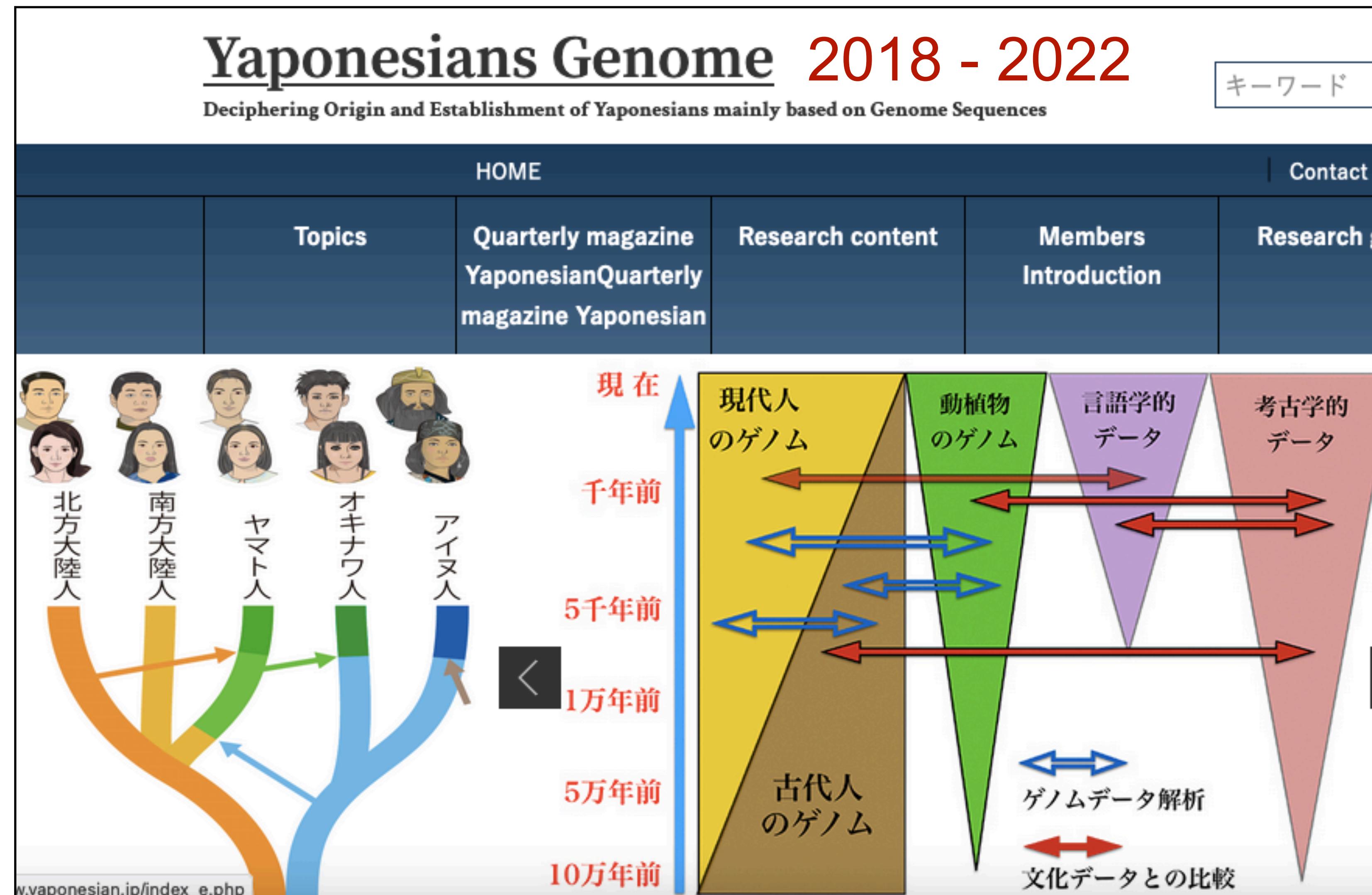
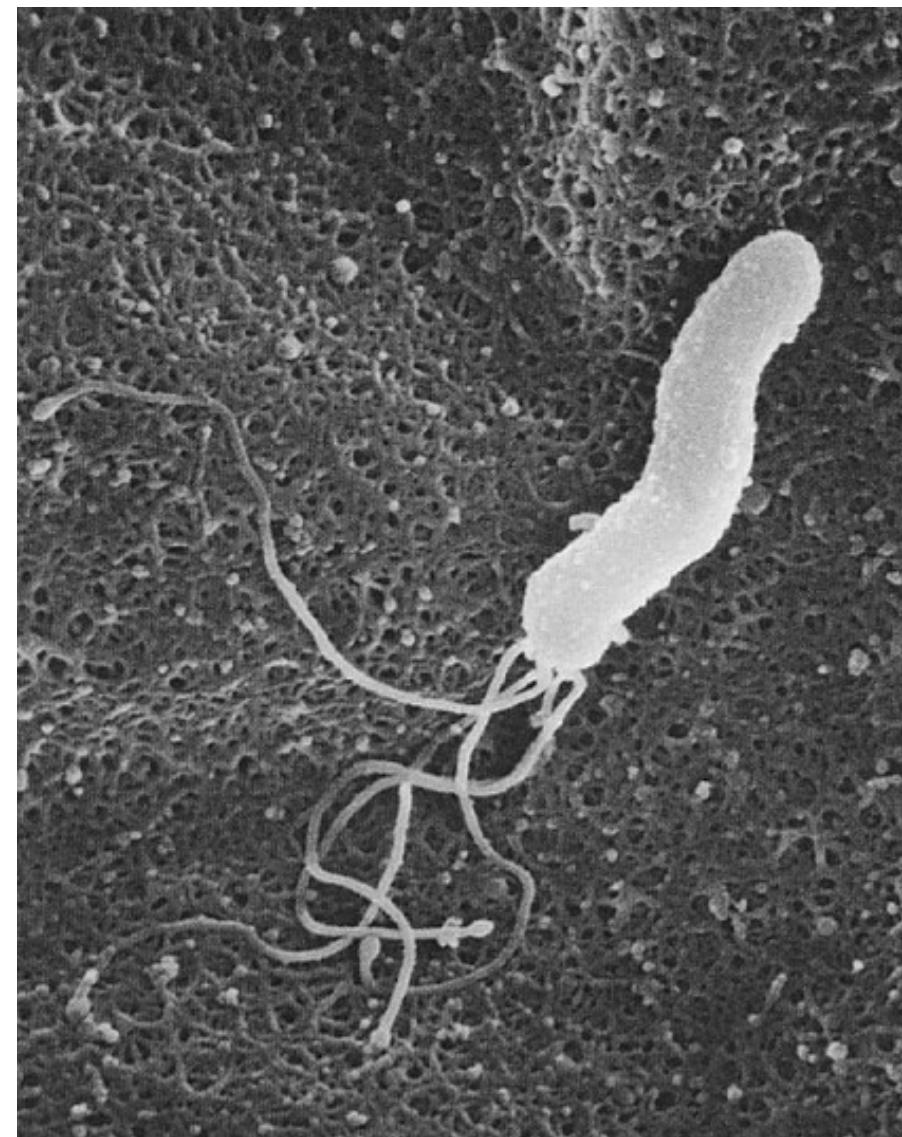


How to analyze bacterial population

2023.12.19
Rumiko Suzuki
National Institute of Genetics



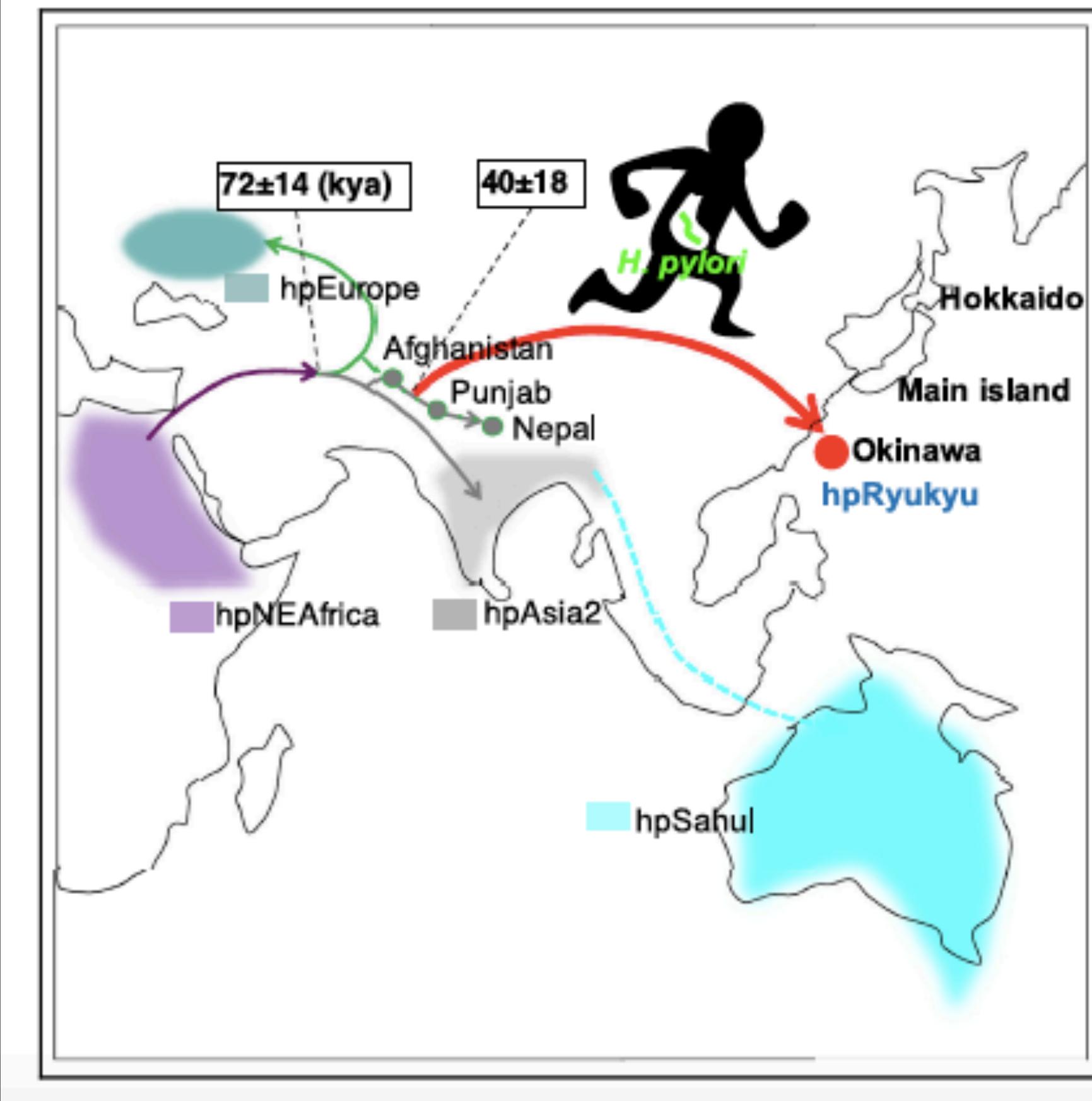
Helicobacter pylori



Wikipedia

Article

Helicobacter pylori genomes reveal Paleolithic human migration to the east end of Asia



Rumiko Suzuki,
Naruya Saitou,
Osamu Matsuo,
..., Kirill Kryukov,
Yoshan Moodley,
Yoshio Yamaoka

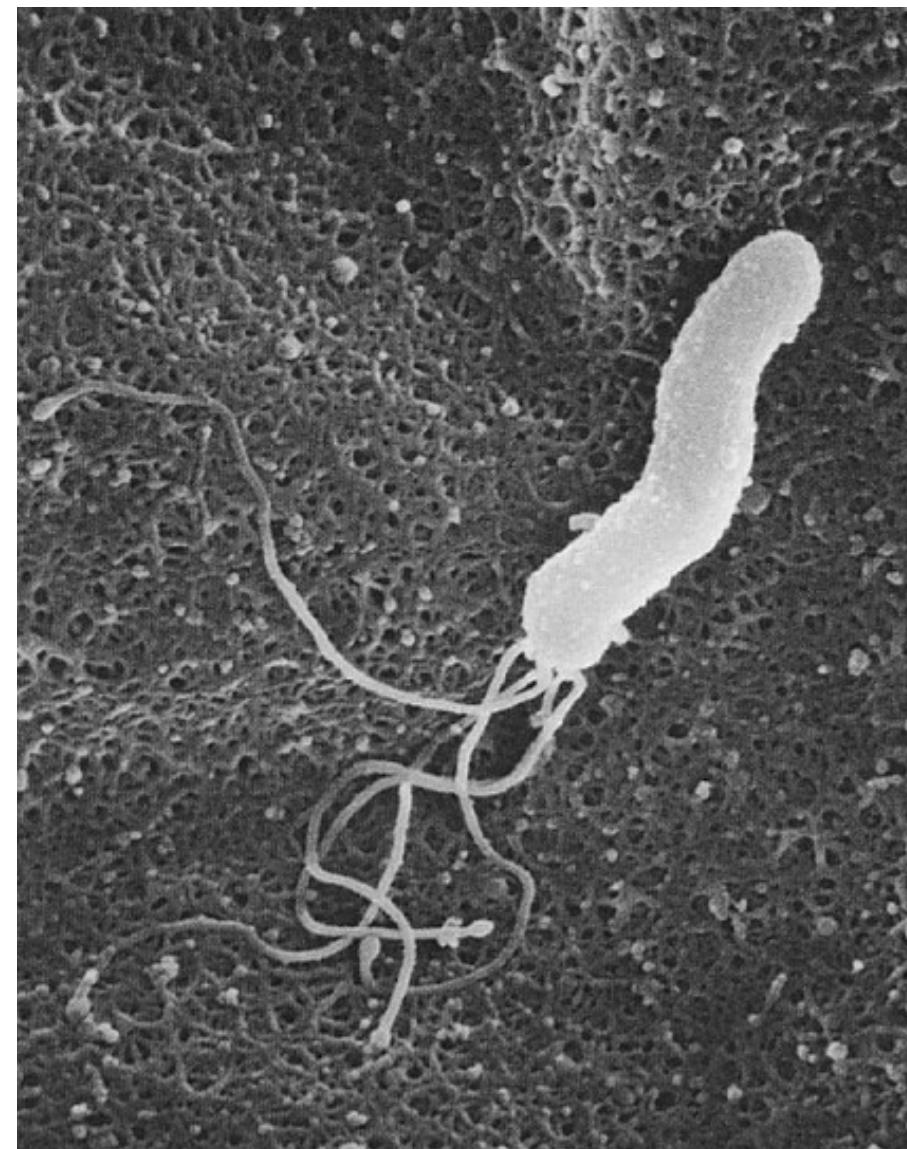
yyamaoka@oita-u.ac.jp

Highlights
H. pylori strains in
Okinawa, Japan, have a
different origin from the
main island

One of the Okinawa-
specific strains originate in
Western Asia in the
Paleolithic era

Our results suggest
ancient human migration
from west to the east end
of Eurasia

Helicobacter pylori



Wikipedia

Gram-negative bacterium that infects human stomach. World wide infection rate is around 50%.

Induces gastritis, stomach and duodenal ulcer/cancer.

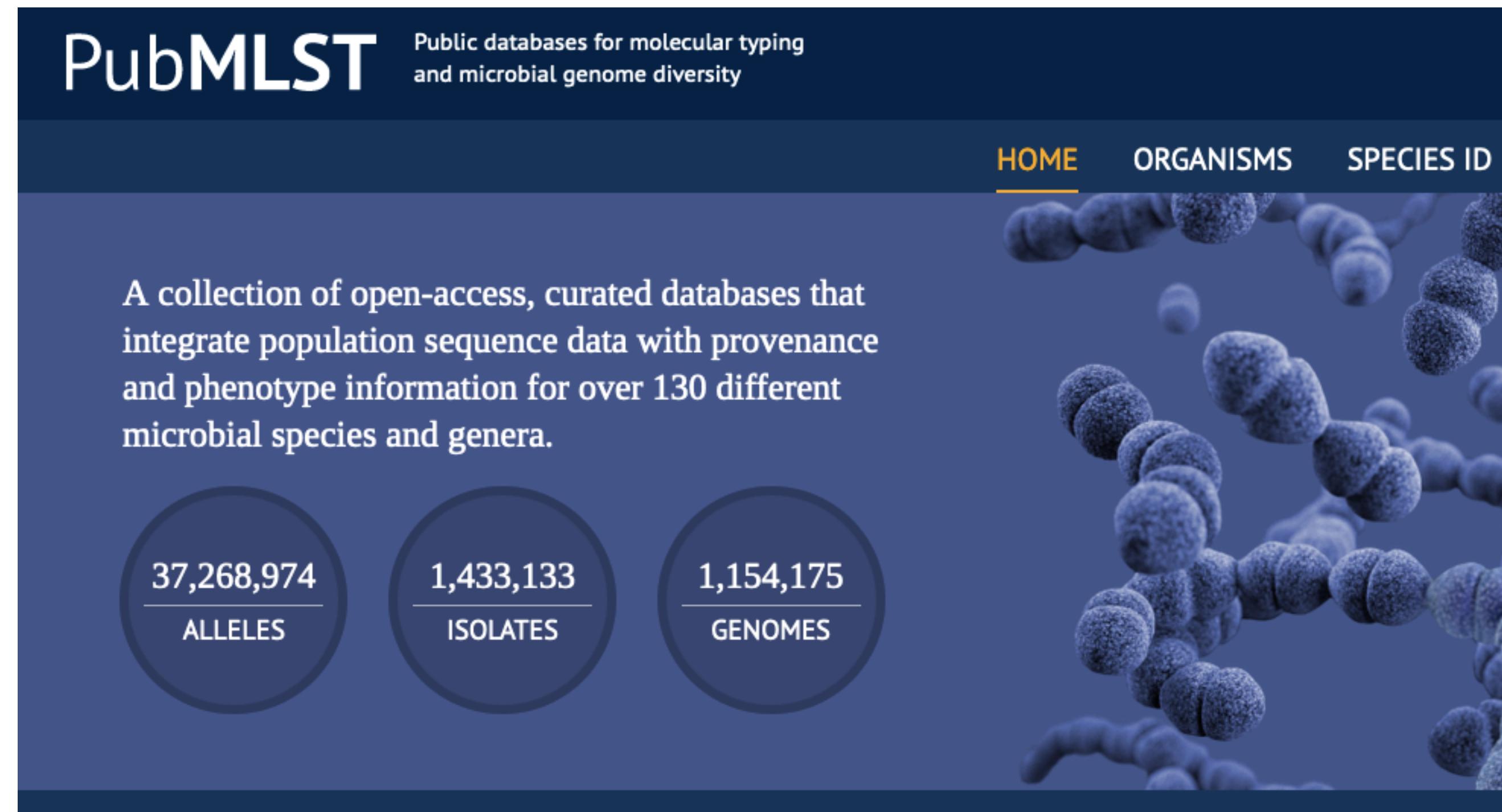
Mainly transmits within families in childhood while the immune system is immature.

Genome size 1.6 Mb, GC \sim 40%, high mutation rate and frequent recombination between different strains.

Conventional typing measure for bacteria

MLST (Multi Locus Sequence Typing)

Classify bacterial strains based on the DNA sequences of multiple housekeeping genes



The image shows the homepage of the PubMLST website. The header features the logo "PubMLST" and the subtitle "Public databases for molecular typing and microbial genome diversity". Below the header, there are three navigation links: "HOME" (which is highlighted in yellow), "ORGANISMS", and "SPECIES ID". The main content area contains a descriptive text: "A collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 130 different microbial species and genera." To the right of this text is a 3D rendering of a bacterial cell. At the bottom left, there are three circular icons with the following statistics: "37,268,974 ALLELES", "1,433,133 ISOLATES", and "1,154,175 GENOMES".

PubMLST

Public databases for molecular typing and microbial genome diversity

HOME ORGANISMS SPECIES ID

A collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 130 different microbial species and genera.

37,268,974 ALLELES

1,433,133 ISOLATES

1,154,175 GENOMES

Conventional typing measure for bacteria

MLST (Multi Locus Sequence Typing)

Partial or full sequences of 7 housekeeping genes are used for
H. pylori

Genes that exists every strain

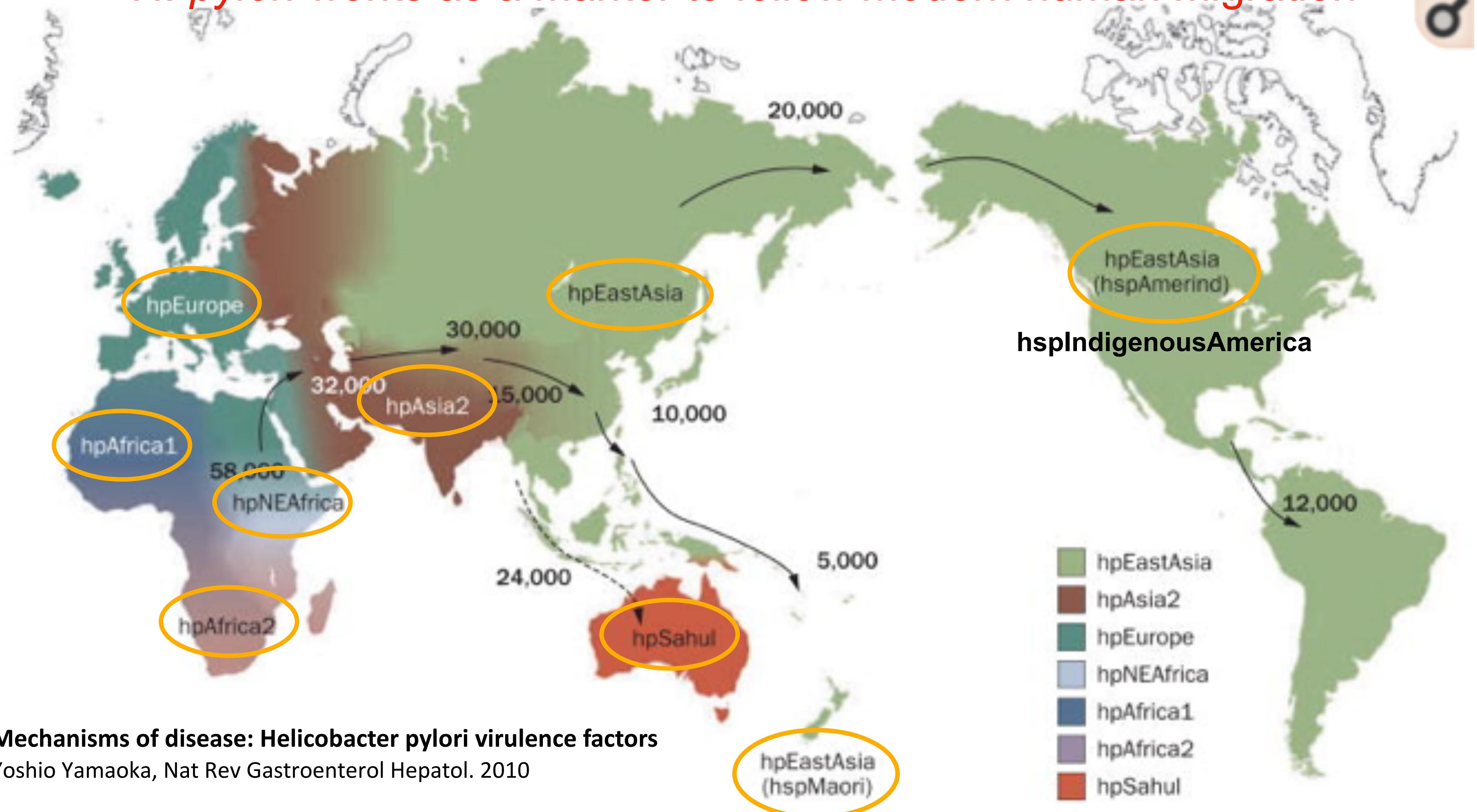
Readable by PCR

Not under positive or negative selection

Locus	Download	Type	Alleles	Length (setting)	Min length	Max length
atpA		DNA	3584	Fixed: 627 bp	626	628
efp		DNA	3346	Fixed: 410 bp	410	410
mutY		DNA	3691	Variable: (419 min; 420 max)	419	421
ppa		DNA	3389	Fixed: 398 bp	398	398
trpC		DNA	3722	Fixed: 456 bp	456	456
ureI		DNA	3706	Variable: (535 min; 585 max)	535	585
yphC		DNA	3818	Variable: (504 min; 531 max)	504	531

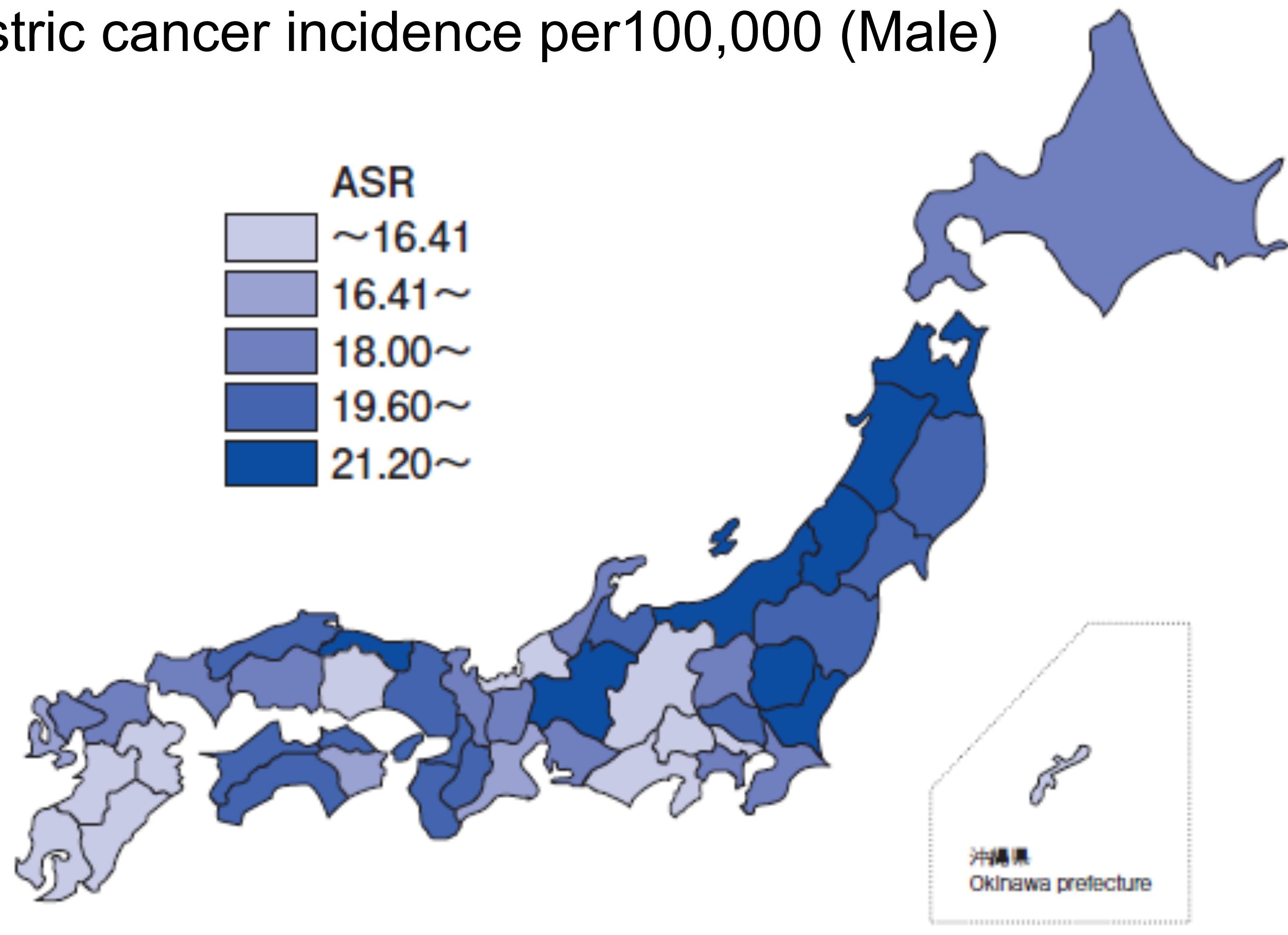
World populations (phylogeographic groups) of *H. pylori*

H. pylori works as a marker to follow modern human migration



H. pylori in Okinawa was different from other areas in Japan

Gastric cancer incidence per100,000 (Male)

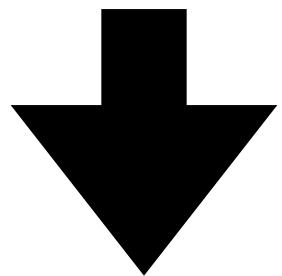


Gastric cancer incidence is low in Okinawa

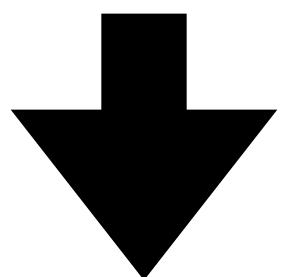
Japanese average	12.7
Okinawa	6.8
Kumamoto	9.0
Miyazaki	9.0
Kagoshima	9.0
Nagano	9.7
Yamanashi	9.8
Oita	9.9
Okayama	10.7
Shizuoka	10.8
...	...

Today's program

Construct a phylogenetic tree by MLST sequences of *H. pylori*



Calculate a distance matrix of the sequences



Execute Principle Component Analysis and plot the result

Construct a phylogenetic tree by MLST sequences of *H. pylori*

MEGA
Molecular Evolutionary Genetics Analysis
software

Open MLST sequence data from

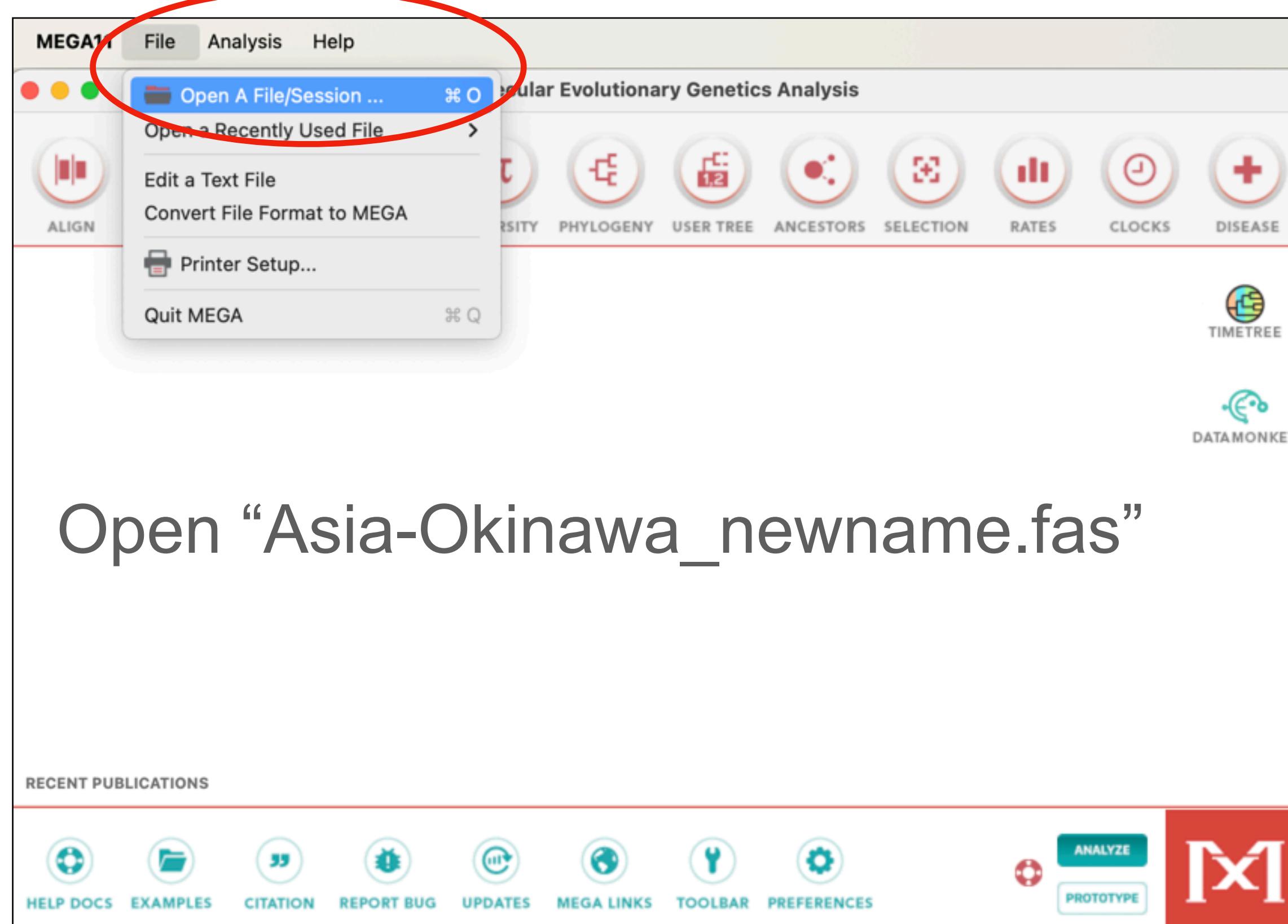
http://sayer.nig.ac.jp/rusuzuki/keio/Asia-Okinawa_newname.fas

Save the data as “**Asia-Okinawa_newname.fas**”

★Extension “.txt” may be automatically added and the file saved as
“Asia-Okinawa_newname.fas.txt.” In that case, delete “.txt” from the file
name.

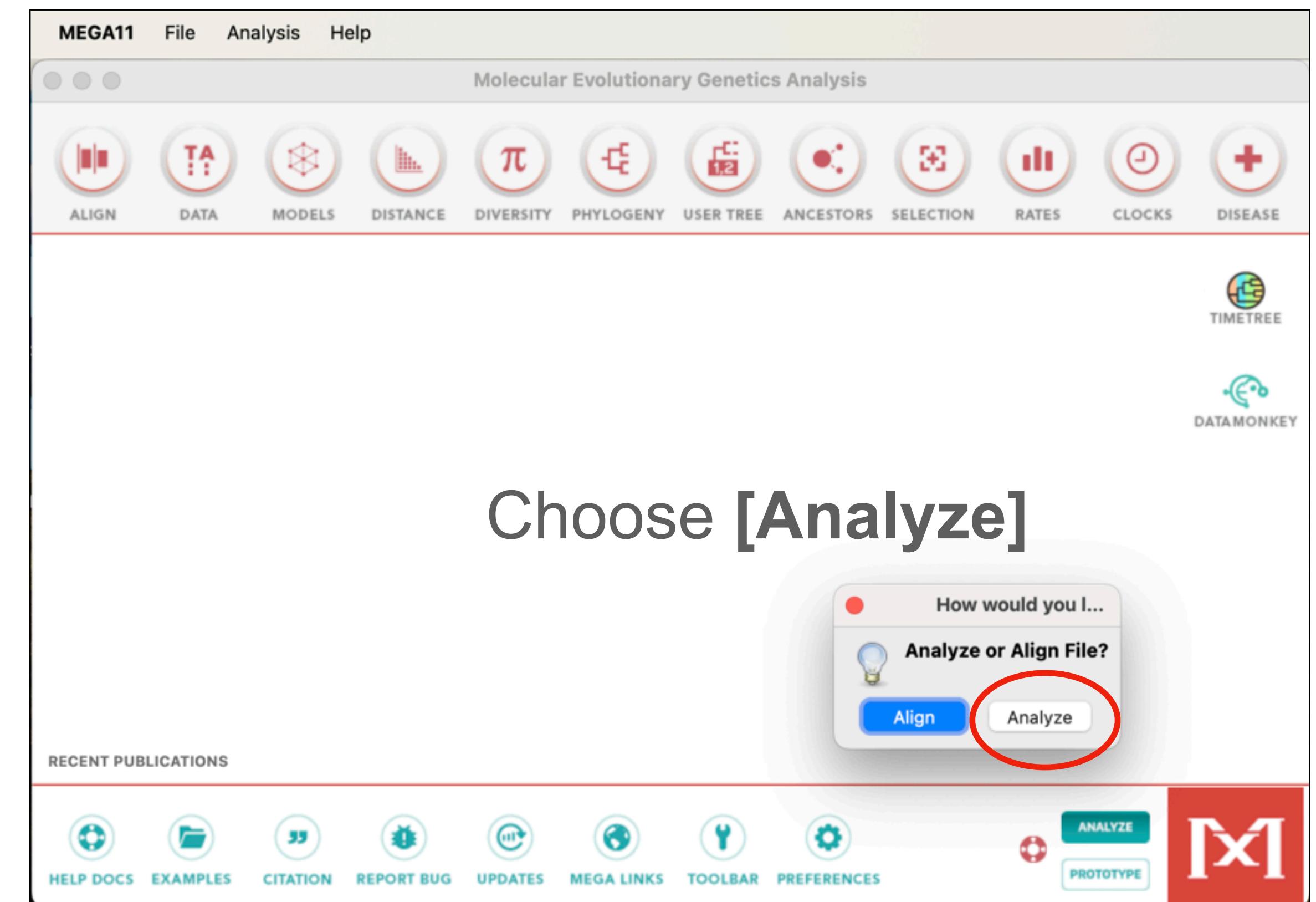
Import the sequence data to MEGA

[File] -> [Open A File/Session]



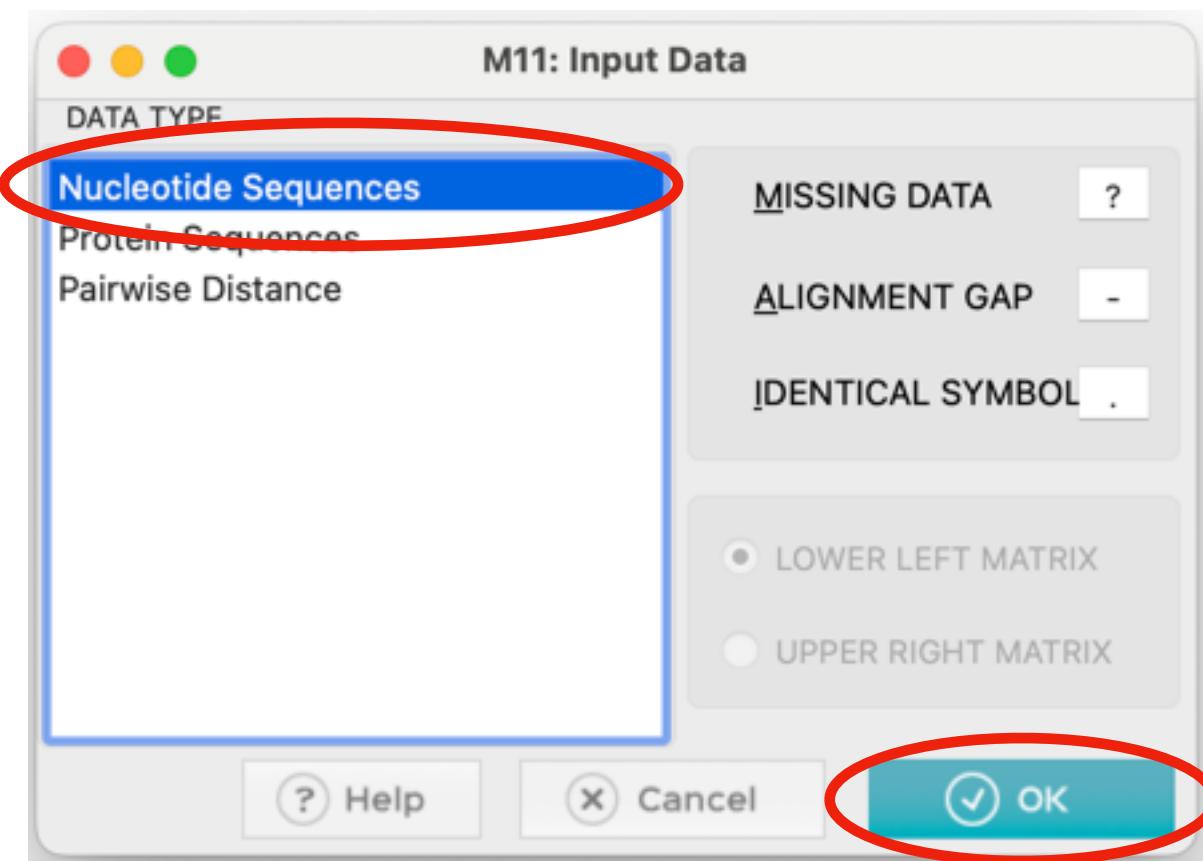
Open “Asia-Okinawa_newname.fas”

The sequences have been already aligned.



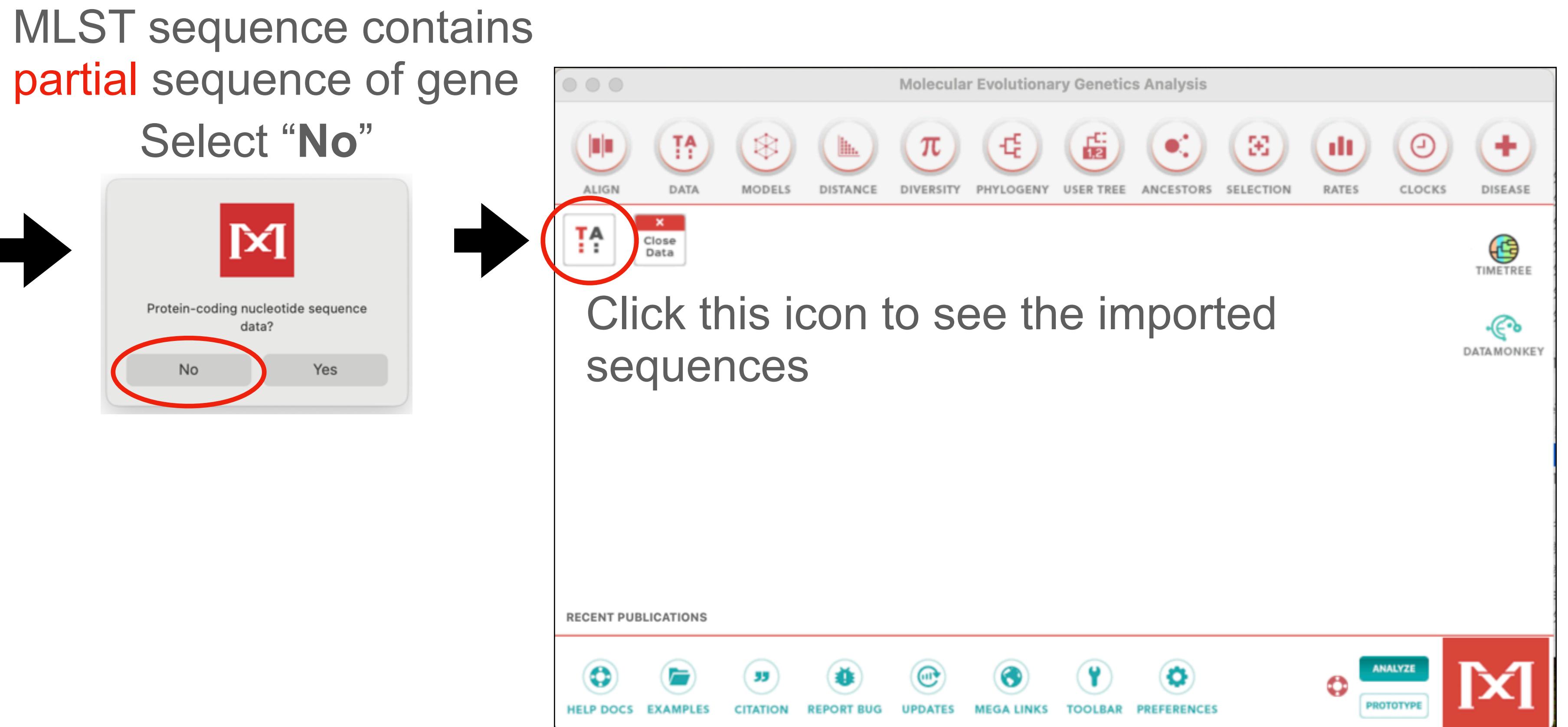
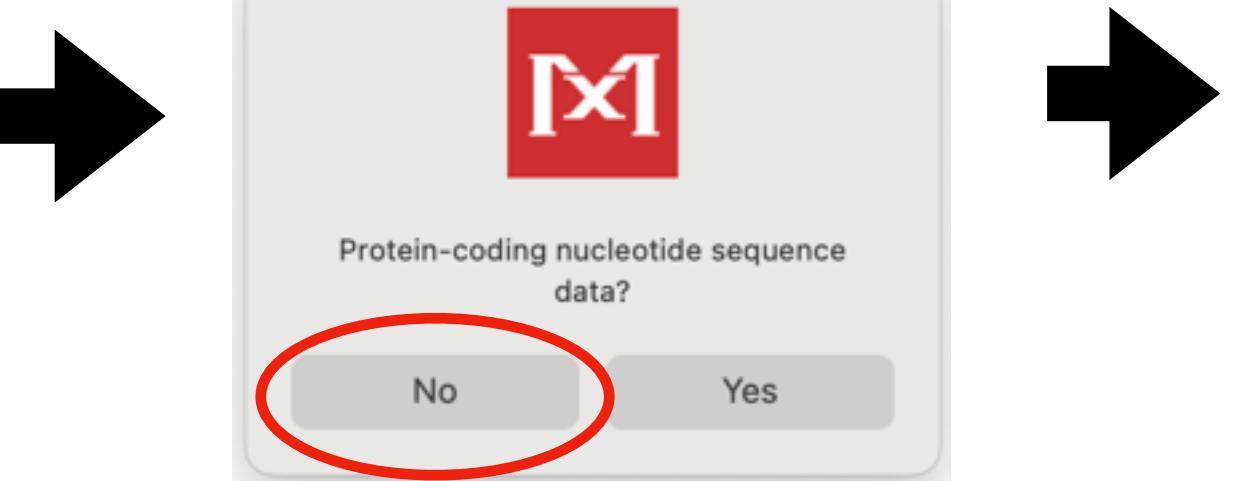
Choose [Analyze]

Choose
[Nucleotide Sequences]



MLST sequence contains
partial sequence of gene

Select “No”



You can choose to use (active) or to exclude sequences by this box

“.” represents the same nucleotide as the first line

The screenshot shows a sequence alignment interface with the following features:

- Header:** M11: Sequence Data Explorer (Asia-Okinawa_newname.fas)
- Left Panel (Taxa List):** A list of 16 samples with checkboxes. The first sample, "1. Af201", has its checkbox circled in red and has a blue vertical bar to its left.
- Sequence View:** A grid of sequence data. The first column is labeled "Site" and shows positions 1 through 3406. The second column is labeled "Name". The third column shows the sequence for "1. Af201" (highlighted with a red box), which starts with "a a t g a a t t a g c c t t a t a g a g c a a a a g g c t c c g g g c a t t a t g g a c". Subsequent columns show the sequence for other samples, with most positions being ".". A red arrow points from the text "“.” represents the same nucleotide as the first line" to the "a" at position 1 of the first sequence.
- Bottom Navigation:** Includes "Site" (set to 1), "Highlighted: None", "326 taxa selected", and "Data".

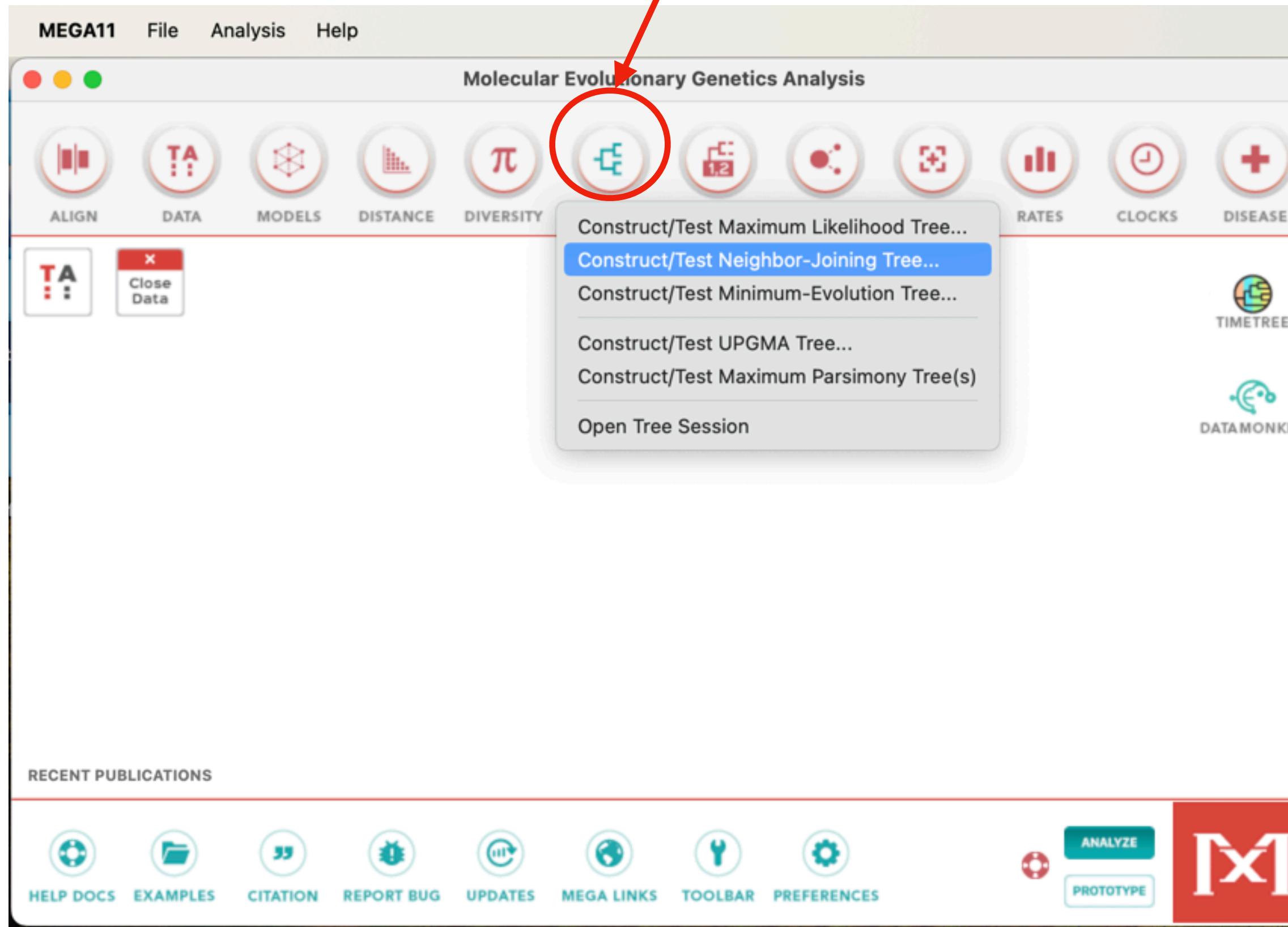
Sequence length

Number of samples

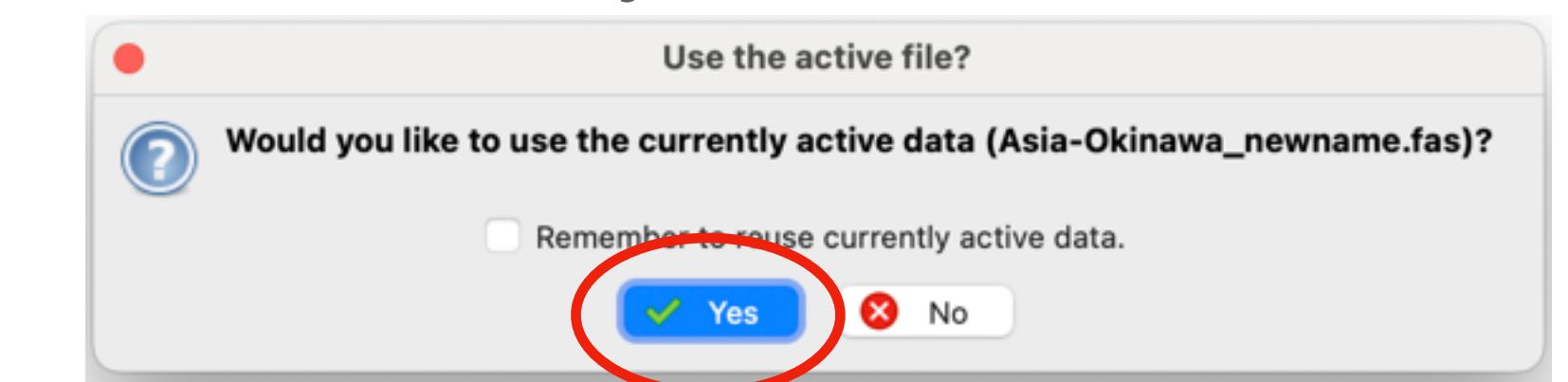
*

Construct a phylogenetic tree

Click the “Phylogeny” icon and choose
[Construct/Test Neighbor-Joining Tree]



Use currently active data : “Yes”



Set the parameters

M11: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Scope	→ All Selected Taxa
Statistical Method	→ Neighbor-joining
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Model/Method	→ Kimura 2-parameter model
Substitutions to Include	→ d: Transitions + Transversions
RATES AND PATTERNS	
Rates among Sites	→ Uniform Rates
Gamma Parameter	→ Not Applicable
Pattern among Lineages	→ Same (Homogeneous)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Complete deletion
Site Coverage Cutoff (%)	→ Not Applicable
SYSTEM RESOURCE USAGE	
Number of Threads	→ Not Applicable

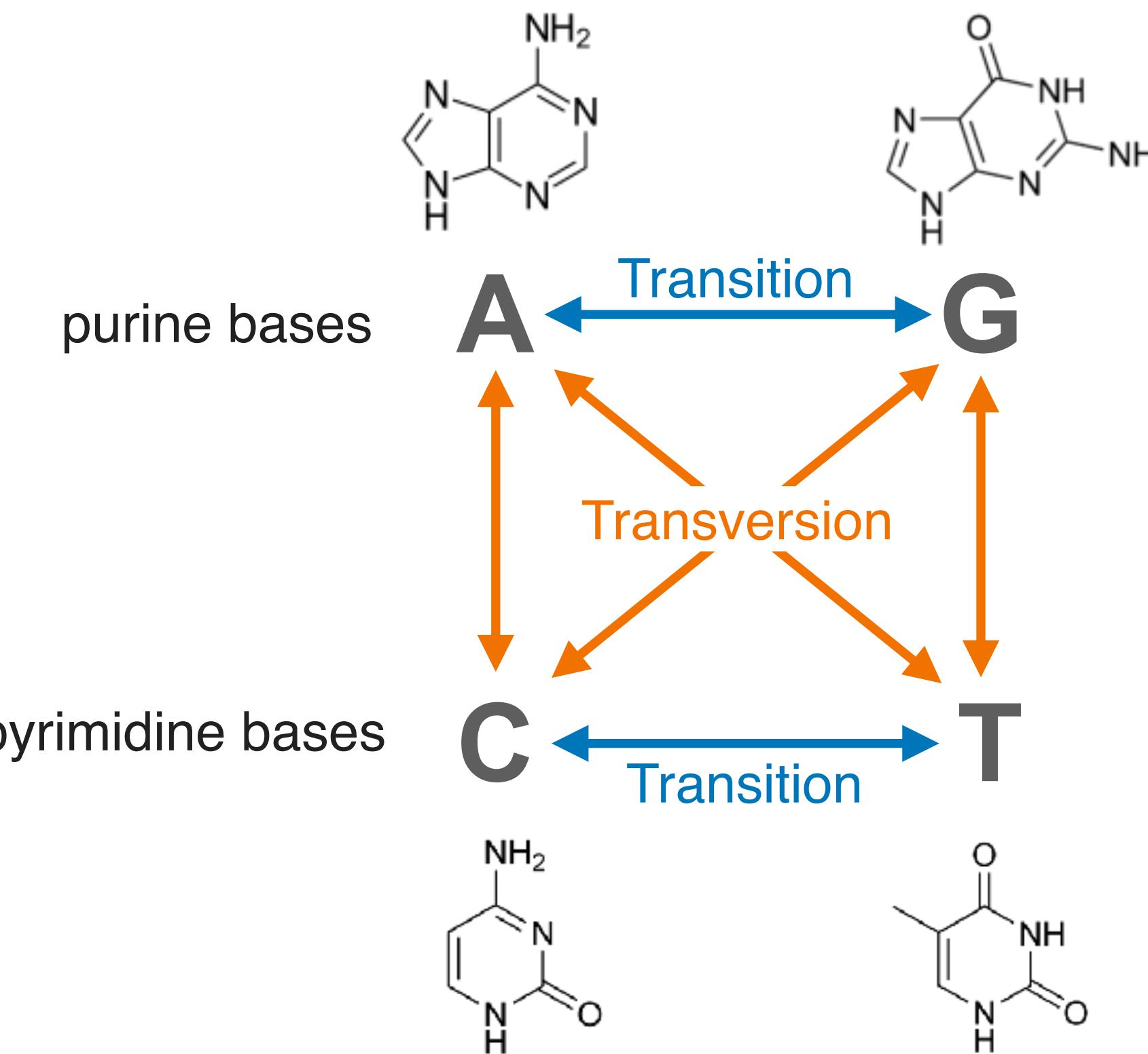
?

Help

X Cancel

OK

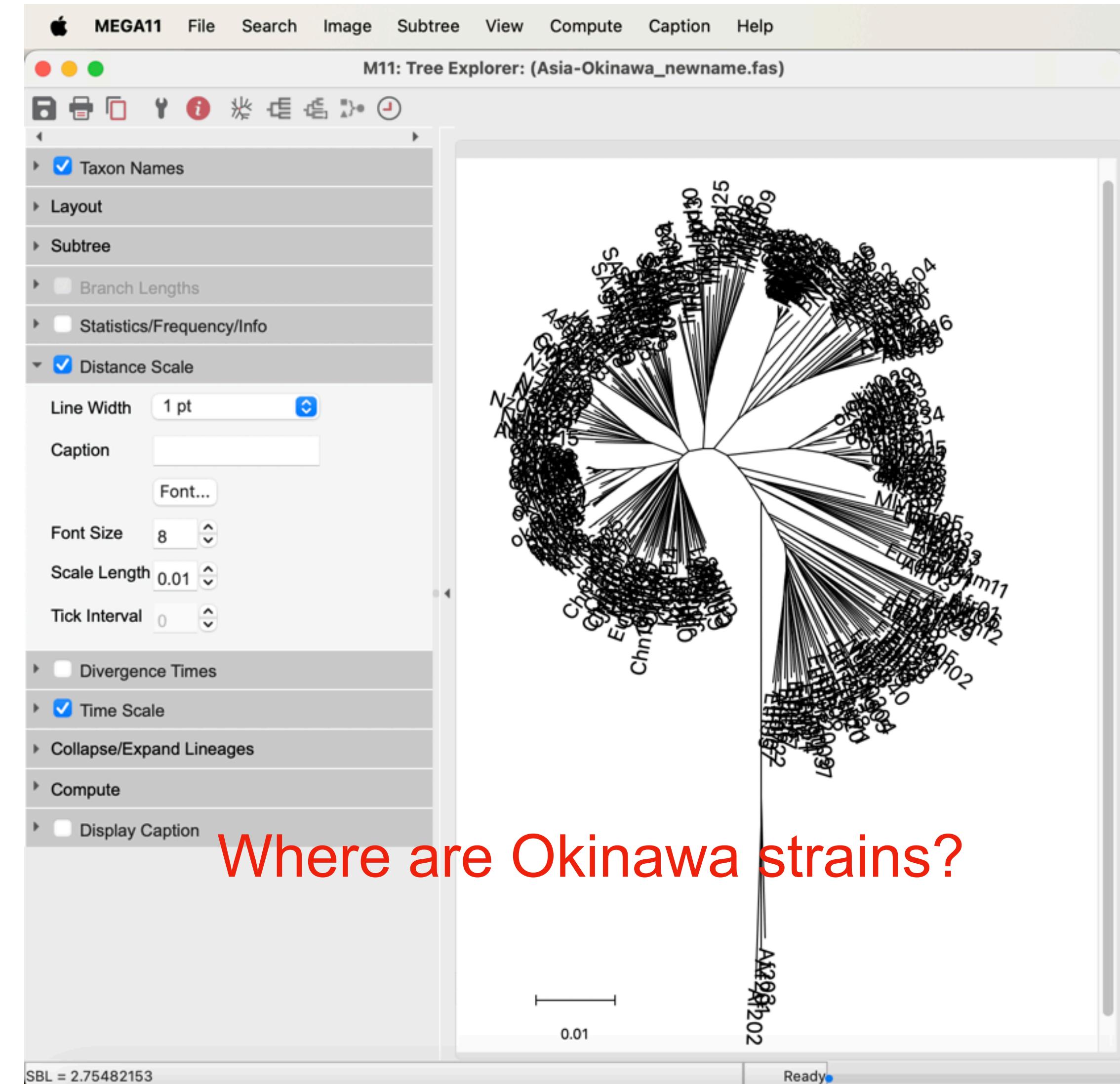
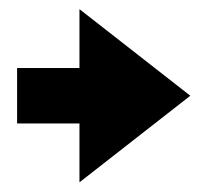
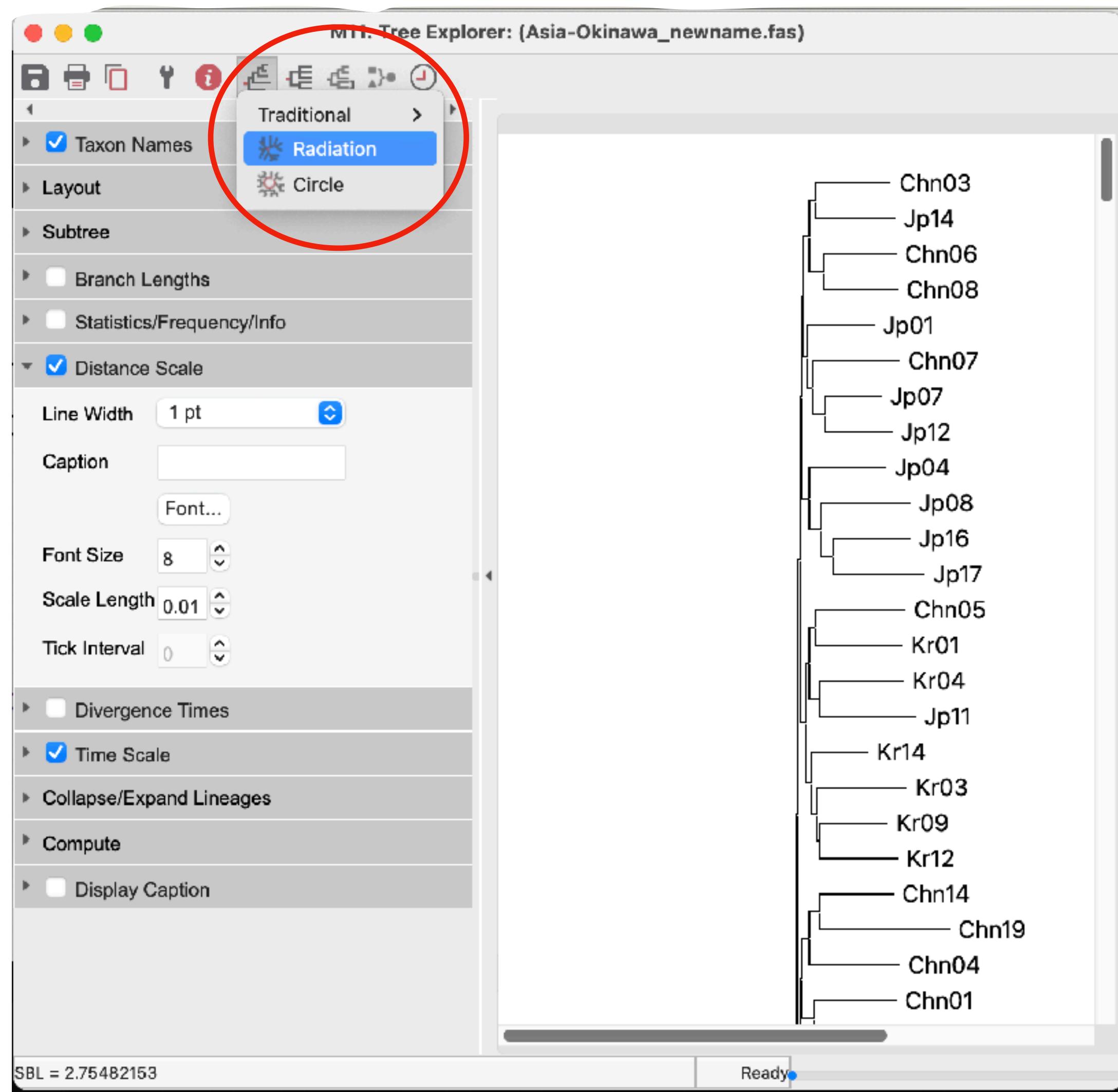
SUBSTITUTION MODEL



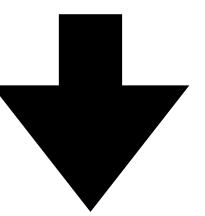
Kimura 2-parameter (= transition and transversion) model assigns longer distances for transvehrsion substitutions than transition substitutions.

Click the 6th icon from the left to change the tree format

Choose [Radiation]



Download Okinawa group data from
[http://sayer.nig.ac.jp/rusuzuki/keio/
group_Okinawa.txt](http://sayer.nig.ac.jp/rusuzuki/keio/group_Okinawa.txt)

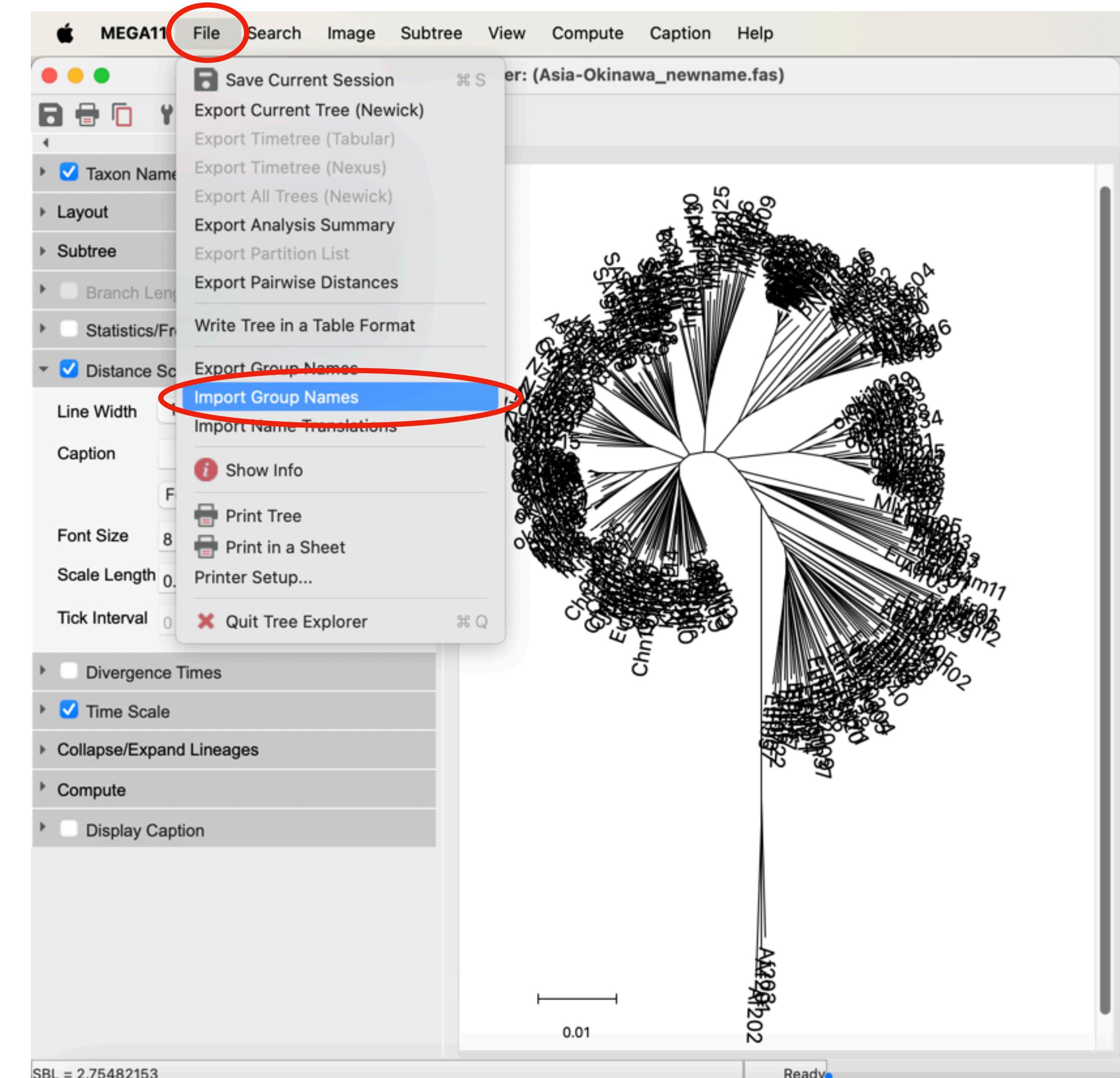


group_Okinawa.txt

oki01=Okinawa
oki10=Okinawa
oki102=Okinawa
oki104=Okinawa
oki106=Okinawa
oki109=Okinawa
oki11=Okinawa
oki112=Okinawa
oki114=Okinawa

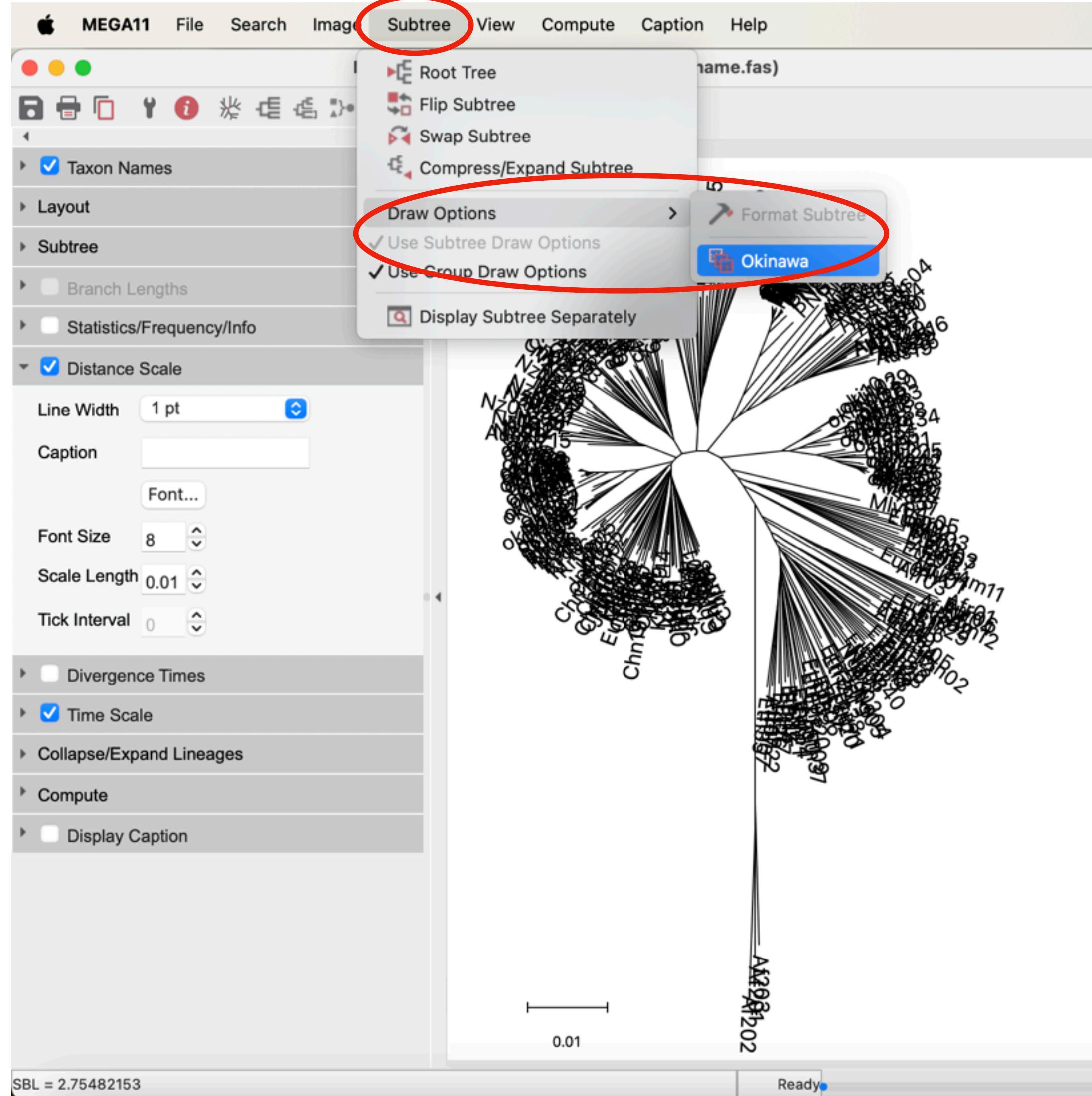
1

Import group data
[File] -> [Import Group Names]

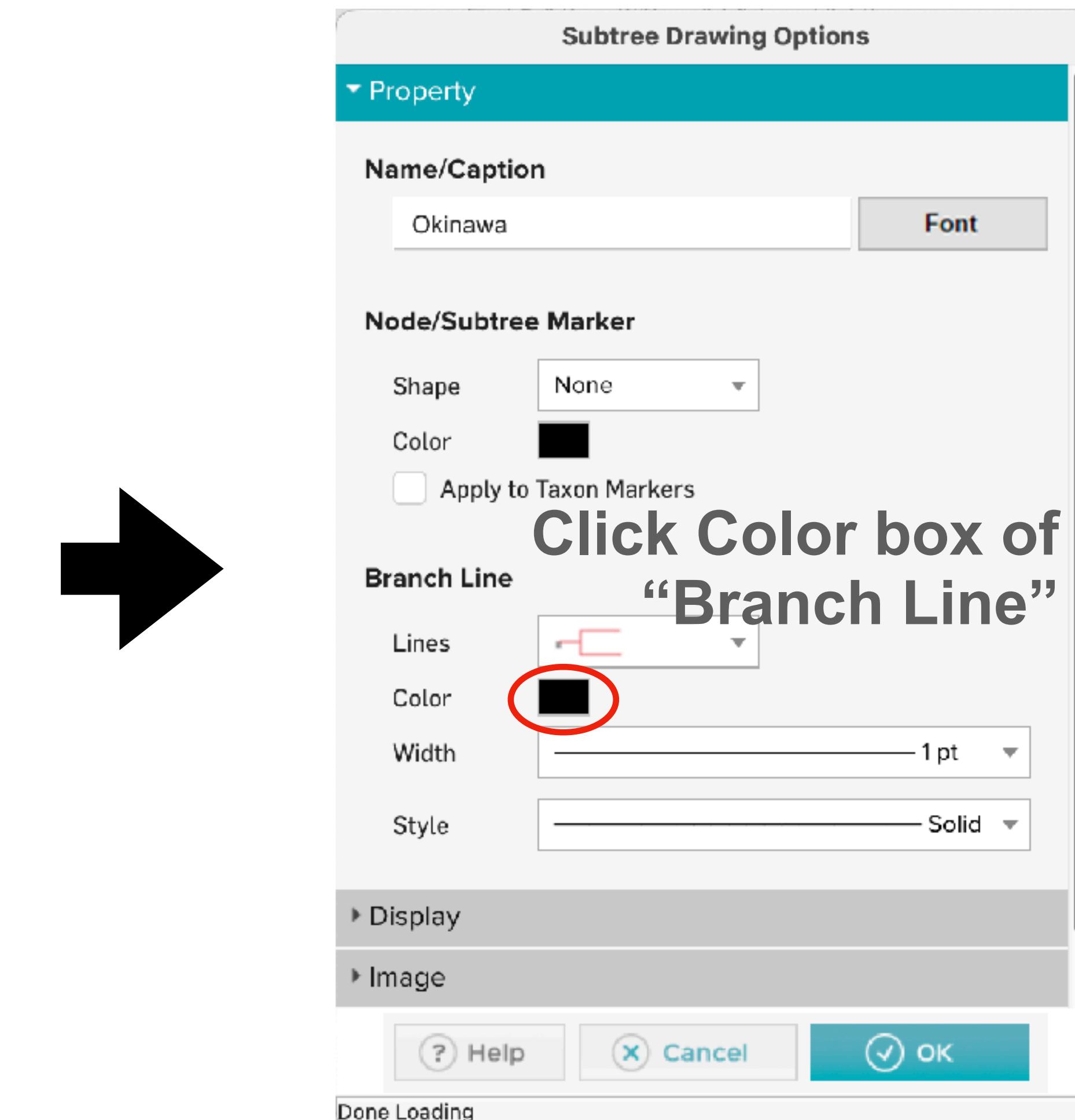


Color the Okinawa branches

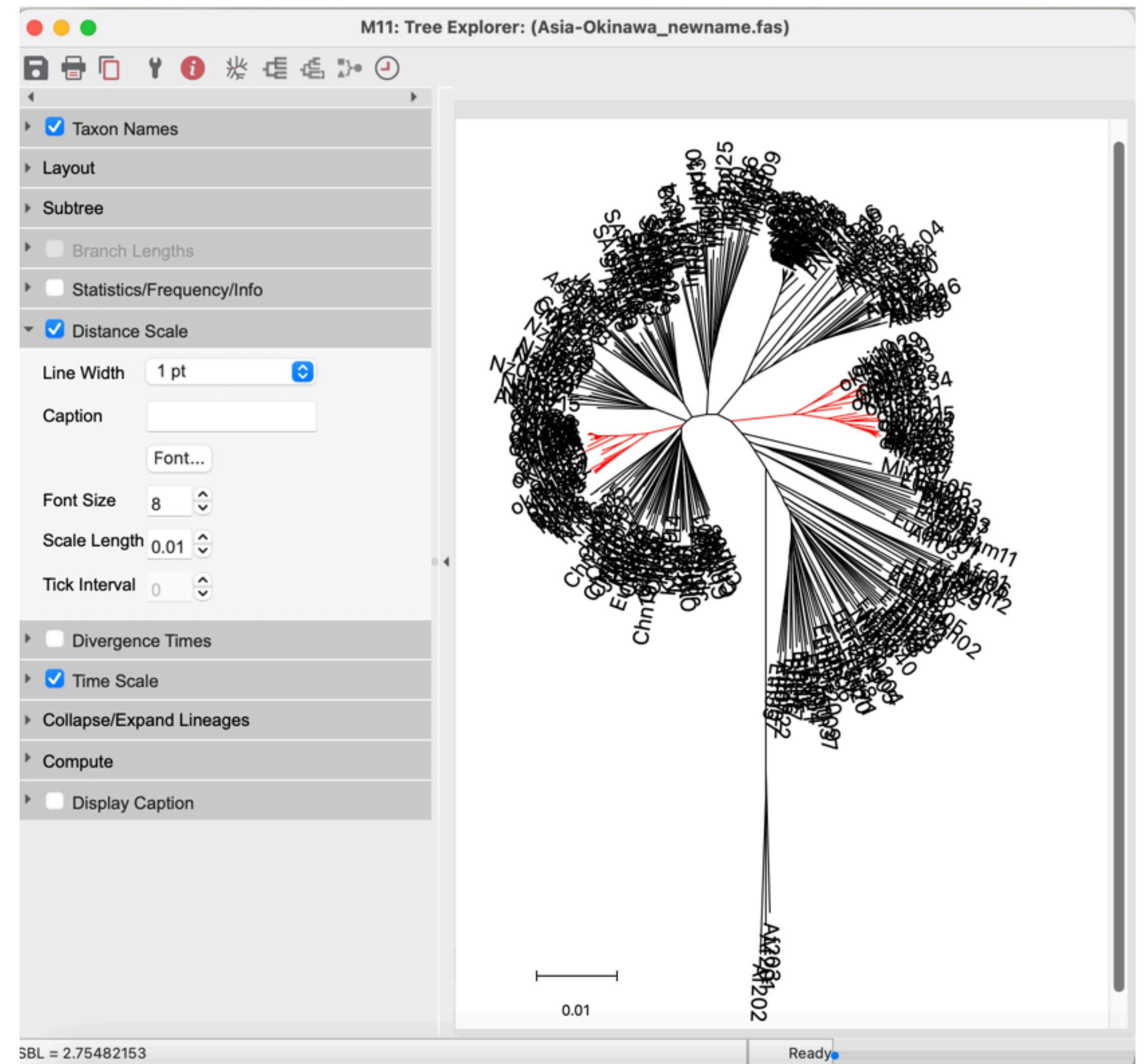
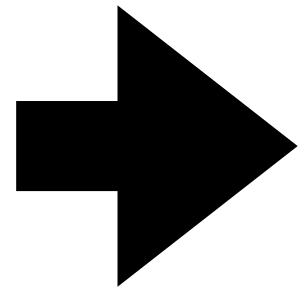
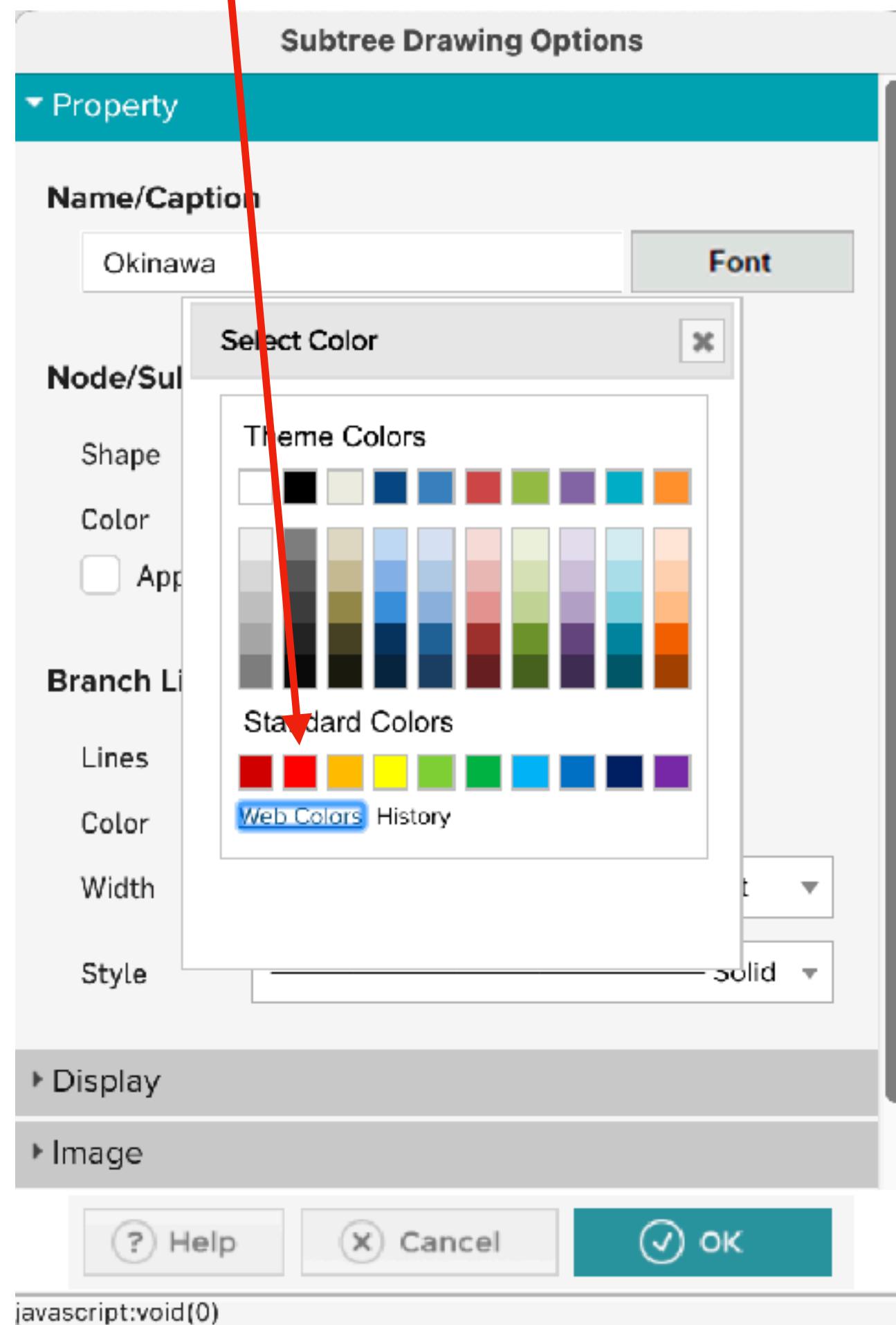
[Subtree] -> [Draw Options] -> [Okinawa]



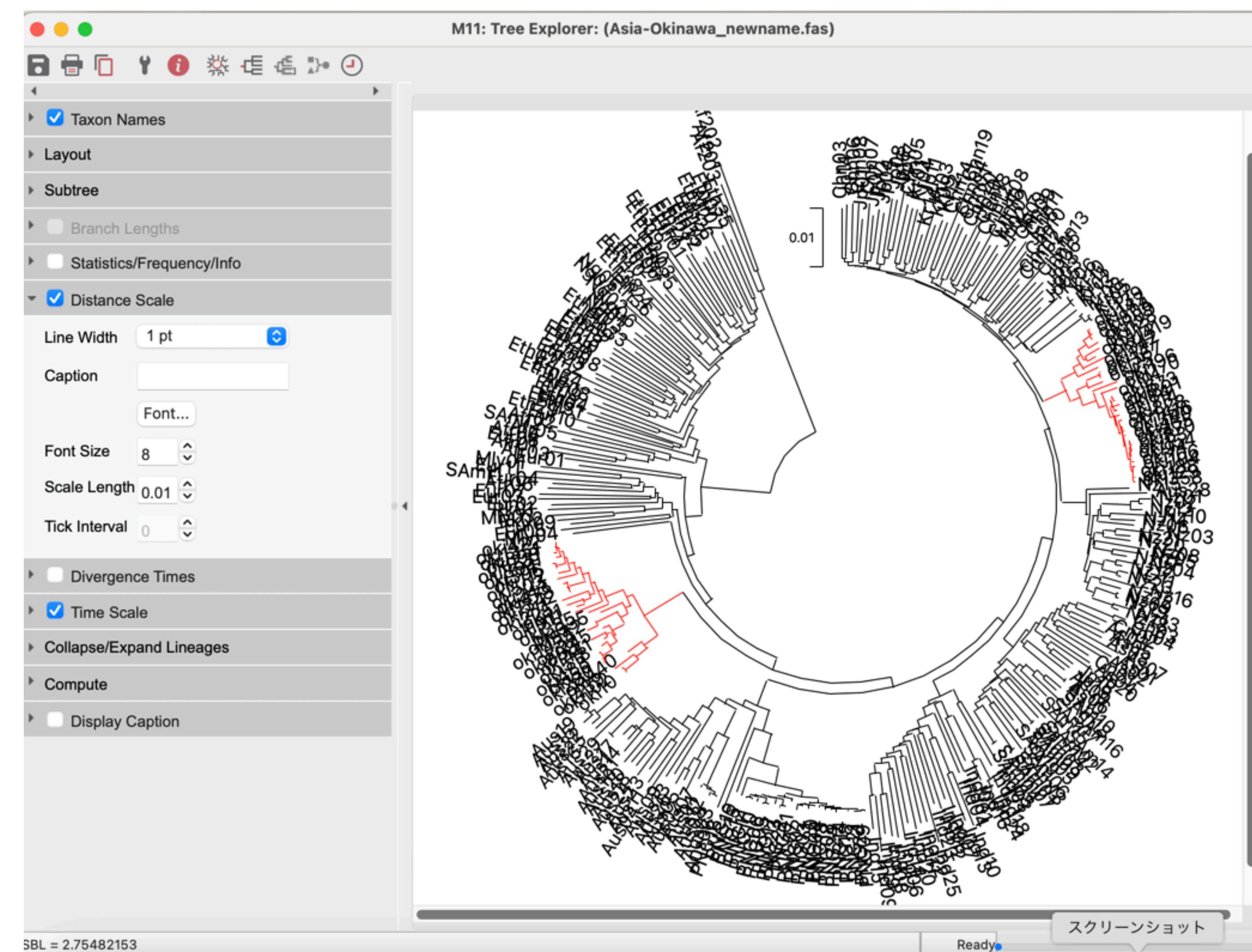
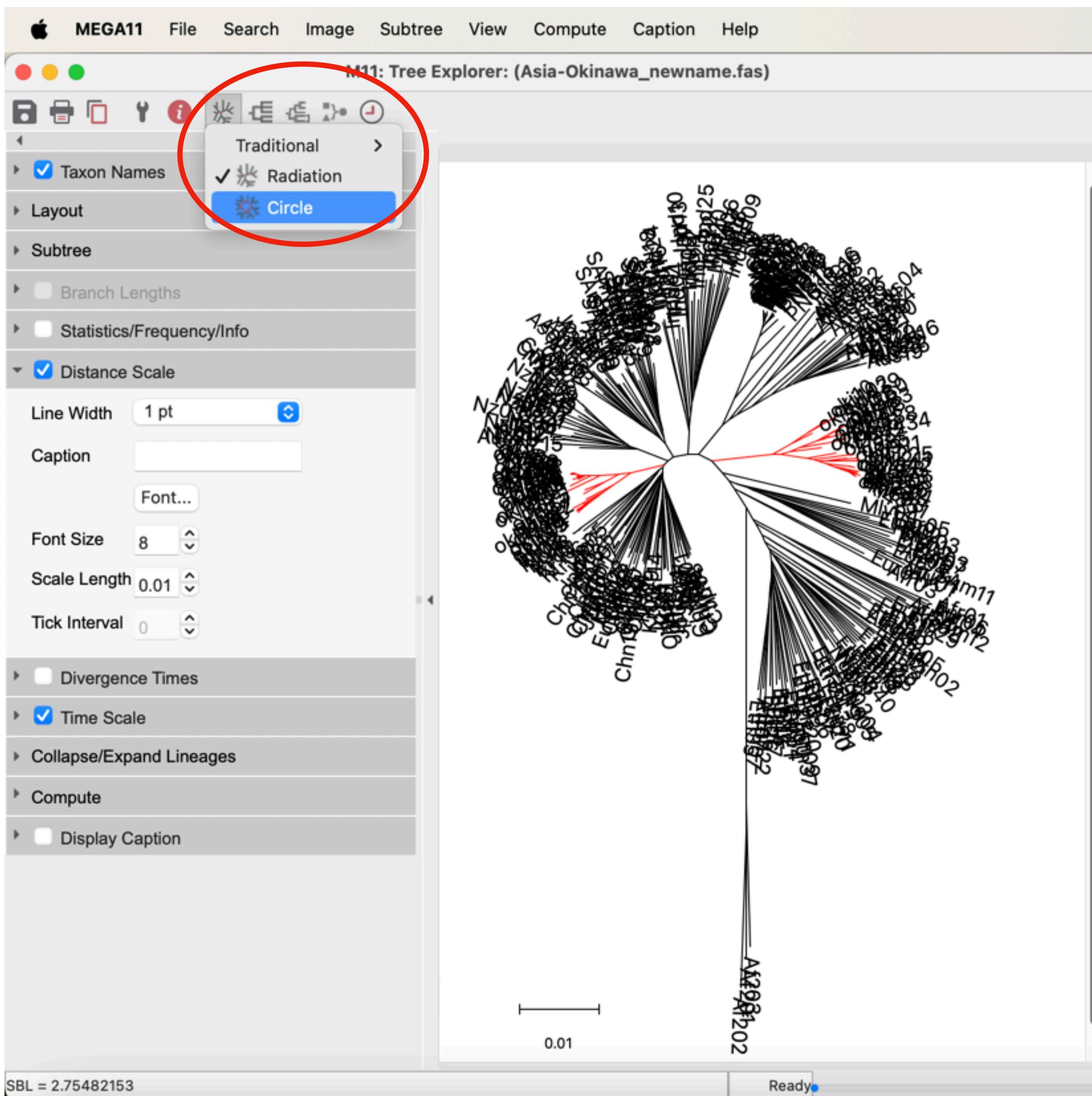
Change the branch color



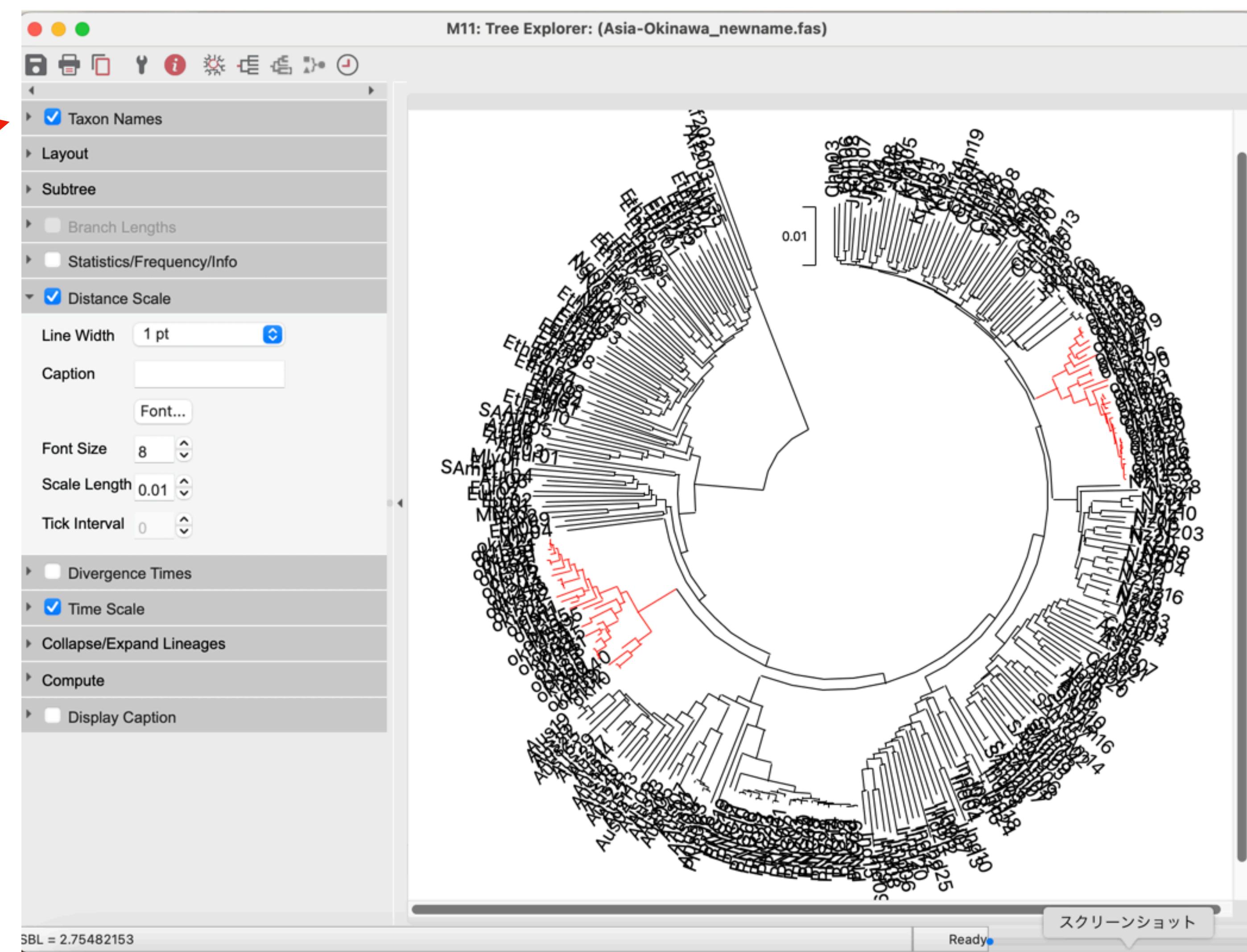
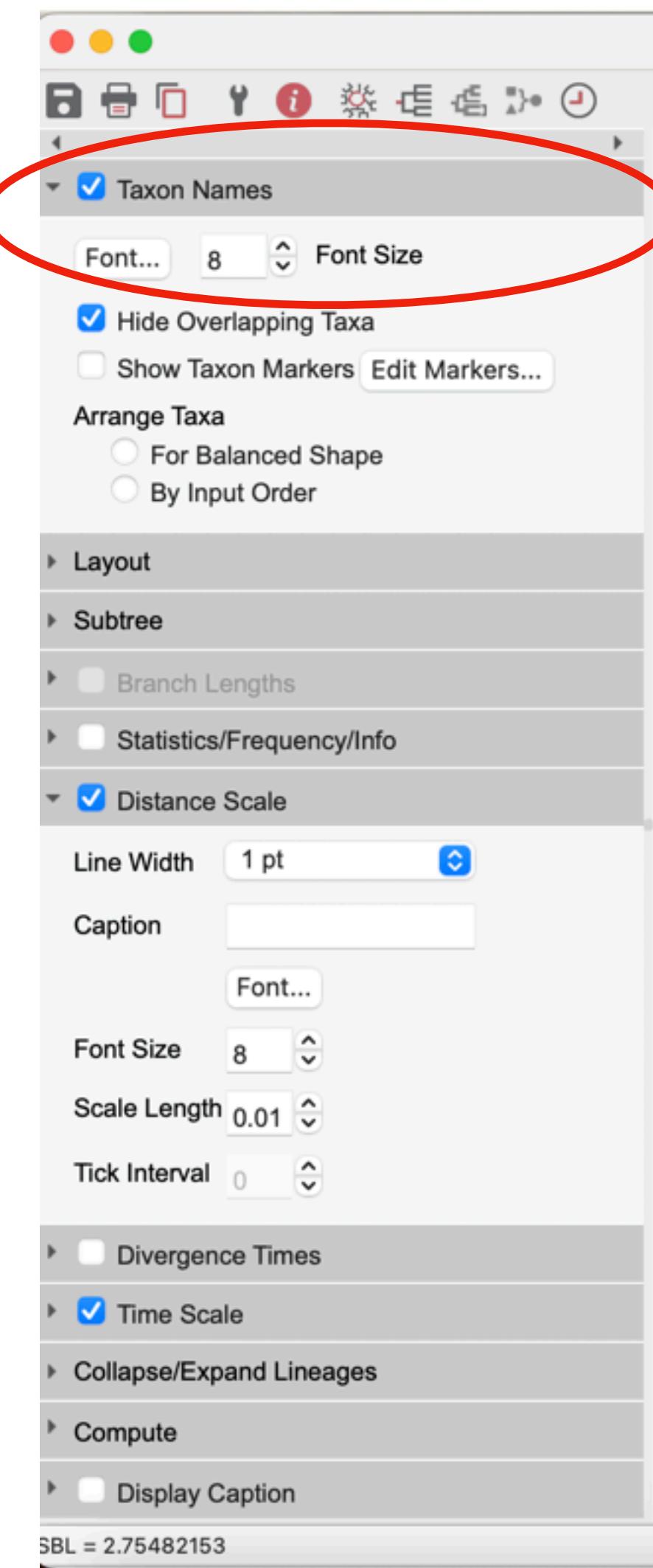
Choose a color
Click the red box (or any color you like)

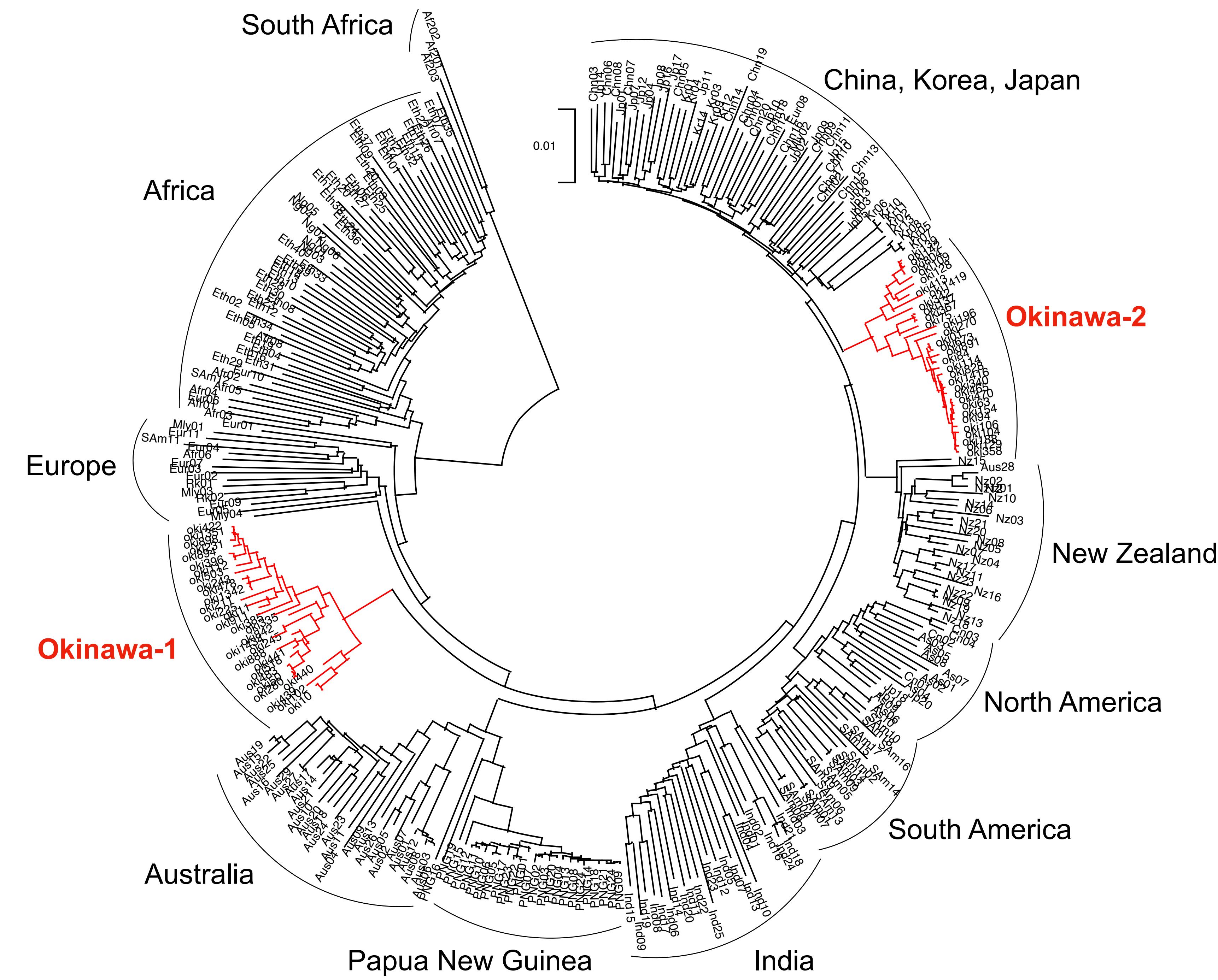


Change the tree shape to circle



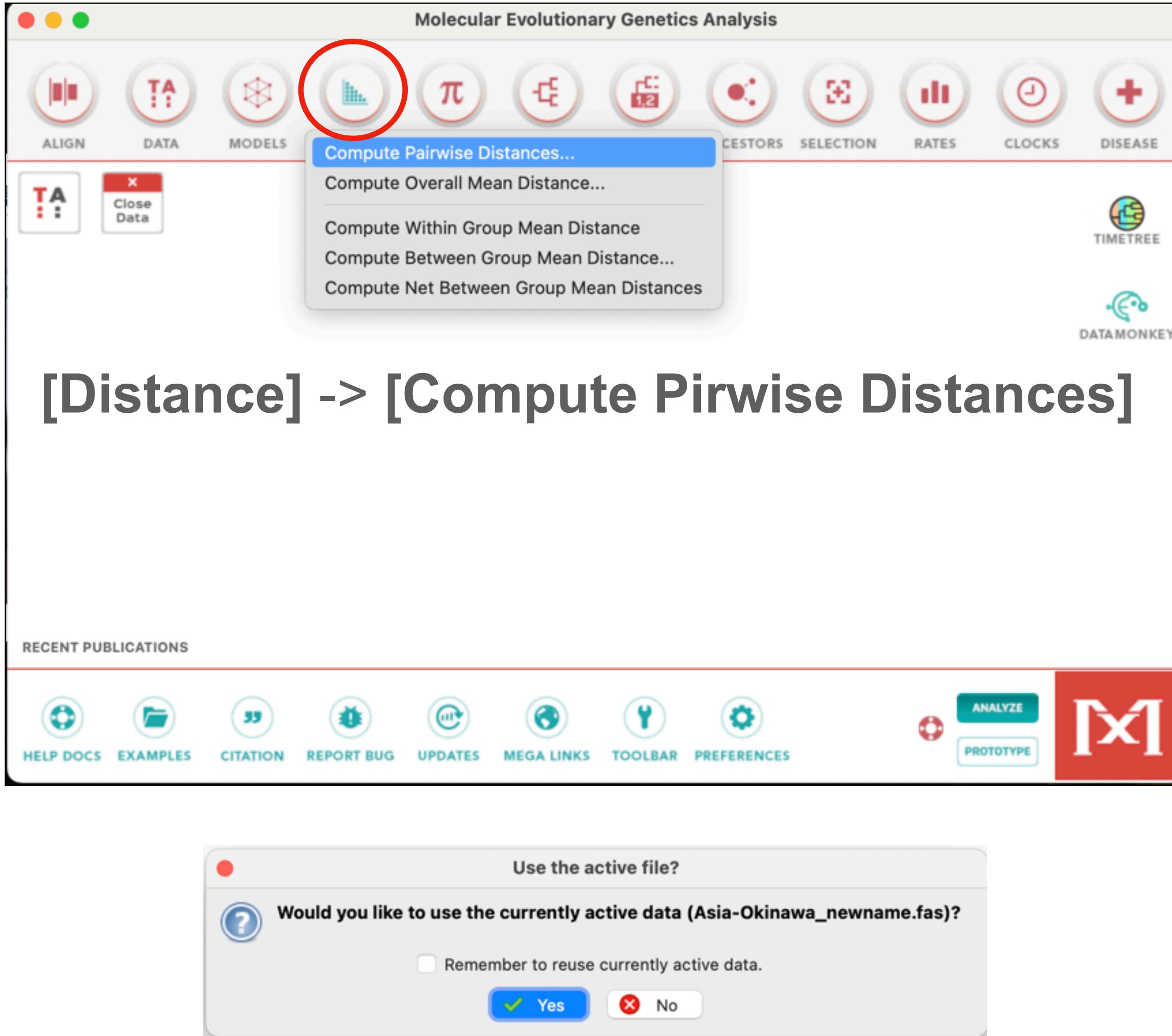
You can modify the font size of the taxon names



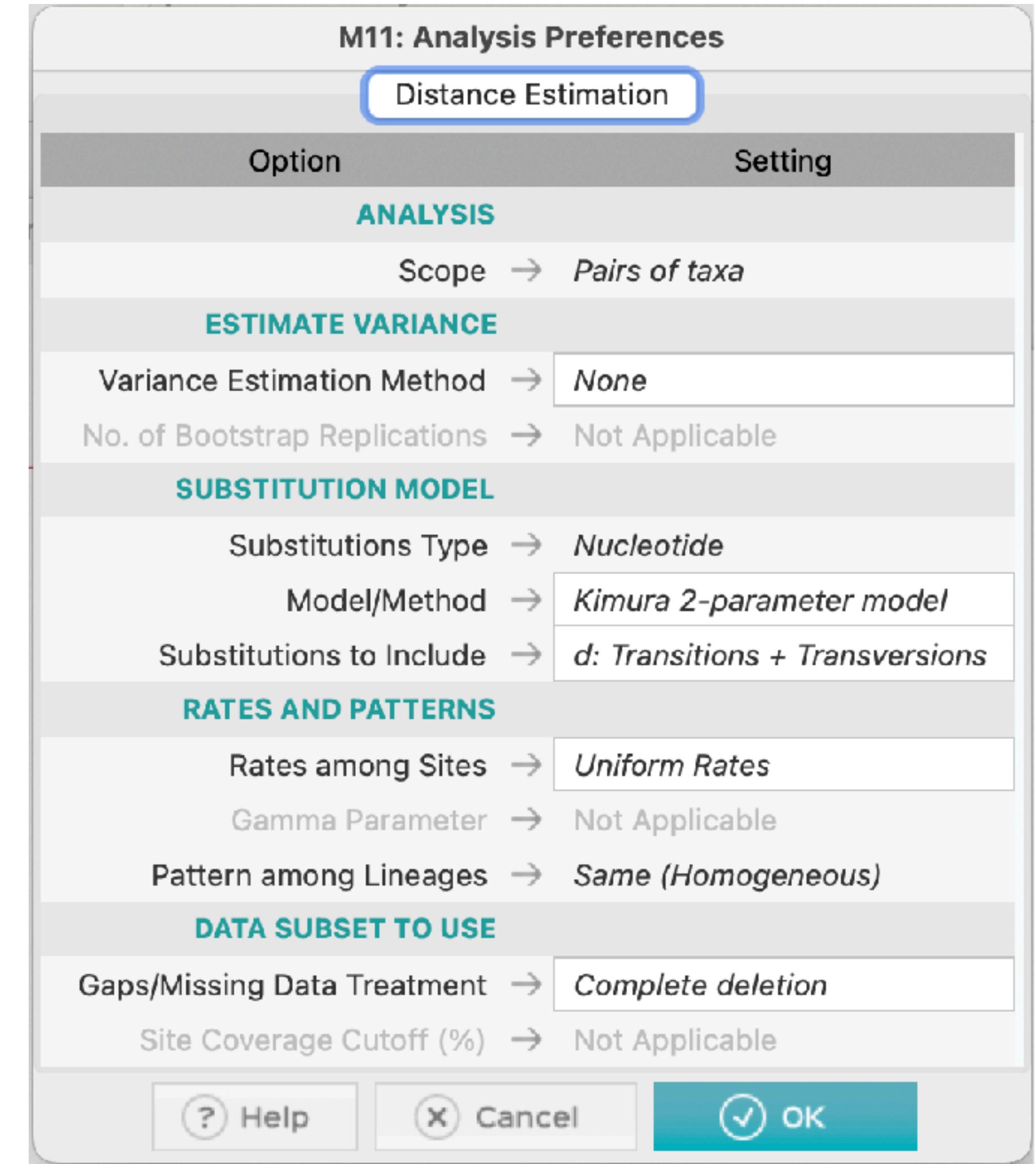


Calculate Distant matrix

Return to the main panel



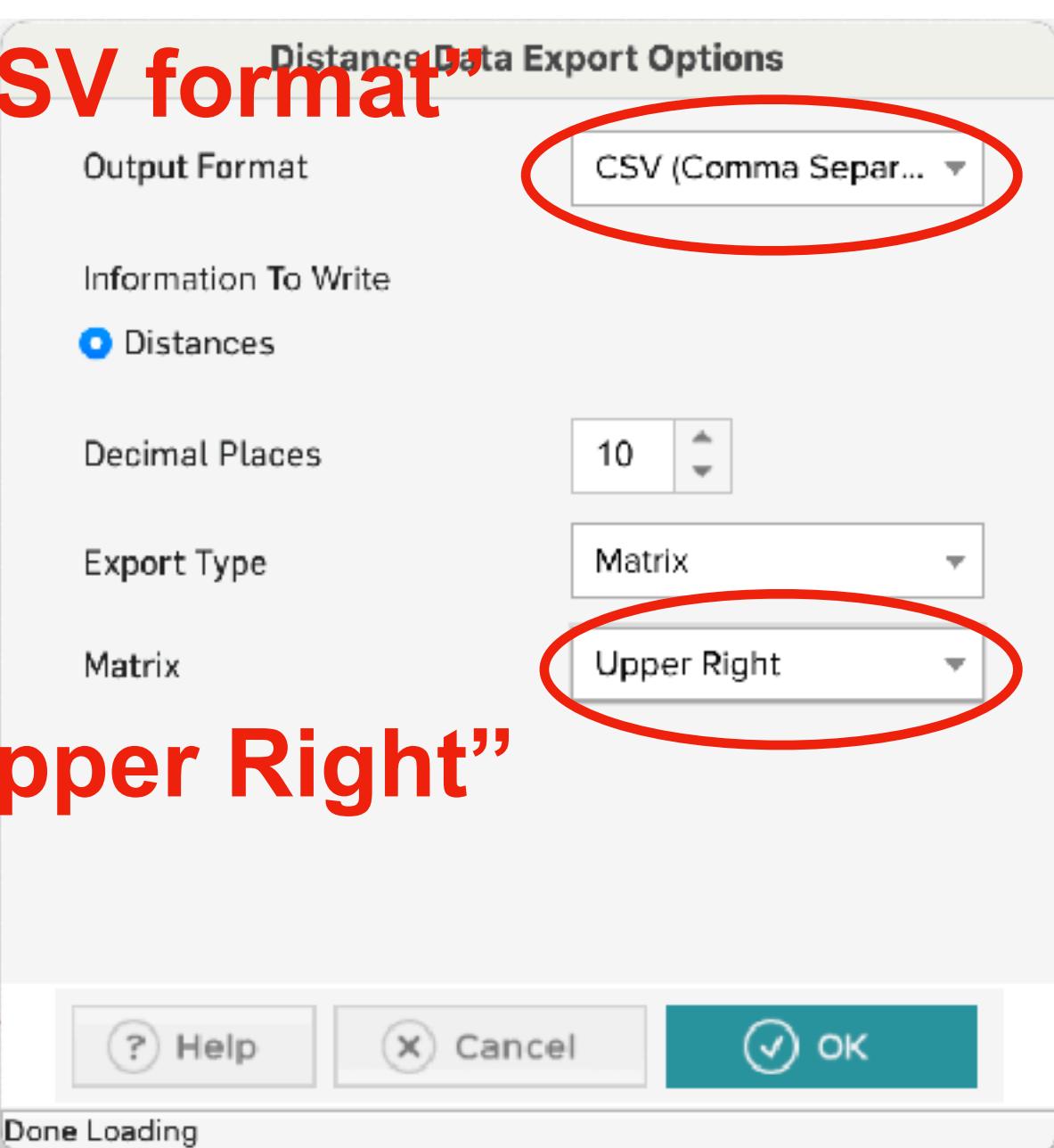
Set the parameters



Click “Save” icon

	1	2	3
1. Nz01			
2. Aus01	0.0362049010		
3. Aus02	0.0434369756	0.0209863349	
4. Aus03	0.0377749719	0.0014758632	0.0212882528
5. PNG01	0.0333826645	0.0290309744	0.0399193852
6. Aus04	0.0450574994	0.0290037747	0.0418057650
7. Aus05	0.0396664412	0.0134070549	0.0092077340
8. PNG02	0.0355832339	0.0293423070	0.0402363424
9. Aus06	0.0361791817	0.0308450209	0.0379904555
10. PNG03	0.0362049010	0.0296466343	0.0405446419
11. Aus07	0.0393491823	0.0206930144	0.0164443074
12. Aus08	0.0380913275	0.0017715462	0.0215943440

Select “CSV format”



Select “Upper Right”

Click “Save” icon and save as “Distance_matrix.txt”

	*Distance Data
1	,Nz01,Aus01,Aus02,Aus03,PNG01,Aus04,PNG02,Aus05,PNG03,Aus06,PNG04,Aus07,PNG05,Aus08,Aus09,Aus10,Aus11,Aus12,Nz02,Aus13,PNG06
2	Nz01,,0.0362049010,0.0434369756,0.0377749719,0.0333826645,0.0450574994,0.0396664412,0.0355832339,0.0361791817,
3	Aus01,,,0.0209863349,0.0134070549,0.0092077340,0.0293423070,0.0402363424,0.0355832339,0.0361791817,0.0296466343,
4	Aus02,,,0.0402363424,0.0355832339,0.0361791817,0.0308450209,0.0405446419,0.0361791817,0.0308450209,0.0418057650,
5	Aus03,,,0.0399193852,0.0361791817,0.0308450209,0.0379904555,0.0399193852,0.0361791817,0.0308450209,0.0396466343,
6	PNG01,,,0.0296466343,0.0361791817,0.0308450209,0.0379904555,0.0296466343,0.0361791817,0.0308450209,0.0396466343,
7	Aus04,,,0.0418057650,0.0361791817,0.0308450209,0.0379904555,0.0418057650,0.0361791817,0.0308450209,0.0405446419,
8	Aus05,,,0.0396664412,0.0361791817,0.0308450209,0.0379904555,0.0396664412,0.0361791817,0.0308450209,0.0396664412,
9	PNG02,,,0.0308450209,0.0361791817,0.0308450209,0.0379904555,0.0308450209,0.0361791817,0.0308450209,0.0308450209,
10	Aus06,,,0.0355832339,0.0361791817,0.0308450209,0.0379904555,0.0355832339,0.0361791817,0.0308450209,0.0355832339,
11	PNG03,,,0.0361791817,0.0308450209,0.0379904555,0.0361791817,0.0361791817,0.0308450209,0.0308450209,0.0308450209,
12	Aus07,,,0.0296466343,0.0361791817,0.0308450209,0.0379904555,0.0296466343,0.0361791817,0.0308450209,0.0296466343,
13	Aus08,,,0.0308450209,0.0361791817,0.0308450209,0.0379904555,0.0308450209,0.0361791817,0.0308450209,0.0308450209,
14	Aus09,,,0.0379904555,0.0361791817,0.0308450209,0.0379904555,0.0379904555,0.0361791817,0.0308450209,0.0379904555,
15	Aus10,,,0.0361791817,0.0308450209,0.0379904555,0.0361791817,0.0361791817,0.0308450209,0.0308450209,0.0308450209,
16	Aus11,,,0.0296466343,0.0308450209,0.0379904555,0.0321145341,0.0296466343,0.0308450209,0.0308450209,0.0321145341,
17	Aus12,,,0.0308450209,0.0379904555,0.0361791817,0.0321145341,0.0308450209,0.0379904555,0.0308450209,0.0321145341,
18	Nz02,,,0.0363834854,0.0355832339,0.0361791817,0.0363834854,0.0363834854,0.0355832339,0.0355832339,0.0361791817,
19	Aus13,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
20	PNG04,,,0.0379904555,0.0361791817,0.0355832339,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
21	Aus14,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
22	Aus15,,,0.0379904555,0.0361791817,0.0355832339,0.0379904555,0.0379904555,0.0361791817,0.0361791817,0.0379904555,
23	PNG05,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
24	Aus16,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
25	PNG06,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
26	Aus17,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
27	PNG07,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
28	Aus18,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
29	PNG08,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
30	Aus19,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
31	PNG09,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
32	Aus20,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
33	PNG10,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
34	Aus21,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,
35	PNG11,,,0.0361791817,0.0355832339,0.0363834854,0.0361791817,0.0361791817,0.0355832339,0.0355832339,0.0361791817,
36	Aus22,,,0.0308450209,0.0361791817,0.0355832339,0.0308450209,0.0308450209,0.0361791817,0.0361791817,0.0308450209,

Principal Component Analysis (PCA)

The basics of PCA

Distant matrix

	D ₁	D ₂	D ₃	D ₄	S
strain1	0	0.036	0.043	0.037	0.116
strain2	0.036	0	0.021	0.0015	0.0585
strain3	0.043	0.021	0	0.021	0.085
strain4	0.037	0.0015	0.021	0	0.0595

D_N : distance between strain N

Summation of distances: S = D₁ + D₂ + D₃ + D₄

S_N : summation for strain N

Average: a = (S₁ + S₂ + S₃ + S₄)/4

Variance = {(S₁-a)² + (S₂-a)² + (S₃-a)² + (S₄-a)²}/4

The basics of PCA

	W_1	W_2	W_3	W_4	
	D ₁	D ₂	D ₃	D ₄	S
strain1	0	0.036	0.043	0.037	S ₁
strain2	0.036	0	0.021	0.0015	S ₂
strain3	0.043	0.021	0	0.021	S ₃
strain4	0.037	0.0015	0.021	0	S ₄

Set weights when taking the summation

$$S = W_1 D_1 + W_2 D_2 + W_3 D_3 + W_4 D_4$$

Restriction: $(W_1)^2 + (W_2)^2 + (W_3)^2 + (W_4)^2 = 1$

Average: $a = (S_1 + S_2 + S_3 + S_4)/4$

Variance = $\{(S_1-a)^2 + (S_2-a)^2 + (S_3-a)^2 + (S_4-a)^2\}/4$

Maximize the variance by changing the weights

The basics of PCA

	D ₁	D ₂	D ₃	D ₄	S _m
strain1	0	0.036	0.043	0.037	-0.047
strain2	0.036	0	0.021	0.0015	0.018
strain3	0.043	0.021	0	0.021	0.011
strain4	0.037	0.0015	0.021	0	0.018

Principal component 1 (PC1)

The weighted summation that has the largest variance

$$S_m = W_{1m}D_1 + W_{2m}D_2 + W_{3m}D_3 + W_{4m}D_4$$

Restriction: $(W_{1m})^2 + (W_{2m})^2 + (W_{3m})^2 + (W_{4m})^2 = 1$

The basics of PCA

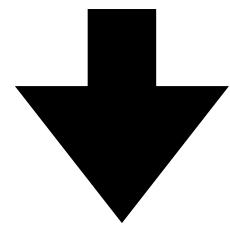
	D ₁	D ₂	D ₃	D ₄	S _m
strain1	0	0.036	0.043	0.037	-0.047
strain2	0.036	0	0.021	0.0015	0.018
strain3	0.043	0.021	0	0.021	0.011
strain4	0.037	0.0015	0.021	0	0.018
Variance	0.000383	0.000294	0.000308	0.000309	0.00101

Proportion of variance

0.00101

$$\frac{0.00101}{0.000383 + 0.000294 + 0.000308 + 0.000309} \div 0.781 \quad (78.1\%)$$

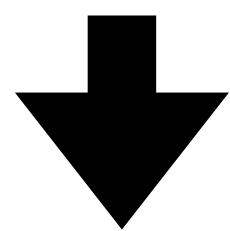
Weight	W_{1m}	W_{2m}	W_{3m}	W_{4m}	
	D ₁	D ₂	D ₃	D ₄	S _m
PC1	strain1	0	0.036	0.043	0.037
	strain2	0.036	0	0.021	0.0015
	strain3	0.043	0.021	0	0.021
	strain4	0.037	0.0015	0.021	0



Make a new matrix

	D ₁	D ₂	D ₃	D ₄	S
strain1	$0 - W_{1m} * (-0.047)$	$0.036 - W_{2m} * (-0.047)$			
strain2					
strain3					
strain4					

Apply the same procedure as PC1



PC2

Execute Principle Component Analysis and plot the result



Confirm your working directory

`getwd()`

Put ‘Distance_matrix.csv’ in your working directory.

You can change the working directory pushing [command]+[D] keys.

Read the matrix data.

`dm <- read.csv(“Distance_matrix.csv”, header=T)`

Filename

Data contains a header

This matrix contains ‘Not Available (NA)’ values.

Change NA to 0.

`dm[is.na(dm)] <- 0`

The first column of ‘dm’ are strain names. Eliminate the first column.

`dm2 <- dm[,-1]`

Convert the triangle matrix to the symmetric matrix
sdm <- dm2 + t(dm2)

dm

0	0.036	0.043	0.037	...
0	0	0.021	0.0015	...
0	0	0	0.021	...
0	0	0	0	...
...

+

t(dm)

0	0	0	0	...
0.036	0	0	0	...
0.043	0.021	0	0	...
0.037	0.0015	0.021	0	...
...

symdm

0	0.036	0.043	0.037	...
0.036	0	0.021	0.0015	...
0.043	0.021	0	0.021	...
0.037	0.0015	0.021	0	...
...

Execute Principle Component Analysis

```
dmpca <- prcomp(sdm)
```

See the proportion of variances

```
summary(dmpca)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	0.1461	0.0643	0.05723	0.05399	0.02775	0.02516	0.02099	0.01773	0.01497	0.01071	0.009385
Proportion of Variance	0.6063	0.1174	0.09302	0.08281	0.02187	0.01798	0.01251	0.00893	0.00637	0.00326	0.002500
Cumulative Proportion	0.6063	0.7237	0.81672	0.89953	0.92141	0.93938	0.95189	0.96082	0.96719	0.97045	0.972950
	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	
Standard deviation	0.007651	0.006705	0.006612	0.006221	0.005957	0.005431	0.005369	0.00520	0.004882	0.004695	
Proportion of Variance	0.001660	0.001280	0.001240	0.001100	0.001010	0.000840	0.000820	0.00077	0.000680	0.000630	
Cumulative Proportion	0.974610	0.975890	0.977130	0.978230	0.979240	0.980070	0.980890	0.98166	0.982340	0.982960	

.....

PC scores are stored in **dmpca\$x**

If it didn't work well, download “Distance_matrix.csv” from
http://sayer.nig.ac.jp/rusuzuki/keio/Distance_matrix.csv

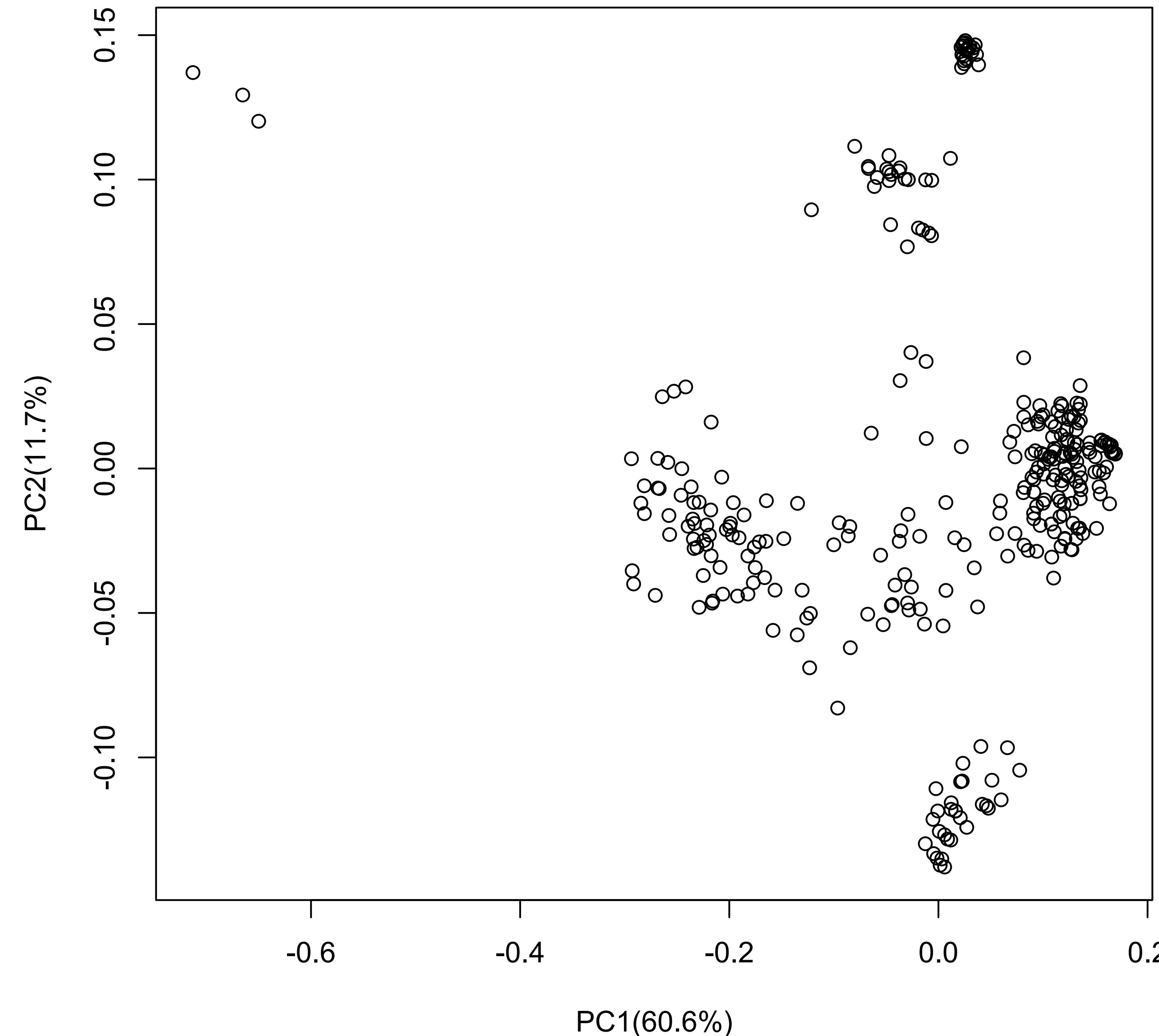
Plot PC1 and PC2

```
PC1 score          PC2 score   x-axis label  
plot(dmpca$x[,1], dmpca$x[,2], xlab="PC1 (60.6%)", ylab="PC2 (11.7%)")
```

Or just

```
plot(dmpca$x[,1], dmpca$x[,2])
```

	PC1	PC2
Standard deviation	0.1461	0.0643
Proportion of Variance	0.6063	0.1174
Cumulative Proportion	0.6063	0.7237



Close the plot
`dev.off()`

Modify the PCA plot

Download color data from

<http://sayer.nig.ac.jp/rusuzuki/keio/dotcolors.csv>

Import the color data

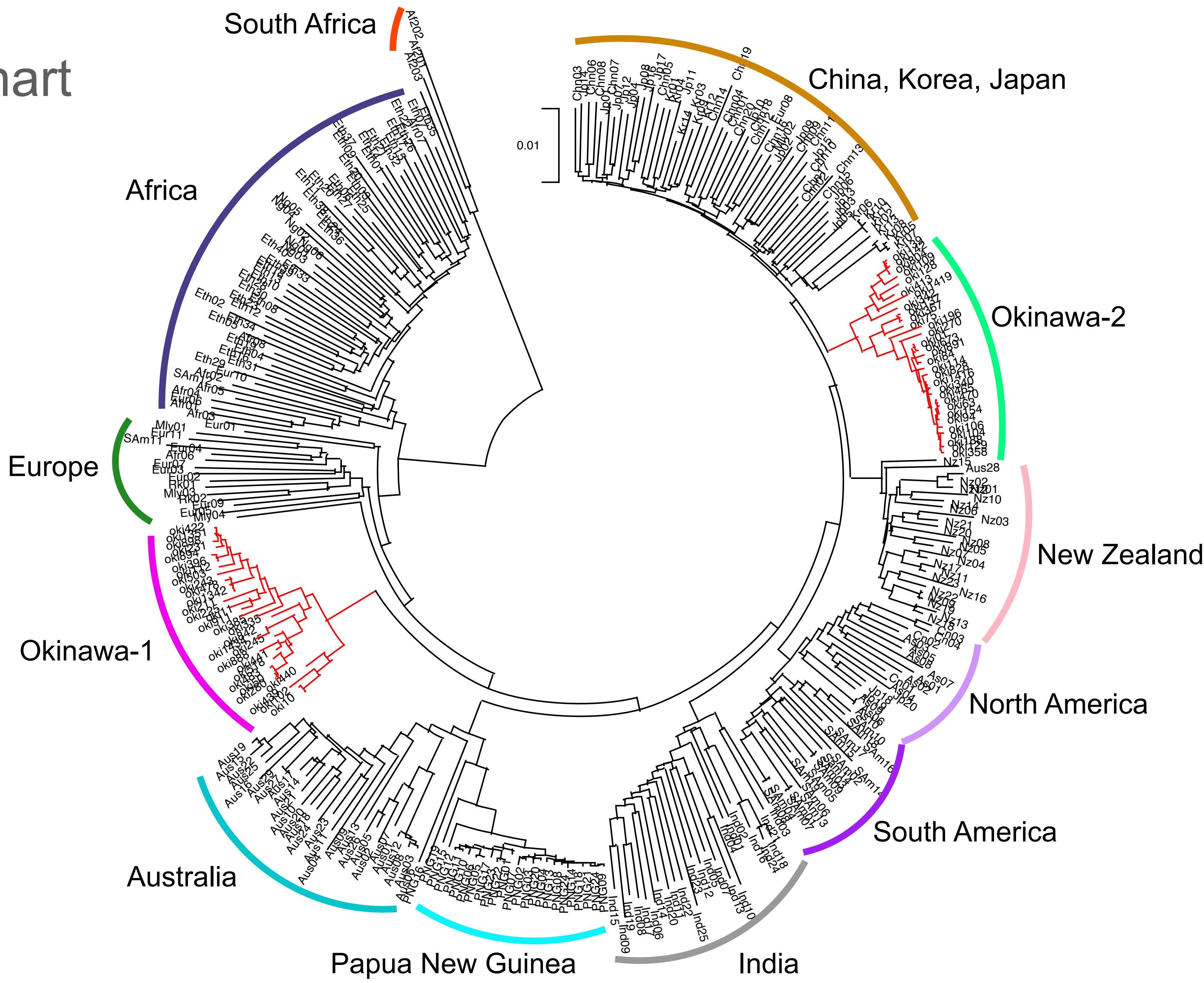
file name

Character data
(color names)

Separated by commas

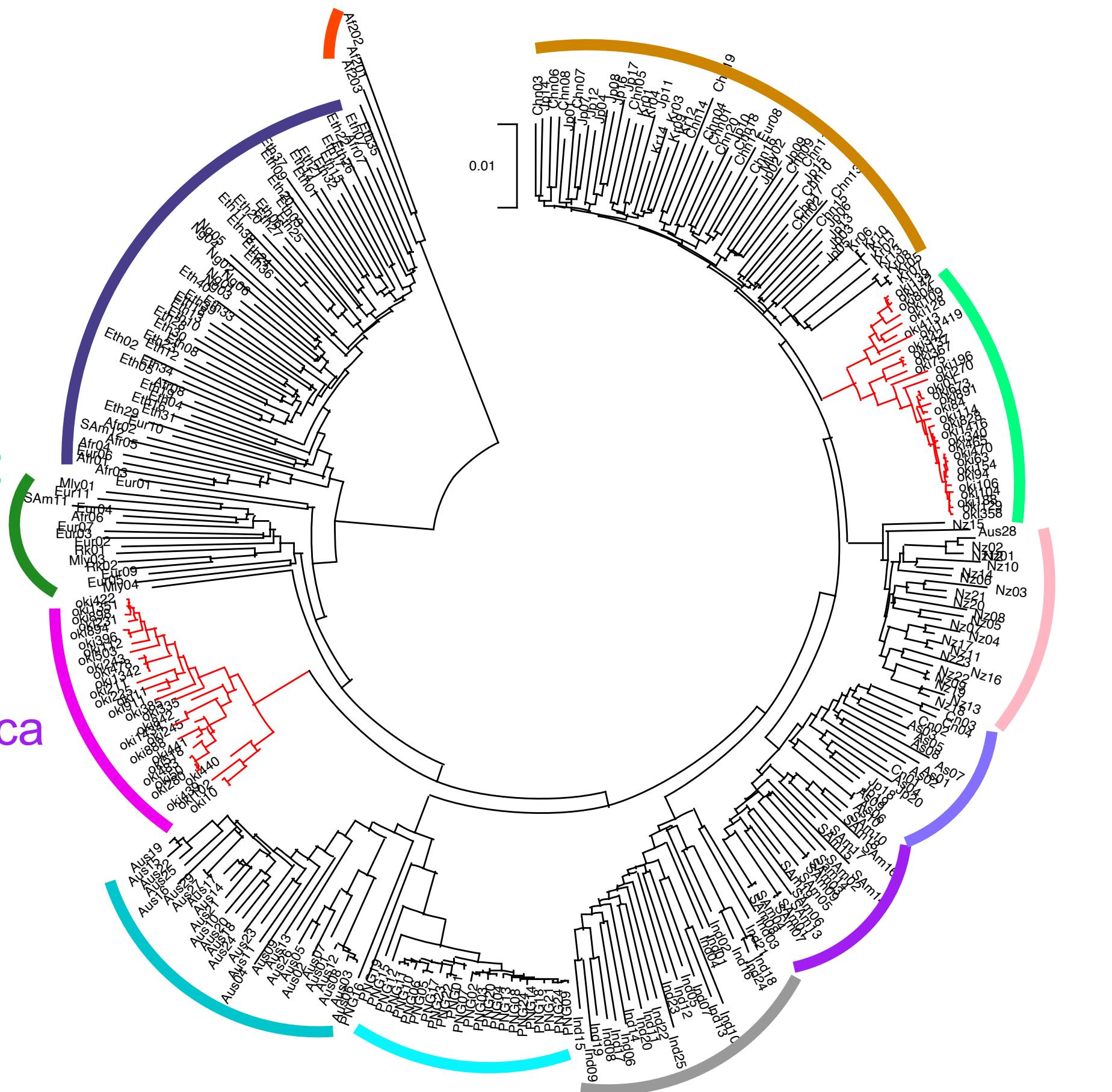
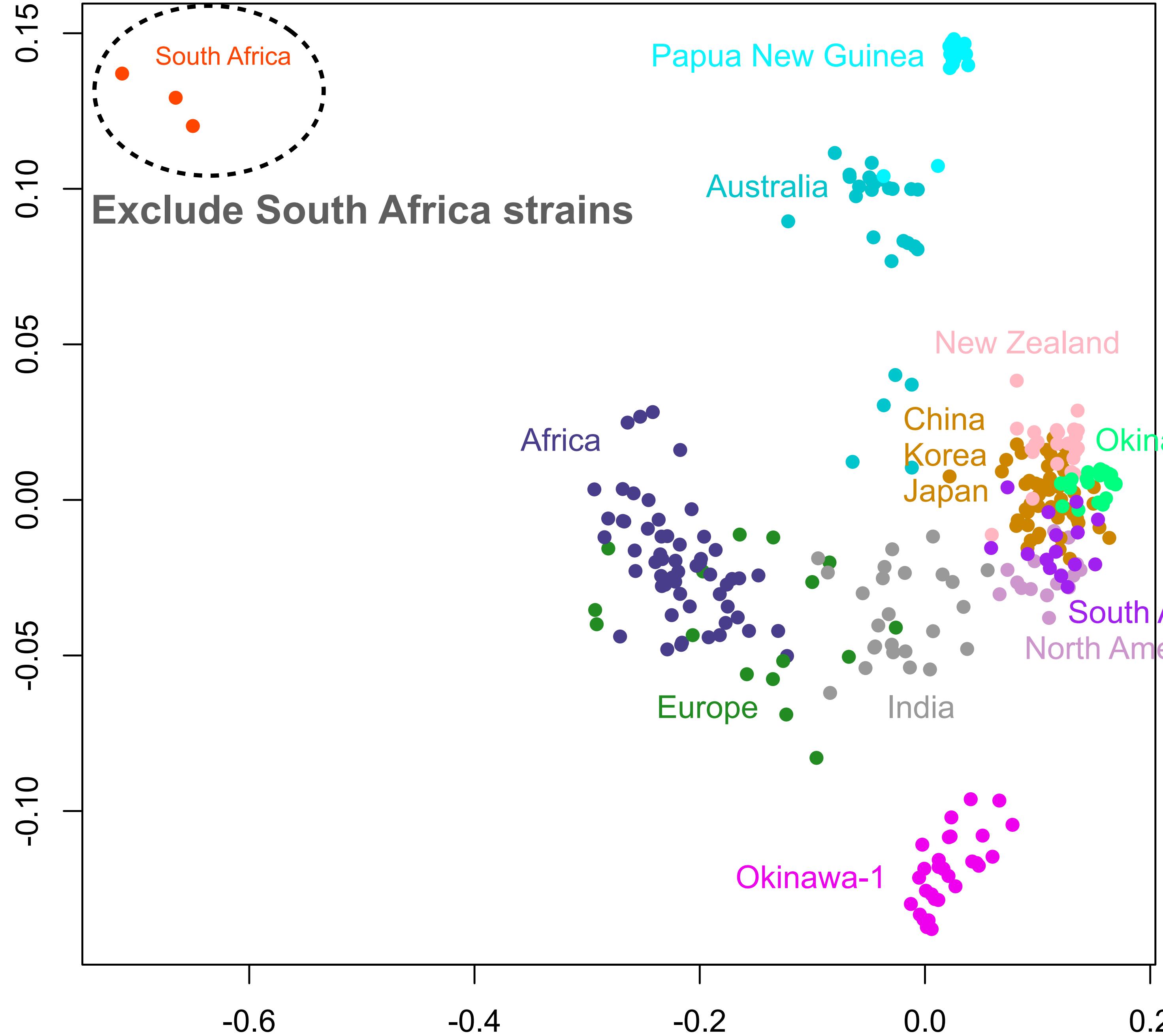
```
dotcol <- scan("dotcolors.csv", what=character(), sep=",")
```

Color chart



Plot PC1 and PC2

```
plot(dmpca$x[,1], dmpca$x[,2], col=dotcol, pch=16)
```



The first 3 rows and columns of the distance matrix are South African strains.

Eliminate the 3 rows and columns from the distance matrix “symdm.”

```
sdm2 <- sdm[-c(1:3), -c(1:3)]
```

Eliminate the first 3 items from the color data “dotcolors”

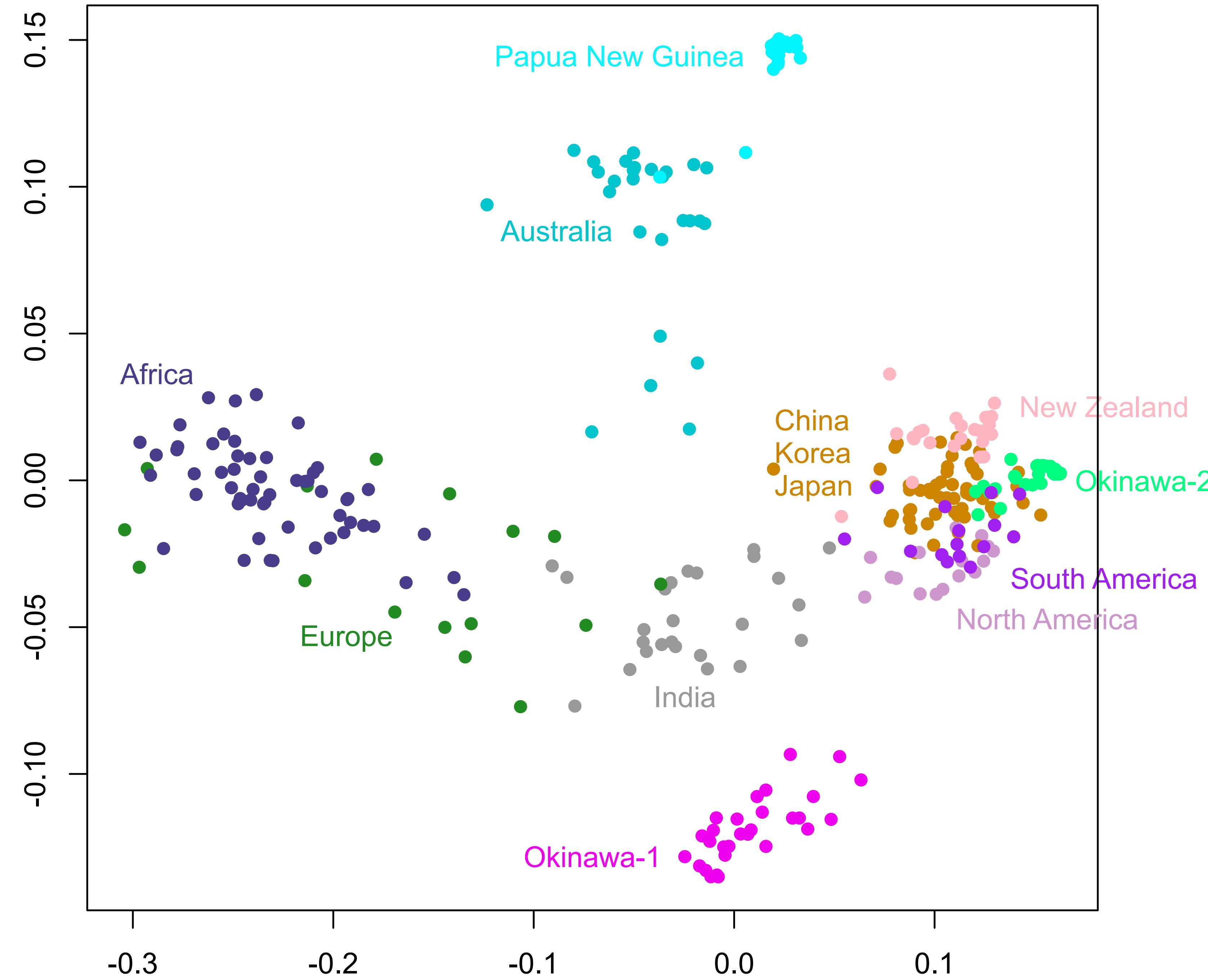
```
dotcol2 <- dotcol[-c(1:3)]
```

Execute PCA

```
dmpca2 <- prcomp(sdm2)
```

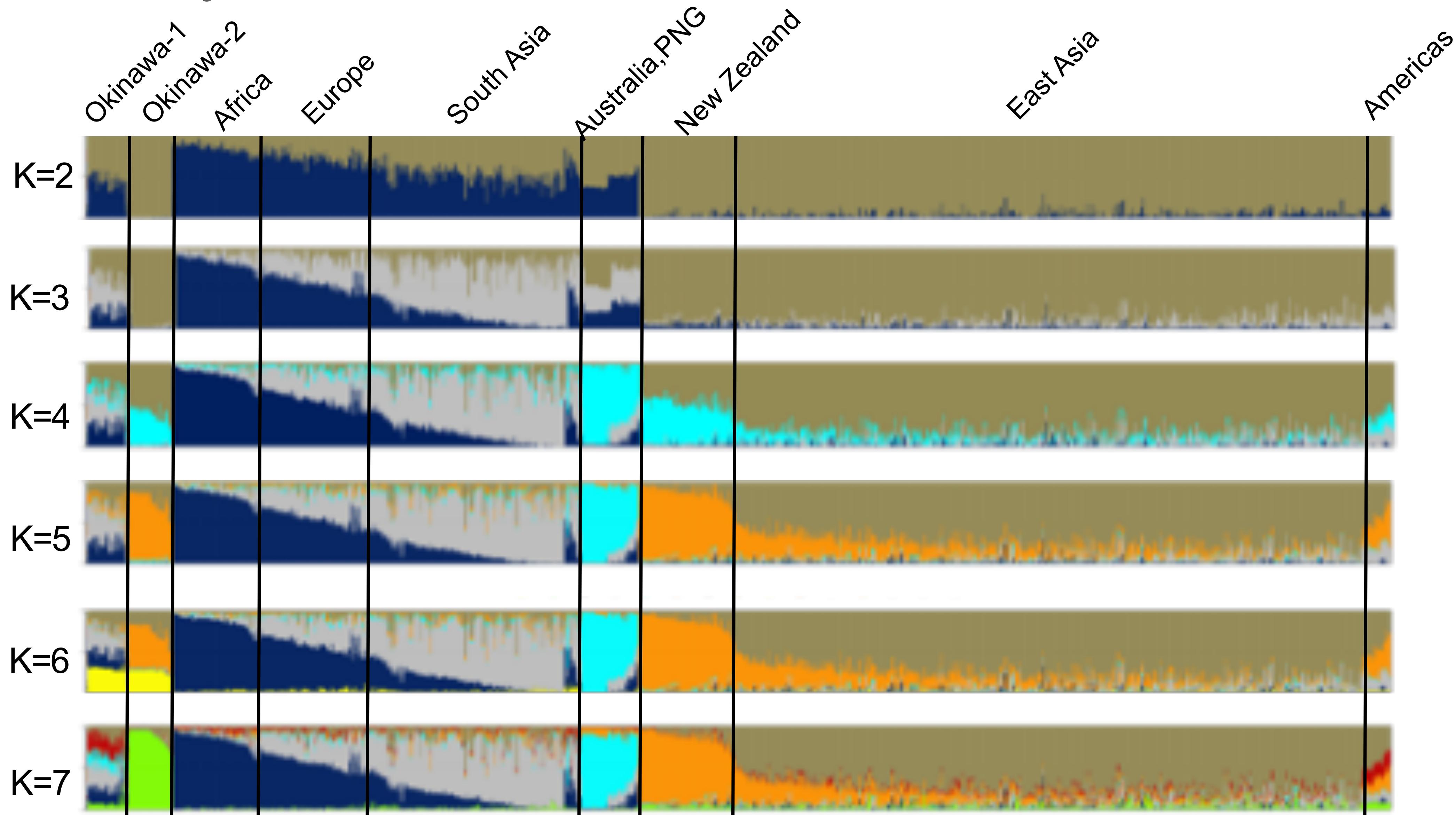
Plot PC1 and PC2

```
plot(dmpca2$x[,1], dmpca2$x[,2], col=dotcol2, pch=16)
```

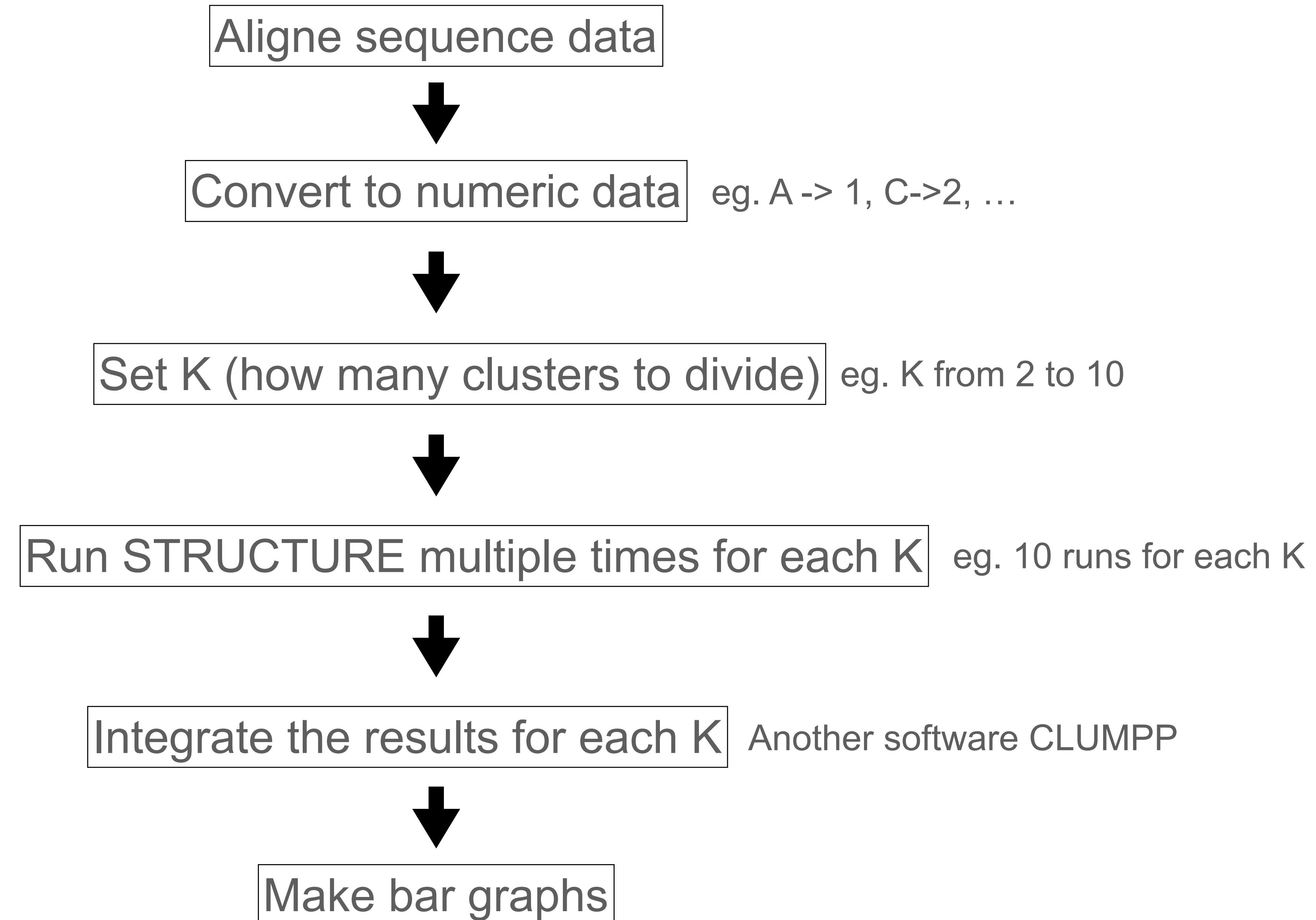


Other methods for population analysis

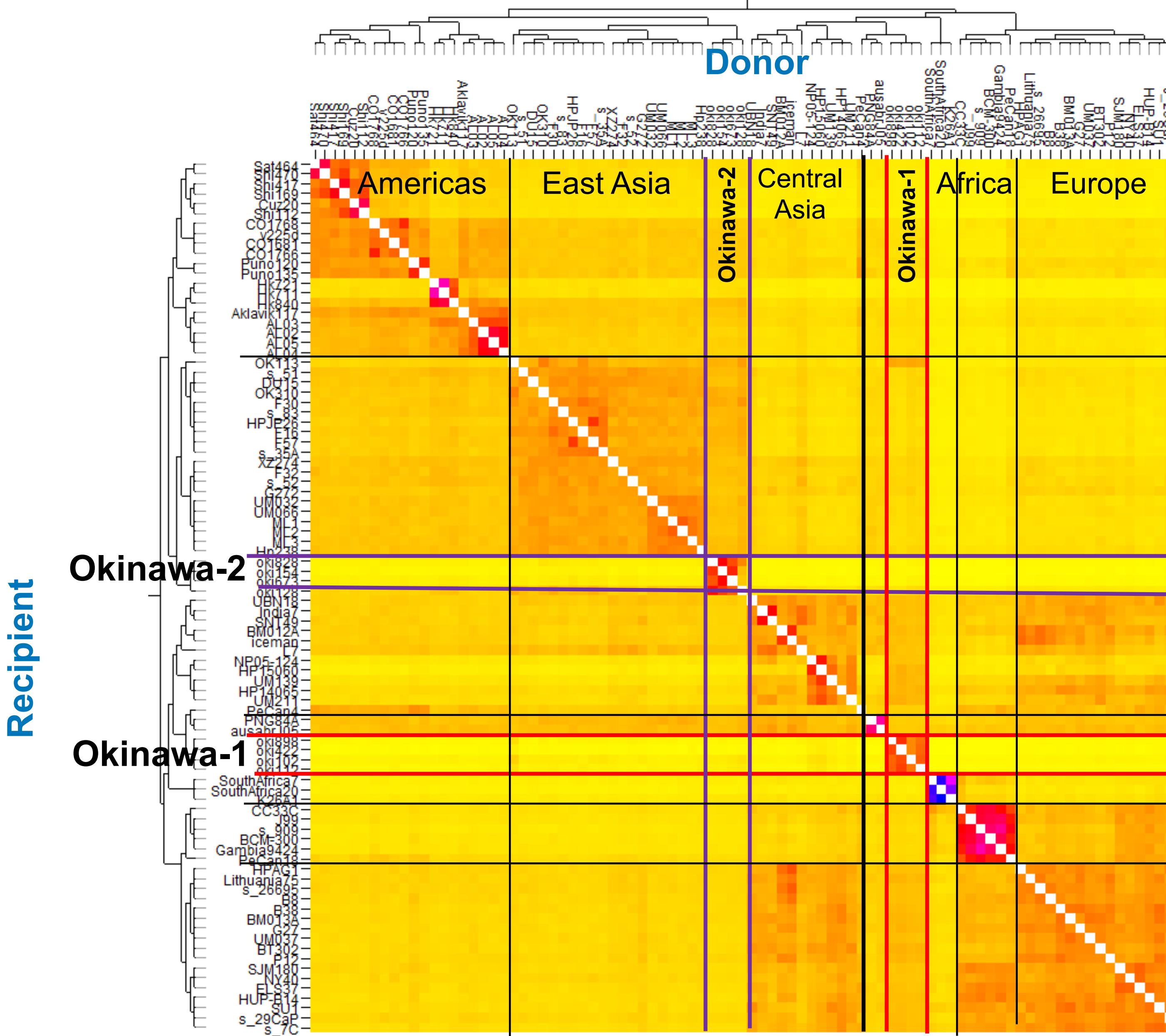
Baysian inference: STRUCTURE, Admixture



Procedure for analysis by STRUCTURE

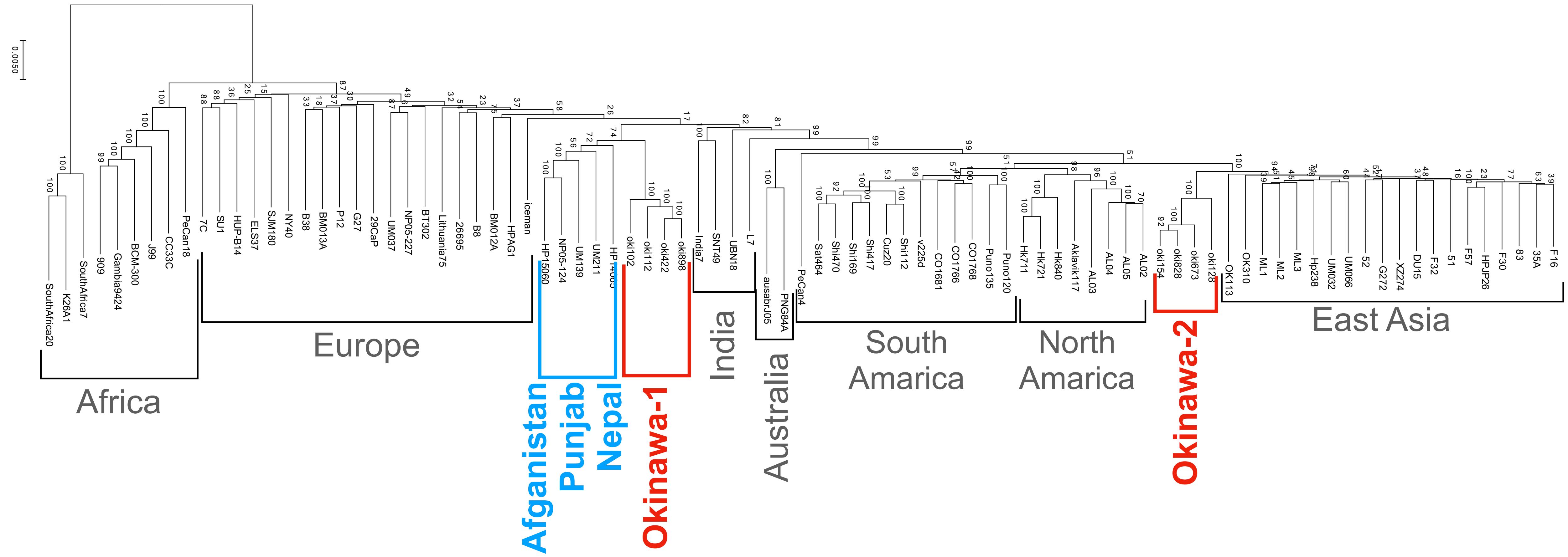


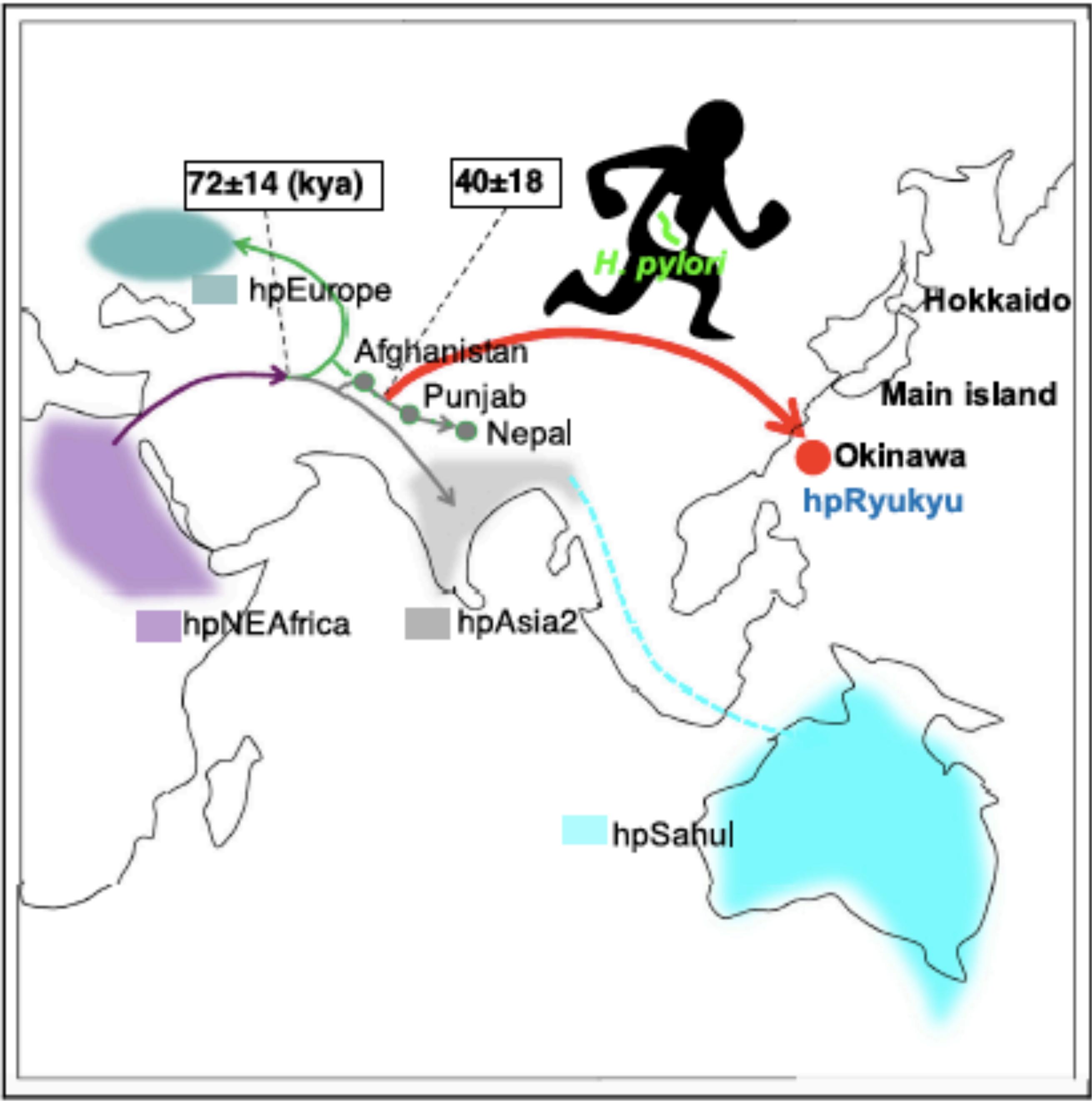
Baysian inference: fineSTRUCTURE



From where the Okinawa strains came?

The phylogenetic tree based on the genome data





The whole genome project for *H. pylori* 1011 strains from 50 countries

nature communications

8

Article

<https://doi.org/10.1038/s41467-023-43562-y>

The *Helicobacter pylori* Genome Project: insights into *H. pylori* population structure from analysis of a worldwide collection of complete genomes

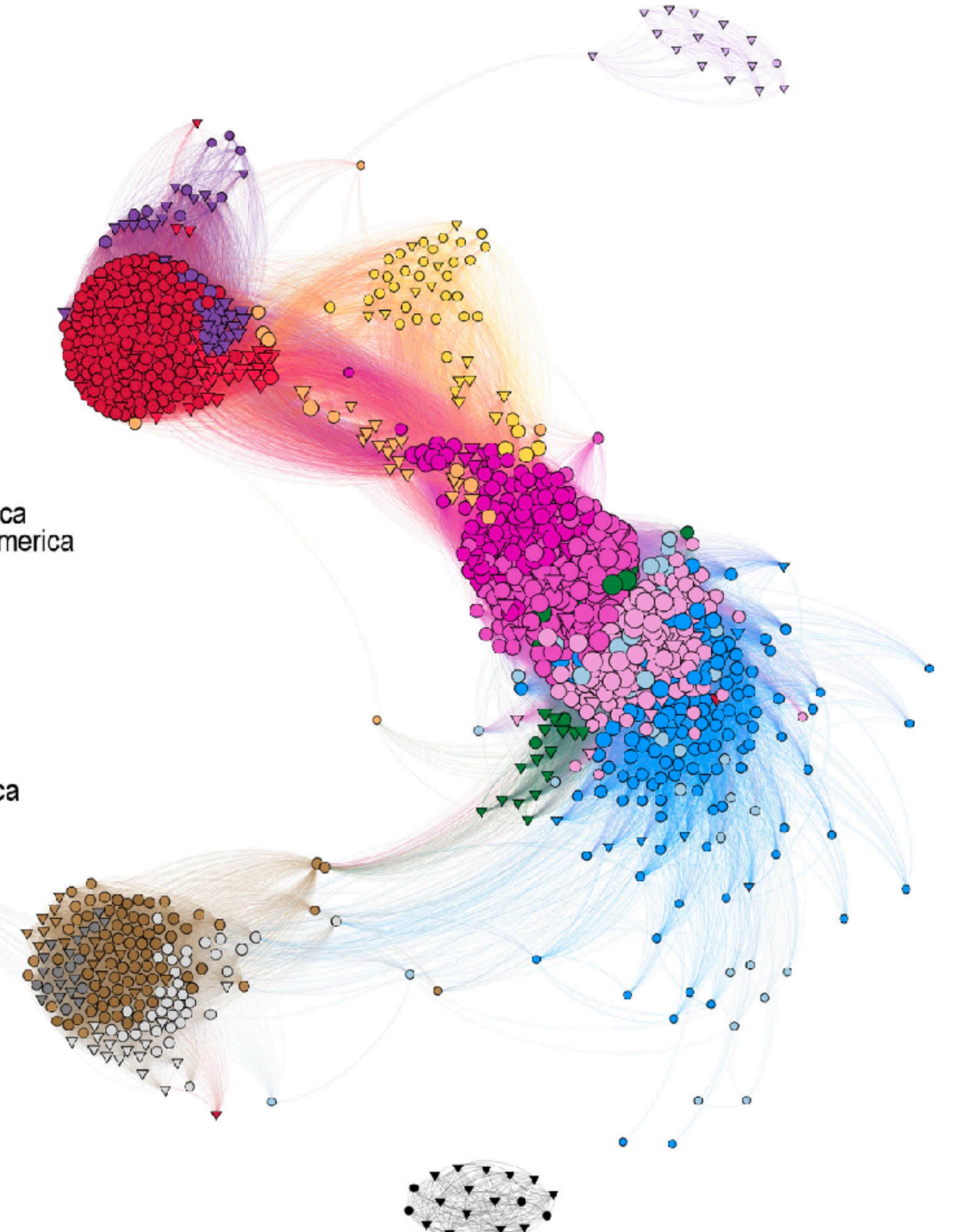
Received: 5 September 2023

Accepted: 13 November 2023

Published online: 11 December 2023

Kaisa Thorell  ^{1,204}, Zilia Y. Muñoz-Ramírez  ^{2,204}, Difei Wang ^{3,4},
Santiago Sandoval-Motta ^{5,6,7}, Rajiv Boscolo Agostini ⁸, Silvia Ghirotto ⁸,
Roberto C. Torres  ⁹, HpGP Research Network ^a, Daniel Falush  ⁹,
M. Constanza Camargo  ^{4,205} & Charles S. Rabkin  ^{4,205}

H. pylori population (hp)
assigned by fineSTRUCTURE

- 
- hpAfrica2
 - hpNEAfrica
 - hspAfrica1SAfrica
 - hspAfrica1WAfrica
 - hspAfrica1NAmerica
 - hspAfrica1MiscAmerica
 - hspSWEuropeLatinAmerica
 - hspSWEuropeChile
 - hspSWEurope
 - hspEurasia
 - hspNEurope
 - hpAsia2
 - hspUral
 - hpSahul
 - hpNorthAsia
 - hsplndigenousAmerica
 - hspEAsia

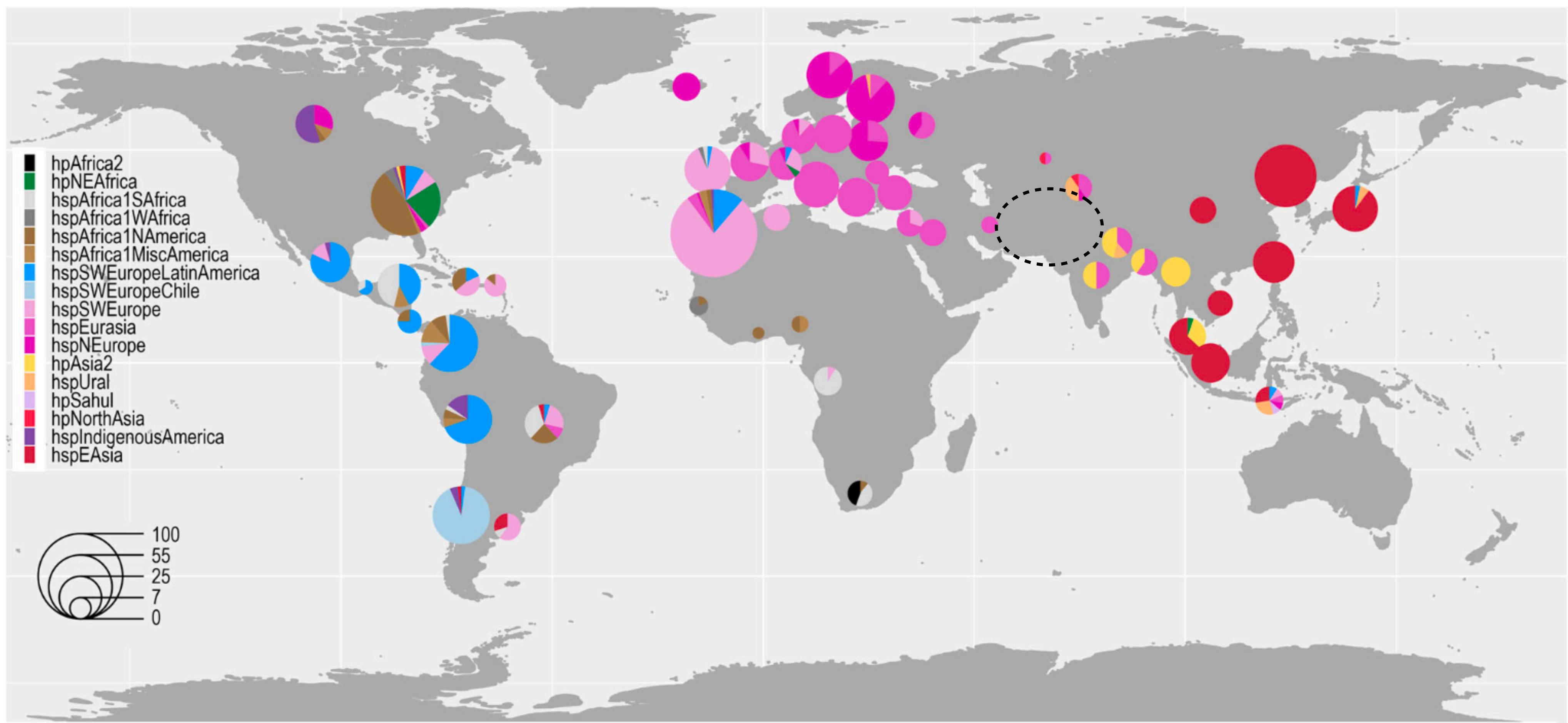


Fig. 1 | World map of *HpGP* strain origins and population assignments. The area of each pie is proportional to the number of *HpGP* genomes from each country and colored by the *H. pylori* population (hp) and subpopulation (hsp) as assigned by fineSTRUCTURE (Supplementary Figs. 1 and 2).