

The phylogenetic pipeline guide

Marc W Cadotte

Spring 2025



This guide will cover the basics of constructing and manipulating phylogenies. There are ways to build the entirety of these pipelines in programming language or with APIs, but we will use web interfaces in a step-by-step approach that aids in understanding how to construct a phylogeny.

You can make a phylogeny if you have a backbone tree by matching taxonomic labels (family or genus) using phylo.maker in R

<https://rdr.io/github/jinyizju/V.PhyloMaker/man/phylo.maker.html>

This is particularly robust for plants and birds, but less so for other taxa.

To do these steps in R, you will need to download separate executables that R packages will call to.

Step 1: Download nucleotide sequences

We can generate sequence data in a number of ways, but we will simply download sequences for two genes, a conserved one and a neutral one. We will do this for a small group of plants for *rbcl* (conserved gene) and ITS1 (neutral marker) – we will discuss some subtle issues as we look at the genes.

Here are the list of species (I often use taxa codes, I'll explain why):

Rudbeckia hirta (RUHI)

Solidago canadensis (SOCA)

Helianthus divaricatus (HEDI)

Leucanthemum vulgare (LEVU)

Rudbeckia laciniata (RULA)

Monarda fistulosa (MOFI)

Bromus inermis (BRIN) -outgroup

For each species, we will copy an *rbcl* and ITS1 sequence into a text file (you should always record accession numbers for repeatability), like this:

Taxa	Taxa code	rbcl	Its1
<i>Rudbeckia hirta</i>	RUHI	AY215173.1	KX671869.1
<i>Monarda fistulosa</i>	MOFI	MK526203.1	AF369200.1

Go to <https://www.ncbi.nlm.nih.gov/> and change “All databases” to “nucleotides” (Fig. 1). Then, to keep it simple, enter one species name at a time with one gene like this:

Rudbeckia hirta and *rbcl*

Then click on Genomes: Nucleotide (Fig. 2). You will then usually see more than one *rbcl* sequence, and you can click on any one of them, I will select the first (Fig. 3), but we can discuss subtle issues. The you get the full accession record and we can click on “FASTA”

(Fig. 4). This gives you the sequence in a FASTA format. Copy the text from >....to the end of the sequence and copy it and paste it into a text file, and remove the descriptive text except for the species name (Fig. 5).

We will be using a species code, like RUHI, because some software cuts names down to a minimum number of characters.

Now repeat this for the other species, and then for ITS1 and in a separate text file. Now that we have two files, one for each gene, and each file contains the sequences for five species, we can move on to the next step.

Step 2: Align sequences

The individual sequences in our text files are of different lengths and can also contain repeats and deletions, so we need to align the sequences so that analyses can determine probabilities of changing A to G, etc. There are a number of algorithms to do this, but we will use a classic one, Muscle:

<https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein>

Here we just copy and paste our sequences (separate for each gene) and let the algorithm run. I prefer the output “Fasta”. Save each aligned file with a meaningful name like “rbcl_aligned_fasta”.

Step 3: Determine nucleotide substitution model

A nucleotide substitution model describes the process by which one nucleotide in a sequence is replaced by another over time (e.g., A to T). These models are essential for understanding evolutionary relationships and constructing phylogenetic trees. They assume that nucleotide substitutions happen at different rates depending on factors like the type of substitution (e.g., purine to purine) and the underlying evolutionary pressures on the DNA sequence. The analyses in the next step will do this automatically, but some phylogenetic analyses want/allow you to enter a preferred model.

Common models include:

- **Jukes-Cantor (JC69):** Assumes equal substitution rates for all nucleotides. (A-T = T-A = C-G = G-C)
- **Kimura two-parameter (K2P):** Differentiates between transitions (purine-to-purine or pyrimidine-to-pyrimidine) and transversions (purine-to-pyrimidine or vice versa).
- **General Time Reversible (GTR):** A more complex model that allows for different rates of substitution between each pair of nucleotides.
- **Transition model (TIM):** Variable frequencies and transition rates and two transversion rates

But there are many many versions of models (Fig 6). Choosing the right model helps in accurately estimating evolutionary distances and constructing reliable phylogenies.

We will use IQ-Tree server: <http://iqtree.cibiv.univie.ac.at/>

And we upload the alignment file and run it. The output (Fig. 7) lists model in order of support.

For rbcl, my best model was HKY+F+G4; and for ITS1 it was JC (not surprising, fewest parameters). These are now used for the tree. Importantly, yours could be different if you include different sequences.

Step 4: Run trees

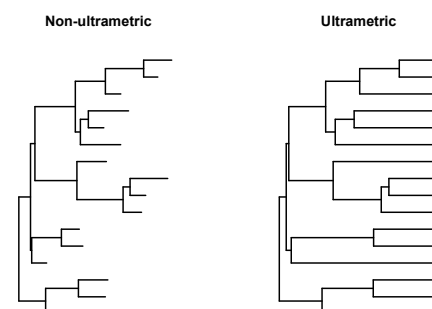
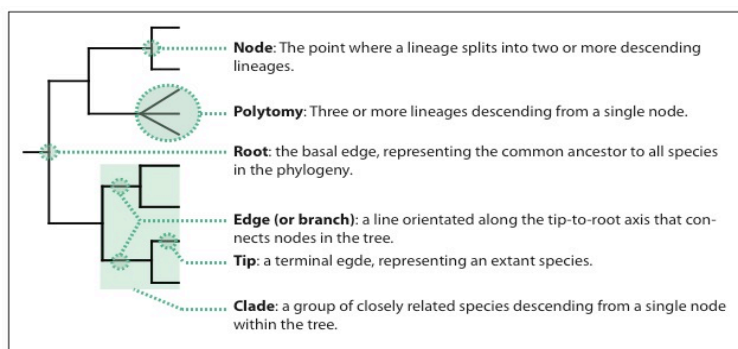
We will run maximum likelihood analysis to determine phylogenies. Here we will do it separately on the two sequences. We would optimally want to combine the sequences (concatenate), but we'll skip this for now so we can compare.

Again, we will use IQ-Tree server: <http://iqtree.cibiv.univie.ac.at/> but there are many other options out there (NJ, ML, Bayesian, etc.). The default settings are fine, but I'll explain some options, we just need to upload aligned files (Fig. 8).

Step 5: Examine and manipulate trees in R

We can root the trees and make them ultrametric (Fig. 9 & 10). We will switch to R script, then concatenate and redo tree in iqtree and do rooting etc in R (Fig. 11).

Then we can prune the outgroup (Fig. 12)



Step 6: do analyses!

We will do this in R with example files.

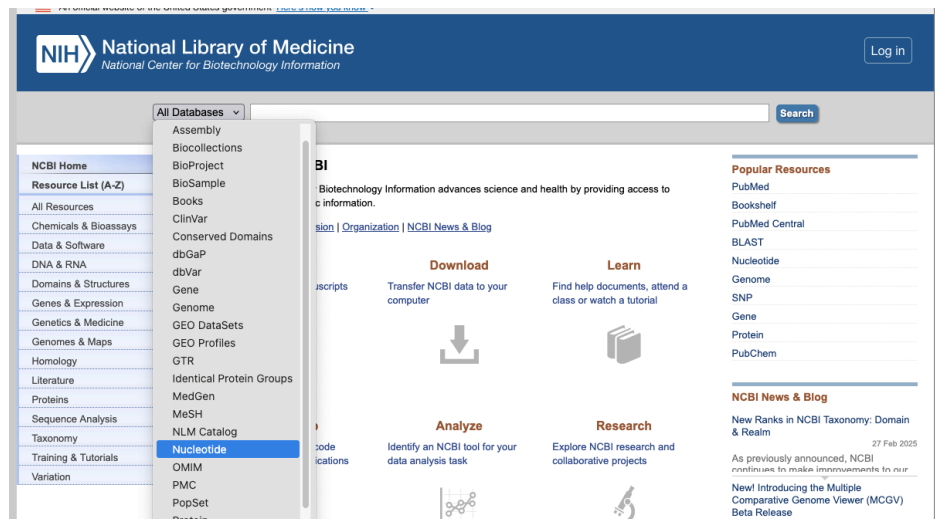


Fig. 1

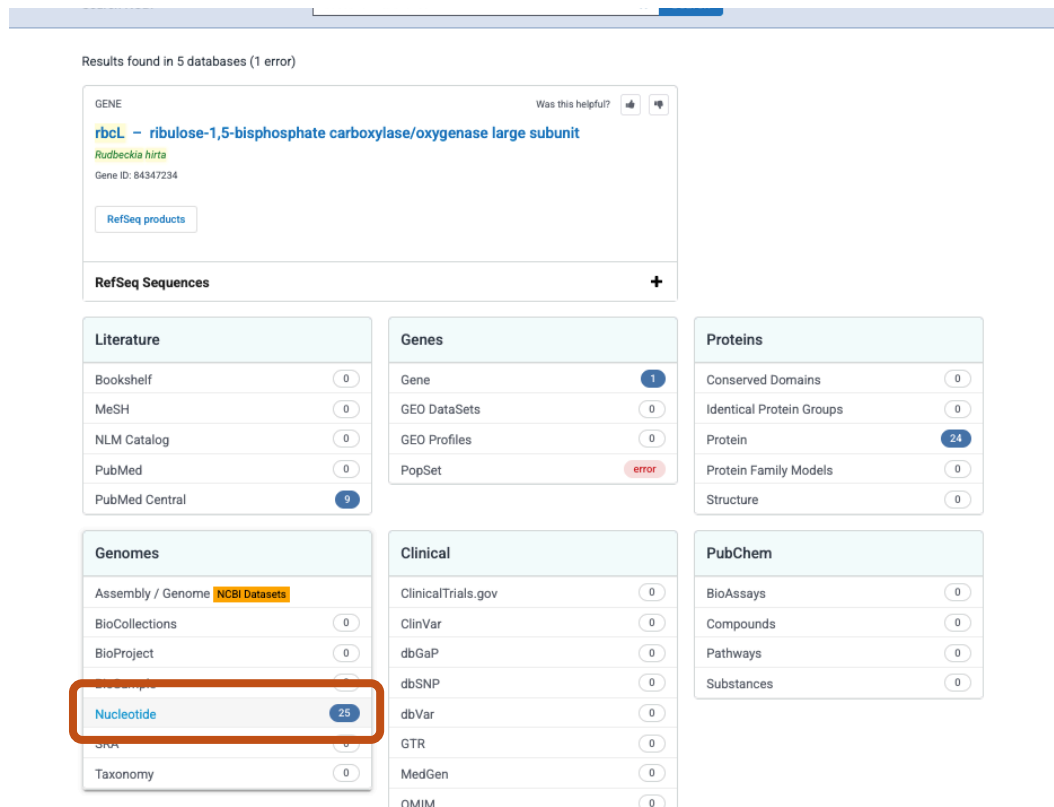


Fig. 2

National Library of Medicine
National Center for Biotechnology Information

Log in

Nucleotide

Nucleotide

rudbeckia hirta and rbcl

Search

Create alert Advanced

Help

Species
Plants (25)
Customize ...

Molecule types
genomic DNA/RNA (25)
Customize ...

Source databases
INSDC (GenBank) (24)
RefSeq (1)
Customize ...

Sequence Type
Nucleotide (25)

Genetic compartments
Chloroplast (24)
Plastid (25)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

[Clear all](#)
[Show additional filters](#)

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾ [Filters: Manage Filters](#)

GENE

Was this helpful?

[rbcl - ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit](#)
[Rudbeckia hirta](#)
Gene ID: 84347234

RefSeq products

RefSeq Sequences

Items: 1 to 20 of 25

☐
[Rudbeckia hirta ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit \(rbcl\) gene, partial cds; chloroplast gene for chloroplast product](#)
1,409 bp linear DNA
Accession: AY215173.1 GI: 34765448
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

☐
[Rudbeckia hirta voucher AP153 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit \(rbcl\) gene, partial cds; chloroplast](#)
607 bp linear DNA
Accession: HQ590248.1 GI: 313684438
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

☐
[Rudbeckia hirta voucher Abbott 25339 \(FLAS\) ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit \(rbcl\) gene, partial cds; chloroplast](#)
517 bp linear DNA
Accession: KC397942.1 GI: 1162229192
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

Results by taxon

Top Organisms [Tree](#)
Rudbeckia hirta (25)

Find related data

Database:

Select ▾

Find items

Search details

{"Rudbeckia hirta"[Organism] OR
rudbeckia hirta[All Fields]) AND
rbcl[All Fields]

Search

See more...

Recent activity

Turn Off

Clear

Q

rudbeckia hirta and rbcl (25)

Nucleotide

Rudbeckia hirta ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit

Nucleotide

Rudbeckia hirta chloroplast, complete genome

Nucleotide

rbcl [Rudbeckia hirta]

Gene

Rudbeckia hirta voucher MT00186528 maturase K (matK) gene, partial cds

Nucleotide

See more...

Fig. 3

GenBank Send to: ▼

Rudbeckia hirta ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast gene for chloroplast product

GenBank: AY215173.1

FASTA [Graphics](#)

[Go to:](#) ☐

LOCUS AY215173 1409 bp DNA linear PLN 31-MAY-2013

DEFINITION Rudbeckia hirta ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast gene for chloroplast product.

ACCESSION AY215173

VERSION AY215173.1

KEYWORDS .

SOURCE chloroplast Rudbeckia hirta

ORGANISM [Rudbeckia hirta](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetales; asterids; campanulids; Asterales; Asteraceae; Asteroideae; Heliantheae alliance; Heliantheae; Rudbeckia.

REFERENCE 1 (bases 1 to 1409)
AUTHORS Panero, J.L., Baldwin, B.G., Schilling, E.E. and Clevinger, J.A.
TITLE A Chloroplast Phylogeny of Tribe Heliantheae (Asteraceae)
JOURNAL Unpublished

REFERENCE 2 (bases 1 to 1409)
AUTHORS Panero, J.L., Baldwin, B.G., Schilling, E.E. and Clevinger, J.A.
TITLE Direct Submission
JOURNAL Submitted (07-JAN-2003) Section of Integrative Biology, University of Texas, 141 Patterson Bldg., 24th St and Speedway, Austin, TX 78712, USA

FEATURES

source	Location/Qualifiers
	1..1409
	/organism="Rudbeckia hirta"
	/organelle="plastid:chloroplast"
	/mol_type="genomic DNA"
	/specimen_voucher="Panero 2002-14 (TEX)"
	/db_xref="taxon:52299"
	/note="authority: Rudbeckia hirta L."
gene	<1..1409
	/gene="rbcL"
CDS	<1..1409
	/gene="rbcL"
	/codon_start=3
	/transl_table=11
	/product="ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit"
	/protein_id="AAR11731.1"
	/translation="KDYKLTYYTPEYETKDTDILAAFRVTPQPGVPPEEAGAAVAES STGTWTTWTDGLTSLDRYKGRCYGIEPVPGEDNQFIAYVAYPLDLFEEGSVTNMFTS IVGNVFGKALRALRLDLRIPTAYVKTFDGPPIQVERDKLNKYGRPLLGCITKPK LGSLAKNYGRACYECLRGGLDFTKDDENVNSQPFMRWRDRFLFCAEAIYKSAQETGEI KGHYLNATAGTCEDMMKRAAFARELGVPIVMHDYLTGGFTANTSLSHYCRDNGLLHI HRAMHAVIDRQKNHGMHFRVLAKALRMSGGDHHSHTVVGKLEGEREITLGFVDLLRD DFIEKDRSRGIYFTQDWVSLPGVLPVASGGIHWHPALTEIFGDDSVLQFGGGTLGH PWGNAPGAVANRVALEACVQARNEGRDLATEGNEIIREATKWSPELAAACEVWKEIKF EFQAMDLTLDKDKDKKR"

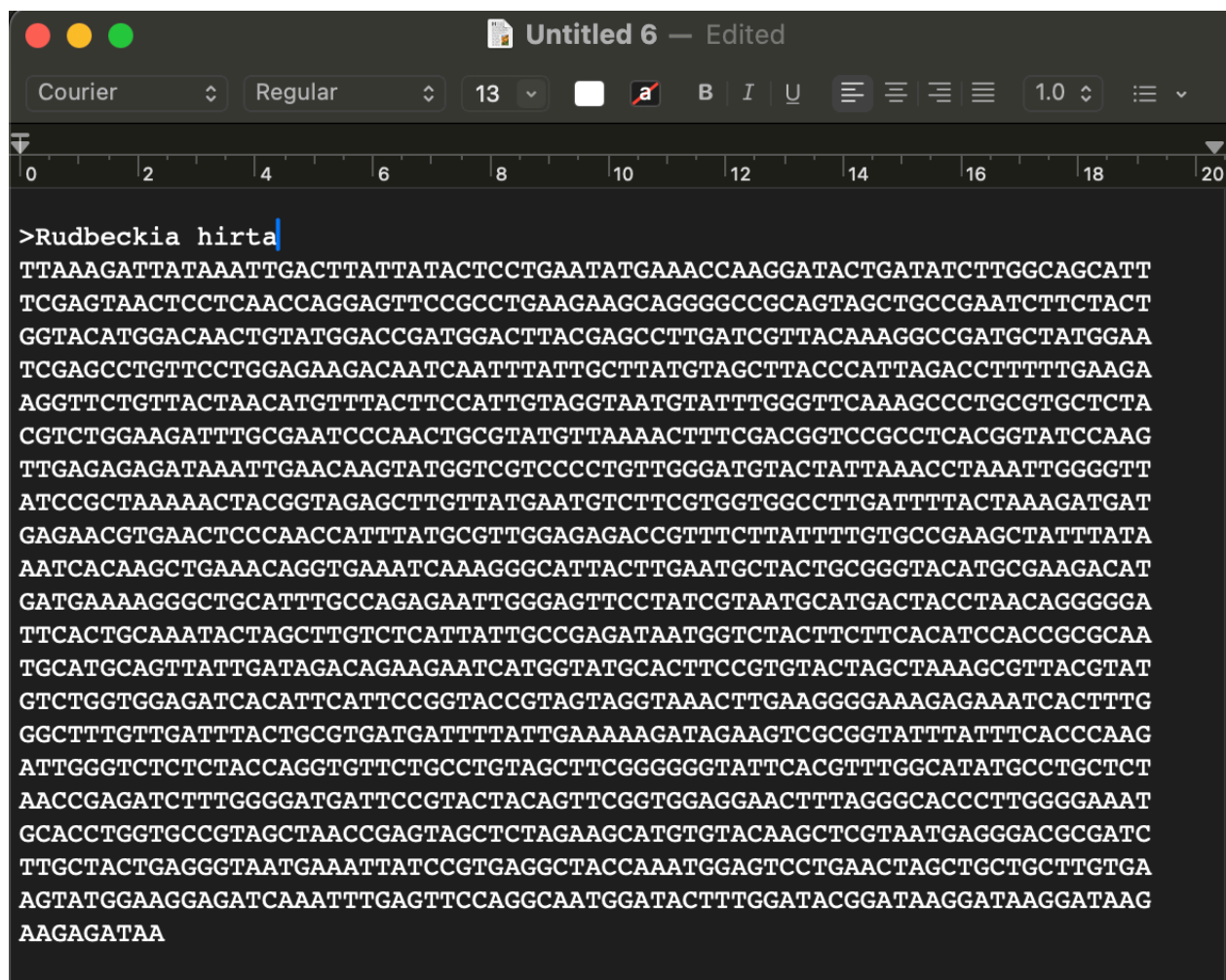
ORIGIN

```

1 ttaaagatta taaattgact tattatactc ctgaatatga aaccaaggat actgatattc
61 tggcagcatt tcgagtaact cctcaaccag gagttccgcc tgaagaagca ggggccgcag
121 tagctgccga atcttctact ggtacatgga caactgtatg gaccgatgga cttacgagcc
181 ttgatcggtt caaaggccga tgctatggaa tcgagcctgt tcctggagaa gacaatcaat
241 ttattgccta tgtagcttac ccattagacc tttttgaaga aggttctgtt actaacatgt
301 ttacttccat tqtaqqtaat qtatttqqqt tcaaaqccct qcqtqctcta cqtctqaaq

```

Fig. 4



The image shows a screenshot of a text editor window titled "Untitled 6 - Edited". The editor has a dark background and a light-colored text. The text is a DNA sequence for *Rudbeckia hirta*. The sequence is displayed in a monospaced font, with line numbers 0 through 20 visible on the left side. The sequence is as follows:

```
>Rudbeckia hirta
TTAAAGATTATAAATTGACTTATTATACTCCTGAATATGAAACCAAGGATACTGATATCTTGGCAGCATT
TCGAGTAACTCCTCAACCAGGAGTTCCGCCTGAAGAAGCAGGGGCCGAGTAGCTGCCGAATCTTCTACT
GGTACATGGACAACCTGTATGGACCGATGGACTTACGAGCCTTGATCGTTACAAAGGCCGATGCTATGGAA
TCGAGCCTGTTCTCGGAGAAGACAATCAATTTATTGCTTATGTAGCTTACCCATTAGACCTTTTTGAAGA
AGGTTCTGTTACTAACATGTTTACTTCCATTGTAGGTAATGTATTTGGGTTCAAAGCCCTGCGTGCTCTA
CGTCTGGAAGATTTGCGAATCCCAACTGCGTATGTTAAACTTTTCGACGGTCCGCCTCACGGTATCCAAG
TTGAGAGAGATAAATTGAACAAGTATGGTTCGTCCTGTTGGGATGTACTATTAAACCTAAATTGGGGTT
ATCCGCTAAAACTACGGTAGAGCTTGTTATGAATGTCTTCGTGGTGGCCTTGATTTTACTAAAGATGAT
GAGAACGTGAACTCCCAACCATTTATGCGTTGGAGAGACCGTTTCTTATTTTGTGCCGAAGCTATTTATA
AATCACAAGCTGAAACAGGTGAAATCAAAGGGCATTACTTGAATGCTACTGCGGGTACATGCGAAGACAT
GATGAAAAGGGCTGCATTTGCCAGAGAATTGGGAGTTCCTATCGTAATGCATGACTACCTAACAGGGGGA
TTCCTGCAAATACTAGCTTGTCTCATTATTGCCGAGATAATGGTCTACTTCTTCACATCCACCGCGCAA
TGCATGCAGTTATTGATAGACAGAAGAATCATGGTATGCACTTCCGTGTACTAGCTAAAGCGTTACGTAT
GTCTGGTGGAGATCACATTCATTCCGGTACCGTAGTAGGTAAACTTGAAGGGGAAAGAGAAATCACTTTC
GGCTTTGTTGATTTACTGCGTGATGATTTTATTGAAAAAGATAGAAGTCGCGGTATTTATTTACCCAAG
ATTGGGTCTCTCTACCAGGTGTTCTGCCTGTAGCTTCGGGGGGTATTCACGTTTGGCATATGCCTGCTCT
AACCGAGATCTTTGGGGATGATTCCGTACTACAGTTCGGTGGAGGAACTTTAGGGCACCTTGGGGAAAT
GCACCTGGTGCCGTAGCTAACCAGTAGCTCTAGAAGCATGTGTACAAGCTCGTAATGAGGGACGCGATC
TTGCTACTGAGGGTAATGAAATTATCCGTGAGGCTACCAAATGGAGTCCGTAAGCTAGCTGCTGCTTGTGA
AGTATGGAAGGAGATCAAATTTGAGTTCAGGCAATGGATACTTTGGATACGGATAAGGATAAGGATAAG
AAGAGATAA
```

Fig. 5

Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like HKY but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like TN but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like K81 but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like TPM2 but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like TPM3 but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIME	3	Like TIM but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like TIM2 but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like TIM3 but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like TVM but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

Fig. 6

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 50%

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

Tree Inference
Model Selection
Analysis Results

User name or Email:

QUERY STATUS

☒ No. Submission Time Status

☒ 1 2025-03-05 16:56 Waiting

Summary

Run Log

Full Result

IQ-TREE multicore version 1.6.12 for Linux 64-bit built Aug 15 2019
 Developed by Bui Quang Minh, Nguyen Lam Tung, Olga Chernomor,
 Heiko Schmidt, Dominik Schrempf, Michael Woodhams.

Host: cox (AVX, FMA3, 997 GB RAM)
 Command: ../../iqtree -s rbcl_aligned.aln-fasta -m TESTONLY
 Seed: 389497 (Using SPRNG - Scalable Parallel Random Number Generator)
 Time: Wed Mar 5 16:56:16 2025
 Kernel: AVX+FMA - 1 threads (48 CPU cores detected)

HINT: Use -nt option to specify number of threads because your CPU has 48 cores!
 HINT: -nt AUTO will automatically determine the best number of threads to use.

Reading alignment file rbcl_aligned.aln-fasta ... Fasta format detected
 Alignment most likely contains DNA/RNA sequences
 WARNING: 2 sites contain only gaps or ambiguous characters.
 Alignment has 7 sequences with 1440 columns, 90 distinct patterns
 31 parsimony-informative, 64 singleton sites, 1345 constant sites

	Gap/Ambiguity	Composition	p-value
1	BRIN	63.96%	passed 94.13%
2	MOFI	61.94%	passed 91.18%
3	LEVU	57.85%	passed 90.97%
4	HEDI	62.01%	passed 99.40%
5	SOCA	64.65%	passed 96.36%
6	RUHI	2.15%	passed 69.94%
7	RULA	61.60%	passed 97.63%

WARNING: 6 sequences contain more than 50% gaps/ambiguity
 **** TOTAL 53.45% 0 sequences failed composition chi2 test (p-value<5%; df=3)

Create initial parsimony tree by phylogenetic likelihood library (PLL)... 0.000 seconds
 NOTE: ModelFinder requires 0 MB RAM!
 ModelFinder will test 88 DNA models (sample size: 1440) ...

No. Model	-LnL	df	AIC	AICc	BIC	
1	JC	2609.296	11	5240.592	5240.776	5298.588
2	JC+I	2596.950	12	5217.900	5218.119	5281.169
3	JC+G4	2596.573	12	5217.146	5217.365	5280.415
4	JC+I+G4	2596.298	13	5218.596	5218.851	5287.137
5	F81+F	2595.749	14	5219.497	5219.792	5293.311
6	F81+F+I	2583.195	15	5196.391	5196.728	5275.477
7	F81+F+G4	2582.809	15	5195.618	5195.955	5274.704
8	F81+F+I+G4	2582.504	16	5197.008	5197.390	5281.366

Fig. 7

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 50%

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: [10.1093/nar/gkv](https://doi.org/10.1093/nar/gkv)

Tree Inference | **Model Selection** | **Analysis Results**

For a quick start, take a look at the [tutorial](#) for the IQ-TREE web server.

Please visit the [IQ-TREE homepage](#) for more information or if you want to download the main software.

Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file :

Browse...

Show example >

Use example alignment: ☐ Yes

?

Sequence type:

☒ Auto-detect ☐ DNA ☐ Protein ☐ Codon
☐ DNA->AA ☐ Binary ☐ Morphology

?

Partition file:

This field is optional.

Browse...

Show example >

Partition type:

☒ Edge-linked
☐ Edge-unlinked

?

Substitution Model Options

Substitution model:

Auto

?

FreeRate heterogeneity: ☐ Yes [+R]

Rate heterogeneity:

☐ Gamma [+G] ☐ Invar. sites [+I]

?

#rate categories:

4

State frequency:

☒ Empirical (from data) ☐ AA model (from matrix) ☐ ML-optimized
☐ Codon F1x4 ☐ Codon F3x4

Ascertainment bias correction:

☐ Yes [+ASC]

?

Branch Support Analysis

Bootstrap analysis:

☐ None ☒ Ultrafast ☐ Standard

?

Fig. 8

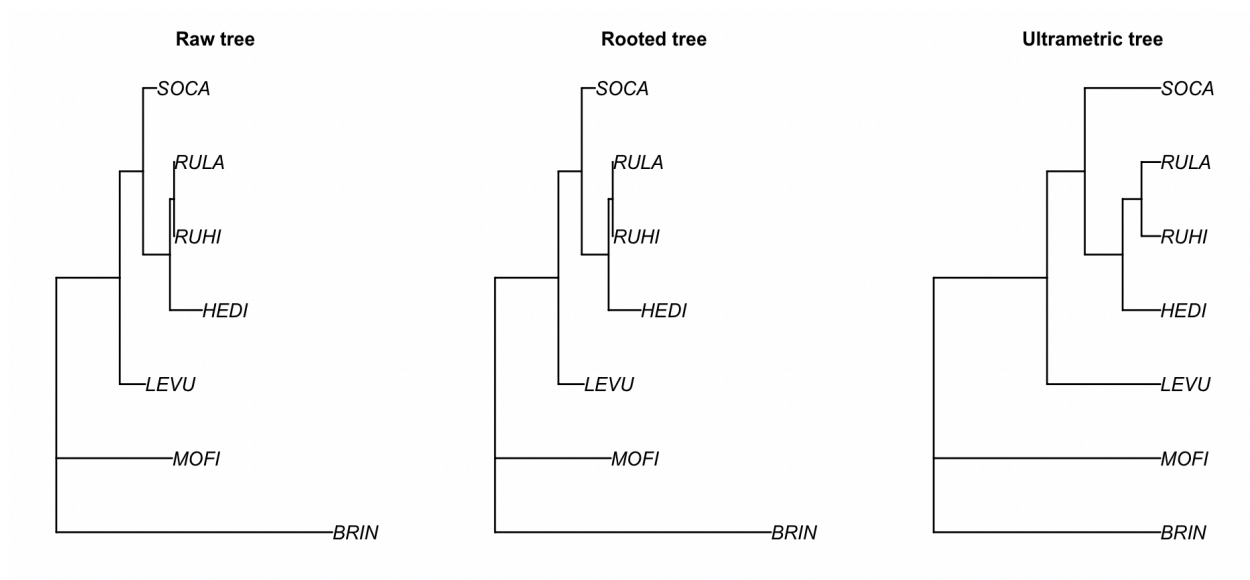


Fig. 9 *rbcl* tree

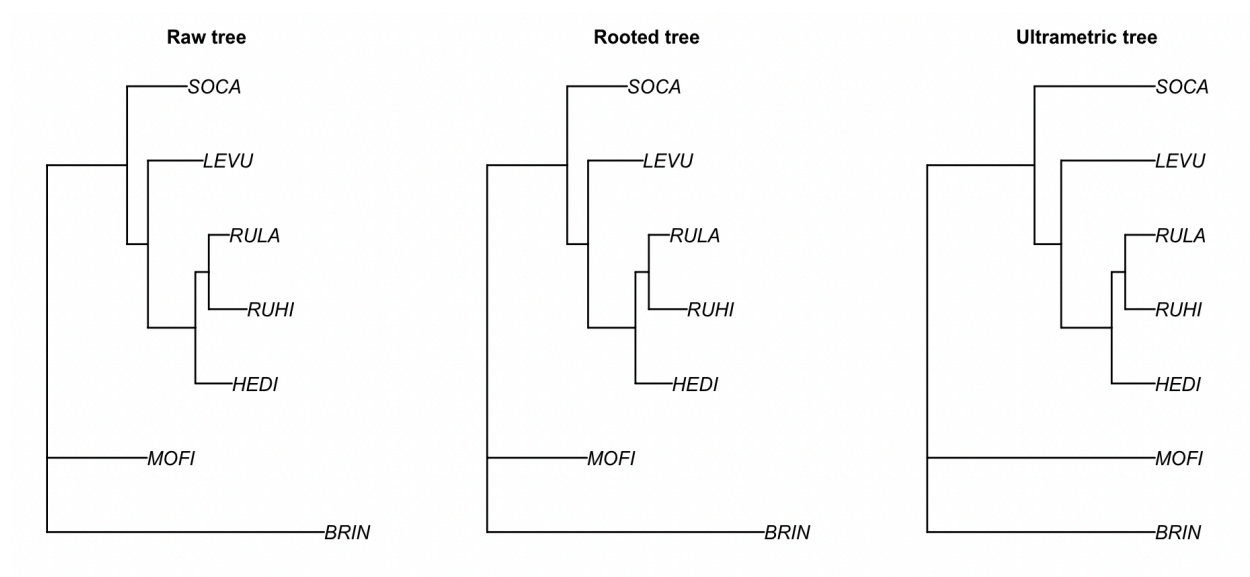


Fig. 10 *its1* tree

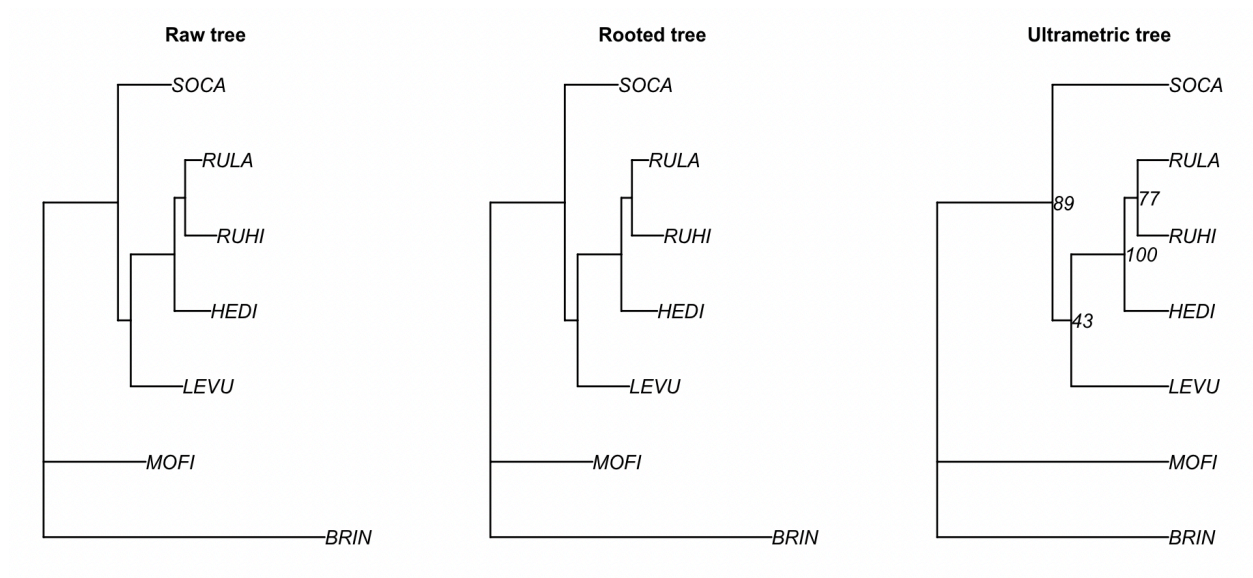


Fig. 11 full tree

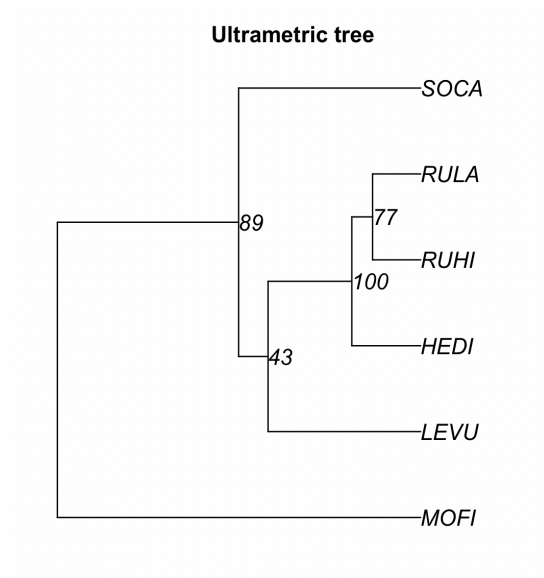


Fig. 12: outgroup pruned