into MEGA, aligned, and then used to build a phylogenetic tree.

#### 46 Helpful Prior Knowledge and Potential Context of this Exercise:

- 47 Students should have some introductory level knowledge of the purpose of evolutionary trees
- 48 and have some experience interpreting simple phylogenetic trees. This exercise would be ideal
- 49 as a final project for evolution units at varying levels, including AP Biology.

#### 50 Learning Objectives:

- 51 By the end of this project, students will:
- 52 1. Learn how to obtain molecular data from GenBank
- 53 (<u>http://www.ncbi.nlm.nih.gov/genbank/</u>).
- 54 2. Learn to build evolutionary trees using freely available software MEGA.
- 55 3. Understand the meaning of ancestral vs. recent species, clades, and be able to
- 56 interpret evolutionary relationship among species.
- 57 As students perform the exercise, they should consider the following questions:
- 58 1. Why was the rcbL gene used?
- 59 2. Which organelle does the rbcL gene originate from?
- 60 3. What function does the protein product of the rcbL gene have in the plant?

#### 61 System Requirements:

- 62 1. Internet access to use the GenBank data base and to download MEGA.
- 63 2. A text editor program, Notepad (Windows PC) or Texteditor (Linux/Mac).
- 64 (http://www.megasoftware.net/mega.php)

### 65 *Getting Started*:



67 Figure 1. Google search for <u>GenBank</u> and click on the result for <u>GenBank Home</u>

68 (http://www.ncbi.nlm.nih.gov/genbank/).

69

66

<ul> <li>⇐ → C</li> <li>➡ www.ncbi.nlm.nih.go</li> <li>➢ NCBI Resources </li> <li>➡ How To </li> <li>➡ GenBank</li> <li>➡ GenBank</li> <li>➡ Submit</li> <li>➡ GenBank</li> <li>➡ Submit</li> <li>➡ Mucleotide</li> <li>All Databas</li> <li>➡ All Databas</li> <li>➡ Submit</li> <li>➡ All Databas</li> <li>➡ Soloriget</li> <li>➡ Books</li> <li>➡ ClinVar</li> <li>➡ ClinVar</li> <li>➡ Cline</li> <li>➡ Conserved</li> <li></li></ul>	e HTGs V EST/GSS V Metag	Search			
← → C	e HTGs V EST/GSS V Metag	Sign in to NCBI			
<ul> <li>➢ NCBI Resources ♥ How To ♥</li> <li>GenBank ♥ Submit</li> <li>GenBank ♥ Submit</li> <li>INSDC ♥</li> <li>GenBank Overviet</li> <li>What is GenBank?</li> <li>GenBank ® is the NIH gen DNA sequences (<i>Nucleic /</i> International Nucleotide Se</li> <li>Cline Conserved three organizations exchan</li> <li>The complete release note three organizations exchan</li> </ul>	• HTGs • EST/GSS • Metag	Sign in to NCBI			
GenBank ▼ Submit INSDC ▼ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■	e HTGs V EST/GSS V Metag	Search			
GenBank v Submt INSDC v CenBank Overviev What is GenBank? GenBank <sup>®</sup> is the NIH gen DNA sequences ( <i>Nucleic A</i> Japan (DDBJ), the Europet three organizations exchan The complete <u>release note</u> Describer of the second	e HTGs ▼ EST/GSS ▼ Metag				
GenBank Overvieu What is GenBank? GenBank <sup>®</sup> is the NIH gen DNA sequences ( <u>Mucleic 4</u> ClinVar International Nucleotide Se Japan (DBJ), the Europe Conserved three organizations exchan The complete <u>release note</u> DNA sequences in and a		enomes 🔻 TPA 💌 TSA 💌			
GenBank <sup>®</sup> is the NIH gen DNA sequences ( <u>Mucleic 4</u> ClinVar Japan (DB-J), the Europe Conserved three organizations exchan The complete <u>release note</u> Epigenomit	ases t e	GenBank Resources			
three organizations exchan dbGaP The complete <u>release note</u> dbVar Epigenomic	ns notated collection of all publicly available <u>11D36-42</u> ). GenBank is part of the 1, which comprises the DNA DataBank of v (EMBL), and GenBank at NCBI. These	Submission Tools Search GenBank Under Cooperate			
divisions and the WGS div	Bank are available on the NCBI ftp site. A statistics for both the traditional GenBank lease.	<u>opuale Genbalik Records</u>			
An <u>annotated sample Gent</u> the features of the GenBar GEO Datas GEO Profile	es cerevisiae gene demonstrates many of asets				
Access to GenBank There are several ways to search and retrieve data from GenBank.					
<ul> <li>Search GenBank for sequence iden divided into three divisions: <u>CoreNu</u> Sequence Tags), and <u>dbGSS</u> (Gence.</li> <li>Search and align GenBank sequenc Alignment Search Tool), BLAST sei independently; see <u>BLAST info</u> for n</li> </ul>	entifiers and annotations with Entrez Nucleotide, which is				

- 71 Figure 2. Use the dropdown next to the word <u>GenBank</u> to change from default
- 72 <u>Nucleotide</u> and select <u>Gene</u>.

S GenBank Home	×				
	:bi.nlm.nih.gov/genbank/	<b>ک</b> =			
S NCBI Resources 🕑	How To 🕑	Sign in to NCBI			
GenBank	Gene   RuBisCO large subunit	Search			
GenBank ▼ Submit	▼ Genomes ▼ WGS ▼ HTGs ▼ EST/GSS ▼	Metagenomes ▼ TPA ▼ TSA ¥			
GenBank Overvie	w	GenBank Resources			
What is GenBank?		GenBank Home			
ConReals ® is the NUL and	ation and the second and an an advected and the stick of all multiply and	Submission Types			
DNA sequences (Nucleic	Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the	Submission Tools			
International Nucleotide S	equence Database Collaboration , which comprises the DNA DataB	ank of Search GenBank			
Japan (DDBJ), the Europe three organizations excha	an Molecular Biology Laboratory (EMBL), and GenBank at NCBI. T	hese Update GenBank Records			
The complete <u>release notes</u> for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank growth <u>statistics</u> for both the traditional GenBank divisions and the WGS division are available from each release.					
An <u>annotated sample Ger</u> the features of the GenBa	<u>Bank record</u> for a Saccharomyces cerevisiae gene demonstrates n nk flat file format.	nany of			
Access to GenBank					
There are several ways to search and retrieve data from GenBank.					
<ul> <li>Search GenBank fr divided into three d Sequence Tags), a</li> <li>Search and align G Alignment Search independently; see</li> </ul>	r sequence identifiers and annotations with Entrez Nucleotide, whit ivisions: <u>CoreNucleotide</u> (the main collection), <u>dbEST</u> (Expressed ad <u>dbCSS</u> (Genome Survey Sequences). enBank sequences to a query sequence using <u>BLAST</u> (Basic Loca fool). BLAST searches CoreNucleotide, dbEST, and dbCSS <u>BLAST info</u> for more information about the numerous BLAST data!	ch is Il bases.			

74 Figure 3. Type **RuBisCO large subunit** in the entry box to the right of dropdown and click

75 <u>Search</u>.

RußisCO large subunit - G X							
← → C 🗋 www.ncbi.nlm.nih.gov/gene/?term=RuBisCO+large+subunit							
운 NCBI Resources ⓒ How To ⓒ Sign in to NCBI							
Gene	Gene	RuBisCO large     Create alert A	ge subunit Advanced		Search Help		
Gene sources Genomic Mitochondria Organelles Plasmids Plastids	Tabular → 20 per Search result Items: 1 to 20 c See also 40 c	page - Sort by Relev S of 1204 << Fir iscontinued or replace	vance <del>v</del> st < Prev Page 1 id items.	Send to: ↓ of 61 Next > Last >>	Hide sidebar >> Filters: <u>Manage Filters</u> Results by taxon  ♠ Top Organisms [ <u>Iree</u> ] Phoenix dactylifera (15)		
Categories Alternatively spliced	Name/Gene ID	Description	Location	Aliases	Elaeis guineensis (15) Zea mays (13)		
Annotated genes Protein-coding Pseudogene	D: 2717040	RuBisCO large subunit [Chlamydomonas reinhardtii]	NC_005353.1 (122490123917, complement)	ChreCp049	Brassica napus (13) Oryza sativa (12) All other taxa (1136) More		
content RefSeq	D: 84 12	RuBisCO large subunit [Zea mays]	NC_001666.2 (5687458304)	ZemaCp032	Find related data		
Status clear ✓ Current Chromoso Iocations	D: 844754	RuBisCO large subunit [ <i>Arabidopsis</i> <i>thaliana</i> (thale cress)]	NC_000932.1 (5495856397)	ArthCp030	Database: Select • Find items		
more <u>Clear all</u> <u>Show additional</u>	D: 3131463	RuBisCO large subunit [Oryza sativa Japonica Group]	NC_001320.1 (5409555528)	OrsajCp033	Search details (RuBisCO[All Fields] AND large[All Fields] AND		
<u>tilters</u>	<u>rbcL</u>	RuBisCO large	NC_008155.1	OrsaiCp23	subunit[All Fields]) AND		

77 Figure 4. Select the link: **rbcL** for RuBisCO large subunit for a given species, here we will

### use Zea mays.

😪 rbcL RuBisCO large s	ibuni ×							
- → C' 🗋 www	v.ncbi	.nlm.nih.gov/ge	ne/84521	2				53
S NCBI Resources	H	ow To 🖂					<u>Sign in t</u>	to NCBI
Gene	G	Sene 🔹					Searc	h
			Advanced					Help
Full Report <del>-</del>					Se	nd to: 👻	Hide side	ebar >>
rbcL RuBisCO	large	e subunit [ Ze	ea mays	1			Table of contents Summary	
Gene ID: 845212, upda	ted on	30-Sep-2014					Genomic context	
Summary						2	Genomic regions, transcript product	ts, and
Gene symt	ol rhe	-1					Bibliography	
Gene descripti	on Ru	BisCO large subur	nit				Pathways from BioSystems	3
Locus t	ag Ze	maCp032					General protein information	
Gene ty RefSeq stat	pe pro us RE	otein coding EVIEWED					NCBI Reference Sequences (RefSeq)	5
Organis	m <u>Ze</u>	<u>a mays</u>	Charles	-huter Easternahu	a. Taashaankata		Related sequences	
Linea	ge ⊑u Sn	ermatophyta: Mag	iae, Strepto noliophyta:	iliopsida: Poales:	a, Tracheophyta, Poaceae: PACMAD	)	Additional links	
	cla	ade; Panicoideae; A	Andropogon	ae; Zea				
Genomic conte	xt					≈ ?	Related information BioProjects	
							BioSystems	
Annotation S	tatus	Assembly		Chr	Location		Conserved Domains	
historia in a final de la companya de la	/9/15/21/	#genomic-regions-	transcripts-pr	aducts onlast (non	NC 001666.2		Full text in PMC	

Figure 5. Click on Genomic regions, transcripts, and products in the table on contents.

S rbcL RuBisCO large subuni ×	
← → C 🗋 www.ncbi.nlm.nih.gov/gene/845212#genomic-regions-transcripts-produc	ts 😒 \Xi
Genomic regions, transcripts, and products	PubMed (GeneRIF)
Genomic Sequence:       NC_001666.2 Chromosome Chloroplast Refere®tre1878/1848/Geng.eq.uom.eukitais.         Go to nucleotide:       Graphics         FASTA       GenBank         Nucleotide:       Graphics         Yould fast       Nucleotide FAST         Nucleotide:       Trabins         Yould fast       Trabins	PubMed(nucleotide/PMC) RefSeq Proteins Taxonomy IA report
300         55,800         57 K         157,200         57,400         57,800         157,200         158,200<	General information About Gene FAQ FTP site Help
Genes, INSDC annotation provided by Maire Genere Sequencing Project x RNA-seq exon coverage, aggregate (filtered), NCBI Zea mays Annotation Release 100X RNA-seq intron-spanning reads, aggregate (filtered), NCBI Zea mays Annotation Release X RNA-seq intron features, aggregate (filtered), MCBI Zea mays Annotation Release 100 x	My NCBI help NCBI Handbook Statistics
Bibliography	Related sites  BLAST
Related articles in PubMed 1. Ectopic expression of Rubisco subunits in maize mesophyll cells does not overcome barriers to	Genome BioProject GEO
<ul> <li>cell type-specific accumulation.</li> <li>Wostrikoff K, et al. Plant Physiol, 2012 Sep. PMID 22744982, Free PMC Article</li> <li>Nucleotide sequence of maize chloroplast rpl32: completing the apparent set of plastid ribosomal protein genes and their tentative operon organization.</li> </ul>	HomoloGene Map Viewer OMIM
Wenlähner IV. et al. Plant Mol Biol. 1003 Eeh. PMID.84/3346 ww.ncbi.nlm.nih.gov/nuccore/NC_001666.2?report=fasta&from=56874&to=58304	Probe

# Figure 6. Click <u>FASTA</u>

S Zea mays chloroplast, com X		
← → C 🗋 www.ncbi.nlm.nih.gov/nuccore/NC	_001666.2?report=fasta&from=56874&to=58304	☆ =
Zea mays chloroplast, complete g	jenome	<ul> <li>Whole sequence</li> <li>Selected region</li> </ul>
CBI Reference Sequence: NC_001666.2 GenBank Graphics		from: 56874 58304
<pre>&gt;gi 11994090:56874-58304 Zea mays chloroplast, co ATGTCACCACAAACAGAAACTAAAGCAAGTGTTGGATTTAAAGCTGGTG</pre>	mplete genome TTAAGGATTATAAATTGACTT	
ACTACKCCCGGAGTACGAAACCAAGGATACTGATATCTTGCAGCAGCATT	CCGAG TAACTCCTCAGCTCGG GGTACATGGACAACTGTTTGG	Customizo view
ACTGATGGACTTACCAGTCTTGATCGTTACAAAGGACGATGCTATCACA CAGATCAATATATCTGTTATGTAGCTTATCCATTAGACCATTTGAAGA	GGGTTCTGTTACTGGGGACC GGGTTCTGTTACTAACATGTT	Customize view
TACTTCCATTGTGGGTAACGTATTTGGTTTCAAAGCCTTACGCGCTC1 CCCCCTGCTTATTCAAAAACTTTCCAAGGTCCGCCTCACGGTATCCA4 AGTACGTCGTCCTTATTGGGATGTACTATTAAACCAAAATTGGGAT	Copy Search Google for ">gil11994090:56874-58304 Zea mays chloroj	Ctrl+C
AGCGTGTTATGAGTGTCTACGCGGTGGACTTGATTTTACCAAAGATGA TTTATGCGCTGGAGAGACCGTTTCGTCTTTTGTGCCGAAAGCAATTTAT	Print	Ctrl+P
AAATCAAGGGGCATTACTTGAATGCGACTGCAGGTACATGCGAAGAA AAGGGAATTAGGGGTTCCTATTGTAATGCATGACTACTTAACAGGAGG	Inspect	Ctrl+Shift+I
TCTCATTATTGCCGCGACAACGGCCTACTTCTCACATTCACCGAGCAA AGAAAAATCATGGTATGCATTTCCGTGTATTAGCTAAAGCATTGCGTAT CTCCGGTACAGTAGTAAGTAGAAGGGGAACGCCGAAATAACTTA GATGGTTATTGATAAAAAGGTCGTTAGGAGGGAACGCCGAAATAACTTAA	TGCATGCAGTTATTGATAGAC GTCGGGGGAGATCATATCCA GGTTTTGTTGATTATTGCGC ACTGGGTATCATGCGC	Find in this Sequence
TTATACCGGTGGCTTCTGGGGGTATCATGTTTGGCATATGCCAGCTCT	GACCGAAATCTTTGGAGATGA	LinkOut to externa
TTCCGTATTACAATTTGGTGGAGGAACTTTAGGACATCCTTGGGGAAAT CGTGTGGCTTTAGAAGCCTGTGTACAAGCTCGTAACGAAGGGCGCGATC	GCACCTGGTGCAGCAGCTAAT TTGCTCGTGAAGGTAATGAAA	Dryad Digital Reposito [Dryad
TATCAAAGLAGC FIGLAAA IGGAD GC IGAACTAGCCGCAGCTIGIGA TGATGGTTTCAAAGCGATGGATACCATATAA	AATATGGAAGGAGATCAAATT	Dryad Digital Repositc [Dryad
		Dryad Digital Reposito [Dryad
		Dryad Digital Reposito [Dryad
4		Dryad Digital Repositc IDrvad ▼

Figure 7. Copy the sequence of rcbL gene from Zea mays.



- 86 Figure 8. Paste the sequence data into a Notepad(PC) or Texteditor(Mac/Linux)
- 87 (To find Notepad on a PC with Windows go to the <u>Start menu</u>, <u>All Programs</u> then click
- 88 <u>Accessories</u> and you should see <u>Notepad</u>.)

Untitled - Notepad	
File Edit Format View Help	
>f111994090:56874-58304       Zea mays_chloroplast_complex         ATGTCACCACAAACAGAAACTAAAGCAAGTGTTGGATTTAAAGCTGGTGTTA         ACTACACCCCGGAGTACGAAACCAAGGCTGCATGTGGATTTAAAGCTGGTGTTA         ACTACACCCCCGAAAGCAGAAACCAAGGCTGCACTATCTGATATCTTGGCAGCATCTCCG         GGTTCCCCCTGAAAGACAGAGAGCTGCAGTAGCTGCAGAATCTTCTACTGGT         ACTGATGGACTTACCAGCTGTAGTGACTTAGCTGCGGAAATCTTCTACTGGT         ACTGATGGACTTACCAGTTTGGTATTGATCGTTACAAAGGACGATGCTTACCATCG         CAGATCAATTATTCTGTTATGTAGCTTATCCATTAGACCTATTCACATCG         AGTACGGTGCTCCTTTATTGGAACGTTATTCCAAAGGCGCTCTACCAAGTTG         AGTACGGTGGTCGTTATTGGAACGTATTTCCAAAGGCGCTCATTCCAAGTGC         AGTACGGTGGCCTTTATTGGAACTGATTATAAACAAAATTGGGATTATCC         AGGGAATTAGGGGTGCTACCGCGGTGGACTTGCAGCAAAATTGGGAAAATGAT         AAGGGAATTAGGGGGTCCTATTGTAATGCAGGACACTGCAGGAAATGACAAAGGGGGCAATTACCAAGGGGGCCTATTCCGTGTATTAGCAAAGCATGCGAAAGGCAAATGAT         AAGGGAAATTAGGGGGTCCTATTGTAAGCAGGGAACGCGAAATGACTACTACAGGAGGATTC         CTCCCATATTGGAGAGGGACCTTAGCTGAAGCAGGGAAATGCCAAGGCCAAATGCC         AAGGGAAATAGGGGGCCTATTCCGGGTATTCCACGAGCAATGCCAAGGCCAATGCC         AGGGGAACTTAGTGAGGAACGTTAGGCAAAGCACGAAAAAAATGCATTGGGAAATGCCCGGAAACGCAAATGCC         AGGGAAAAAACATTGGAGGGAAACTTAGGCAAATGCATTAGGAAATGCATTGGCATATGCCAGCCTCTGACCACTTTGAAAAGCATTGGGGCTTCTGGGGTATTCATGTTTGGCATATGCCAGCCA	Cut Cut Copy Paste Delete Select All Right to left Reading order Show Unicode control characters Insert Unicode control character Open IME Reconversion

90 Figure 9. After pasting into *Notepad,* leave the prompt sign > and delete text before the

91 DNA sequence, then replace deleted text with <u>Corn, the common name for *Zea mays*</u>.

File	Edit	Format	View	Help	
>Cor	'n				
ATG	CACC	ACAAAC	AGAAA	CTAAAGCAAGTGTTGGATTTAAAGCTGGTGTTAAGGATTATAAATT	GAC
АСТА	CACC	CCGGAG	TACGA	AACCAAGGATACTGATATCTTGGCAGCATTCCGAGTAACTCCTCAG	сто
GGTT	CCGC	CTGAAG	AAGCA	GGAGCTGCAGTAGCTGCGGAATCTTCTACTGGTACATGGACAACTG	TTT
ACTO	ATGG	ACTTAC	CAGTC	TTGATCGTTACAAAGGACGATGCTATCACATCGAGCCCGTTCCTGG	GGA
CAG/	TCAA	TATATC	TGTTA	TGTAGCTTATCCATTAGACCTATTTGAAGAGGGTTCTGTTACTAAC	ATG
TACT	тсса	TTGTGG	GTAAC	GTATTTGGTTTCAAAGCCTTACGCGCTCTACGTTTGGAGGATCTAC	GAA
cccd	CTGC	TTATTC	<b>AAAA</b> A	CTTTCCAAGGTCCGCCTCACGGTATCCAAGTTGAAAGGGATAAGTT	GAA
AGTA	CGGT	CGTCCT	TTATT	GGGATGTACTATTAAACCAAAATTGGGATTATCCGCAAAAAATTAC	GGT
AGCO	TGTT	ATGAGT	GTCTA	CGCGGTGGACTTGATTTTACCAAAGATGATGAAAACGTAAACTCAC	AAC
TTTA	TGCG	CTGGAG	AGACC	GTTTCGTCTTTTGTGCCGAAGCAATTTATAAAGCACAAGCCGAAAC	TGG
AAAT	CAAG	GGGCAT	TACTT	GAATGCGACTGCAGGTACATGCGAAGAAATGATTAAGAGAGCTGTA	TTT
AAGO	GAAT	TAGGGG	ттсст,	ATTGTAATGCATGACTACTTAACAGGAGGATTCACCGCAAATACTA	стт
тсто	ATTA	TTGCCG	CGACA	ACGGCCTACTTCTTCACATTCACCGAGCAATGCATGCAGTTATTGA	TAG
AGA/	AAAT	CATGGT	ATGCA	TTTCCGTGTATTAGCTAAAGCATTGCGTATGTCGGGGGGGAGATCAT	ATC
стсо	GGTA	CAGTAG	TAGGT	AAGTTAGAAGGGGAACGCGAAATAACTTTAGGTTTTGTTGATTTAT	TGC
GATO	ATTT	TATTGA	AAAAG	ATCGTTCTCGCGGTATCTTTTTCACTCAGGACTGGGTATCCATGCC	AGG
TTAT	ACCG	GTGGCT	TCTGG	GGGTATTCATGTTTGGCATATGCCAGCTCTGACCGAAATCTTTGGA	GAT
TTCC	GTAT	TACAAT	TTGGT	GGAGGAACTTTAGGACATCCTTGGGGAAATGCACCTGGTGCAGCAG	СТА
CGTO	TGGC	TTTAGA	AGCCT	GTGTACAAGCTCGTAACGAAGGGCGCGATCTTGCTCGTGAAGGTAA	TGA
TTAT	CAAA	GCAGCT	TGCAA	ATGGAGTGCTGAACTAGCCGCAGCTTGTGAAATATGGAAGGAGAGC	AAA
TGAT	GGTT	TCAAAG	CGATG	GATACCATATAA	

92

89

93

Figure 10. Notepad should appear as shown above.

Untitled - Notepad		
File Edit Format View	/ Help	
Save As	a "seek asy's "seek "seek" and	×
	9 DNA Barcoding 🕨 151110_Seq 🔹 🍕	Search 151110_Seq
Organize 🔻 New f	older	≣ ▼ 🔞
E Desktop	^ Name	Date modified Type
Jownloads	No items match you	r search.
🖳 Recent Places		
🚔 SpiderOak Hive	E	
Saved Pages		
Lebooks	_	
📗 📗 Non-HIPPA Proje	21	
🥽 Libraries		
Documents		
🚽 Music		
Pictures		
Videos	<b>▼</b>	•
File name: co	vrn.fasta	•
Save as type: Al	l Files (*.*)	•
Te	xt Documents (*.txt)	
Alide Folders		Save Cancel

Figure 11. Save this file as <u>corn.fasta</u> and in the drop down choose <u>All files (\*.\*)</u>.

96

95

97 We now have the *rbcL* gene sequence for one species. We need to collect sequences for nine

98 other species for comparison in MEGA (See Table 1). In addition, we will use Euglena viridis as

99 our outgroup. Repeat the procedure with each species listed below by typing the GenBank

100 Gene ID into the search box shown as shown in Figure 3. Make sure to select <u>Gene</u> to the right

101 of the search box when searching by the Gene ID number.

102 Table 1. List of plants from which the rcbL gene that can be used to create the phylogenetic

103 tree, their GenBank ID numbers (Accession Numbers), and the suggested file names.

Common name	GenBank ID	File name
Corn	845212	corn.fasta

Thale cress	844754	thalecress.fasta
Rice	4126887	rice.fasta
Tobacco	800513	tobacco.fasta
Potato	4099985	potato.fasta
Liverwort	2702554	liverwort.fasta
Sunflower	4055709	sunflower.fasta
Grape	4025045	grape.fasta
Cucumber	3429289	cucumber.fasta
Spinach	2715621	spinach.fasta

# **Obtaining the Outgroup:**

*Euglena viridis* will be used as the outgroup in our evolutionary tree. We will use the NCBI

107 GenBank to locate the sequence for *Euglena viridis*.



109 Figure 12. Starting from GenBank Home we will select the Nucleotide search filter

110 option to the left of the search box.

# 111

112



113 Figure 13. Type **U21010.1** into the search box and click <u>search</u>.

P Euglena viridis ribulose 15 x	
← → C P www.ncbi.nlm.nih.gov/nuccore/U21010.1	∽ =
Nucleotide	S
Advanced	
GenBank 🗸	Send: - Change region sh
Euglena viridis ribulose 1,5-bisphosphate carboxylase/oxygena large subunit (rbcL) mRNA, chloroplast gene encoding chlorop protein, partial cds	ase Customize view
GenBank: U21010.1	Analyze this seque
FA&TA Graphics	Run BLAST
	Pick Primers
<u>Go to:</u> 🕑	Highlight Sequence Fe
LOCUS EVU21010 1310 bp mRNA linear PLN 12-MAR-1996	Find in this Sequence
DEFINITION Euglena viridis ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) mRNA, chloroplast gene encoding chloroplast	
protein, partial cds.	Related informatio
ACCESSION U21010	Full text in PMC
KEYWORDS .	Brotoin
SOURCE chloroplast Euglena viridis	Flotelli
ORGANISM <u>Euglena viridis</u>	PubMed
Eukaryota; Euglenozoa; Euglenida; Euglenales; Euglenaceae; Euglena.	Taxonomy
AUTHORS Thompson,M.D., Copertino,D.W., Thompson,E., Favreau,M.R. and Hallick.R.B.	
TITLE Evidence for the late origin of introns in chloroplast genes from	Recent activity
www.ncbi.nlm.nih.gov/nuccore/699536?report=fasta	· · · ·

# 115 Figure 14. Click FASTA

S Euglena viridis ribulose 1,5 ×	
← → C 🗋 www.ncbi.nlm.nih.gov/nuccore/699536?report=fasta	☆ <b>=</b>
FASTA - Send: -	Change region sh
Euglena viridis ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) mRNA, chloroplast gene encoding chloroplast protein. partial cds	Customize view
GenBank: U21010.1 GenBank: Graphics	Analyze this seque Run BLAST
Salo99536[gb]U21010.1[EVU21010 Euglena viridis ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) mRNA, chloroplast gene encoding chloroplast	Pick Primers Highlight Sequence Fr
protein, partial cds ACTACCGTCTTACGTACTATACGCCTGACTATCAAGTTGCTGAAACTGACATTTTAGCAGCTTTCCGTAT GACACCTCAACCAGGTGTTCCTGCTGAAGAGTGTGGAGCCGCTGTAGCTGCTGAATCTAGTACAGGTACT GGACAACTGTTTGGACTGATGGACTAACACAATTAGATAAATACAAAGGTCGTTGTTATGACTTAGAAC	Find in this Sequence
СТЕПТССАБЕТБАААБСААТСААТАТТАСТТАТЕСТАТССТАТС	Ctrl+C PMC
GAGATAAACTTAATAAATATGGACGGCCCGCTTTTAGG TAAATATTATGGTAAGACTGTTTATGGACAGTGTTTACGA Print	Ctrl+P
ATCAGAACTEGGGAAGTATTATAGGGTCATTATAY Inspect Ctractional Control Contro	rl+Shift+I
GCAAACACTTCTTTATCTATGTTCTGTCGTGATAATGGTTTGTTACTTCATATTCACCGTGCAATGCATG CTGATATCGATCGTCAAAGAAATCATGGGAATTCATTTCCGTGTTCTTGCAAAAAACTCTTCGGATGTGGGG TGGTGATCATTGCATTCAGGAACTGTTGTTGGTAAATTAGAAGGTGGAACGTGGAAGTTACTTAGGTTTT GTAGATCTTATGCGTGATCCTTATATTGAAAAGATCGTTCTAGAGGTATTTATT	Recent activity
GTGGTATGGGTGGAGTAATGCCTGTTGCATCCGGAGGTATTCATGTATGGCATATGCCAGCTCTTACTGA AATCTTTGGTGATGATGCTTGTTTACAATTTGGCGGTGGTACACTGGGGCATCCTTGGGGGAACGCTCCT GGAGCGGTAGCAAACAGGGTAGCGTCAGAAGCTTGTGTACAGGCTAGAAATGAAGGACGTGATTTGTCTC	Euglena viridis rit bisphosphate
	8//75/fuid1 ΔND *

116

117 Figure 15. Copy the sequence and pate into a new Notepad file. Repeat steps in figures

118 10 and 11. Save the file as euglenaviridis.fasta.

# 119Building the Evolutionary Trees:

- 120 The first step in the process of building evolutionary trees with this molecular data is to
- 121 download MEGA and install it on the computers that are going to be used for the
- 122 project. This tutorial features the latest stable version of MEGA at the time of print
- 123 (<u>http://www.megasoftware.net/mega.php</u>).
- 124 Instructions for using MEGA:
- 125 1. Open MEGA.
- 126 2. Click on <u>Align</u> then selected <u>Edit/Build Alignment</u>



- 127
- 128 Figure 16. Click Edit/Build Alignment
- 129 3. Create a new alignment. A secondary menu will appear that requires you to select an
- 130 option. Choose <u>Create a new alignment</u> option and click <u>OK</u>.



132 Figure 17. Select Create a New Alignment and click OK

- 133 4. A second submenu will appear asking you to select the type of sequence data that will
- be used to build the alignment. Select the <u>DNA</u> option (Figure 18). This will open the
- 135 MEGA Alignment Explorer in a new window (Figure 19).



### 137 Figure 18. Click DNA

MEGA 6.06(6	14022	6)											
File Analysis	Help	α.β ¥—	•  <sup>1</sup>	L _	<u> Ti</u>	с <del>і</del>	(1,2)	্ প্ল		. 🜌	<b>.</b> ()	÷	-
Align	🗰 M	l6: Alignm	nent Exp	lorer				-		-	_		
	Data	Edit	Search	Alignm	ent Web	Sequencer	Display	Help					
1984 268 49 8		🖻 🖌	UUC S4F	1	i 👬 🛛 W	<u> </u>	1	n B	% 🖻 >	- 兴   🛉	1 🛅 🛛 📲	D (4	444
	DN/	Sequen	ces Tr	ranslated	l Protein Se	equences							
-					0								
MEGA release #													
	Site #	# 0			with O	w/o Gaps							

138

139 Figure 19. Alignment Explorer opens

- 140 5. At the top of the MEGA Alignment Explorer Window select the Edit menu by clicking on
- 141 it. From this menu, select the <u>Insert Sequence From File</u> option (Figure 20). This will
- 142 open a new window.

Data	Edit	Search	Alignment	Web	Seque	encer	Display	Help										
D	ß	Undo		Ctrl+	۰Z	Ж	*	K) B	1 %	ß	× )	€   ♦	d 🖄	)   •	I D	桷	# #	. 4
DNA	Ēð	Сору		Ctrl+	-C	es												
	*	Cut		Ctrl+	⊦X													
	B	Paste		Ctrl+	۰V													
	$\times$	Delete																
	×	Delete Ga	aps	Ctrl+D	)el													
	B	Insert Bla	nk Sequence	Ctrl+	N													
	콀	Insert Sec	quence From	File Ctrl	+1													
		Select Sit	e(s)		~~													
		Select Se	quence(s)															
		Select All		Ctrl+	A													
	$\checkmark$	Allow Bas	se Editing															
		Mofify Al	ll Bases to Up	percase														

- 143
- 144 Figure 20. Click Insert Sequence from File
- 145 6. In the window of the <u>Open</u>, select the .fasta data files saved earlier

M6: Alignment Explorer	
W Open	✓ 49 Search 151110_Seq
Organize 🔻 New folder	₩ ▼ 1 0
★ Favorites     ▲     Name       ■ Desktop     ▶     No items m       ▶ Downloads     ▶     Finished       ■ Recent Places     ■     Saved Pages       ▶ Ebooks     ▶     Non-HIPPA Proje	Date modified         Type           atch your search.         ABI (".abi;".ab1)           Staden (".s.cf.)         Text (".txt", seq)           XML (".xml)         Supported sequence files (".fas;".fst;".ffa;".fsa,".fasta;".fsta; ".mexus;"           FASTA (".fsa;".fst;".fta;".fsa;".fsa;".fasta;".fsta; ".mexus;"           PAUP/MacClade (".nexus;".nex)           MEGA (".meg]           Aln Session (".mas)
Ibitraries	*.* *.aln *.phylip*.phylip2 *.gcg *.pir *.nofrf*.nbr *.msf *.ig *.ig *.ig *.ig •.dBI (*.abi;*.ab1) ▼ Open Cancel

147 Figure 21. Select Supported sequence files

				• A A A
○○○ <mark>   </mark> ≪ 151109 D	NA Barcoding   151110_Seq	Search 151110_Seq	٩	
rganize 👻 New fold	er			
Favorites	Name	Date modified	Туре	
Nesktop	🕙 corn.fasta	11/10/2015 9:39 PM	FASTA File	
🐌 Downloads	🕙 cucumber.fasta	11/13/2015 3:50 PM	FASTA File	
Finished	🕙 euglenaviridis.fasta	11/13/2015 3:57 PM	FASTA File	
🖳 Recent Places 🛛 🗉	🕮 grape.fasta	11/13/2015 3:49 PM	FASTA File	
🖶 SpiderOak Hive	🕙 liverwort.fasta	11/13/2015 3:47 PM	FASTA File	
Saved Pages	🕙 potato.fasta	11/13/2015 3:46 PM	FASTA File	
Ebooks	🖳 rice.fasta	11/13/2015 3:44 PM	FASTA File	
퉬 Non-HIPPA Proje	👜 spinach.fasta	11/13/2015 3:51 PM	FASTA File	
	🕙 sunflower.fasta	11/13/2015 3:48 PM	FASTA File	
Libraries	🕙 thalecress.fasta	11/13/2015 3:43 PM	FASTA File	
Documents	🖳 tobacco.fasta	11/13/2015 3:45 PM	FASTA File	
J Music				
Pictures				
🛃 Videos				
•				

- 149 Figure 22. Select the .fasta files and click Open
- 150 7. Once you have selected all of the files that you wish to upload, select the <u>Open</u> button
- 151 by clicking on it.
- 152 8. Once the sequences are loaded into MEGA, we want to align them. This is done by going
- 153 to the <u>Alignment</u> menu at the top of the Alignment Explorer window. Click to open the
- dropdown menu and select <u>Align by ClustalW</u> by clicking on it (Figure 23).

👬 M6: Alignment Explorer		
Data Edit Search Alig	nment Web Sequencer Dis	play Help
🖹 🗅 🚅 🔚 📽 🛛 🖤	Align by ClustalW	🗠 🖻 🌡 💼 🗙 💥 椈 🏙 🖉 💶 🕨 🛛 🖓 🥠 🧌
DNA Sequences Tr W	Align by ClustalW (Codons)	
Species (7 Crown N V	Align By Muscle	
1. Corn	Align by Muscle (Codons)	AACTAAAGCAAGTGTTGGATTTAAAGCTGGTGTTAAGGATTATAA
2. Cucumb	Mark/Hannak Site - Chill M	GACTAAAGCAAGTGTTGGATTCAAAGCTGGTGTTAAAGATTATAA
3. Euglen	Mark/Unmark Site Ctri+M	TATACGCCTGACTATCAAGTTGCTGAAACTGACATTTTAGCAGCT
4. Grape	Align Marked Sites Ctrl+L	GACTAAAGCAAGTGTTGGATTCAAAGCCGGTGTTAAAGATTACAA
5. Liverw	Unmark All Sites	GACTAAAGCAGGTGTTGGATTCAAAGCTGGTGTTAAAGATTATCG
6. Potato 🏾 🕺	Delete Gap-Only Sites	GACTAAAGCAAGTGTTGGATTCAAAGCTGGTGTTAAAGAGTACAA
7. Rice 🗸	Auto-Fill Gaps	ACGTATGTCACCACAAACAGAAACTAAAGCAAGTGTTGGATTTAA
8. Spinad		-GACTAAAGCAAGTGTTGGATTTAAAGCTGGTGTTAAAGATTACAA
9. Sunfld	AIGICACCACAAACAG	AGACTAAAGCAAGTGTTGGATTCAAAGCTGGTGTTAAAGATTATAA
10. Inale	ATGICACCACAAACAG	AGACIAAAGCAAGIGIIGGGIICAAAGCIGGIGIIAAAGAGIAIAA
11. IODad	AIGICACCACAAACAG	AGACTAAAGCAAGIGIIGGAIICAAAGCIGGIGIIAAAAGAGIACAA
		۹. ا
Site # 1434	with © w/o Gaps	



### 156 Figure 23. Align By ClustalW

- 157 9. This will open another window that is filled with ClustalW parameters. For our purposes,
- 158 the default settings are adequate. Select the <u>OK</u> option at the bottom of this menu to
- 159 proceed (Figure 24). This will set in motion the alignment algorithm. Aligning the
- 160 sequences may take several minutes depending on the size and number of the
- 161 sequences being examined.

	👫 M6: ClustalW Paramete	ers 🗆 🗆 🕮	
👫 M6: Alignment Explorer	DNA		
Data Edit Search Alignment	Pairwise Alignment		
📄 🗅 🚅 🔚 👯 🗮 🏍 🎆	Gap Opening Penalty	15	🖹 🕕 🕨 🏘 🏘 🏘
DNA Sequences Translated Pro	e Gap Extension Penalty	6.66	
Species/F Group Name *	Multiple Alignment		
1. Corn AT	G Gap Opening Penalty	15	AAGCIGGIGIIAAGGAIIAIAA
2. Cucumb AT	G Gap Extension Penalty	6.66	AAGCIGGIGIIAAAGAIIAIAA
3. Euglen AC			TGAAACTGACATTTTAGCAGCI
4. Grape AT	G DNA Weight Matrix	IUB 👻	AAGCCGGTGTTAAAGATTACAA
5. Liverw AT	G Transition Weight	0.5	AAGCTGGTGTTAAAGATTATCG
6. Potato AT	G		AAGCTGGTGTTAAAGAGTACAA
7. Rice AT	G		CTAAAGCAAGTGTTGGATTTAA
8. Spinac AT	G		AAGCTGGTGTTAAAGATTACAA
9. Sunflo AT	G		AAGCTGGTGTTAAAGATTATAA
10. Thale AT	G		AAGCTGGTGTTAAAGAGTATAA
11. Tobac AT	G		- AAGCIGGIGIIAAAGAGIACAA
	Use Negative Matrix	OFF V	
	Delay Divergent Cutoff (%)	30	
	Keep Predefined Gaps		
<	Specify Guide Tree		•
Site # 1434 💮 🖲 with			
		VOK X Cancel	

163 Figure 24. ClustalW Parameters Dialog. Leave all set to their defaults.

164	10. Now that the sequences are aligned (Figure 10), we may need to trim the ends of the
165	sequences so that they line up nicely and do not contain too much excessive data. To do
166	this, look for an asterisk (*) in the line above the first sequence in the Alignment
167	Explorer window. At the first asterisk from the beginning of the sequences, use Click and
168	Drag to select the nucleotides from there to the beginning of the sequences to be
169	removed. (Figure 25).

DNA Sequences Translated Protein Sequences									
Species/I Group Name		* :	* * * * *	* ** ** ** **	** *				
1. Corn	ATTTAAAGCTGG	TGTTAAGGA <mark>T</mark>	A T A A A T T <mark>G A C</mark> T T	ACTACACCCCGGAG	TACGAAACCAAG				
2. Cucumb	ATTCAAAGCTGG	TGTTAAAGAT <mark>T</mark>	A T A A A T T <mark>G</mark> A C T T	A T T A T A C T C C T G A A	TATGAAACCAA				
8. Euglen		EEEEEEEEEEEE	A C C G T C T T A C G T	ACTATACGCCTGAC	TATCAAGTTGC				
. Grape	ATTCAAAGCCGG	TGTTAAAGAT <mark>T</mark>	ACAAATI <mark>G</mark> ACII	A T T A T A C T C C T G A A	TATGAGACCAA				
. Liverw	ATTCAAAGCTGG	TGTTAAAGAT <mark>T</mark>	A T C G A T T A A C T T	A T T A C A C T C C G G A T	TATGAGACCAA				
. Potato	ATTCAAAGCIGG	TGTTAAAGAG <mark>T</mark>	ACAAATTGACTT	ATTATACTCCTGAG	TACCAAACCAA				
. Rice	ATTTAAAGCIGG	TGTTAAGGAT <mark>T</mark>	A T A A A T T <mark>G A C</mark> T T	ACTACACCCC <mark>GG</mark> AG	TACGAAACCAA				
. Spinac	ATTTAAAGCTGG	TGTTAAAGAT <mark>T</mark>	ACAAATTGACTT	ATTATACTCCTGAG	TATGAAACCCT:				
. Sunflo	ATTCAAAGCIGG	TGTTAAAGAT <mark>T</mark>	A T A A A T T <mark>G A C</mark> T T	ATTATACTCCTGAA	TATGAAACCAA				
10. Thale	GTTCAAAGCIGG	TGTTAAAGAG <mark>T</mark>	A T A A A T T <mark>G A C</mark> T T	ACTATACTCCTGAA	TATGAAACCAA				
11. Tobac	ATTCAAAGCTGG	TGTTAAAGAG <mark>T</mark>	ACAAATTGACTT	A T T A T A C T C C T G A G	TACCAAACCAA				
7. Rice 8. Spinac 9. Sunflo 10. Thale 11. Tobac	ATITAAAGCTGG ATITAAAGCTGG ATICAAAGCTGG GTICAAAGCTGG ATICAAAGCTGG	TGTTAAGGAT TGTTAAAGAT TGTTAAAGAT TGTTAAAGAG TGTTAAAGAG	A TAAA TIGACII A CAAAIIGACII A TAAAIIGACII A TAAAIIGACII A TAAAIIGACII A CAAAIIGACII	ACTACACCCC ATTATACTCC ATTATACTCC ACTATACTCC ACTATACTCC ATTATACTCC	GGAG TGAG TGAA TGAA TGAA				

171 Figure 25. Remove the Excess Nucleotides

- 172 11. This process will need to be repeated for the tail ends of the sequences. Here find the
- 173 last asterisk from the end, then Click and Drag to highlight the nucleotides to the end of
- the sequences.
- 175 12. Now the sequences are aligned and trimmed (Figure 26).

🗰 M6: Alignment Explorer		
Data Edit Search Alignm	nent Web Sequencer Display Help	
🗅 😅 🖬 📽 🗮 🌾	B 🎆   W 💔   💥 🐍   🗠 🖻 🌡 💼 🗙 兴   物 🎒   4 🕨 -	M 🏘 🏘 🐴
DNA Sequences Translate	d Protein Sequences	
Species// Group Name	** *********	* ** **** **
1. Corn	TATAAATIGACITACTACACCCCGGAGTACGAAACCAAGGATACIGATA	I <mark>c I I g g c a g c a I</mark> I
2. Cucumb	TATAAATIGACITATTATACICCIGAATAIGAAACCAAAGATACIGATA	I <mark>C I I G G C A G C A</mark> I I
3. Euglen	TACCGTCTTACGTACTATACGCCTGACTATCAAGTTGCTGAAACTGACA	I T T T <mark>A G C A G C</mark> T T I
4. Grape	TACAAATTGACTTATTATACTCCTGAATATGAGACCAAACCTACTGATA	I <mark>C I I G G C A G C A</mark> I I
5. Liverw	TATCGATTAACTTATTACACTCCGGATTATGAGACCAAGGATACGGATA	I T T T <mark>A G C A G C A</mark> T I
6. Potato	TACAAATTGACTTATTATACTCCTGAGTACCAAACCAAGGATACTGATA	T <mark>attggcag</mark> catt
7. Rice	TATAAATTGACTTACTACACCCCGGAGTACGAAACCAAGGACACTGATA	T <mark>C T T G G C A </mark> G C <mark>A</mark> T T
8. Spinac	TACAAATIGACITATIATACICCIGAGIAIGAAACCCIAGAIACIGAIA	I <mark>C I I G G C A </mark> G C <mark>A</mark> I I
9. Sunflo	TATAAATTGACTTATTATACTCCTGAATATGAAACCAAGGATACTGATA	I <mark>C I I G G C A G C A</mark> I I
10. Thale	TATAAATIGACITACIATACICCIGAATAIGAAACCAAGGATACIGATA	I <mark>C I I G G C A </mark> G C <mark>A</mark> I I
11. Tobac	TACAAATTGACTTATTATACTCCTGAGTACCAAACCAAGGATACTGATA	T <mark>a t t g g c a g c a t</mark> t
<		Þ
Site # 1301 💮 🕷	with 🗢 w/o Gaps	



178 13. Now, we need to save the Alignment Session so that the data is saved in a format that
179 MEGA can use to build the phylogenetic trees. Save the Alignment Session by selecting
180 the <u>Data</u> menu at the top right of the Alignment Explorer window. Click on the <u>Save</u>
181 <u>Alignment</u> option. (Figure 27). After this has been completed and the session saved,
182 close the Alignment Explorer window.



183

### 184 Figure 27. Save Session

185 14. Now, we need to open the saved alignment session by selecting the <u>File</u> menu in the

186 general MEGA window and clicking on the <u>Open a File/Session</u> option (Figure 28).

File	Analysis	Help										
2	Open A File	/Session	Ctrl+O	-	<u><u><u></u><u><u></u><u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u></u></u></u></u>	. <del>6</del> .	(1,2)	8	₹N.	. 🗮	. <u>()</u>	<b>†</b>
	Open a Rece	ently Used F	ile	e	Diversity	Phylogeny	User Tree	Ancestors	Selection	Rates	Clocks	Diagnose
e∰ artis	Edit a Text F Convert File	ile Format to I	MEGA									
B.	Printer Setu	p										
	Exit MEGA		Alt+X									
	(Ch)		õ			=	G	3		Ľ	6	
	•••									· · · ·		_

187

188 Figure 28. Open A File-Session

189 15. Select the saved alignment file (it will be a .MAS file) in the Open a File window and

190 select <u>Open</u> by clicking on it (Figure 29).



191

192 Figure 29. Click Open

- 193 16. This will bring up a second window where you have the choice to open the .MAS to
- analyze it or align it. Select <u>Analyze</u> by clicking on it (Figure 30).



195

- 196 Figure 30. Click Analyze
- 197 17. To construct a phylogenetic tree, select the <u>Phylogeny</u> menu mid-way through the

198 second menu bar at the top of the MEGA window. Here we will continue our example

- 199 with a Neighbor-Joining tree, but the process is the same for other types of phylogenetic
- 200 trees. Select <u>Construct/Test Neighbor-Joining Tree</u> (Figure 31).



- 202 Figure 31. Construct-Test Neighbor-Joining Tree
- 203 18. A menu will appear that asks if you want to use the currently active data sheet, select

## 204 <u>Yes</u> (Figure 32).



- 206Figure 32. Click Yes
- 207 19. This will open a dialogue box called <u>Analysis Preferences</u>. For the <u>Statistical Method</u>
- 208 select <u>Neighbor-Joining</u> and for the <u>Test of Phylogeny</u> select <u>Bootstrap Method</u>. In the
- 209 field entitled <u>No. of Bootstrap Replications</u>, select 1000 to obtain stable estimates of
- 210 reliability of the tree. For the <u>Substitution Type</u> start by selecting <u>Nucleotide</u> followed by
- 211 selecting the <u>Jukes-Cantor Model</u>. Here all other fields are left at their default values. To
- 212 generate the tree, click on <u>Compute</u> (Figure 33).

ſ	M6: Analysis Preferences		
	Options Summary		
	Option	Selection	
MEGA 6.06(6140226	Analysis	Phylogeny Reconstruction	
File Archeis Hele	Scope	All Selected Taxa	
TA TA	Statistical Method	Neighbor-joining	
Alion Data	Phylogeny Test		
	Test of Phylogeny	Bootstrap method	Clocks Diagnose
	No. of Bootstrap Replications	1000	
Data	Substitution Model		2
	Substitutions Type	Nucleotide	
	Genetic Code Table	Not Applicable	G
	Model/Method	Jukes-Cantor model	i Si
	Fixed Transition/Transversion Ratio	Not Applicable	
	Substitutions to Include	All	
🚱 😼	Rates and Patterns		······································
MEGA release #6140226	Rates among Sites	Uniform rates	d MEGA Alignment.mas
	Gamma Parameter	Not Applicable	
	Pattern among Lineages	Same (Homogeneous)	
	Data Subset to Use		
	Gaps/Missing Data Treatment	Complete deletion	
	Site Coverage Cutoff (%)	Not Applicable	
	Select Codon Positions	V 1st V 2nd V 3rd V Noncoding Sites	
	? Help	Compute X Cancel	

### Figure 33. Click Compute.

20. After a few minutes, a tree will be generated. The length of time that this takes will
depend in part on the length and number of sequences that are being used to create the
tree (Figure 34).
21. The numbers on the branches of the tree represents the Bootstrap value, which is the
statistical support that each branch receives by the Bootstrap analysis. Higher numbers
mean that the branch has higher support and is most likely to be a real branch (Figure

221 34).



223 Figure 34. Generated Tree

224 22. To simplify the tree, we now want to condense or cut out the branches that have less

support and are less likely to be true branches. To do this, go to the <u>Compute</u> menu on

the TreeExplorer menu and select <u>Condensed Tree</u> (Figure 35).



### 228 Figure 35. Select Condensed Tree

- 229 23. This opens a new menu, <u>Tree Options</u>. Select the <u>Cutoff</u> submenu and input 50 for the
- 230 <u>Cut-off Value for Condensed Tree</u>, then click <u>OK</u> (Figure 36). Leave all other values at

their defaults.

	<b>#</b>	M6: Tree E	xplore	er (	M M6: Tree Ontions						23		Л
	File	Image	Sub	ree	Troo	Branch	Labele	Scala	Cutoff	1			
		۵ ا	N	Û	nee	Dianch	Labels	Scale	outon				
MEGA 6.06(6:	R	Original	tree	Boot	Cut-of	f Value fo	r Conden	sed Tree	50		%		
File Analysis	GLE -				Cut-of	f Value fo	r Consen	sus Tree	50		%	Potato	*
Align Da	12											pacco	Diagnose
TA 🗮	J.											unflower	
Close Data	<b>4</b> -4											- Spinach	2
	1260											cumber	
	垕			г								e	<b>Q</b>
	<b>#</b>											Thale_cress	
	E)											0	
First Time User?	Ē,			l								0	+ 85
MEGA release #6	H.											- Euglena_Viridis	Alignment.mas
	۲°												
	11												
	SBL	. = 0.71138	3161			?	Help	✓ <u>o</u> l	<u>_</u>	( <u>C</u> a	ncel		



234 24. Now, the tree in the TreeExplorer will reflect the changes. All branches that have less

than 50% support will have been removed. (See Figure 37)







238 25. The last thing that we need to do is to set our out group. In our example here it is
239 Euglena viridis. To do this, right click on the branch that has Euglena viridis. This brings
240 up a submenu. In this submenu, select the <u>Place Root</u> option. This provides us with a
241 rooted phylogenetic tree. (See Figure 38)



243 Figure 38. Place Root



245 Figure 39. *Euglena viridis* as the Root

The tree can be saved as a PDF or printed out. To save the tree as a PDF, go to the <u>Image</u> menu

and select Export as PDF. A window will pop up and you can save the file there. To Print the

tree, there is a <u>Printer</u> Icon that you can use just below the upper menu.

#### 249 Assessment Questions

In order to ensure that students understand the output of this exercise, questions along the following lines may be asked: Which species are most closely related? Give an example of sister taxa. Why do you think that corn and rice are so closely related? How many base pairs were included in your analysis? How do you think the length of the gene sequence used affects the validity/reliability of your results? Why does the liverwort group with the outgroup? These questions, or similar, will let the instructor assess not only if the student has understood the process that they have gone through to create the phylogenetic tree, but in conjunction with the questions that the students should consider as they build the tree, the student'sunderstanding of the process.

259

#### 260 Additional Background Information:

261 The NJ and ML methods for building evolutionary trees rely on different statistical principles 262 (Nei and Kumar 2000; Tamura et al. 2011). In NJ, the least squares method is used along with 263 pairwise evolutionary distances (Nei and Kumar 2000). In ML, the maximum likelihood is 264 optimized such that the inferred tree is the most likely tree (Nei and Kumar 2000). Generally, 265 they will produce very similar results, but NJ is much faster. Despite slight differences in the 266 branching patterns between NJ and ML trees, they both are robust methods for building 267 evolutionary trees. The Jukes-Cantor model is simply a mathematical model that describes the 268 change of one the nucleotides in the DNA sequence to another one, over time (Nei and Kumar 269 2000). All of its parameters are automatically estimated by MEGA. 270 Both NJ and ML produce trees that are unrooted, even though they are frequently drawn from 271 left to right. In this case, if one knows the outgroup then it can be used to properly root the 272 tree. Choosing a proper out group can be a difficult task and may require some trial and error 273 (Nei and Kumar 2000). A good out group should be similar to the sequences in question, but 274 different enough so that the computer program can see the differences (Nei and Kumar 2000). 275 Acknowledgements: 276 The authors acknowledge the editors and Dr. Sudhir Kumar, Director of iGEM (Institute for Genomics 277 and Evolutionary Medicine at Temple University for their thoughtful assistance with this article.

278 *References:* 

- Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol.
- 280 30:1229-1235.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA 5L: Molecular
- evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum
- 283 parsimony methods. Mol. Biol. Evol. 28:2731-2739
- Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2013) MEGA6: Molecular evolutionary
- 285 genetics analysis version 6.0. Mol. Biol. Evol. 30:2725-2729
- 286 Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford,
- 287 UK. pp 333.
- 288 Newmaster, SG, Fazekas, JA, and Ragupathy S (2006) DNA barcoding in land plants: evaluation
- rbcL in a mulitigene tiered approach. Can. J. Bot. 84:355-341.
- 290 Ryan, MP, Adley, CC, Pembroke, JT (2013) The use of MEGA as an educational tool for
- 291 examining the phylogeny of antibiotic resistance genes. In. Microbial pathogens and strategies
- for combating them: science, technology and education. Ed. A Menendez-Vilas. pp. 736-743.