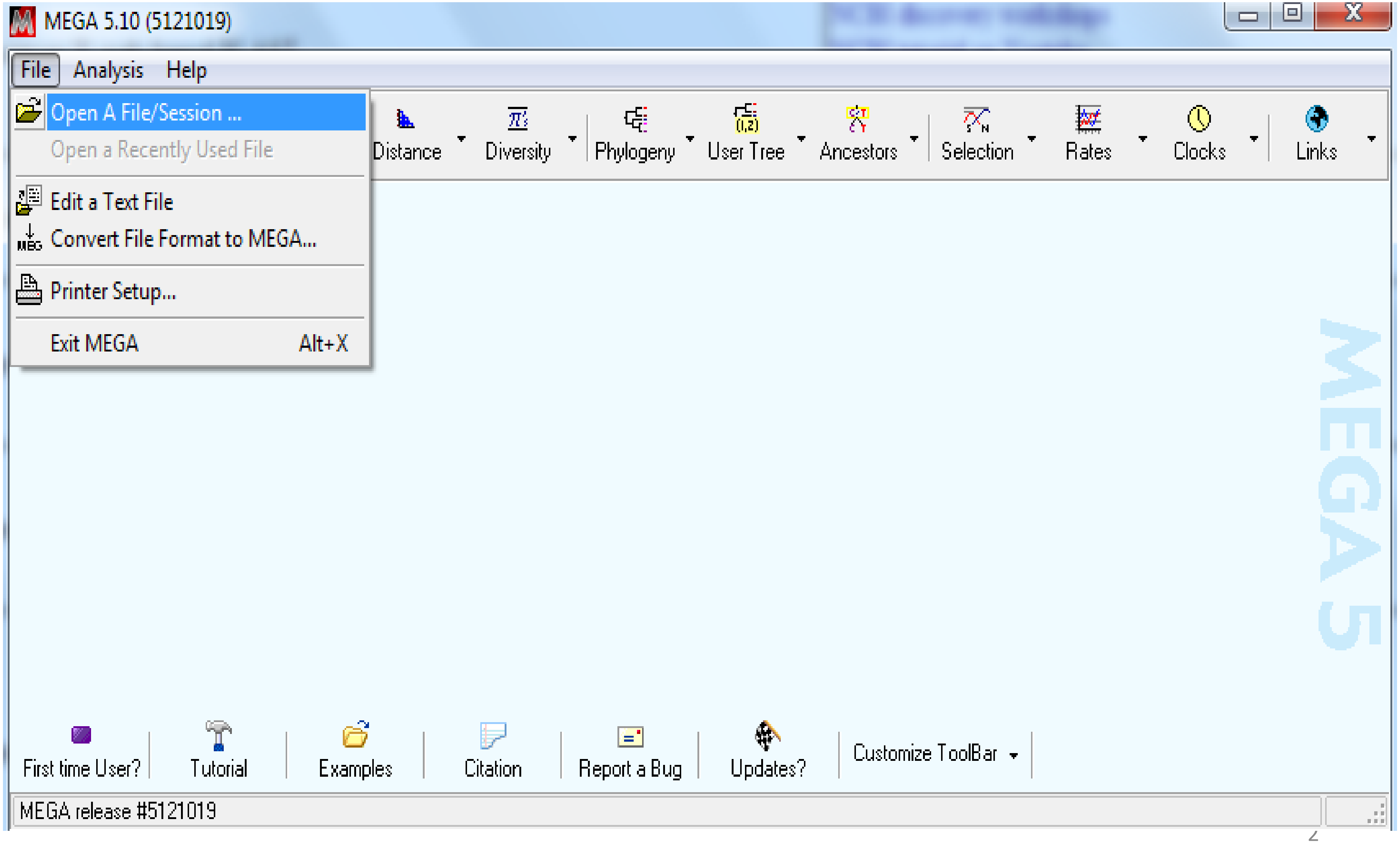
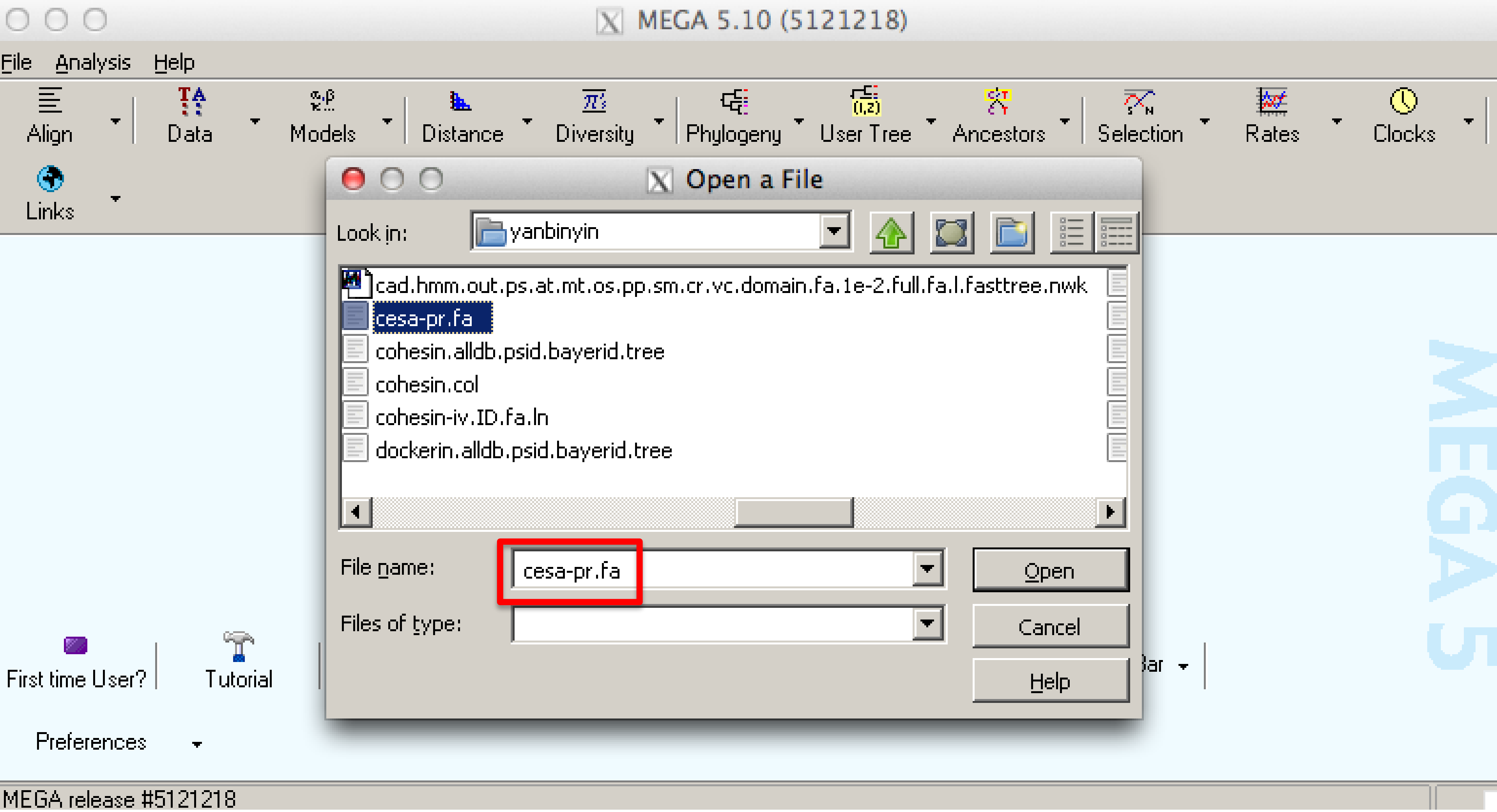
Phylogeny and visualization: MEGA and iTOL

Yanbin Yin

Spring 2013

If we want to start from unaligned sequences:





Look in: yanbinyin

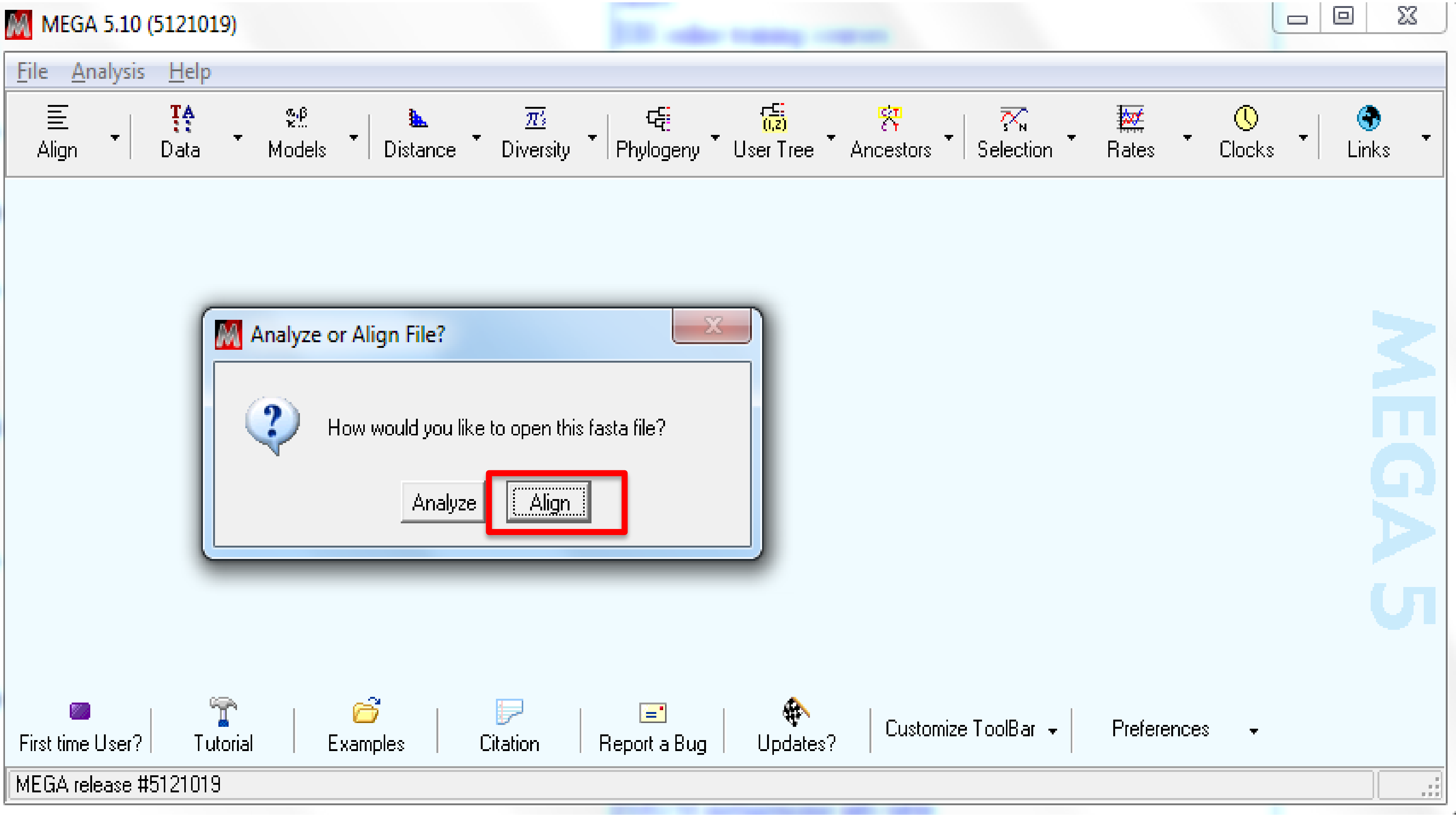
cad.hmm.out.ps.at.mt.os.pp.sm.cr.vc.domain.fa.1e-2.full.fa.l.fasttree.nwk
cesa-pr.fa
cohesin.alldb.psid.bayerid.tree
cohesin.col
cohesin-iv.ID.fa.ln
dockerin.alldb.psid.bayerid.tree

File name: cesa-pr.fa

Files of type:

MEGA release #5121218

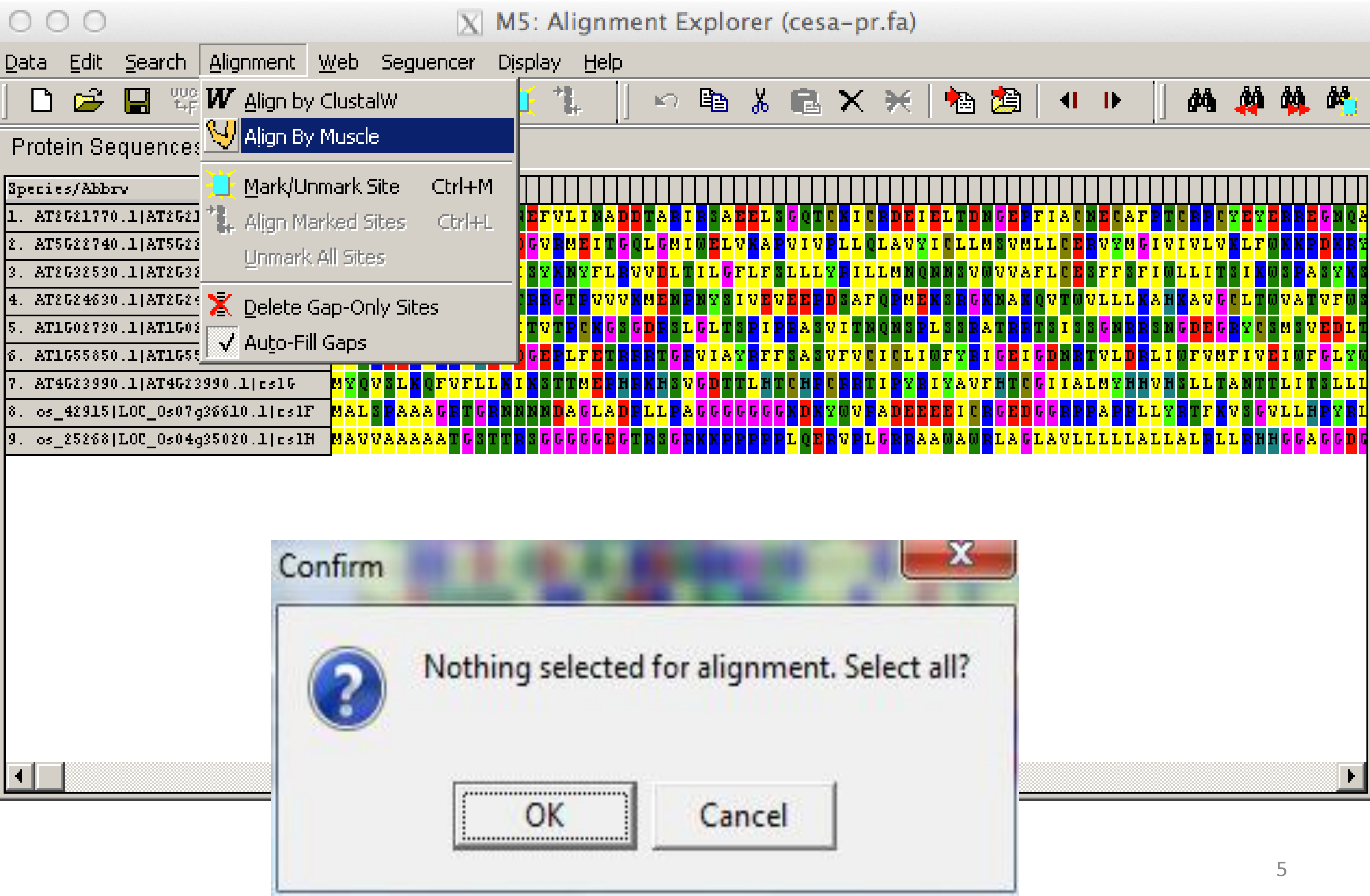
http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa



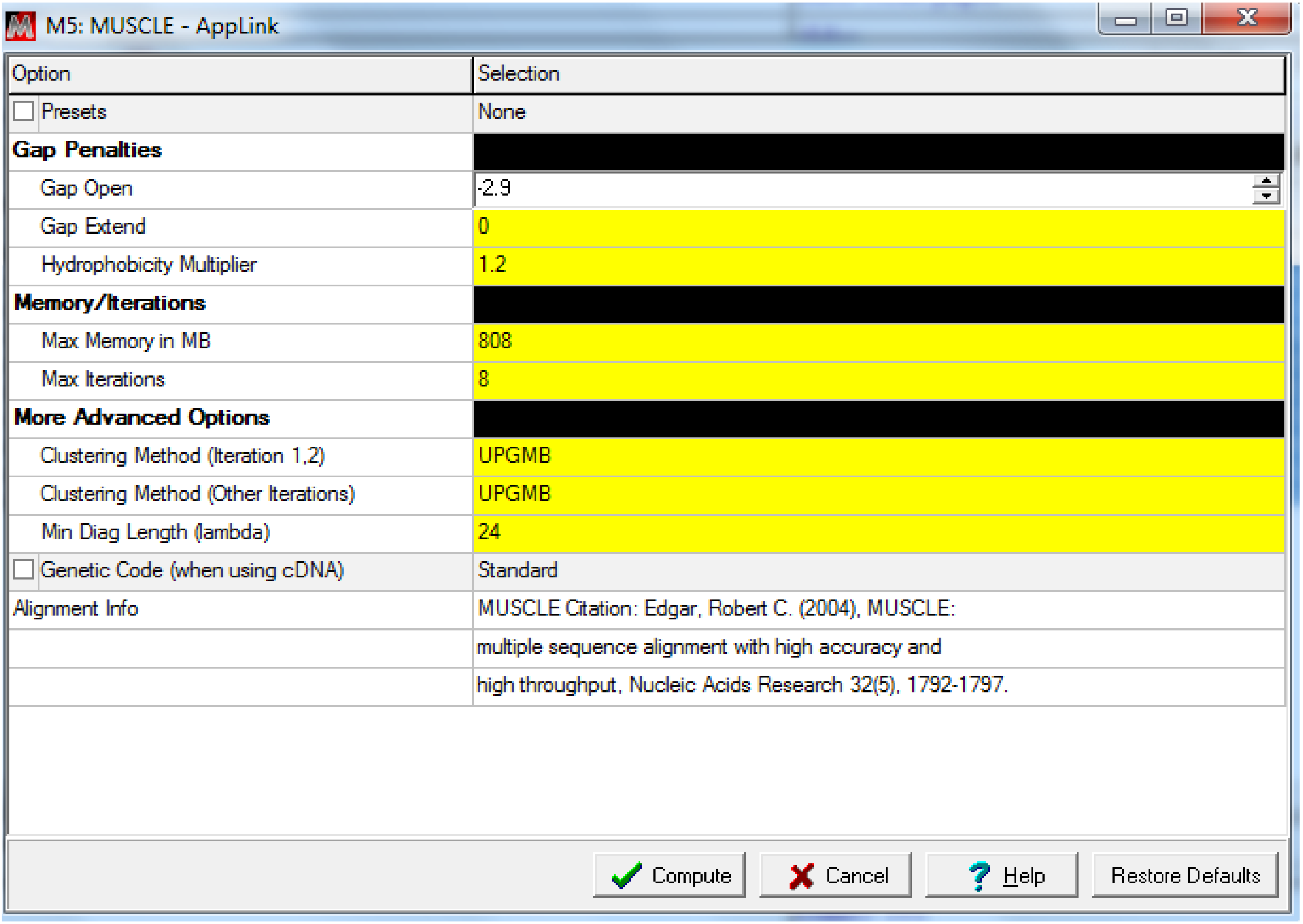
How would you like to open this fasta file?

MEGA release #5121019

Alignment editor

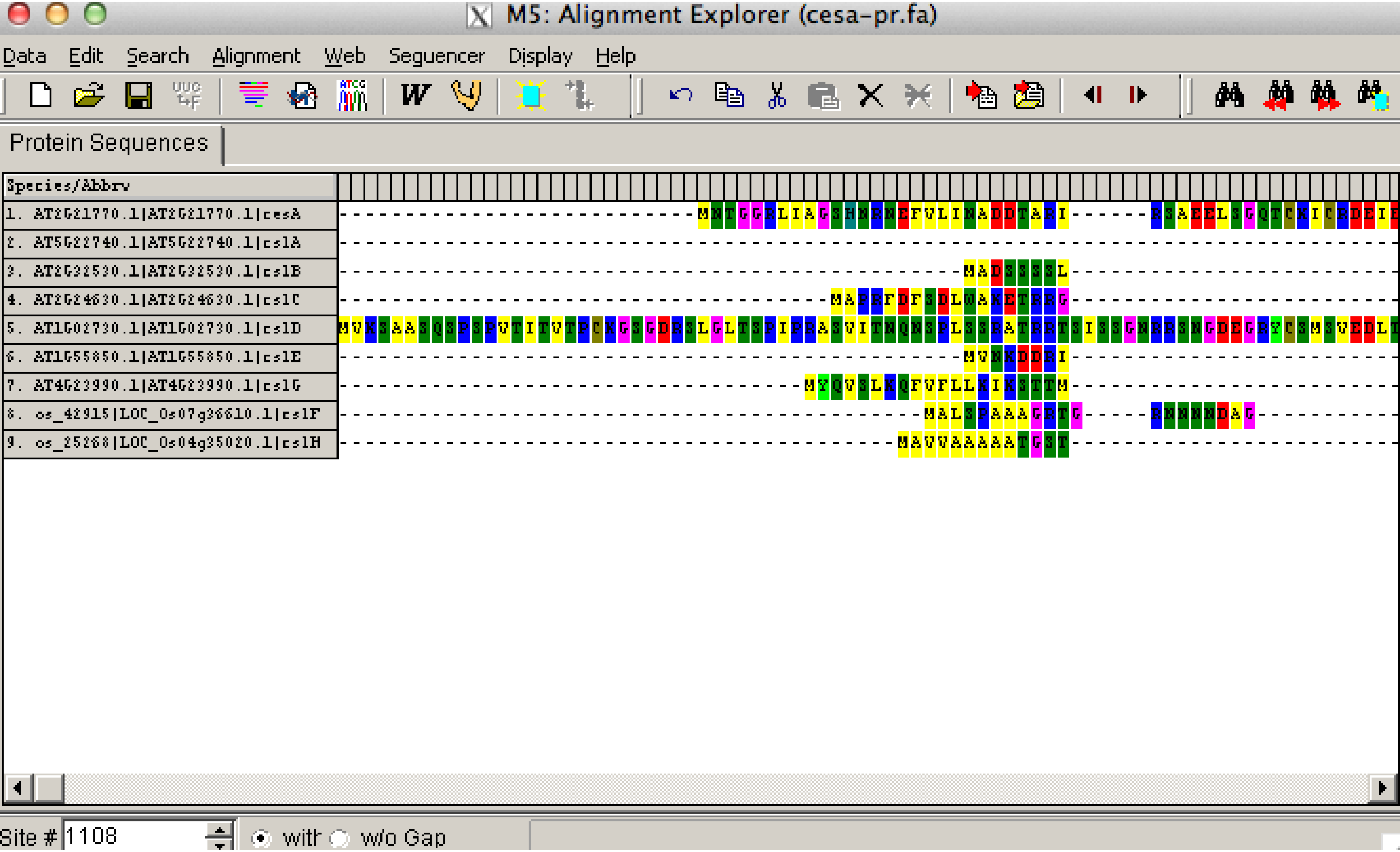


Nothing selected for alignment. Select all?



M5: MUSCLE - AppLink

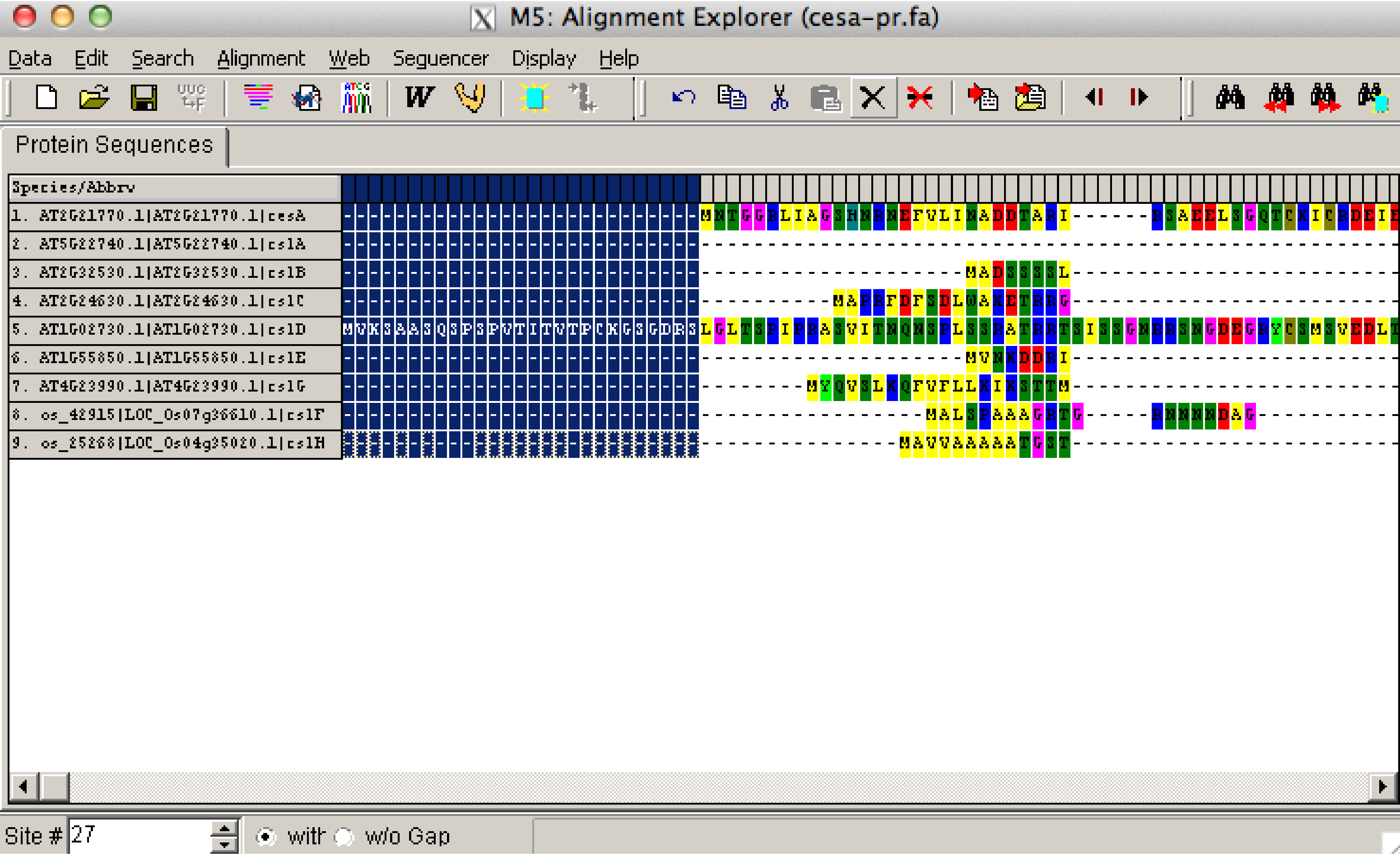
Option	Selection
☐ Presets	None
Gap Penalties	
Gap Open	-2.9
Gap Extend	0
Hydrophobicity Multiplier	1.2
Memory/Iterations	
Max Memory in MB	808
Max Iterations	8
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other Iterations)	UPGMB
Min Diag Length (lambda)	24
☐ Genetic Code (when using cDNA)	Standard
Alignment Info	MUSCLE Citation: Edgar, Robert C. (2004), MUSCLE:
	multiple sequence alignment with high accuracy and
	high throughput, Nucleic Acids Research 32(5), 1792-1797.



```
Species/Abbrv
1. AT2G21770.1|AT2G21770.1|cesA       ---------------------------MNTGGRLIAGSHNRNEFVLINADDTARI------RSAEELSGQTCKICRDEIE
2. AT5G22740.1|AT5G22740.1|cslA       --------------------------------------------------------------------------------
3. AT2G32530.1|AT2G32530.1|cslB       ------------------------------------------------MADSSSL-------------------------
4. AT2G24630.1|AT2G24630.1|cslC       -------------------------------------MAPRFDFSDLWAKETRRG-------------------------
5. AT1G02730.1|AT1G02730.1|cslD       MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSGNRRSNGDEGRYCSMSVEDLT
6. AT1G55850.1|AT1G55850.1|cslE       -----------------------------------------------MVNKDDRI-------------------------
7. AT4G23990.1|AT4G23990.1|cslG       -----------------------------------MYQVSLKQFVFLLKIKSTTM-------------------------
8. os_42915|LOC_Os07g36610.1|cslF     --------------------------------------------MALSPAAAGRTG-----RNNNNDAG-----------
9. os_25268|LOC_Os04g35020.1|cslH     ------------------------------------------MAVVAAAAATGST-------------------------
```

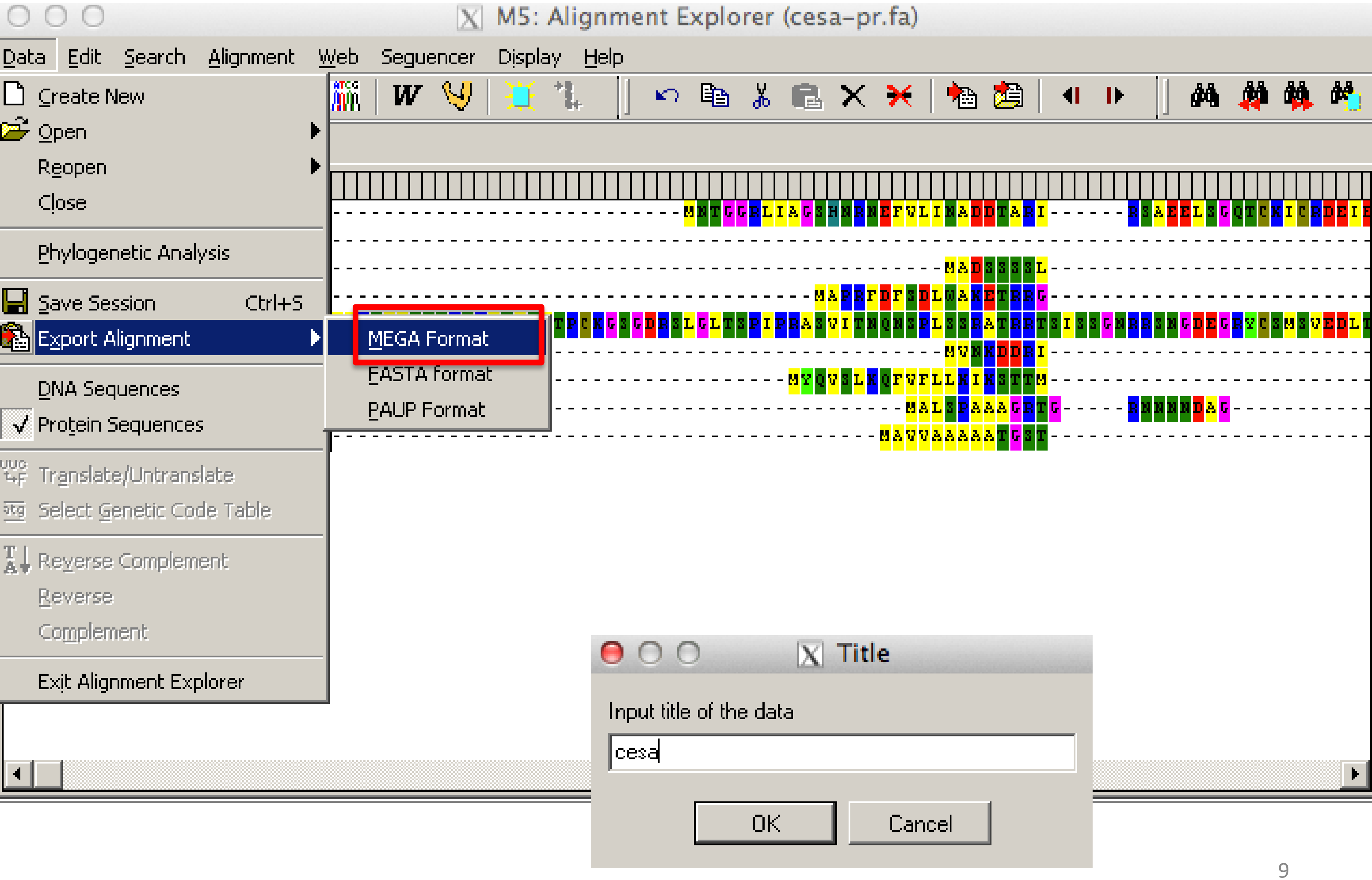
Site # 1108 ● with ○ w/o Gap

Alignment can be edited, e.g. delete long gaps



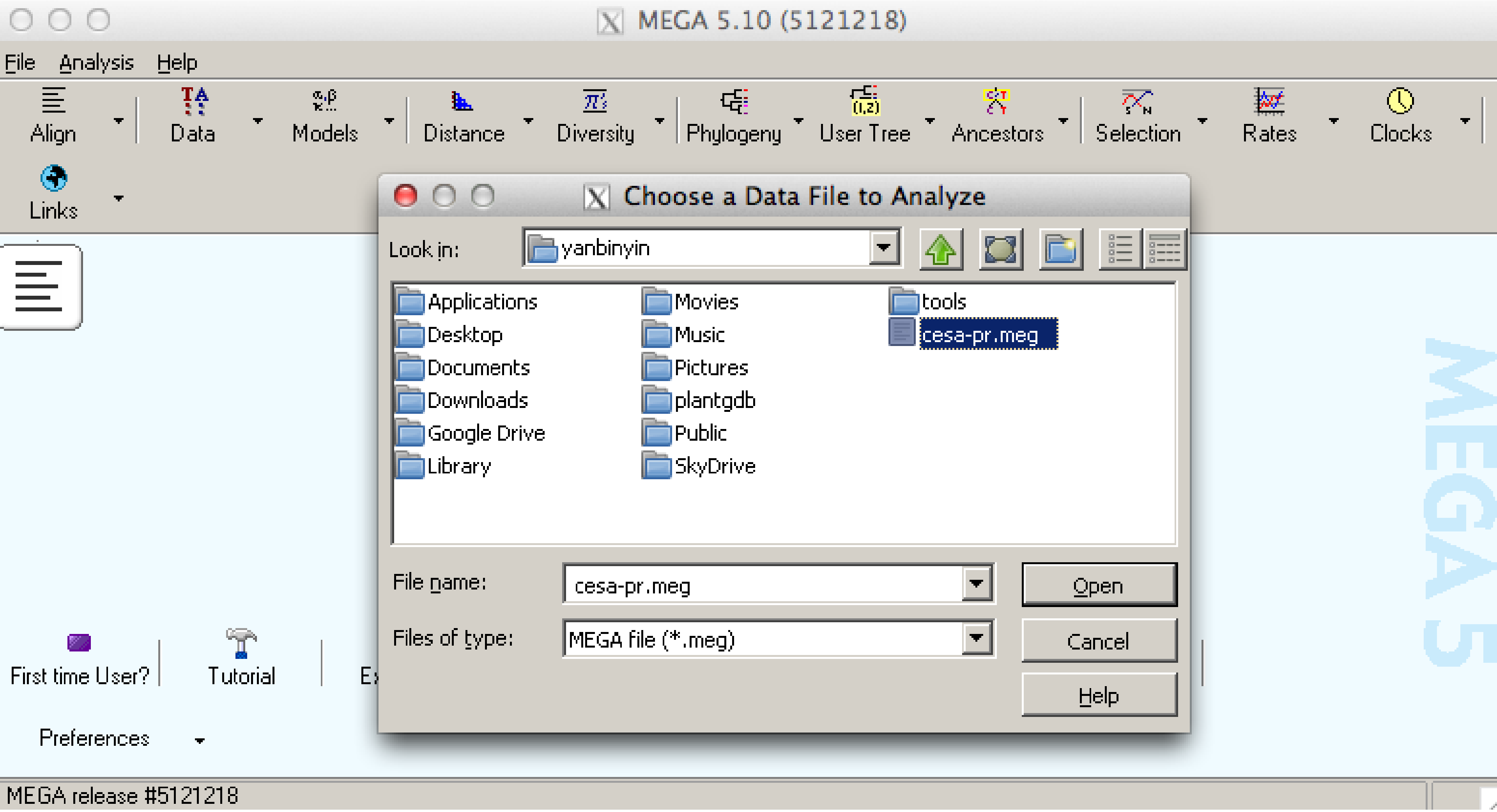
```
Species/Abbrv
1. AT2G21770.1|AT2G21770.1|cesA    ---------------------------MNTGGRLIAGSHNRNEFVLINADDTARI------RSAEELSGQTCKICRDEIE
2. AT5G22740.1|AT5G22740.1|cslA    --------------------------------------------------------------------------------
3. AT2G32530.1|AT2G32530.1|cslB    -----------------------------------------------MADSSSSL-------------------------
4. AT2G24630.1|AT2G24630.1|cslC    -------------------------------------MAPRFDFSDLWAKETRRG-------------------------
5. AT1G02730.1|AT1G02730.1|cslD    MVKSAASQSPSPVTITVTPCKGSGDRSLGLTSPIPRASVITNQNSPLSSRATRRTSISSGNRRSNGDEGRYCSMSVEDLT
6. AT1G55850.1|AT1G55850.1|cslE    -----------------------------------------------MVNKDDRI-------------------------
7. AT4G23990.1|AT4G23990.1|cslG    -----------------------------------MYQVSLKQFVFLLKIKSTTM-------------------------
8. os_42915|LOC_Os07g36610.1|cslF  --------------------------------------------MALSPAAAGRTG-----RNNNNDAG-----------
9. os_25268|LOC_Os04g35020.1|cslH  ------------------------------------------MAVVAAAAATGST-------------------------
```

Site # 27 ● with ○ w/o Gap



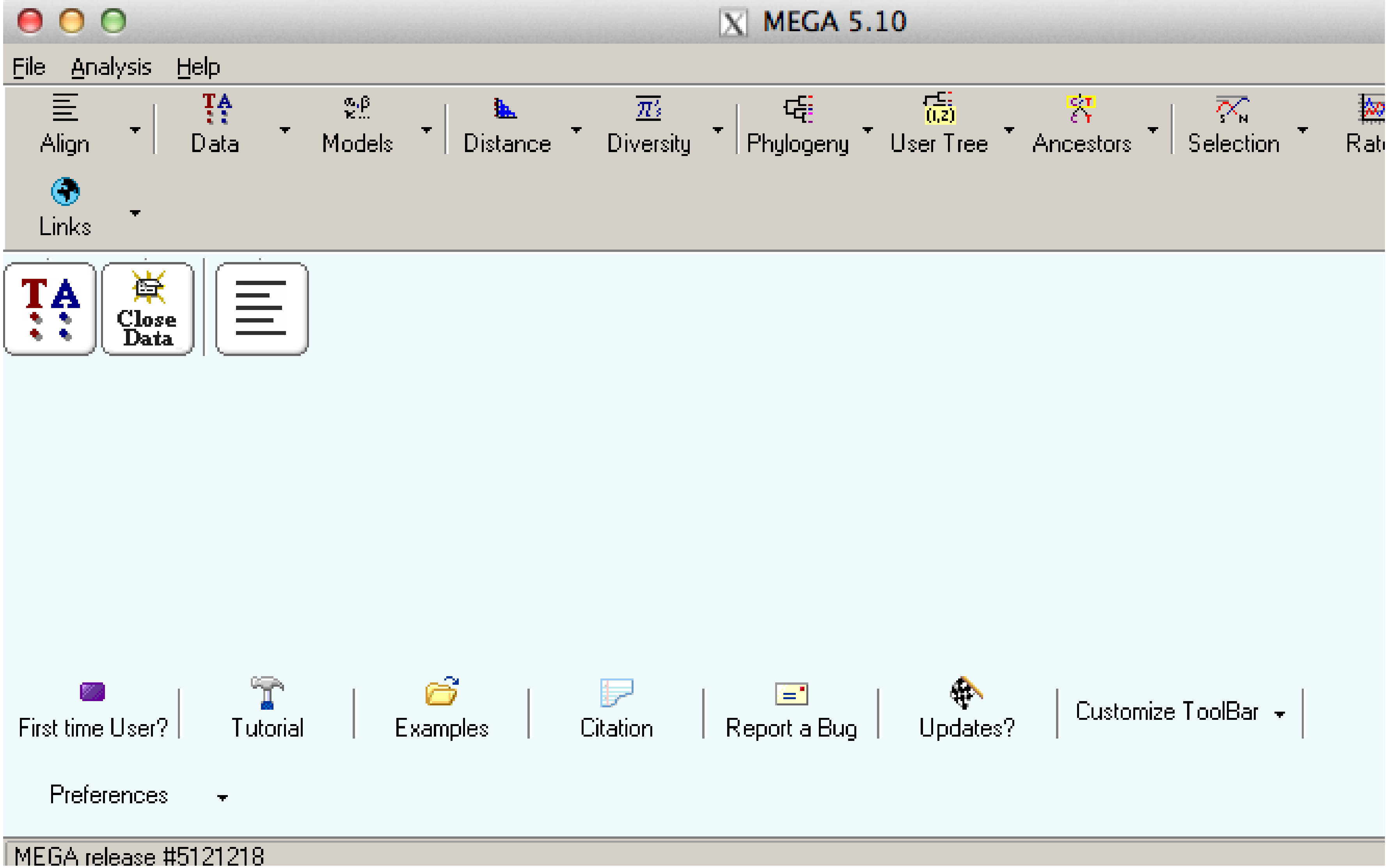
Input title of the data: cesa

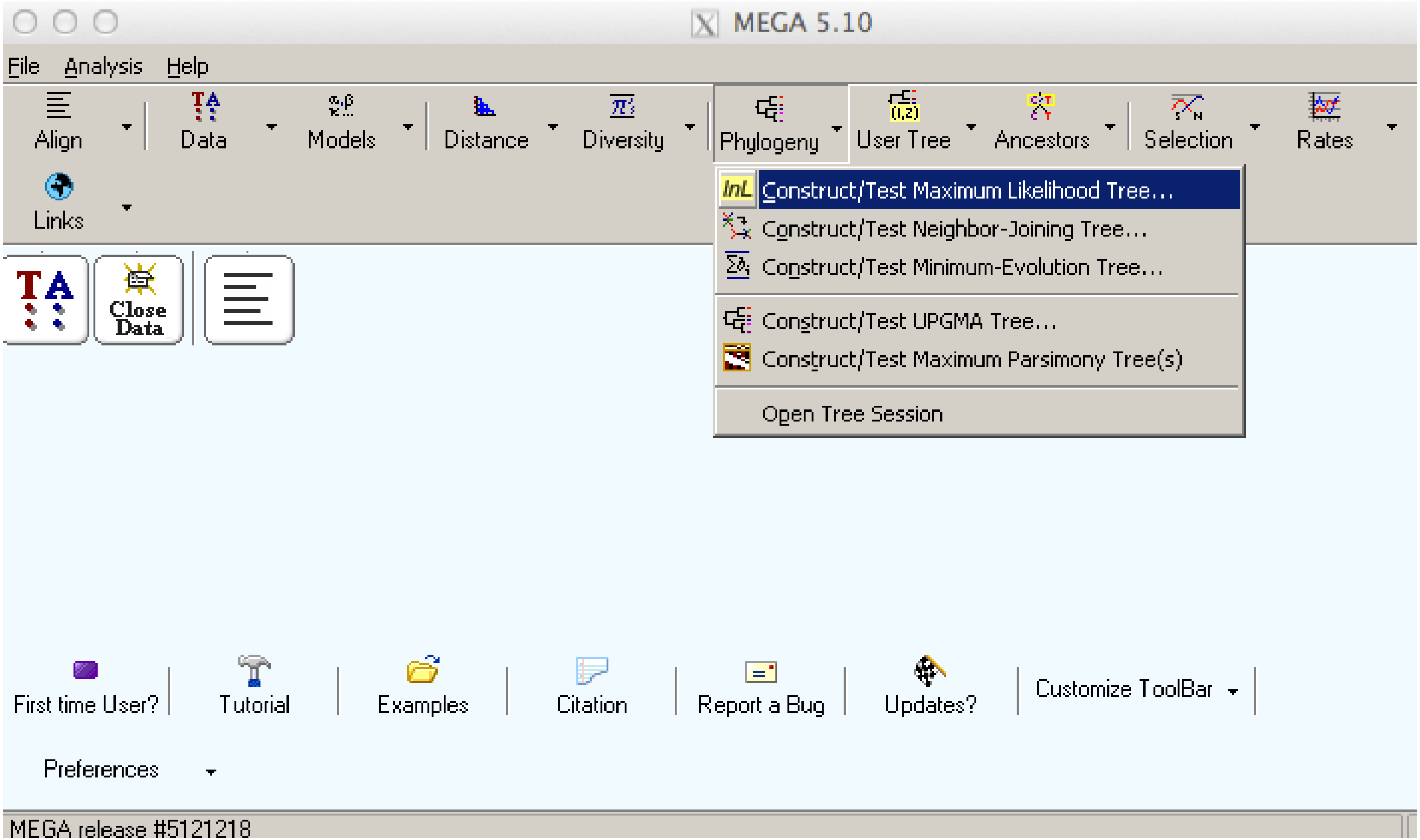
Open the meg format alignment for phylogeny building



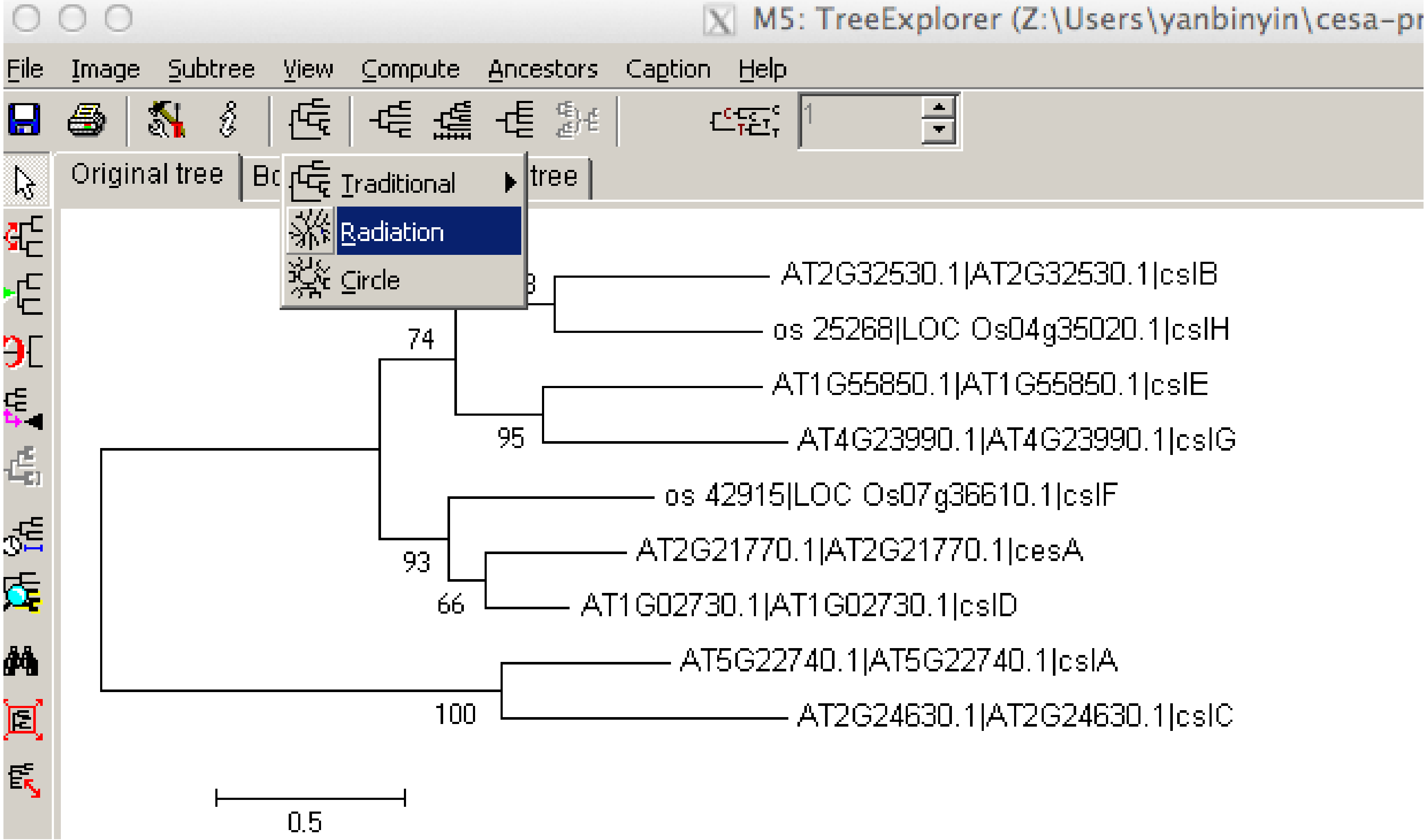
File name: cesa-pr.meg

Files of type: MEGA file (*.meg)

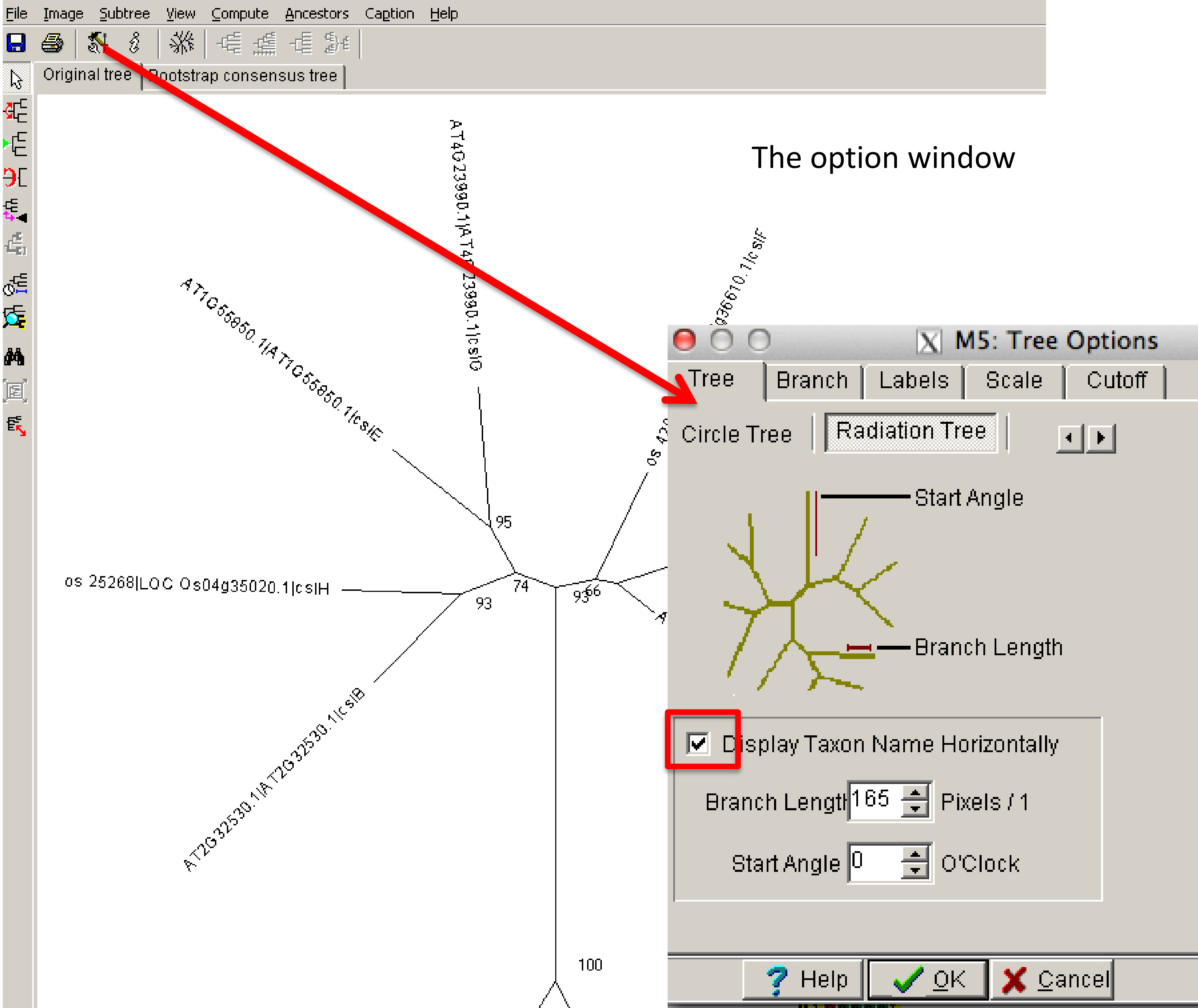

MEGA 5.10
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rat
Links
Close Data
First time User?
Tutorial
Examples
Citation
Report a Bug
Updates?
Customize ToolBar
Preferences
MEGA release #5121218


MEGA 5.10
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Links
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
Close Data
First time User?
Tutorial
Examples
Citation
Report a Bug
Updates?
Customize ToolBar
Preferences
MEGA release #5121218

Different presentation views of phylograms



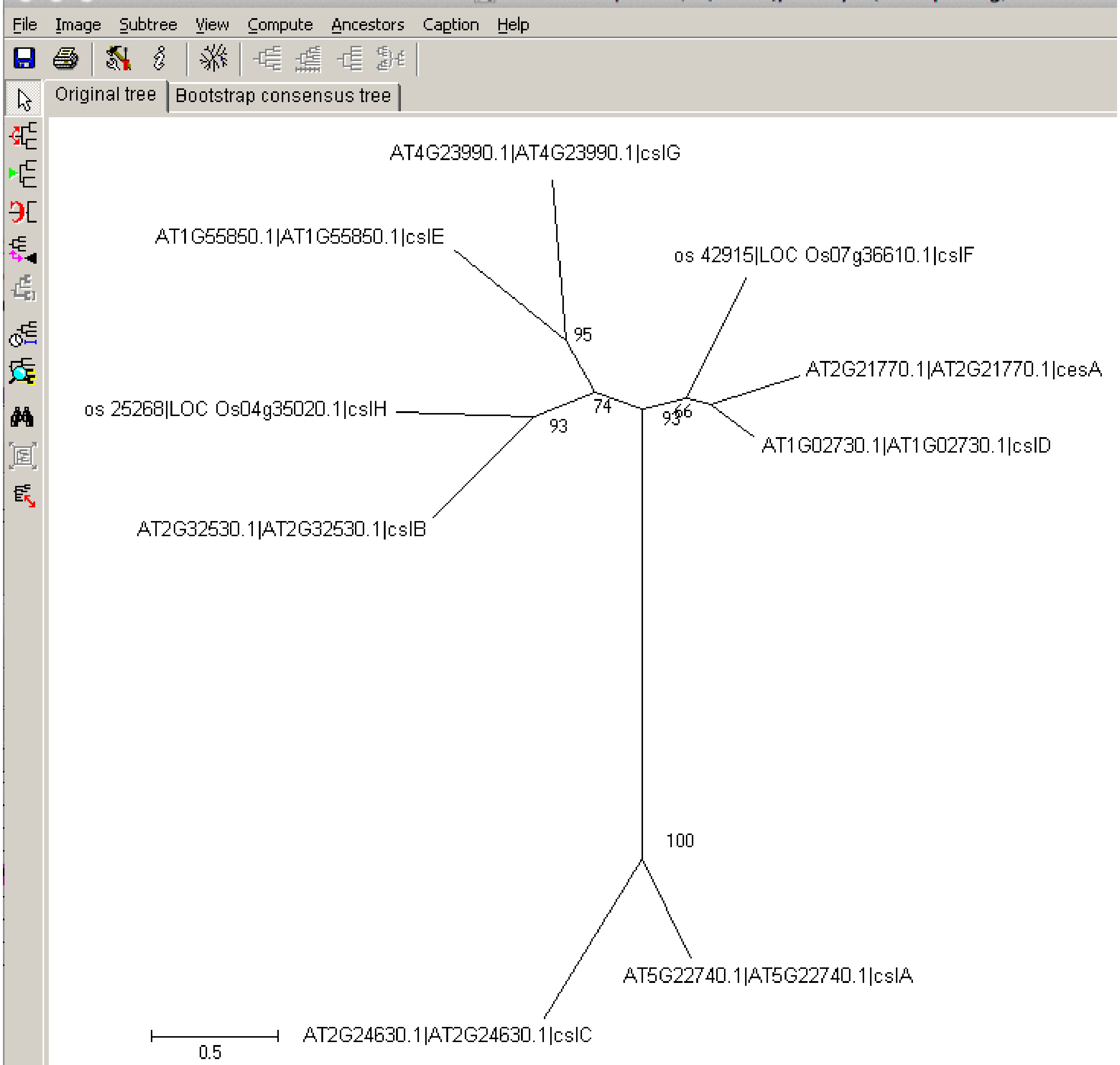
The option window



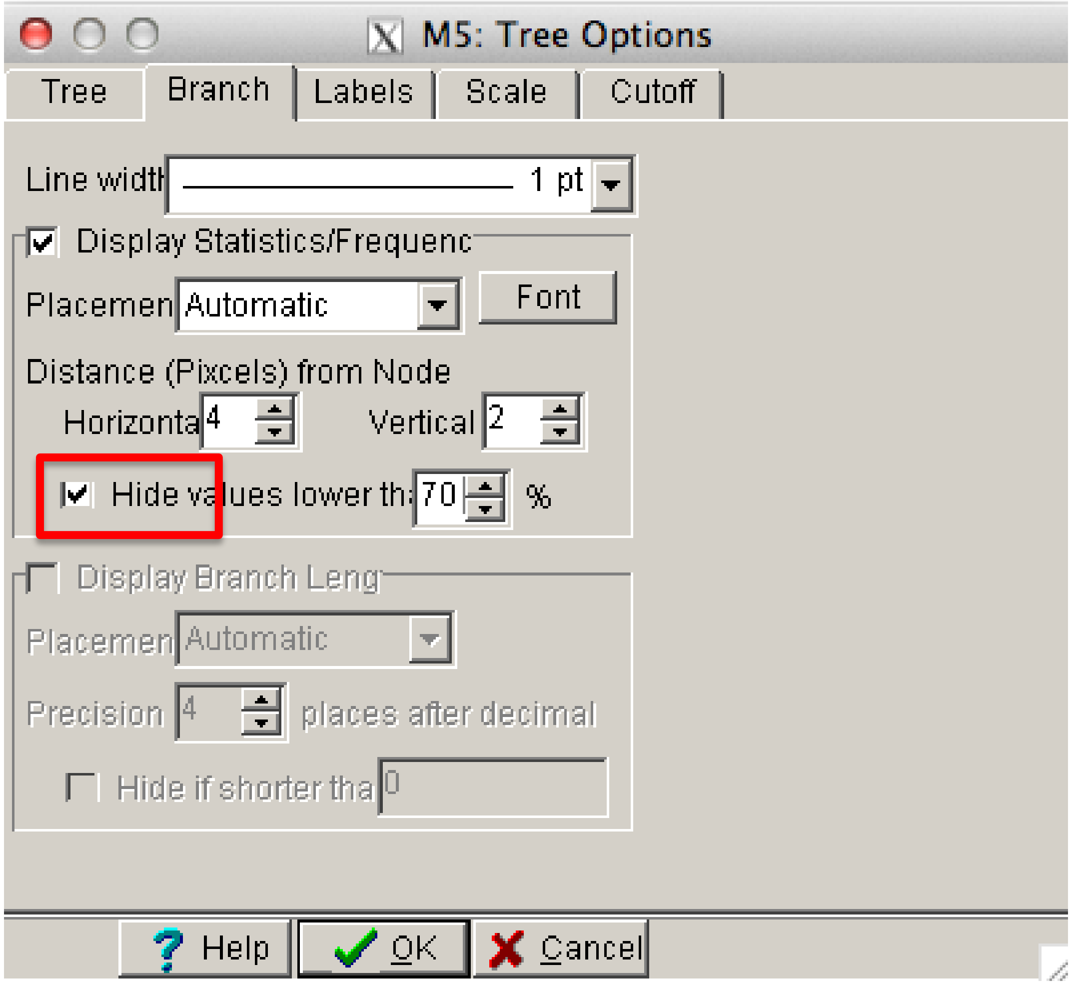
☑ Display Taxon Name Horizontally

Branch Length 165 Pixels / 1

Start Angle 0 O'Clock


File
Image
Subtree
View
Compute
Ancestors
Caption
Help
Original tree
Bootstrap consensus tree
AT4G23990.1|AT4G23990.1|cslG
AT1G55850.1|AT1G55850.1|cslE
os 42915|LOC Os07g36610.1|cslF
AT2G21770.1|AT2G21770.1|cesA
AT1G02730.1|AT1G02730.1|cslD
os 25268|LOC Os04g35020.1|cslH
AT2G32530.1|AT2G32530.1|cslB
AT5G22740.1|AT5G22740.1|cslA
AT2G24630.1|AT2G24630.1|cslC
95
74
93
93
66
100
0.5

To only show good bootstrap values



Line widtl 1 pt

☑ Display Statistics/Frequenc

Placemen Automatic

Distance (Pixcels) from Node

Horizonta 4 Vertical 2

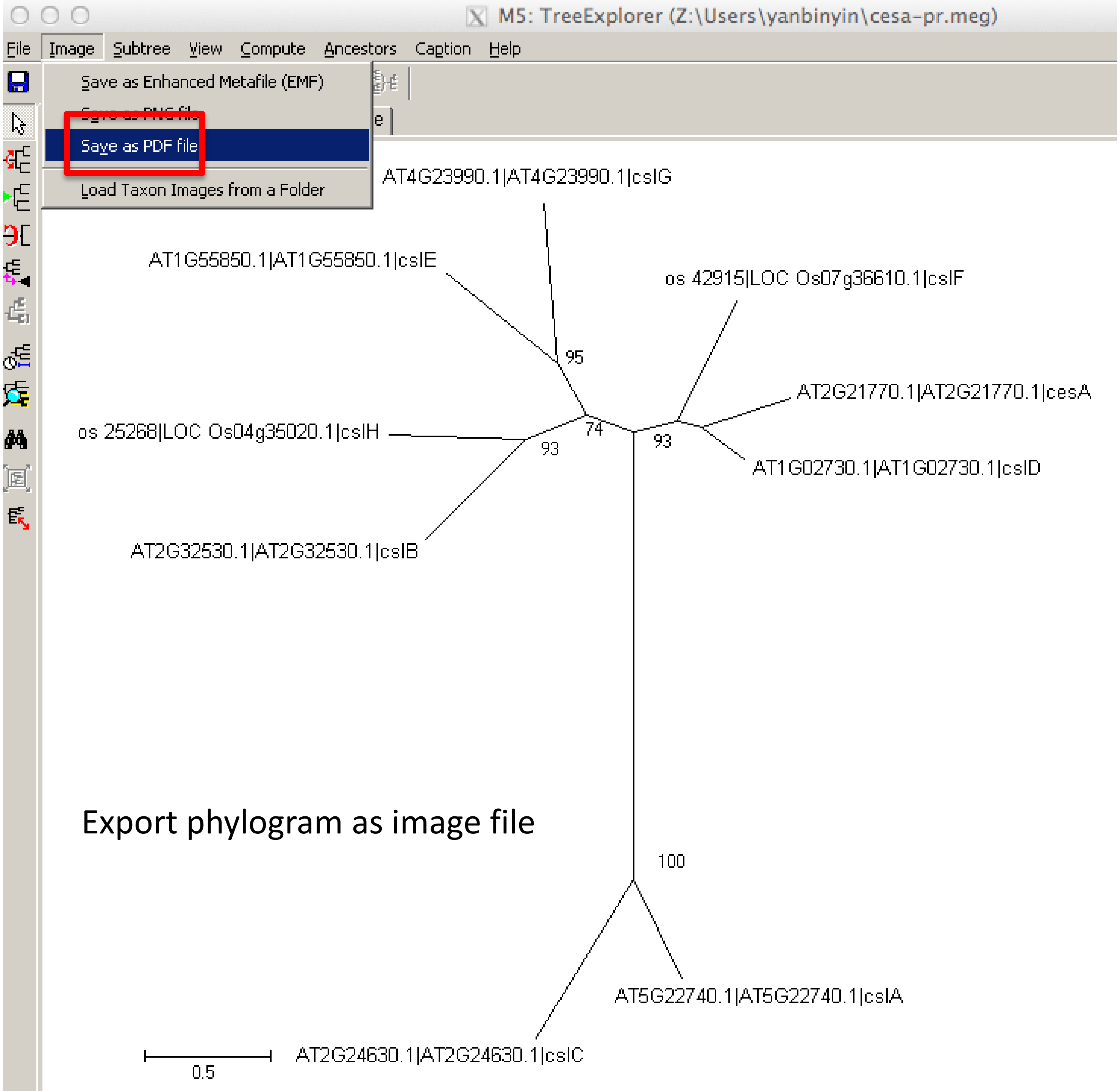
☑ Hide values lower th: 70 %

☐ Display Branch Leng

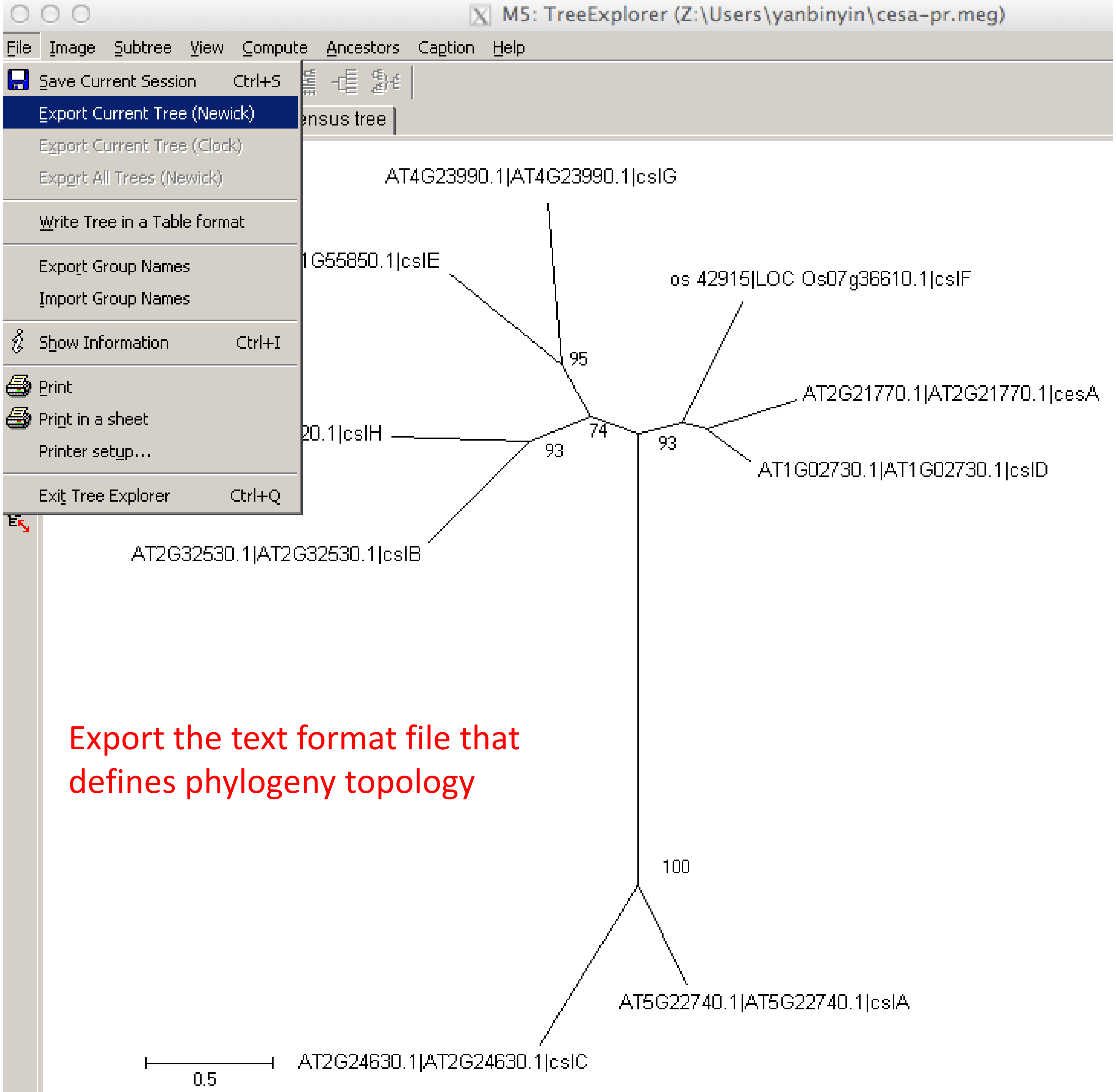
Placemen Automatic

Precision 4 places after decimal

☐ Hide if shorter tha 0



Export phylogram as image file



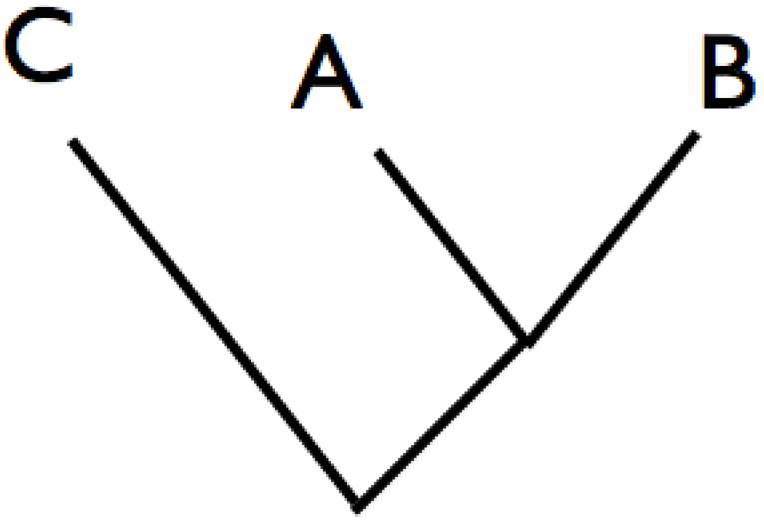
Export the text format file that defines phylogeny topology

Newick format

```
((((AT2G32530.1|AT2G32530.1|cslB:0.57078988,os_25268|LOC_Os04g35020.1|cslH:0.55075714)0.9300:0.26338963,(AT1G55850.1|AT1G55850.1|cslE:0.57830980,AT4G23990.1|AT4G23990.1|cslG:0.64691609)0.9500:0.23352951)0.7400:0.19857786,(os_42915|LOC_Os07g36610.1|cslF:0.54191868,(AT2G21770.1|AT2G21770.1|cesA:0.37516472,AT1G02730.1|AT1G02730.1|cslD:0.22502015)0.6600:0.09521396)0.9300:0.18369951)1.0000:0.73286595,(AT5G22740.1|AT5G22740.1|cslA:0.44848889,AT2G24630.1|AT2G24630.1|cslC:0.75671710)1.0000:1.05517231);
```

Not for human read!!!

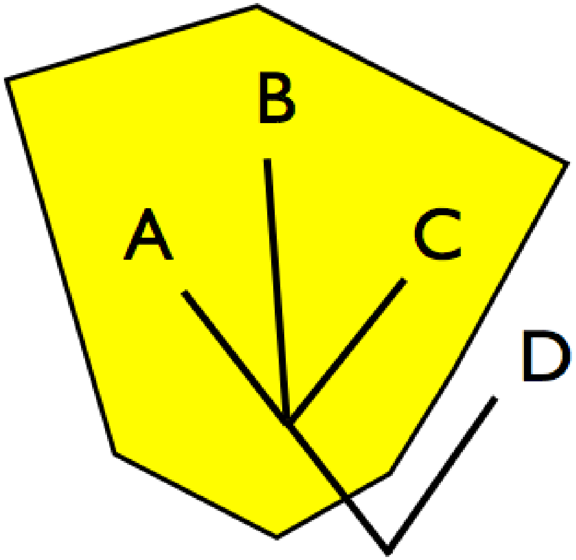
A most simplified example



```
((A,B),C);
```

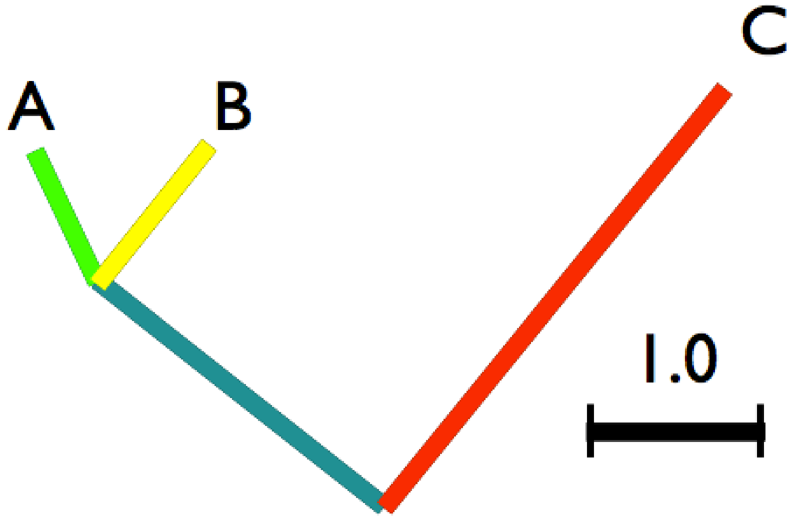
http://www.embl.de/~seqanal/courses/molEvolSofiaMar2012/newickPhylipTreeFormat.pdf polytomy/multifurcation

`((A,B,C),D);`

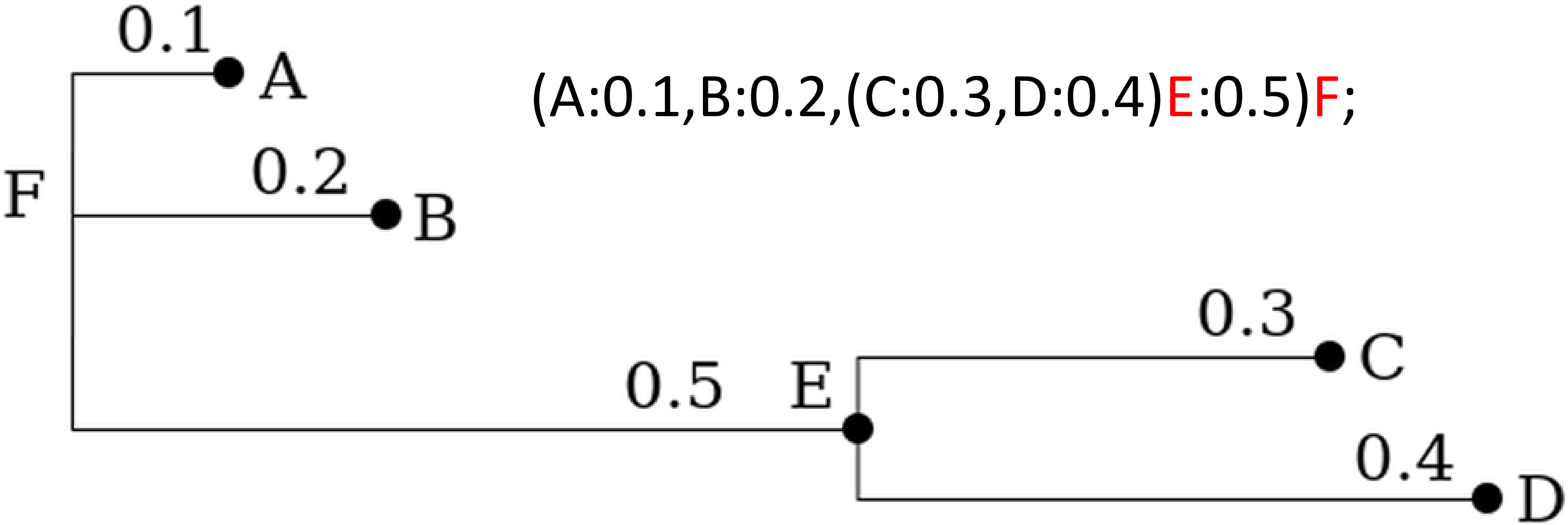


Add the branch length

((A:1,B:1):2,C:3)

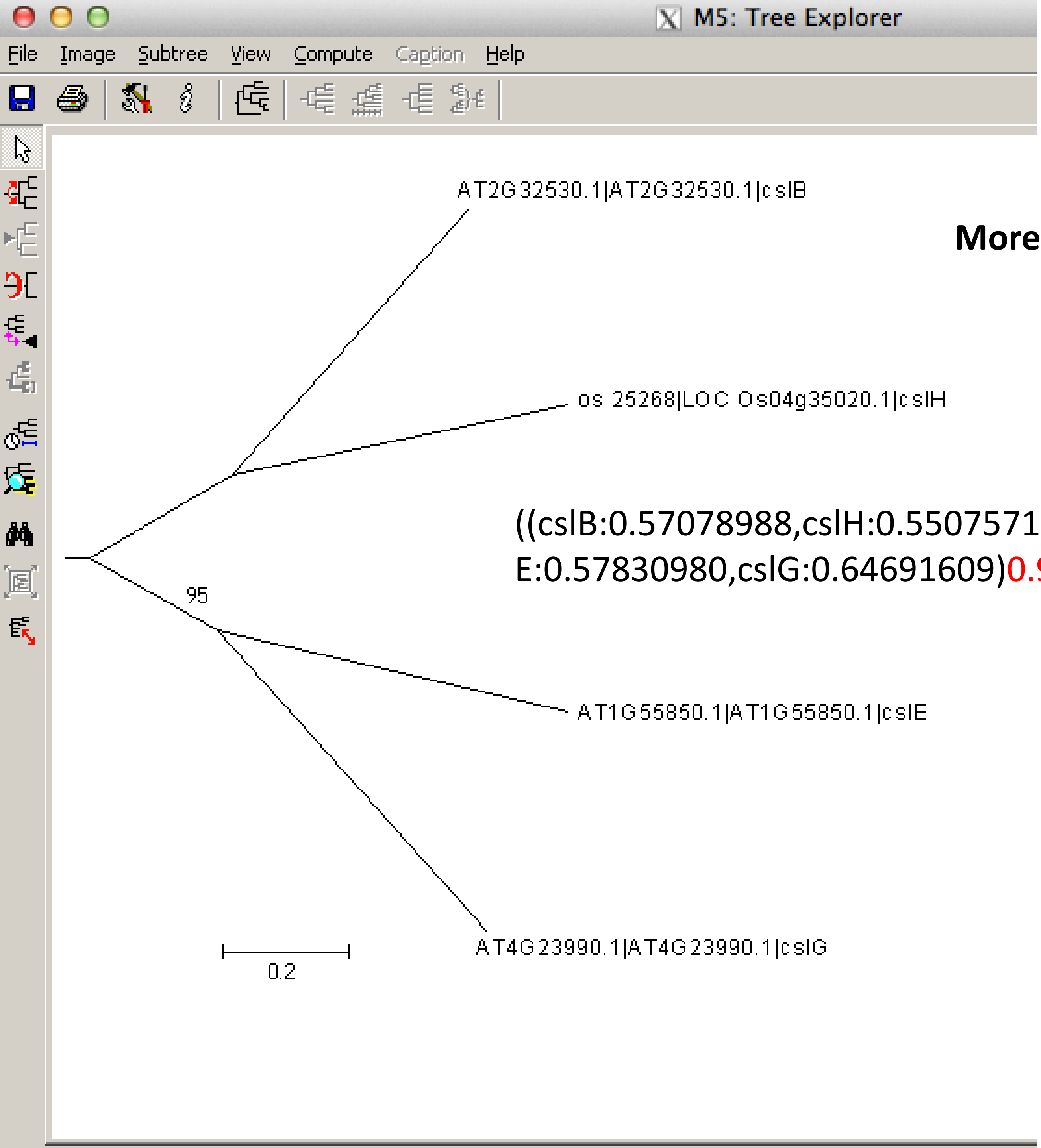


Add the internal node name

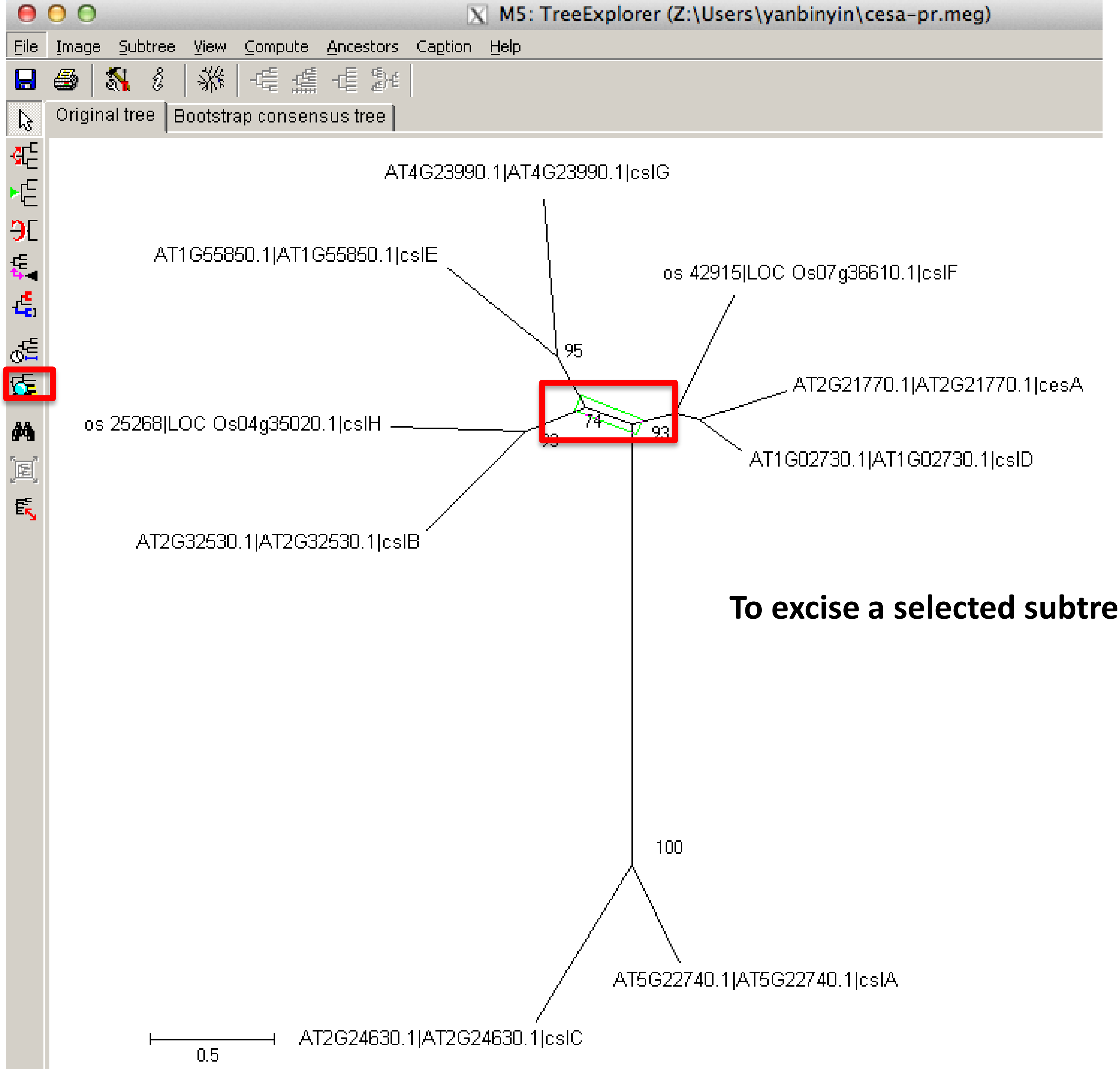


(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;

More often, add bootstrap values

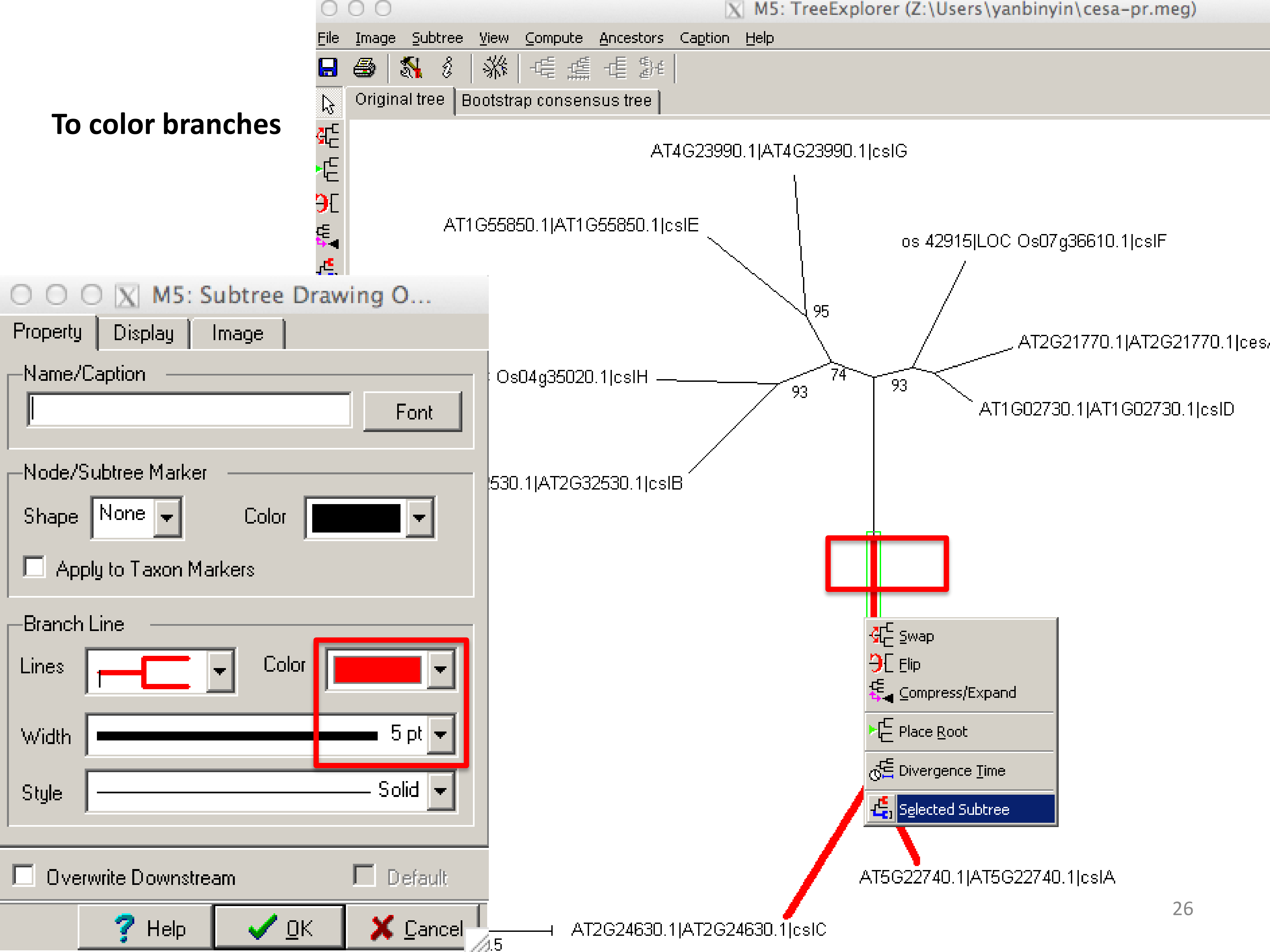


((cslB:0.57078988,cslH:0.55075714)0.9300:0.26338963,(cslE:0.57830980,cslG:0.64691609)0.9500:0.23352951);



To excise a selected subtree (clade)

To color branches



M5: Subtree Drawing O...

Property | Display | Image

Name/Caption

Node/Subtree Marker

Shape: None

Color: ⬛

☐ Apply to Taxon Markers

Branch Line

Lines

Color: 🟥

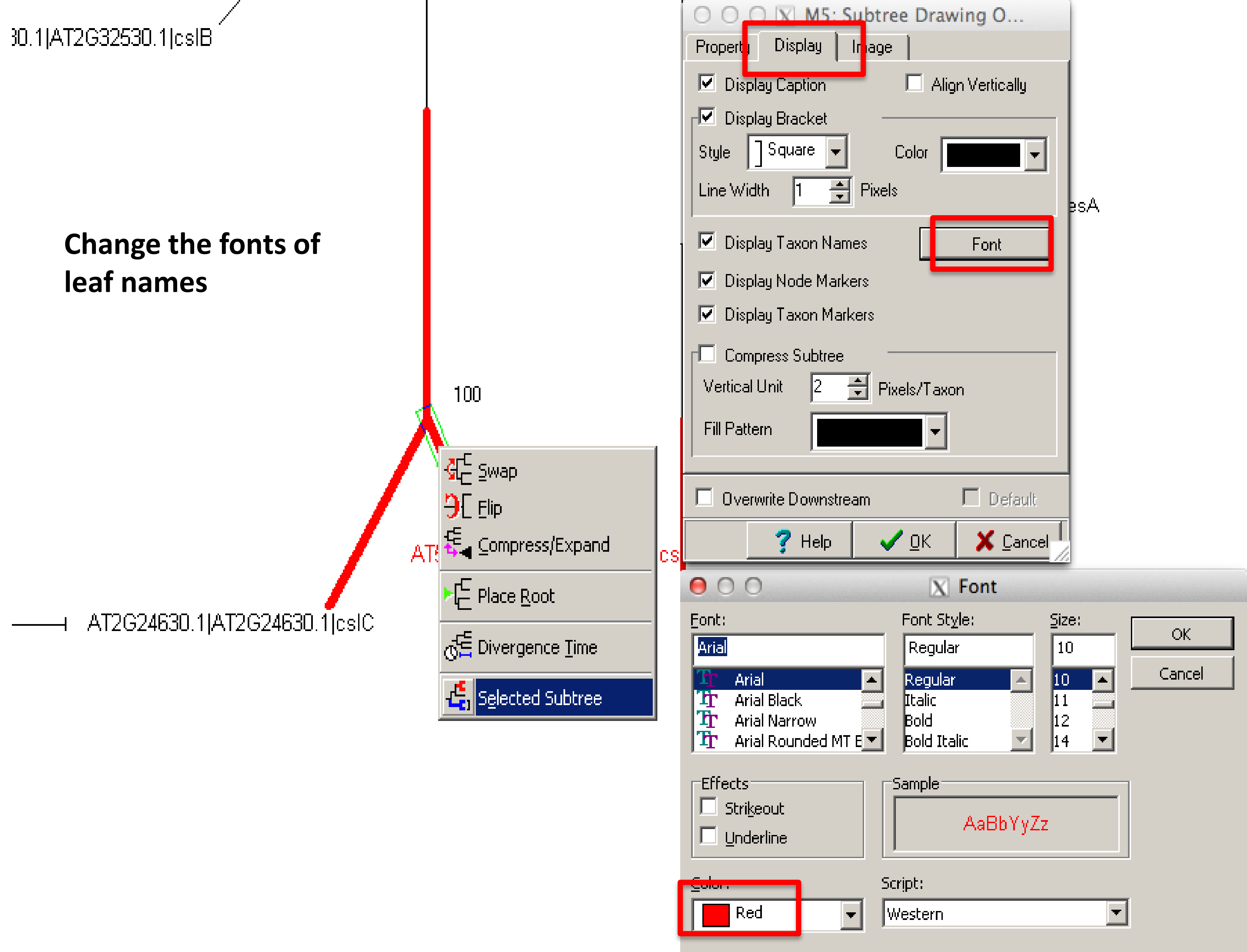
Width: 5 pt

Style: Solid

☐ Overwrite Downstream

☐ Default

Change the fonts of leaf names



M5: Subtree Drawing O...

☑ Display Caption ☐ Align Vertically

☑ Display Bracket

Style:] Square Color: ■

Line Width: 1 Pixels

☑ Display Taxon Names

☑ Display Node Markers

☑ Display Taxon Markers

☐ Compress Subtree

Vertical Unit: 2 Pixels/Taxon

Fill Pattern: ■

☐ Overwrite Downstream ☐ Default

Font

Font: Arial

Font Style: Regular

Size: 10

Effects

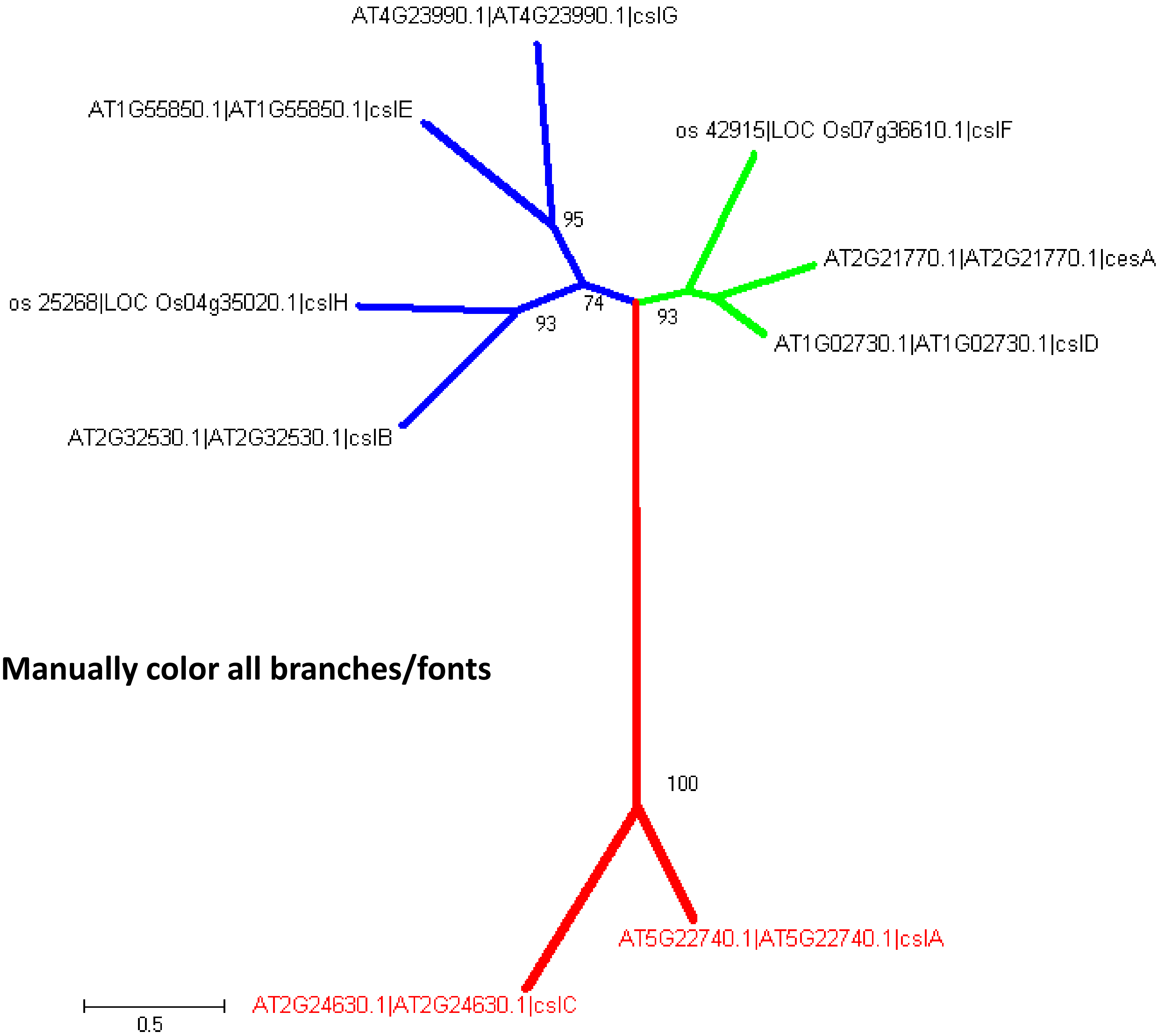
☐ Strikeout

☐ Underline

Sample: AaBbYyZz

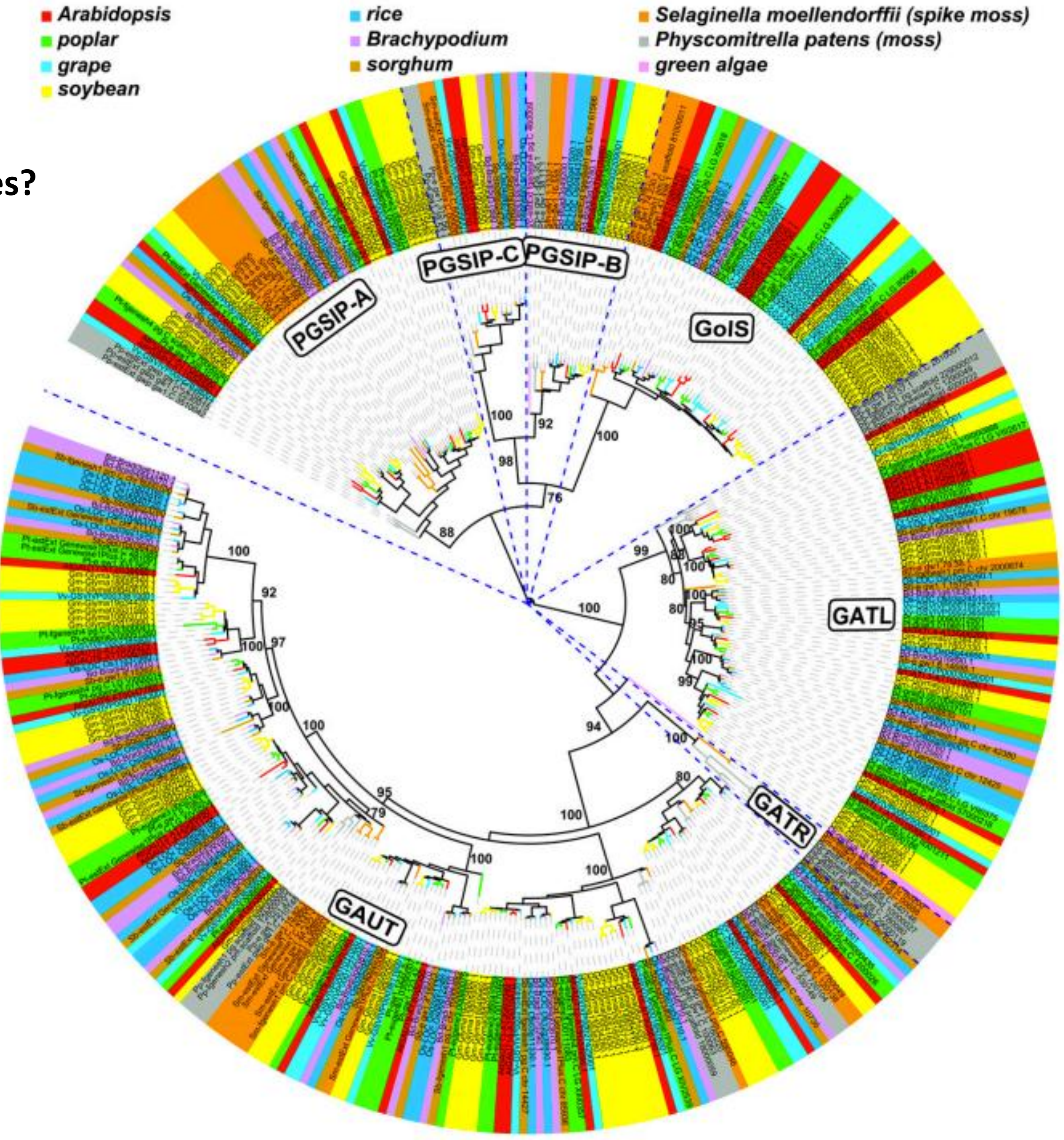
Color: 🟥 Red

Script: Western

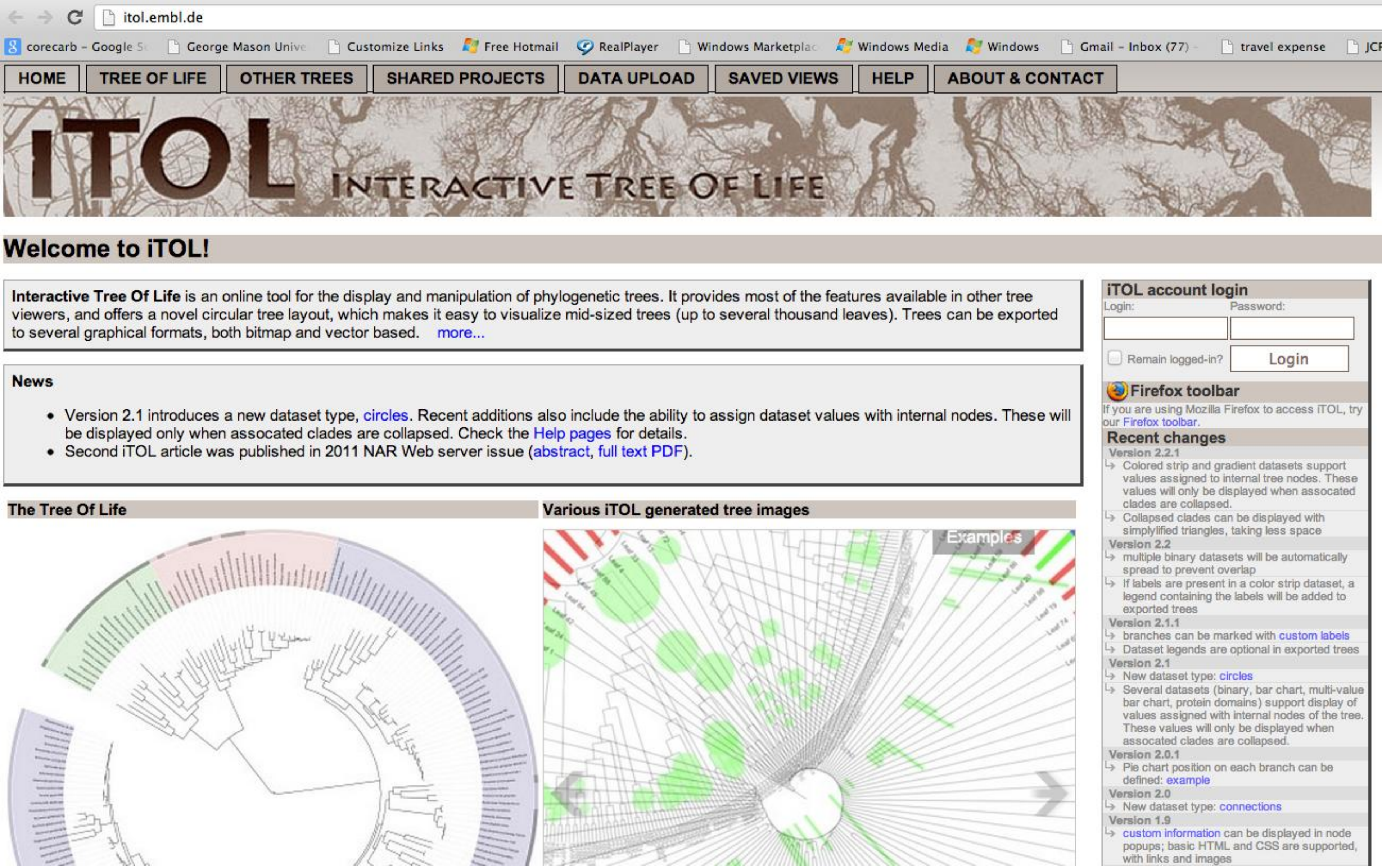


Manually color all branches/fonts

What if we have hundreds of genes?



http://itol.embl.de/



Welcome to iTOL!

Interactive Tree Of Life is an online tool for the display and manipulation of phylogenetic trees. It provides most of the features available in other tree viewers, and offers a novel circular tree layout, which makes it easy to visualize mid-sized trees (up to several thousand leaves). Trees can be exported to several graphical formats, both bitmap and vector based. more...

News

- Version 2.1 introduces a new dataset type, circles. Recent additions also include the ability to assign dataset values with internal nodes. These will be displayed only when assocated clades are collapsed. Check the Help pages for details.
- Second iTOL article was published in 2011 NAR Web server issue (abstract, full text PDF).

The Tree Of Life

Various iTOL generated tree images

iTOL account login

Login:
Password:

☐ Remain logged-in?

Firefox toolbar

If you are using Mozilla Firefox to access iTOL, try our Firefox toolbar.

Recent changes

Version 2.2.1
- Colored strip and gradient datasets support values assigned to internal tree nodes. These values will only be displayed when assocated clades are collapsed.
- Collapsed clades can be displayed with simplyfied triangles, taking less space

Version 2.2
- multiple binary datasets will be automatically spread to prevent overlap
- If labels are present in a color strip dataset, a legend containing the labels will be added to exported trees

Version 2.1.1
- branches can be marked with custom labels
- Dataset legends are optional in exported trees

Version 2.1
- New dataset type: circles
- Several datasets (binary, bar chart, multi-value bar chart, protein domains) support display of values assigned with internal nodes of the tree. These values will only be displayed when assocated clades are collapsed.

Version 2.0.1
- Pie chart position on each branch can be defined: example

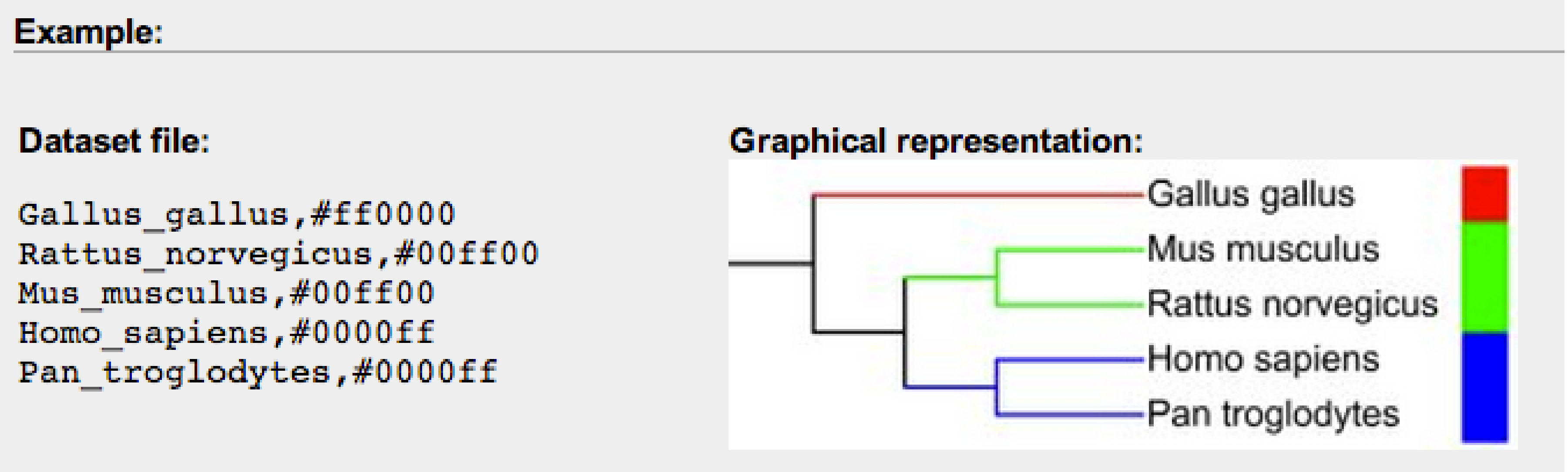
Version 2.0
- New dataset type: connections

Version 1.9
- custom information can be displayed in node popups; basic HTML and CSS are supported, with links and images

Automatically define branch colors by uploading a **color definition file**

You can define your own colors for each branch/leaf separately. Use standard hexadecimal color notation (for example, #ff0000 for red)

http://www.w3schools.com/html/html_colors.asp



Example:

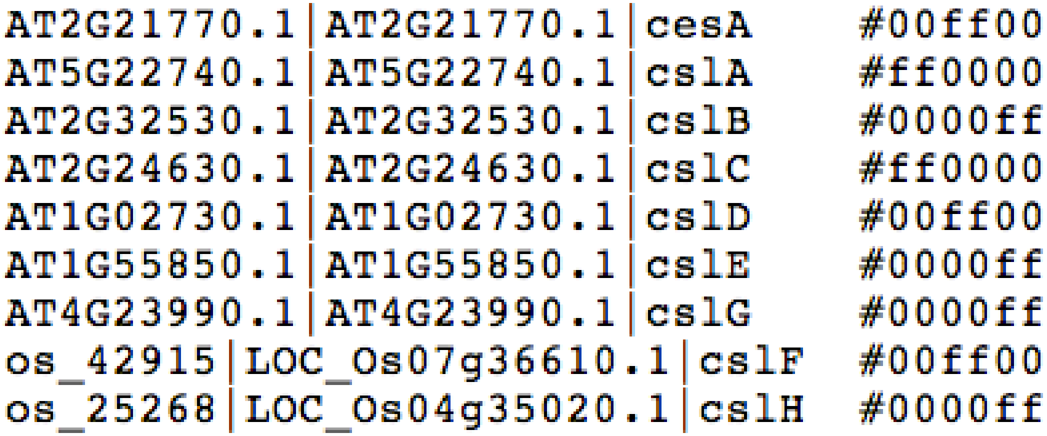
Dataset file:

```
Gallus_gallus,#ff0000
Rattus_norvegicus,#00ff00
Mus_musculus,#00ff00
Homo_sapiens,#0000ff
Pan_troglodytes,#0000ff
```

Graphical representation:

http://itol.embl.de/help/help.shtml

http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.col



```
AT2G21770.1|AT2G21770.1|cesA    #00ff00
AT5G22740.1|AT5G22740.1|cslA    #ff0000
AT2G32530.1|AT2G32530.1|cslB    #0000ff
AT2G24630.1|AT2G24630.1|cslC    #ff0000
AT1G02730.1|AT1G02730.1|cslD    #00ff00
AT1G55850.1|AT1G55850.1|cslE    #0000ff
AT4G23990.1|AT4G23990.1|cslG    #0000ff
os_42915|LOC_Os07g36610.1|cslF #00ff00
os_25268|LOC_Os04g35020.1|cslH #0000ff
```

```
((((AT2G32530.1|AT2G32530.1|cslB:0.57078988,os_25268|LOC_Os04g35020.1|cslH:0.55075714)0.9300:0.26338963,(AT1G55850.1|AT1G55850.1|cslE:0.57830980,AT4G23990.1|AT4G23990.1|cslG:0.64691609)0.9500:0.23352951)0.7400:0.19857786,(os_42915|LOC_Os07g36610.1|cslF:0.54191868,(AT2G21770.1|AT2G21770.1|cesA:0.37516472,AT1G02730.1|AT1G02730.1|cslD:0.22502015)0.6600:0.09521396)0.9300:0.18369951)1.0000:0.73286595,(AT5G22740.1|AT5G22740.1|cslA:0.44848889,AT2G24630.1|AT2G24630.1|cslC:0.75671710)1.0000:1.05517231);
```

Upload a prune list for our tree of life

You can upload a list of NCBI taxonomy IDs and the tree of life will be pruned to include only the species from your file. There should one species I
download.

File with tax IDs: Choose File No file chosen

Upload your own tree

NEW! If you are using iTOL to upload your own trees, try creating a personal account. More info about the iTOL personal accounts is availab

Use this form to upload your own phylogenetic tree. It should be in plain text, in one of supported formats (Newick, Nexus or PhyloXML) . Please c
features available.

You can either paste your tree into the box, or upload a file using the file selector below. Don't forget to select the correct tree format.

Paste or type the tree:	Upload a file which contains your tree:
Paste your tree here...	Choose File No file chosen **Tree format:** Newick Make sure the file is plain text, and contains only trees in the selected format.

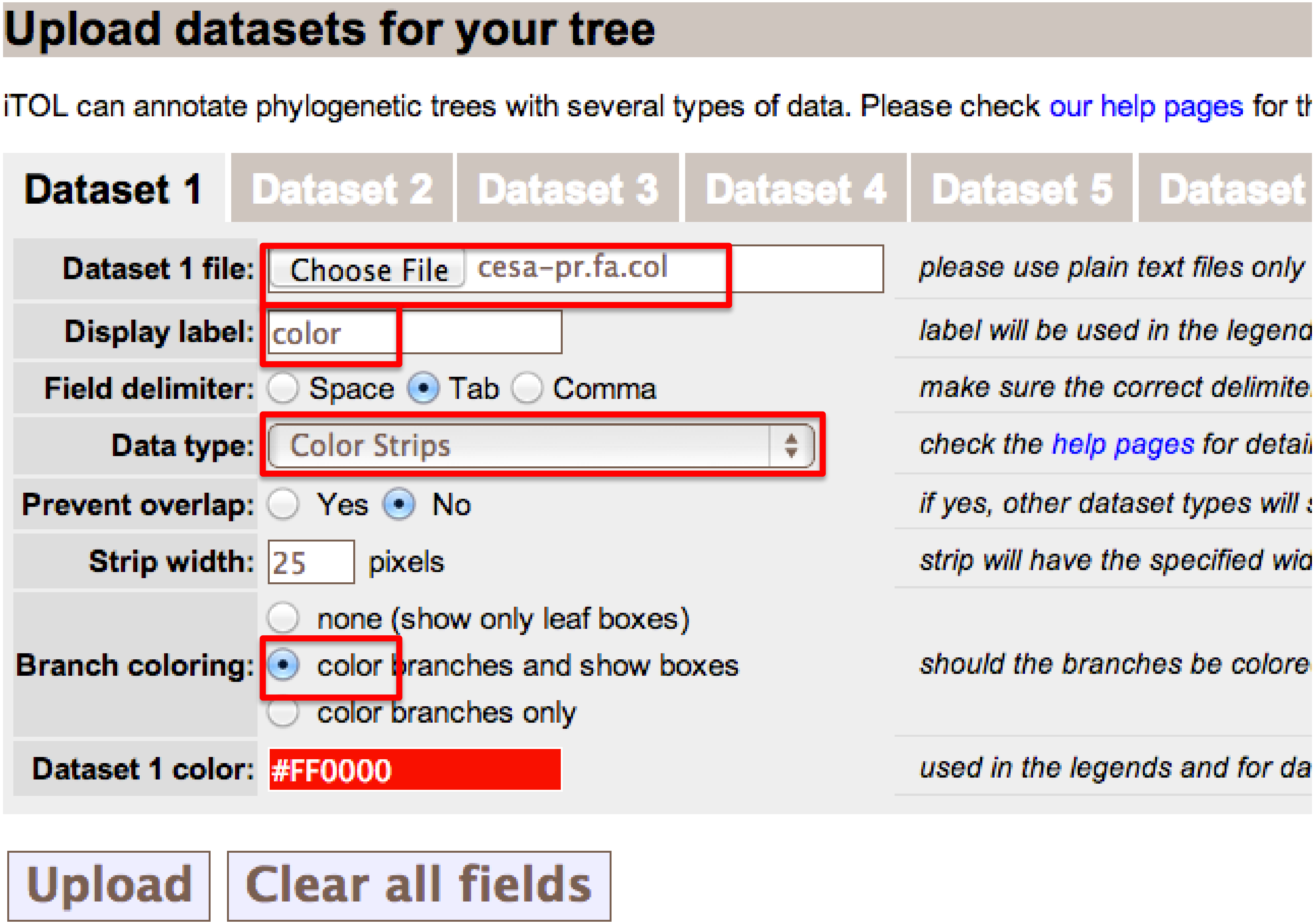
Optional information

Tree name: *if you don't specify a name, a numeric ID will be used instead*

Advanced options (show)

(if you're uploading extra data with your tree, fill the dataset section below before clicking 'Upload')

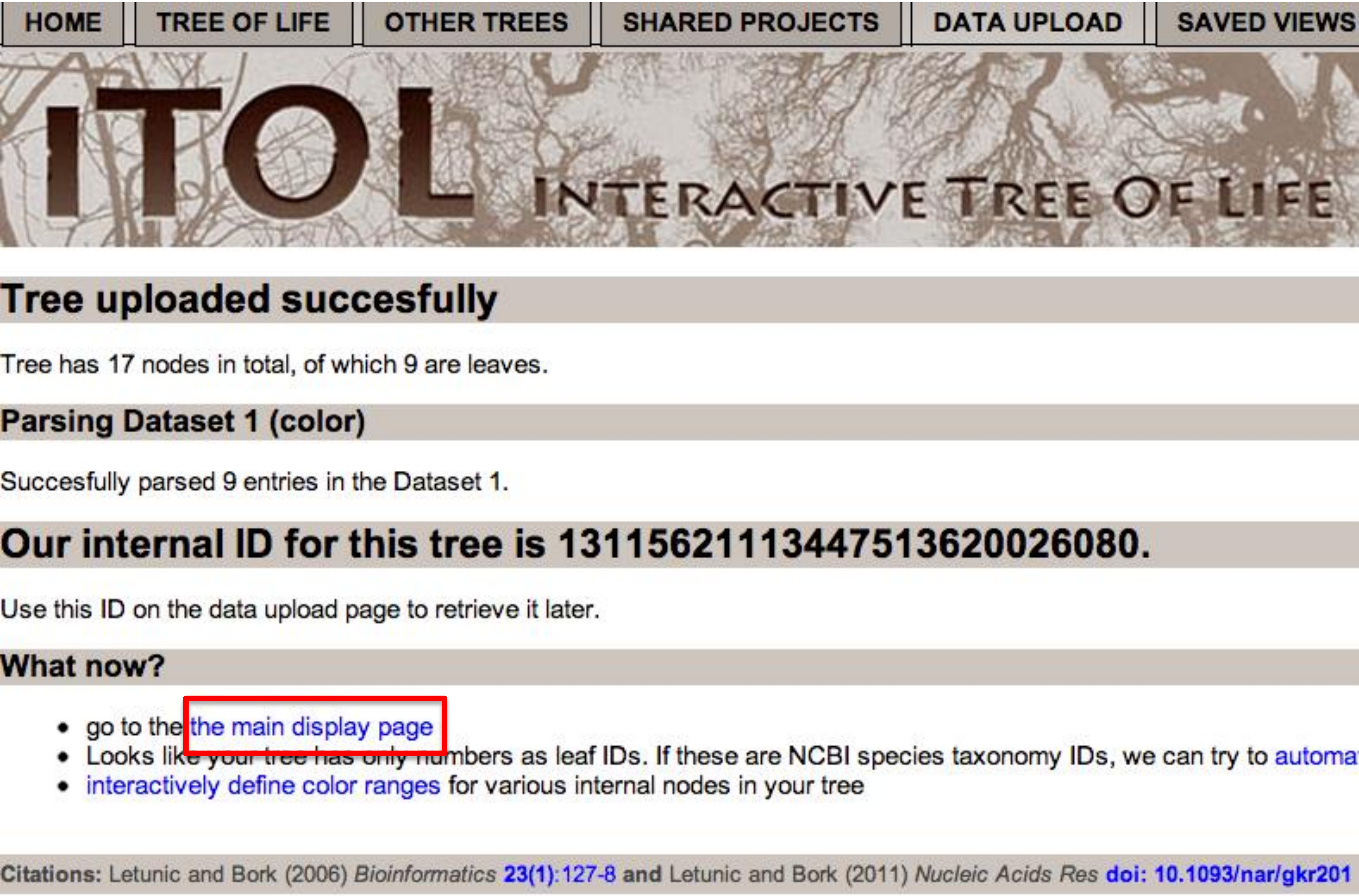
Upload color definition file



Upload datasets for your tree

iTOL can annotate phylogenetic trees with several types of data. Please check our help pages for th

Dataset 1 file:	cesa-pr.fa.col	*please use plain text files only*
Display label:	color	*label will be used in the legend*
Field delimiter:	○ Space ● Tab ○ Comma	*make sure the correct delimite*
Data type:	Color Strips	*check the help pages for detai*
Prevent overlap:	○ Yes ● No	*if yes, other dataset types will*
Strip width:	25 pixels	*strip will have the specified wid*
Branch coloring:	○ none (show only leaf boxes) ● color branches and show boxes ○ color branches only	*should the branches be colore*
Dataset 1 color:	#FF0000	*used in the legends and for da*



Tree uploaded succesfully

Tree has 17 nodes in total, of which 9 are leaves.

Parsing Dataset 1 (color)

Succesfully parsed 9 entries in the Dataset 1.

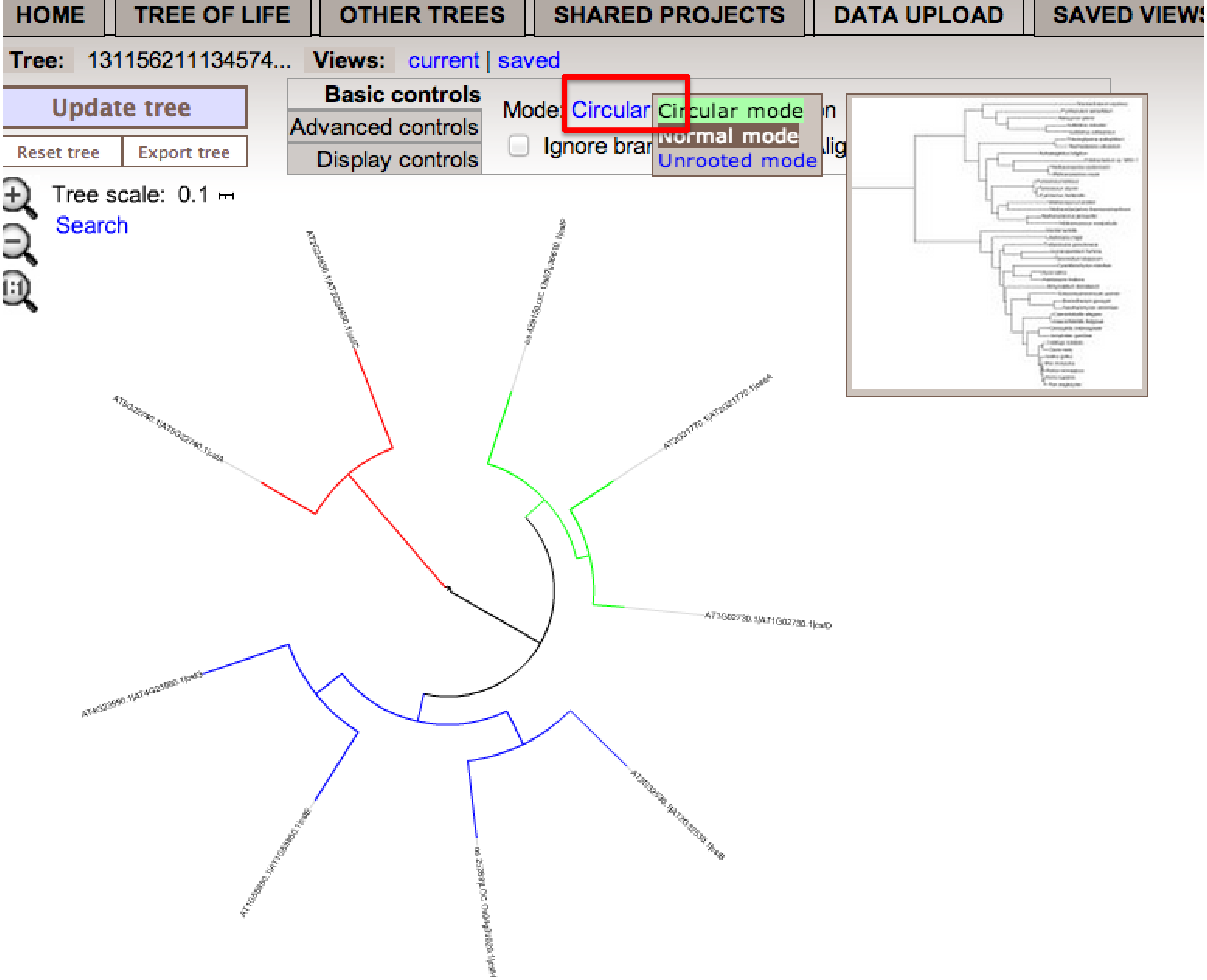
Our internal ID for this tree is 1311562113447513620026080.

Use this ID on the data upload page to retrieve it later.

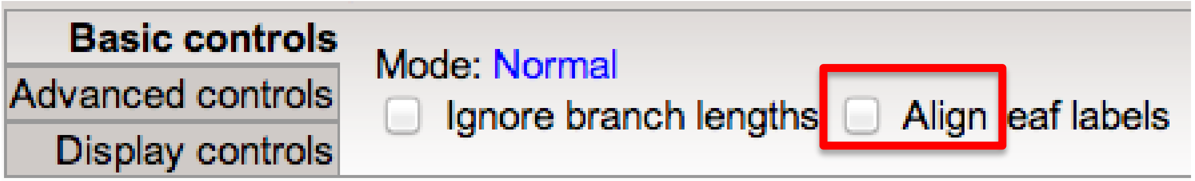
What now?

- go to the the main display page
- Looks like your tree has only numbers as leaf IDs. If these are NCBI species taxonomy IDs, we can try to automa
- interactively define color ranges for various internal nodes in your tree

Citations: Letunic and Bork (2006) *Bioinformatics* **23(1)**:127-8 **and** Letunic and Bork (2011) *Nucleic Acids Res* **doi: 10.1093/nar/gkr201**

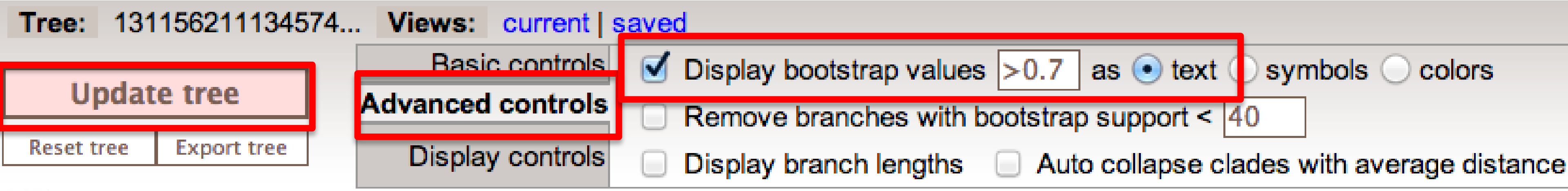

HOME
TREE OF LIFE
OTHER TREES
SHARED PROJECTS
DATA UPLOAD
SAVED VIEWS
Tree: 131156211134574...
Views: current | saved
Update tree
Reset tree
Export tree
Basic controls
Advanced controls
Display controls
Mode: Circular
Circular mode
Normal mode
Unrooted mode
Ignore bran
Tree scale: 0.1
Search

More options to display the phylogram



Mode: Normal

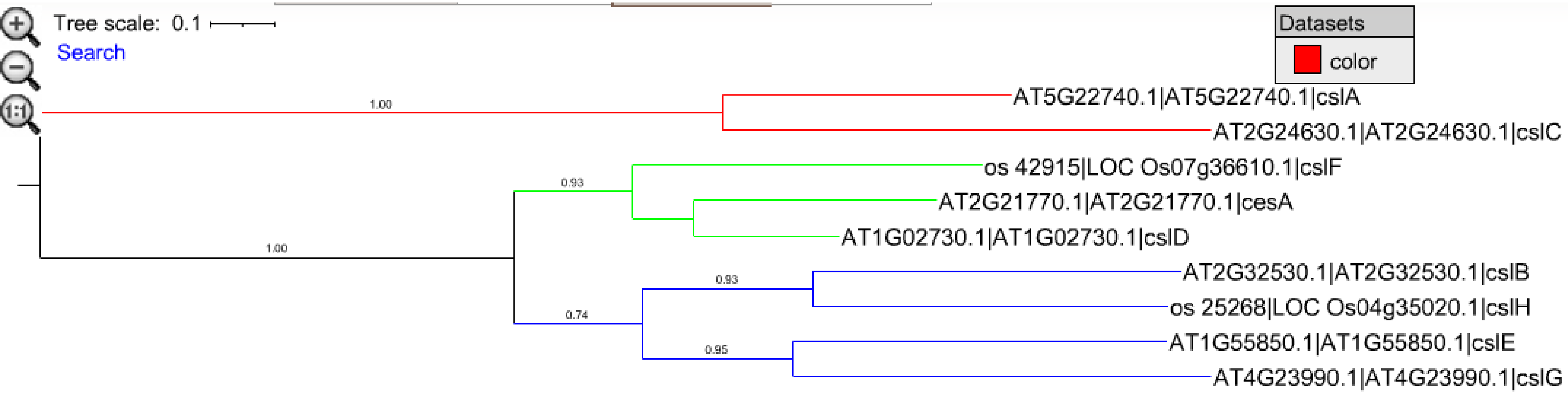
☐ Ignore branch lengths ☐ Align eaf labels

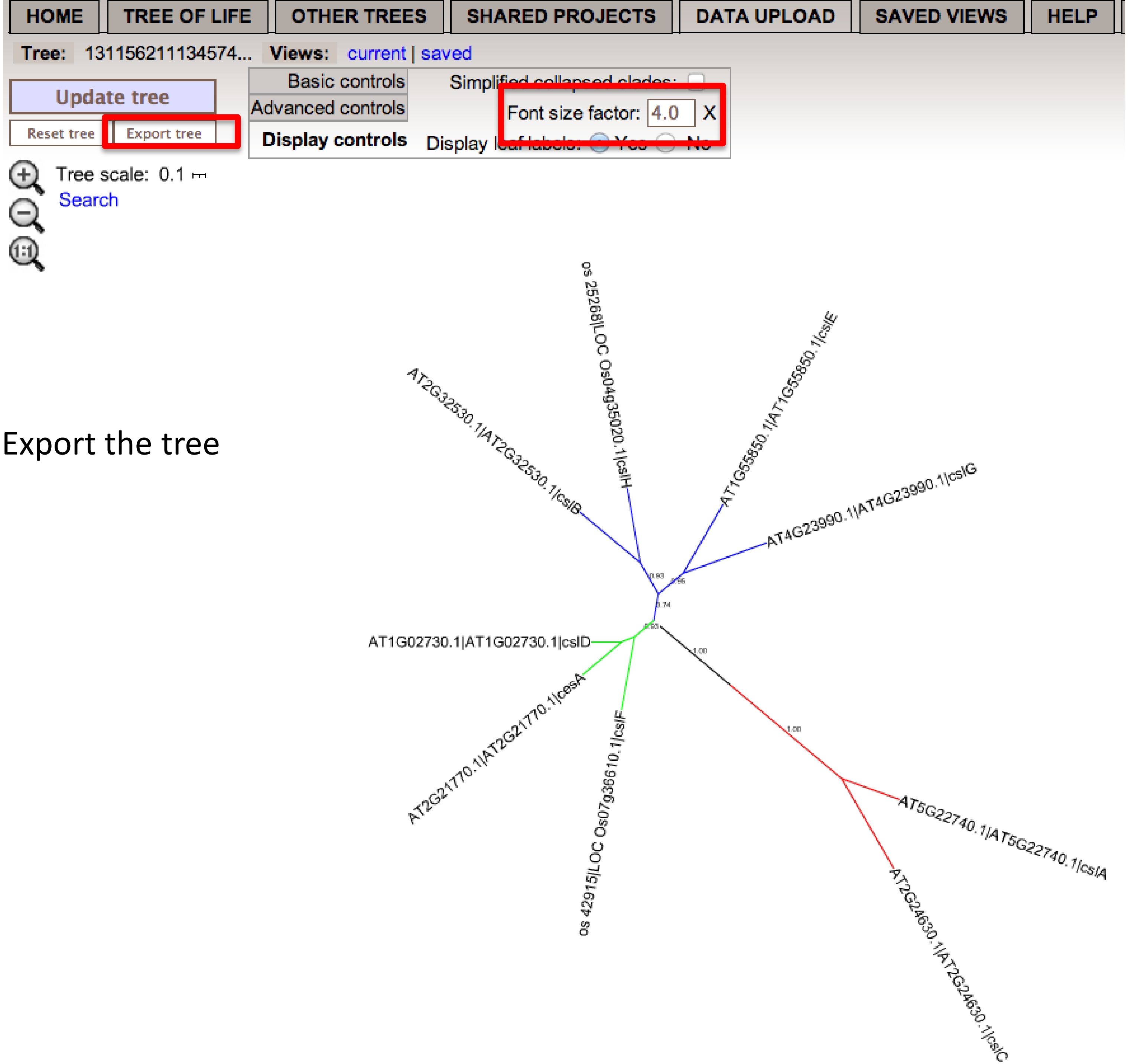

Tree: 131156211134574... Views: current | saved

☑ Display bootstrap values >0.7 as ● text ○ symbols ○ colors

☐ Remove branches with bootstrap support < 40

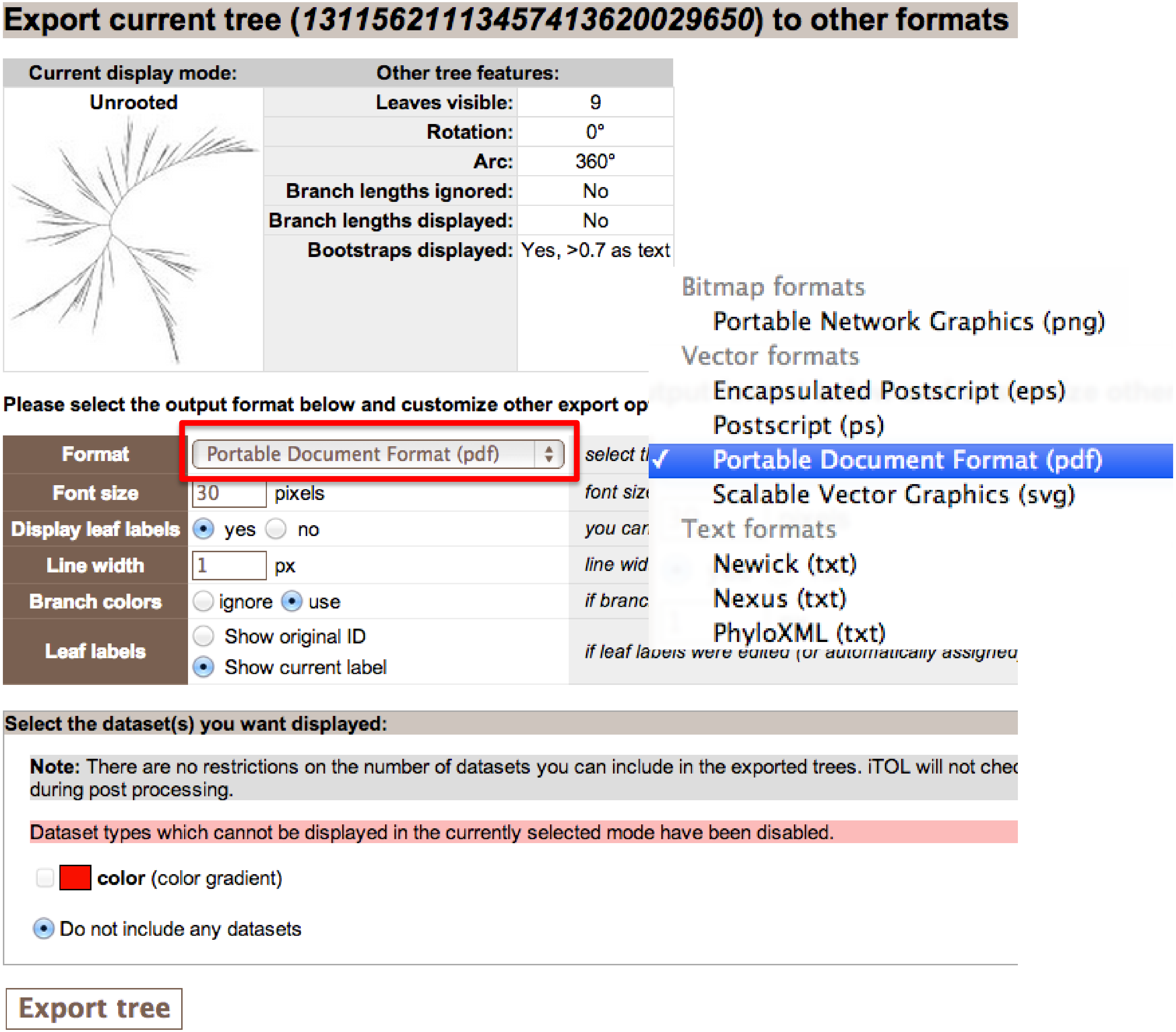
☐ Display branch lengths ☐ Auto collapse clades with average distance





Font size factor: 4.0 X

Export the tree



Export current tree (*13115621113457413620029650*) to other formats

Current display mode:	Other tree features:	
Unrooted	Leaves visible:	9
	Rotation:	0°
	Arc:	360°
	Branch lengths ignored:	No
	Branch lengths displayed:	No
	Bootstraps displayed:	Yes, >0.7 as text

Please select the output format below and customize other export op

Format	Portable Document Format (pdf)	*select t*
Font size	30 pixels	*font siz*
Display leaf labels	● yes ○ no	*you car*
Line width	1 px	*line wid*
Branch colors	○ ignore ● use	*if branc*
Leaf labels	○ Show original ID ● Show current label	*if leaf labels were edited (or automatically assigned*

Select the dataset(s) you want displayed:

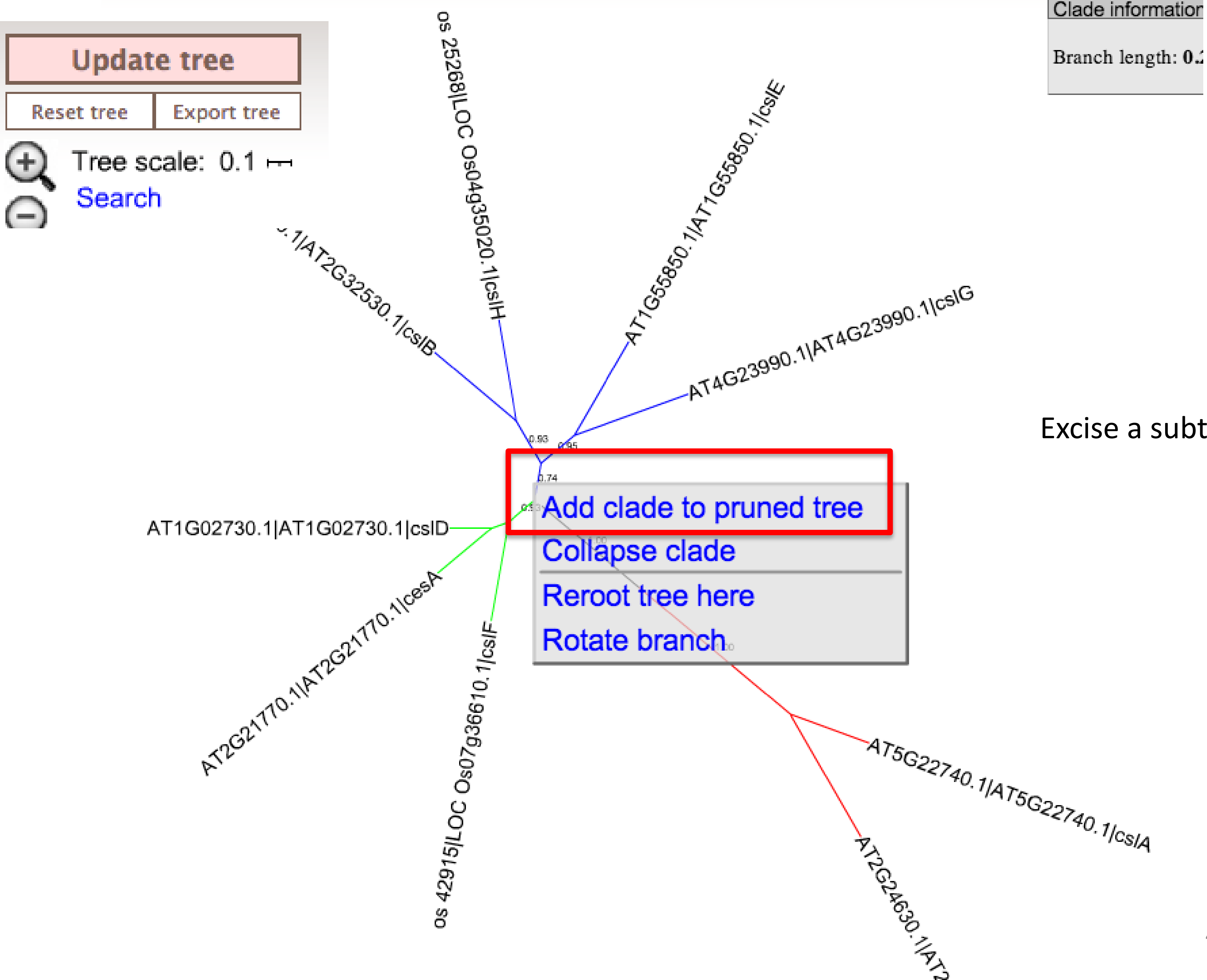
Note: There are no restrictions on the number of datasets you can include in the exported trees. iTOL will not chec during post processing.

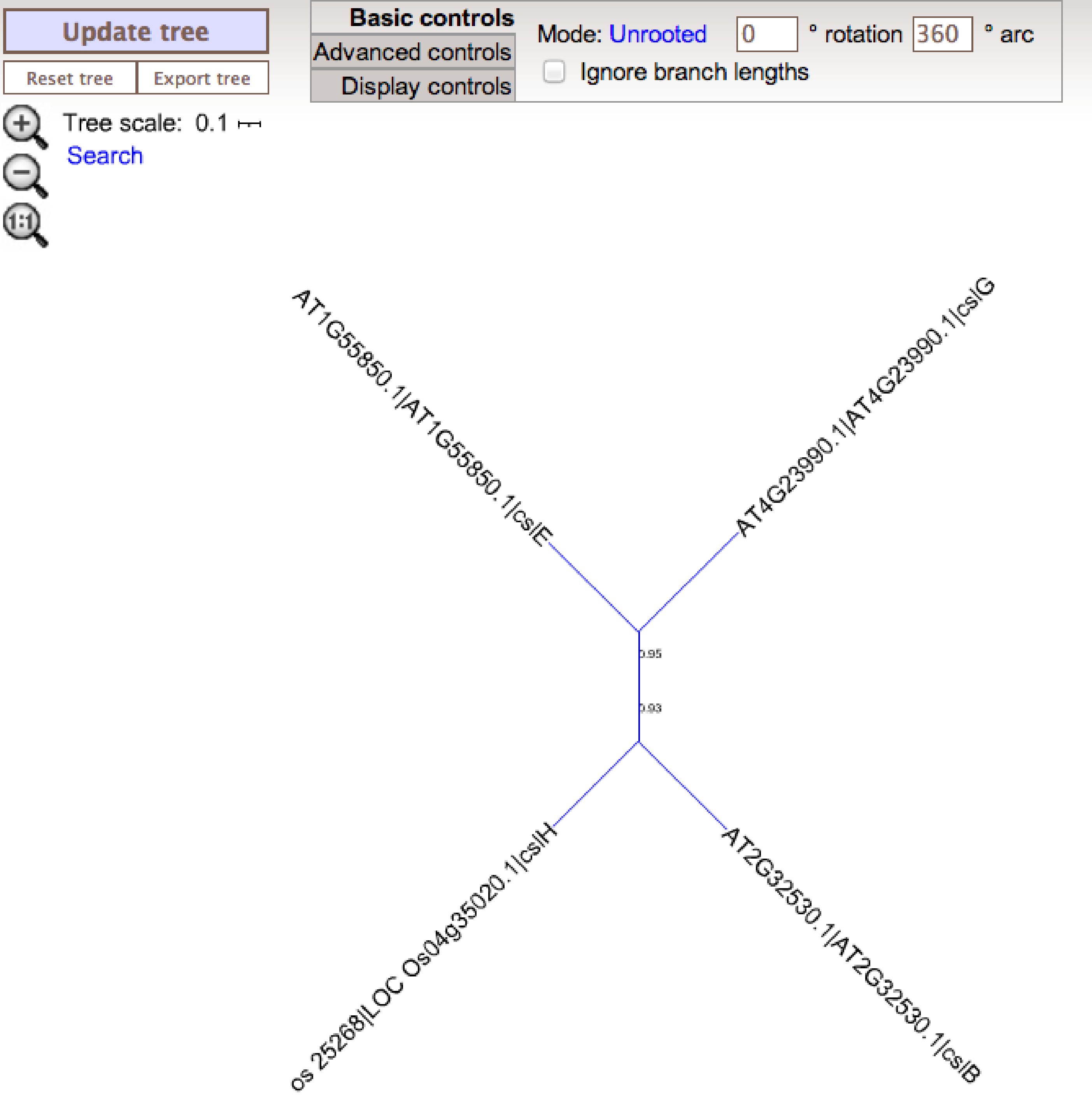
Dataset types which cannot be displayed in the currently selected mode have been disabled.

☐ 🟥 **color** (color gradient)

● Do not include any datasets

Excise a subtree





Mode: Unrooted

0 ° rotation

360 ° arc

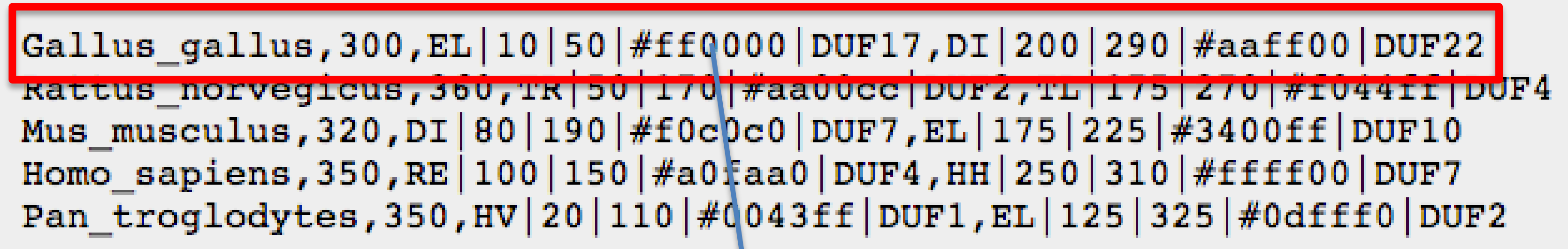
☐ Ignore branch lengths

Tree scale: 0.1

http://itol.embl.de/help/help.shtml → Uploading and working with your own trees

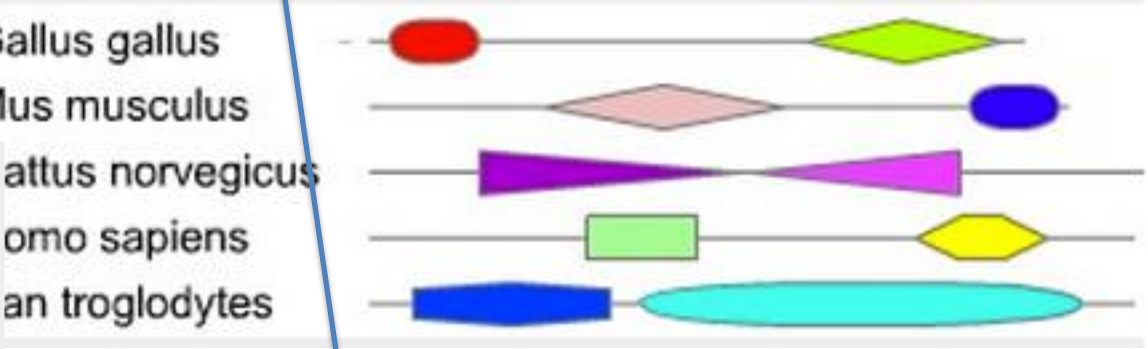
Prepare a domain definition file to show domain structures

Dataset file:

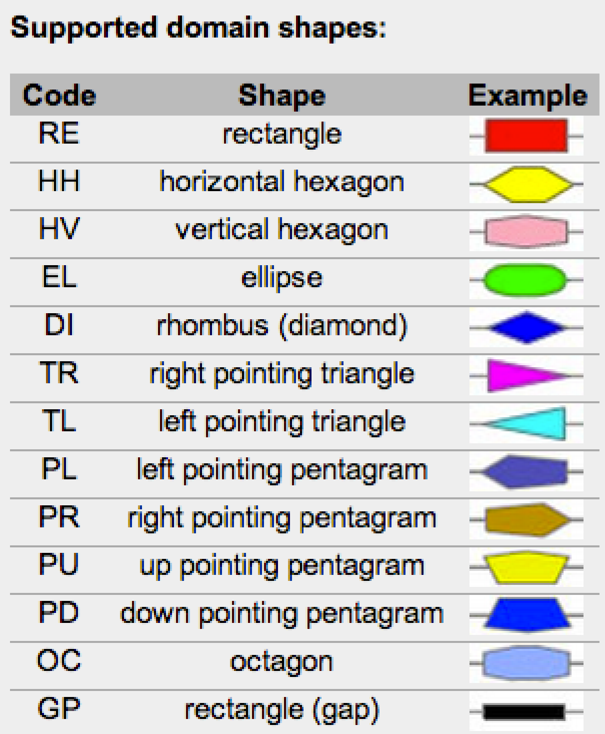


```
Gallus_gallus,300,EL|10|50|#ff0000|DUF17,DI|200|290|#aaff00|DUF22
Rattus_norvegicus,360,TR|50|170|#aa00cc|DUF2,TL|175|270|#f044ff|DUF4
Mus_musculus,320,DI|80|190|#f0c0c0|DUF7,EL|175|225|#3400ff|DUF10
Homo_sapiens,350,RE|100|150|#a0faa0|DUF4,HH|250|310|#ffff00|DUF7
Pan_troglodytes,350,HV|20|110|#0043ff|DUF1,EL|125|325|#0dfff0|DUF2
```

Graphical representation:



Supported domain shapes:



Code	Shape	Example
RE	rectangle	🟥
HH	horizontal hexagon	⬣
HV	vertical hexagon	⬢
EL	ellipse	⬭
DI	rhombus (diamond)	◆
TR	right pointing triangle	▶
TL	left pointing triangle	◀
PL	left pointing pentagram	⬟
PR	right pointing pentagram	⬟
PU	up pointing pentagram	⬟
PD	down pointing pentagram	⬟
OC	octagon	⯃
GP	rectangle (gap)	▬

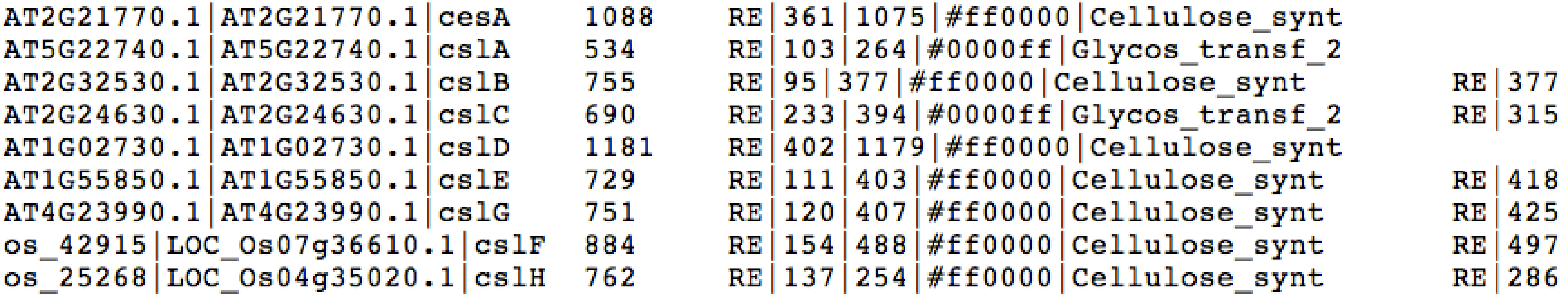
Gallus_gallus,
300,
EL|10|50|#ff0000|DUF17,
DI|200|290|#aaff00|DUF22

↓

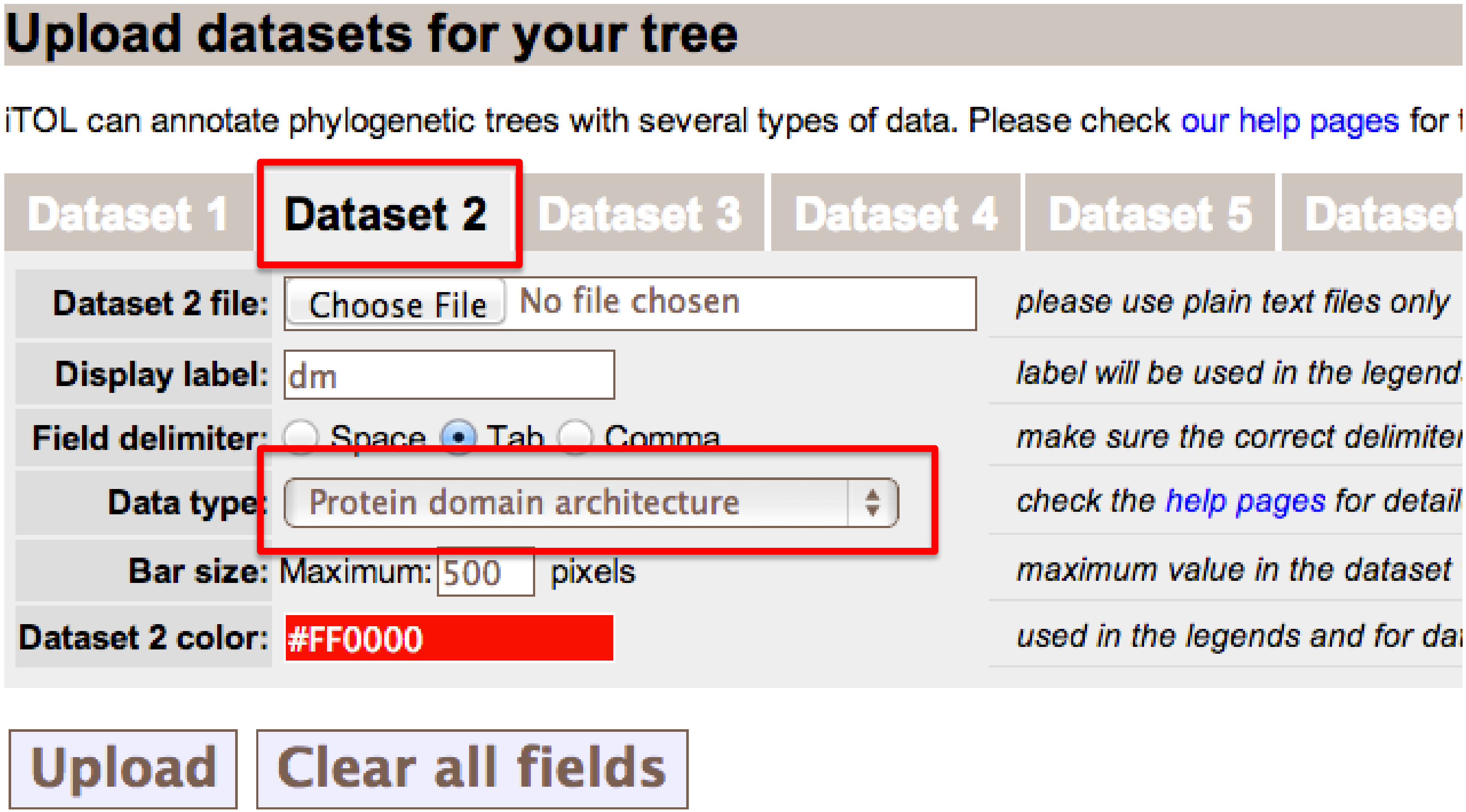
Shape|start|end|color|name

1. ID,
2. Full length,
3. Domain definition,
4. Domain definition,
5. …

http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.dm



```
AT2G21770.1|AT2G21770.1|cesA    1088    RE|361|1075|#ff0000|Cellulose_synt
AT5G22740.1|AT5G22740.1|cslA    534     RE|103|264|#0000ff|Glycos_transf_2
AT2G32530.1|AT2G32530.1|cslB    755     RE|95|377|#ff0000|Cellulose_synt      RE|377
AT2G24630.1|AT2G24630.1|cslC    690     RE|233|394|#0000ff|Glycos_transf_2    RE|315
AT1G02730.1|AT1G02730.1|cslD    1181    RE|402|1179|#ff0000|Cellulose_synt
AT1G55850.1|AT1G55850.1|cslE    729     RE|111|403|#ff0000|Cellulose_synt     RE|418
AT4G23990.1|AT4G23990.1|cslG    751     RE|120|407|#ff0000|Cellulose_synt     RE|425
os_42915|LOC_Os07g36610.1|cslF  884     RE|154|488|#ff0000|Cellulose_synt     RE|497
os_25268|LOC_Os04g35020.1|cslH  762     RE|137|254|#ff0000|Cellulose_synt     RE|286
```



Upload datasets for your tree

iTOL can annotate phylogenetic trees with several types of data. Please check our help pages for t

Dataset 2 file: Choose File No file chosen — *please use plain text files only*

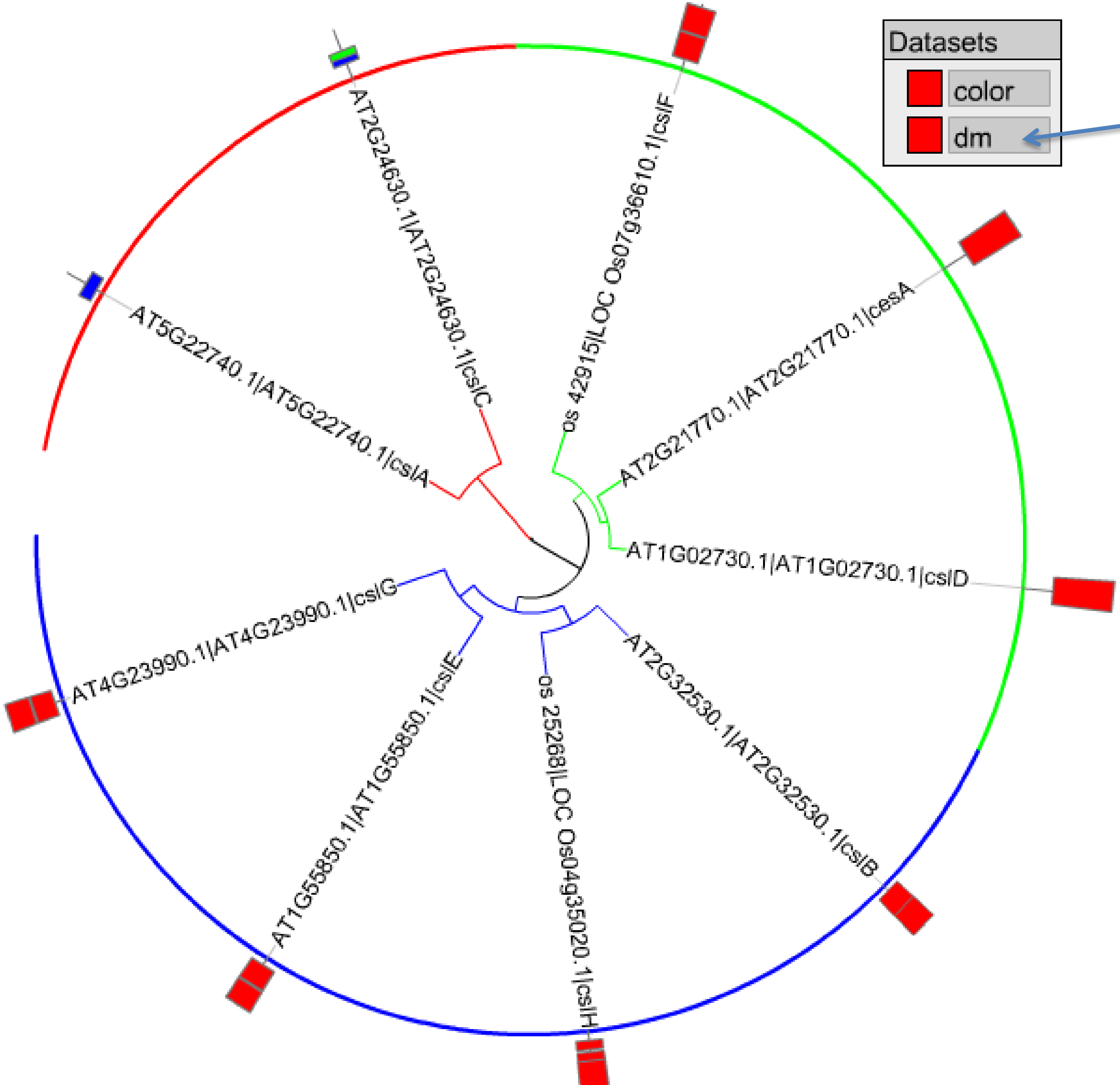
Display label: dm — *label will be used in the legend*

Field delimiter: ○ Space ● Tab ○ Comma — *make sure the correct delimiter*

Data type: Protein domain architecture — *check the help pages for detail*

Bar size: Maximum: 500 pixels — *maximum value in the dataset*

Dataset 2 color: #FF0000 — *used in the legends and for da*


Datasets
🟥 color
🟥 dm
AT2G24630.1|AT2G24630.1|cslC
os 42915|LOC Os07g36610.1|cslF
AT5G22740.1|AT5G22740.1|cslA
AT2G21770.1|AT2G21770.1|cesA
AT1G02730.1|AT1G02730.1|cslD
AT4G23990.1|AT4G23990.1|cslG
AT1G55850.1|AT1G55850.1|cslE
os 25268|LOC Os04g35020.1|cslH
AT2G32530.1|AT2G32530.1|cslB

Click here

Homework assignment 6

1. Take the MAFFT alignment http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fa.aln as input and use MEGA5 to build a phylogenetic tree

2. Try maximum likelihood (ML), neighbor-joining (NJ) and maximum parsimony (MP) algorithms with 100 bootstrap replications and compare the running time and the topology of the resulting trees. If encounter errors, try to use the HELP link to find out and solve it

3. Color the branches and leafs in the resulting ML tree graph using different colors for different gene subfamilies

Homework assignment 6 Cont.

4. Export the tree as a newick format file

5. Use the original sequence file in http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fa to calculate the lengths of C3H/C4H/F5H proteins (try to search "length" in galaxy server) and identify the Pfam domains in the C3H/C4H/F5H protein sequences; with the two results, prepare a domain definition file

6. Prepare a color definition file for different gene subfamilies (see step 3); upload the newick tree file, the color definition file and the domain definition file to iTOL to color display the tree

Write a report (in word or ppt) to include all the operations and screen shots.

Office hour:

Due on March 12 (send by email)

Tue, Thu and Fri 2-4pm, MO325A

Or email: yyin@niu.edu

Next class: Install linux