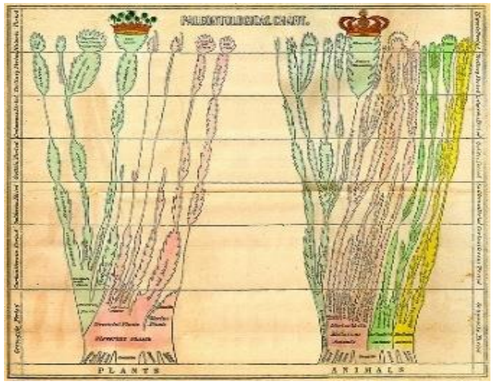
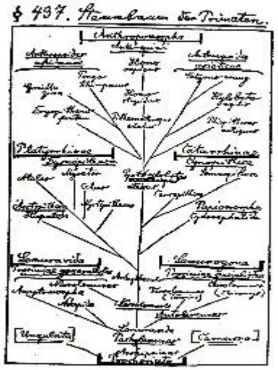
Big Trees Big Problems

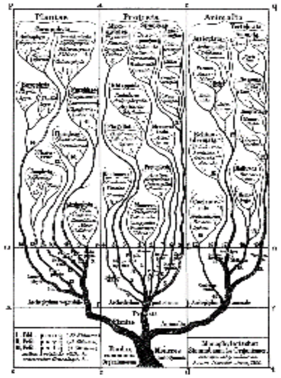
Nomi Hadar

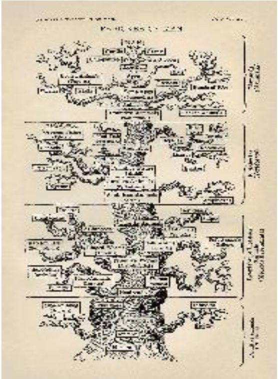
Background

Inferring the evolutionary relationships among species is one of the oldest and most basic tasks of evolutionary research.

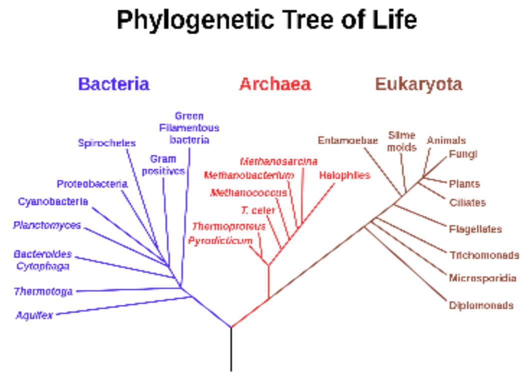


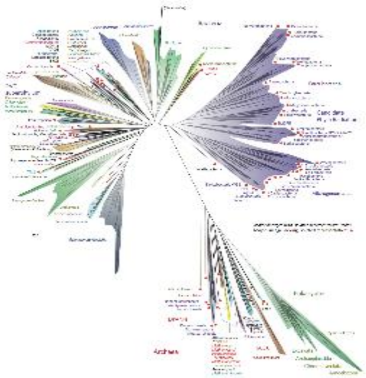


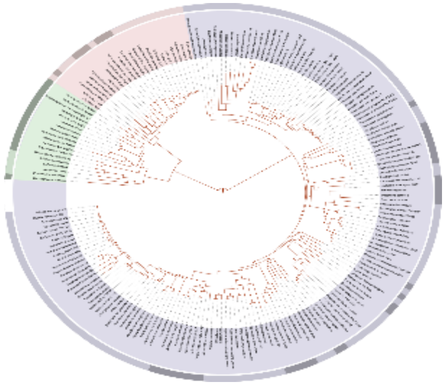




Since 19th

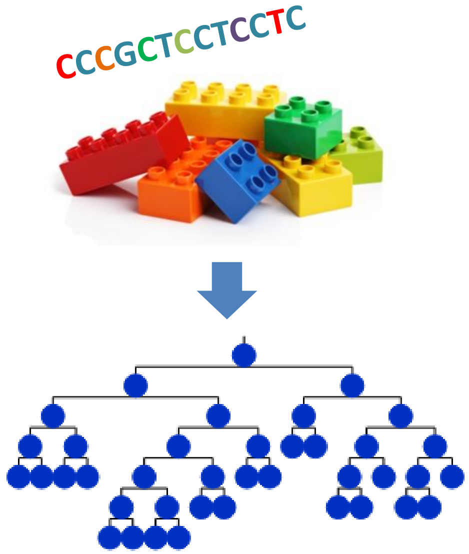


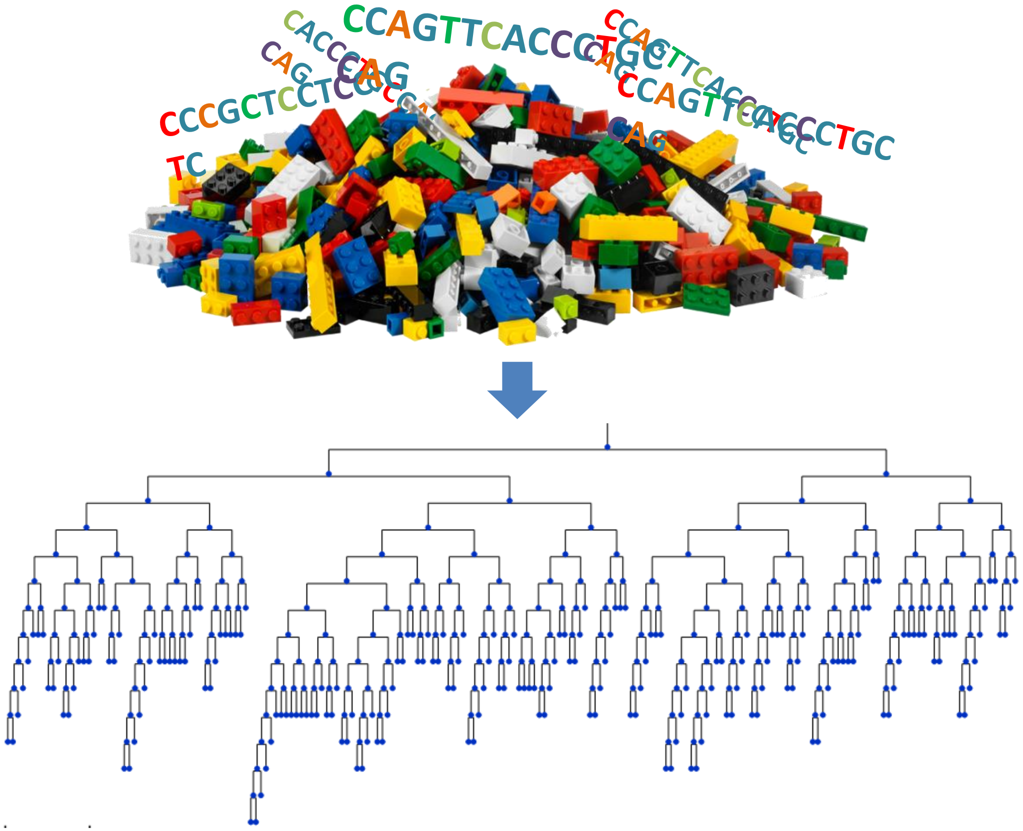




Background

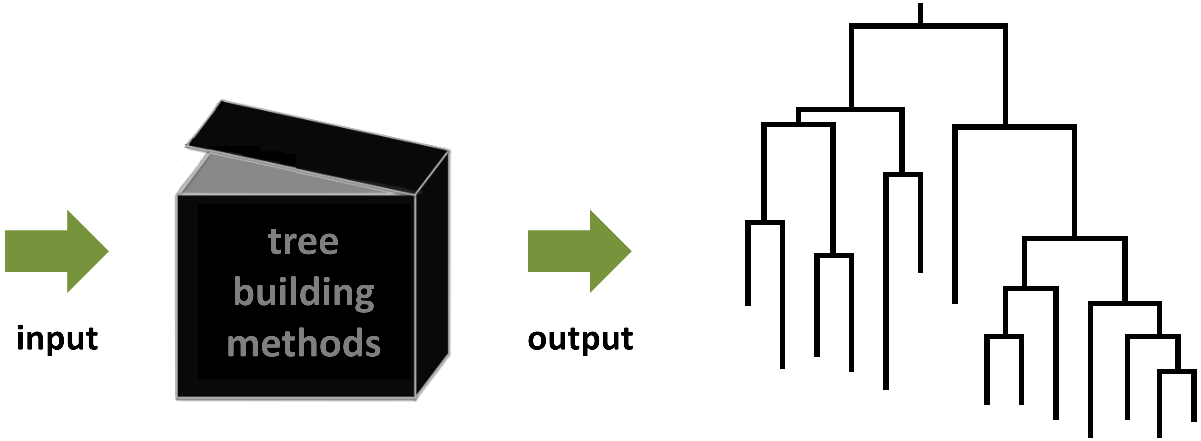
The massive accumulation of sequence data should provide more accurate phylogenies with the hope to resolve the tree of life.





Phylogeny reconstruction with *one* gene

```
Sequence1    -TCAGGA-TGAAC----
Sequence2    ATCACGA-TGAACC---
Sequence3    ATCAGGAATGAATCC--
Sequence4    -TCACGATTGAATCGC-
Sequence5    -TCAGGAATGAATCGCM
```



Phylogeny reconstruction with *many* genes

What if we want to use many loci?

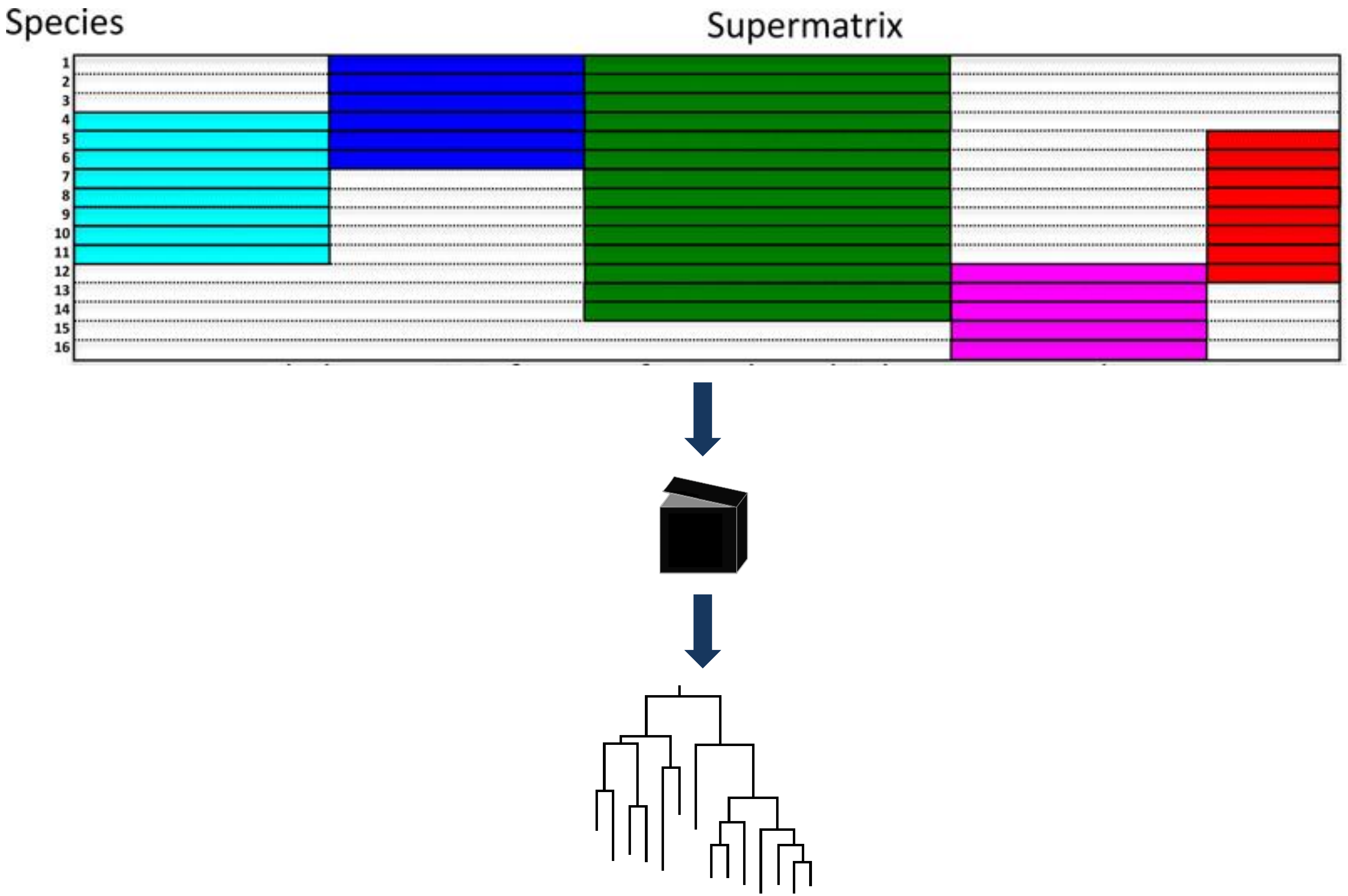
We concatenate genes into a single sequence

gene1		gene2		gene3
GTTCAAAATCACTGCCCGCTCCTCCTCGC	+	CAATGTGAAAGCTGGTG	+	GGCTTGGCGCGACAAAAGCTCCACCTA

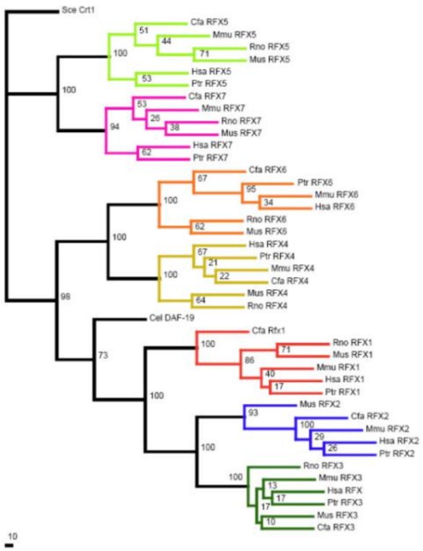
=

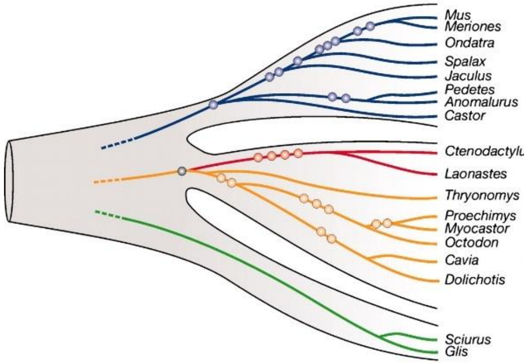
GTTCAAAATCACTGCCCGCTCCTCCTCGCCAATGTGAAAGCTGGTGGGCTTGGCGCGACAAAAGCTCCACCTA

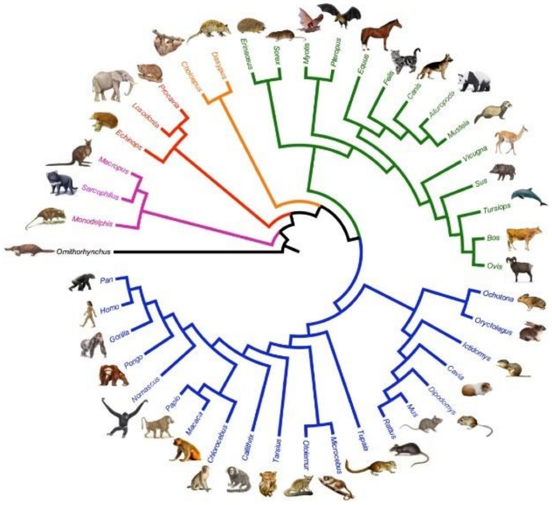
*Super*matrix method

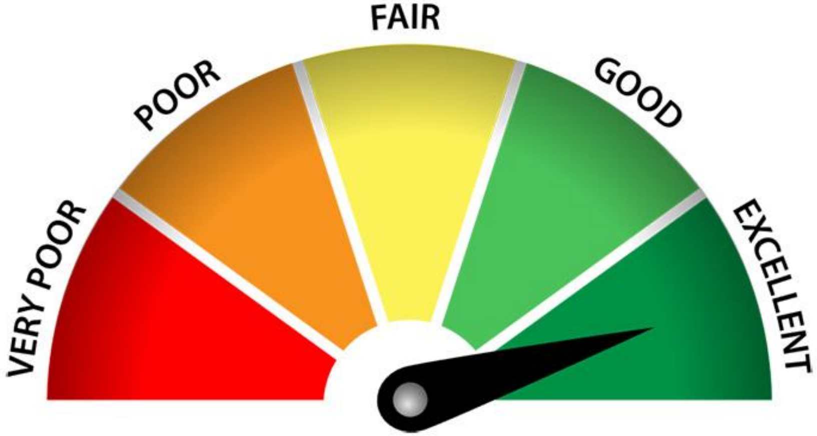


How can we asses the quality of the tree?





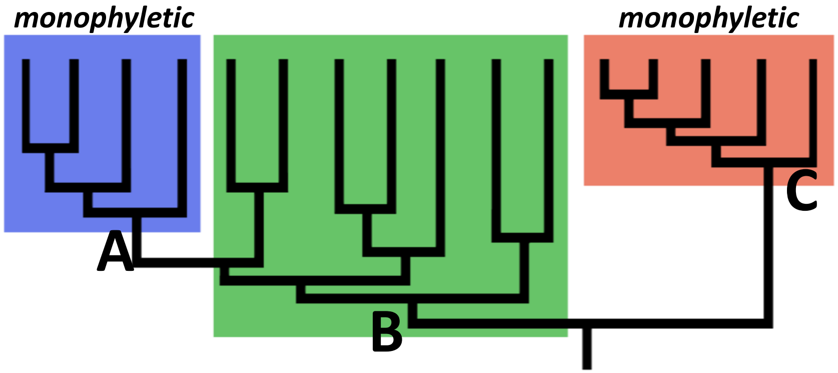


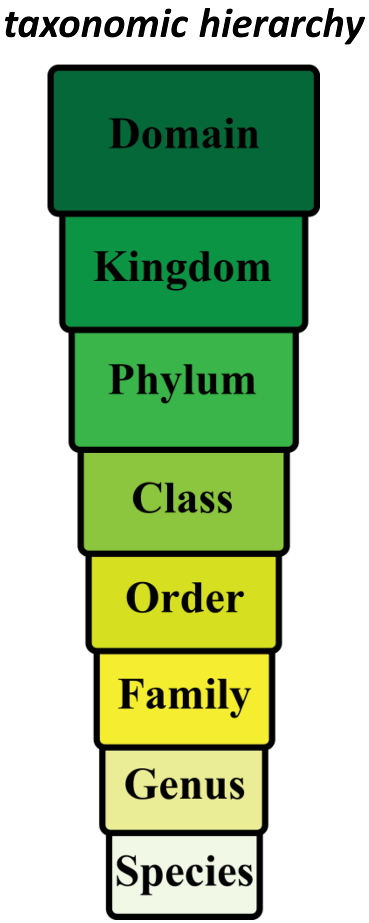


Assessing phylogeny quality

If the tree is correct, we expect the species that are part of the same **taxonomic rank** to form a **monophyletic** group in the tree.

A clade is ***monophyletic*** if it consists of an ancestral species and all its descendants.

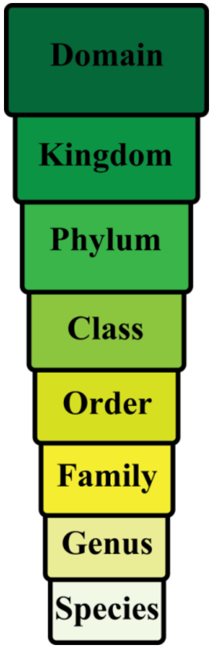




Assessing phylogeny quality

If the tree is correct, we expect the species that are part of the same **taxonomic rank** to form a **monophyletic** group in the tree.

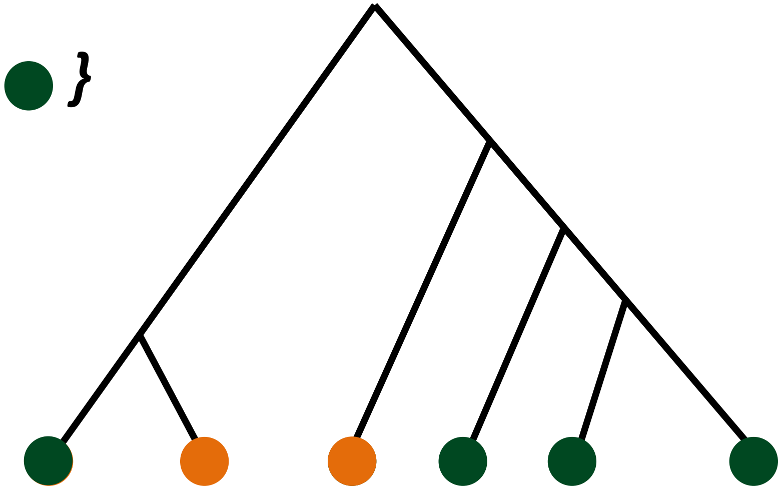
Linnaean taxonomy



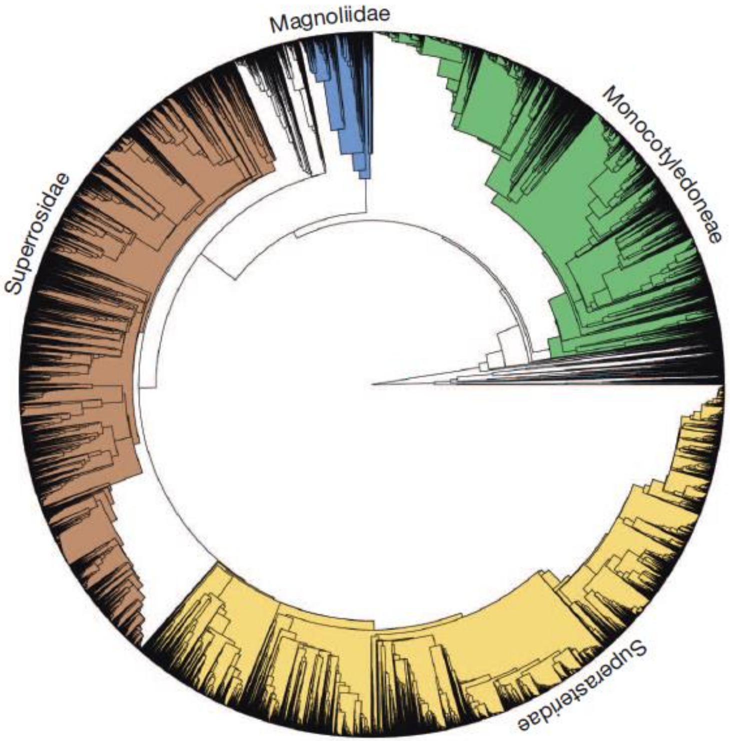
For example:

***Genus1* = { 🟢 , 🟢 , 🟢 , 🟢 }**

***Genus2* = { 🟠 , 🟠 }**



An example of a large phylogeny



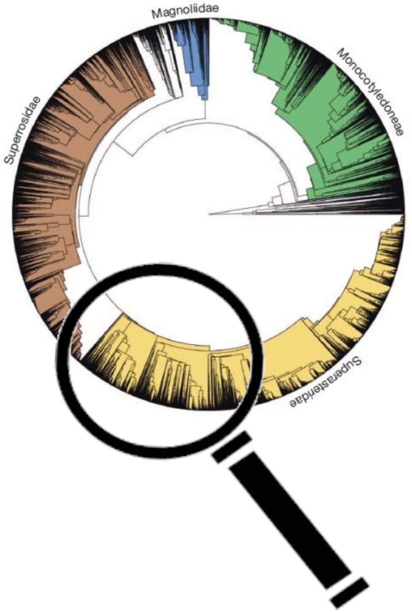
A phylogeny of land plant species

32,223 species

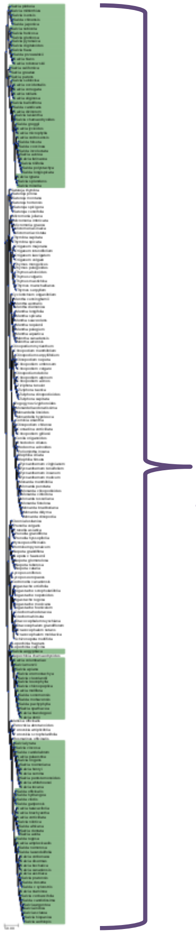
Based on 7 loci

Zanne et al., nature 2014

However, many genera are far from being monophyletic



Genus: Salvia (99 species)

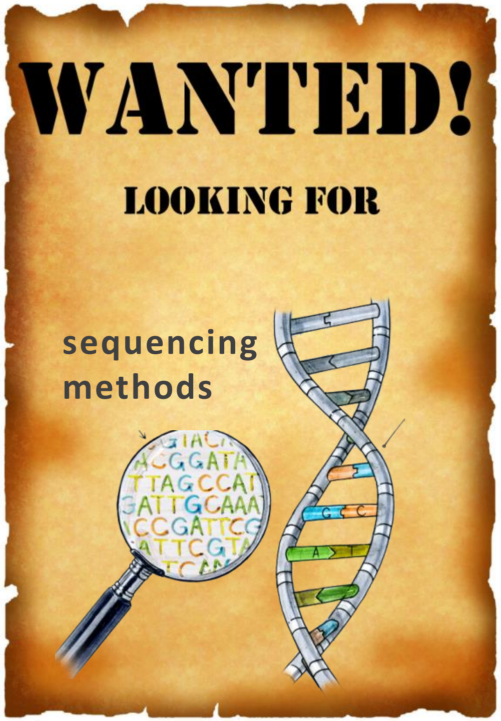


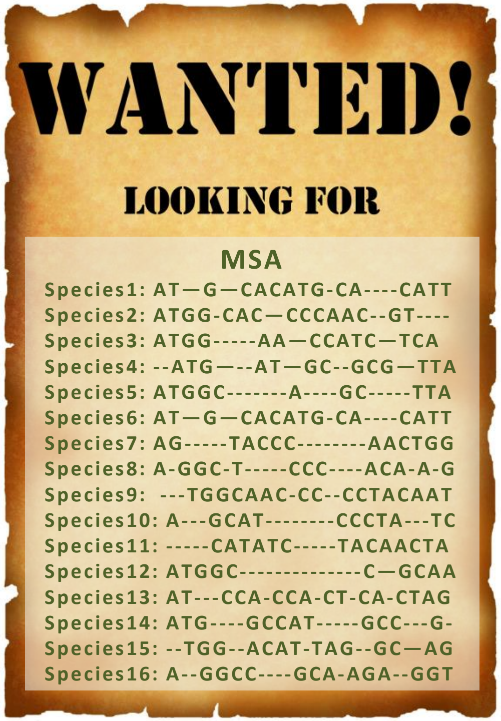
211 species

The average monophyletic score for >1000 genera is only **0.66**

Monophyletic score:
99/211 = 0.47

Who should we blame of distorting the trees?




WANTED!
LOOKING FOR

MSA

```
Species1: AT—G—CACATG-CA----CATT
Species2: ATGG-CAC—CCCAAC--GT----
Species3: ATGG-----AA—CCATC—TCA
Species4: --ATG—--AT—GC--GCG—TTA
Species5: ATGGC-------A----GC-----TTA
Species6: AT—G—CACATG-CA----CATT
Species7: AG-----TACCC--------AACTGG
Species8: A-GGC-T-----CCC----ACA-A-G
Species9:  ---TGGCAAC-CC--CCTACAAT
Species10: A---GCAT--------CCCTA---TC
Species11: -----CATATC-----TACAACTA
Species12: ATGGC--------------C—GCAA
Species13: AT---CCA-CCA-CT-CA-CTAG
Species14: ATG----GCCAT-----GCC---G-
Species15: --TGG--ACAT-TAG--GC—AG
Species16: A--GGCC----GCA-AGA--GGT
```

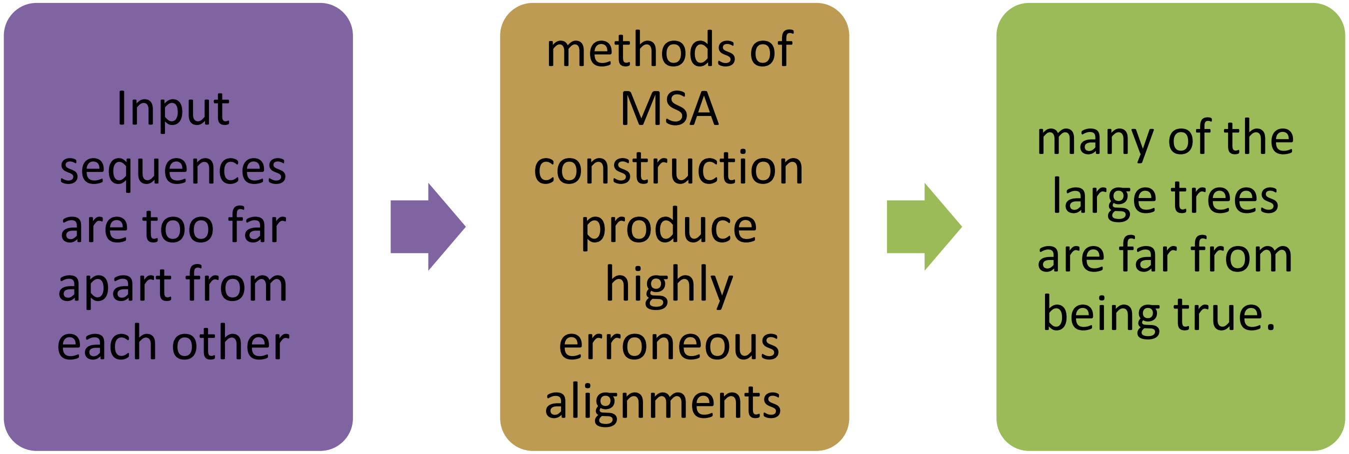
Who should we blame?

Hypothesis:

MSA

```
Species1: AT—G—CACATG-CA----CATT
Species2: ATGG-CAC—CCCAAC--GT----
Species3: ATGG-----AA—CCATC—TCA
Species4: --ATG—--AT—GC--GCG—TTA
Species5: ATGGC--------A----GC-----TTA
Species6: AT—G—CACATG-CA----CATT
Species7: AG-----TACCC---------AACTGG
Species8: A-GGC-T-----CCC----ACA-A-G
Species9:  ---TGGCAAC-CC--CCTACAAT
Species10: A---GCAT---------CCCTA---TC
Species11: -----CATATC-----TACAACTA
Species12: ATGGC---------------C—GCAA
Species13: AT---CCA-CCA-CT-CA-CTAG
Species14: ATG----GCCAT-----GCC---G-
Species15: --TGG--ACAT-TAG--GC—AG
Species16: A--GGCC----GCA-AGA--GGT
```

SUSPECT



Input sequences are too far apart from each other

methods of MSA construction produce highly erroneous alignments

many of the large trees are far from being true.

Aligning of sequences

Designed by Vecteezy

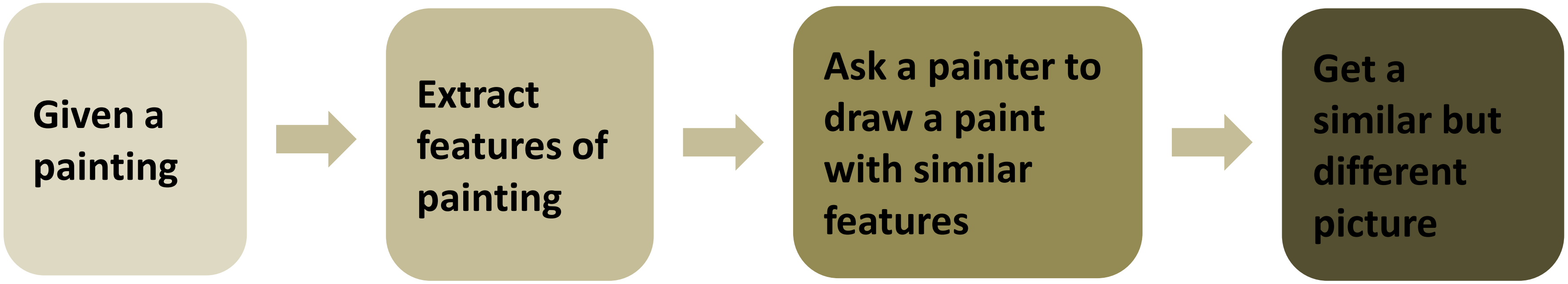
sequences evolves	*slowly*	*rapidly*
aligning is	*easy*	*hard*
good for solving	*ancient divergence events*	*recent divergence events*

How to prove that the MSA is guilty?

By simulating MSAs of different sizes (in terms of number of species) and test our hypothesis that the size of the MSA affect the quality of the tree.

Species1: ATGGCACATG-CA-CT-CATT
Species2: ATGG-CACCCCAAC--GT-TGA
Species3: ATGG-GT-AACCATGTTCATG
Species4: ATG--ATG-GCGTTAATGTTCA
Species5:ATGGCATGTTCAAATGTTCA-TT
Species6: AT—G—CACATG-CA----CATT
Species1: AT—G—CACTG-CA-CT-CATC
Species2: TGG-CACCCCAAC--GT--GT-
GG-AA--AA—CCATCTCACC
TGGT-ATGC--GC—TTATGT
GC--TG-A---GC--TTAGTT
GCACATG-CA-GT--CATT
TACCCAAGTCTAACTGG
C-T--TG-CCC---ACA-A-G
G-TGGCAAC-CC-CCTACAAT
pecies10: A-TGCAT-GTCAT-CCCTAC-T
Species1: AT—G—CACATG-CA----CATT
Species2: ATGG-CAC—CCCAAC--GT----
Spec s3: ATGG-----AA—CCATC—TCA
s4: --ATG---AT—GC--GCG—TTA
s5: ATGGC-------A----GC-----TTA
6: AT—G—CACATG-CA----CATT
7: AG-----TACCC--------AACTGG
8: A-GGC-T-----CCC----ACA-A-G
9: ---TGGCAAC-CC--CCTACAAT
0: A---GCAT--------CCCTA---TC
1: -----CATATC-----TACAACTA
2: ATGGC-------------C—GCAA
3: AT---CCA-CCA-CT-CA-CTAG
: ATG----GCCAT-----GCC---G-
ecies15: --TGG--ACAT-TAG--GC—AG
Species16: A--GGCC----GCA-AGA--GGT

Understanding the concept of simulations



Given a painting

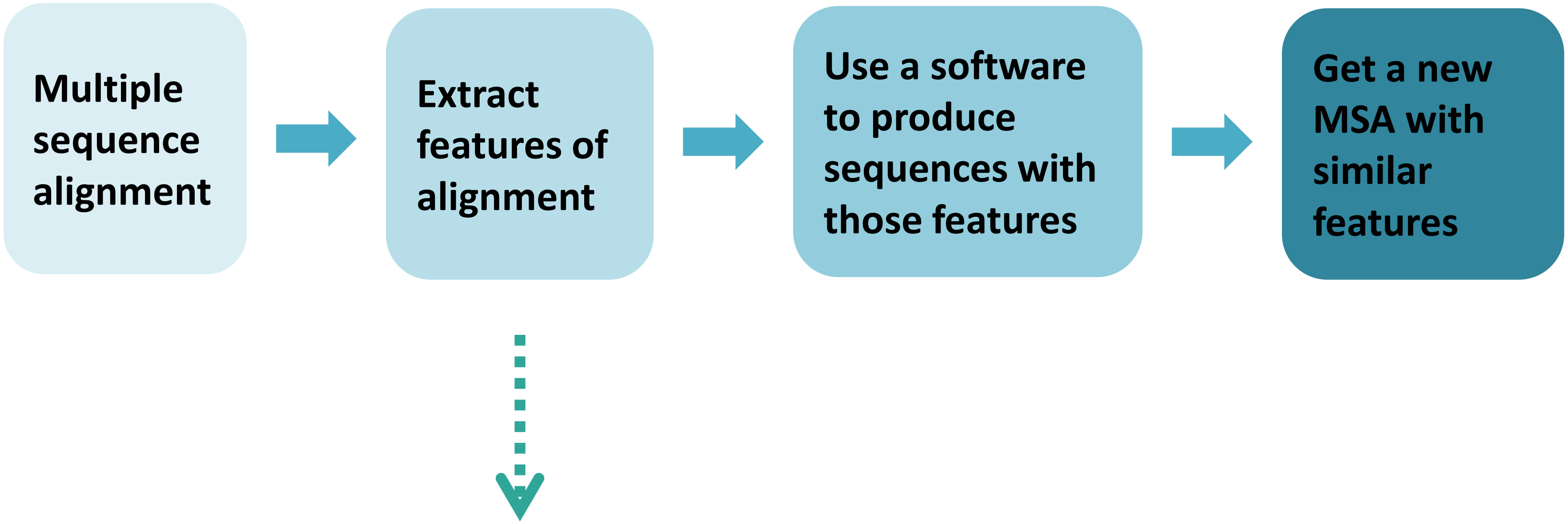
Extract features of painting

Ask a painter to draw a paint with similar features

Get a similar but different picture

- Color scale, shades
- Shapes
- Number of shapes
- Objects
- And more.

What does it mean to simulate sequences?



Multiple sequence alignment

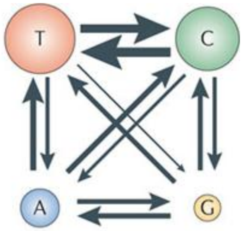
Extract features of alignment

Use a software to produce sequences with those features

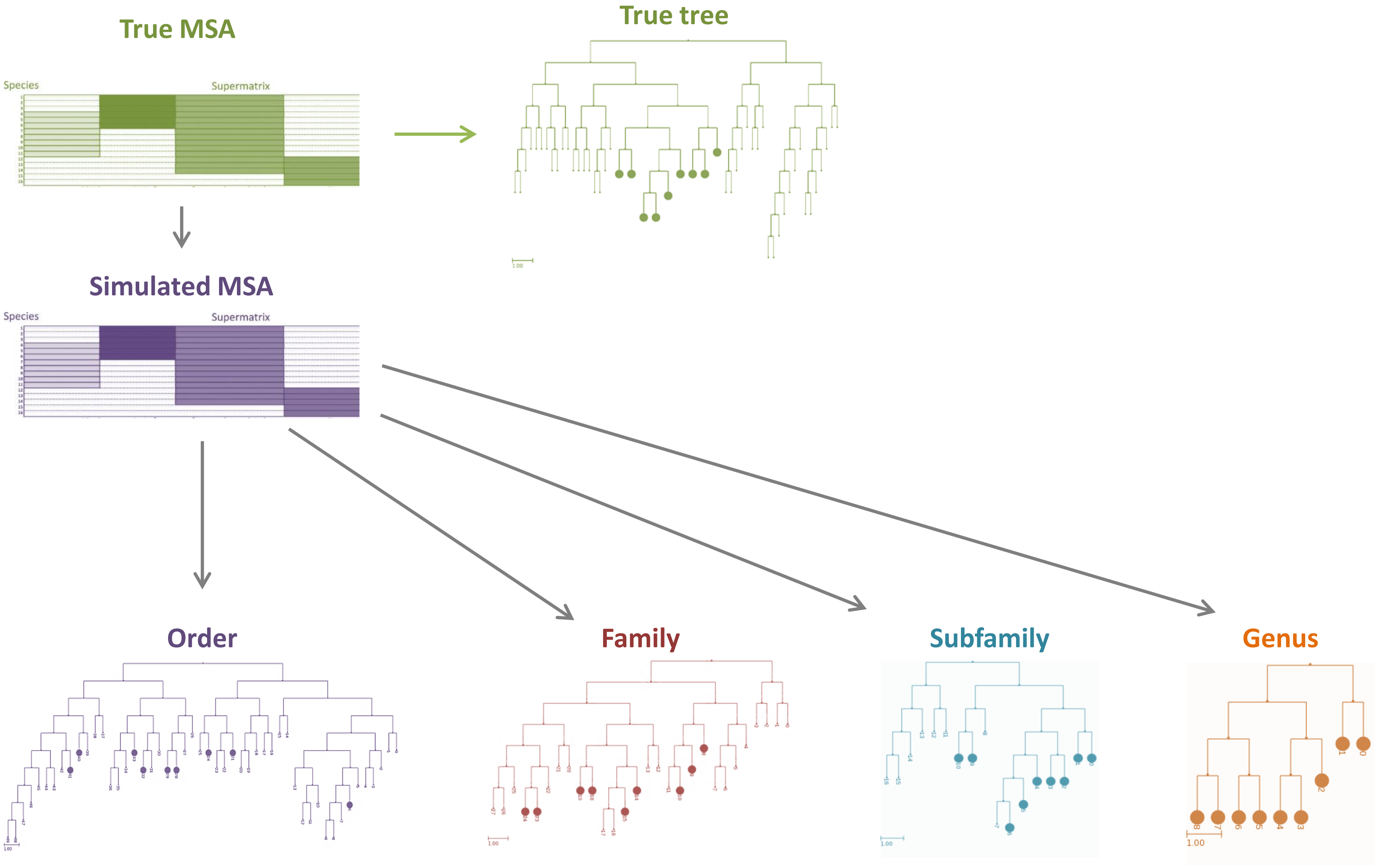
Get a new MSA with similar features

- Length of alignment 10
- Number of gaps 1
- Average lengths of gaps 3
- Nucleotides frequencies A: 30 %
- Substitutions rates (A -> C, G -> T, …)
- And more.

```
ATTGACCTGA
||     |||||
AT - - -CCTGA
```



The effect of MSA size on tree quality



True MSA

True tree

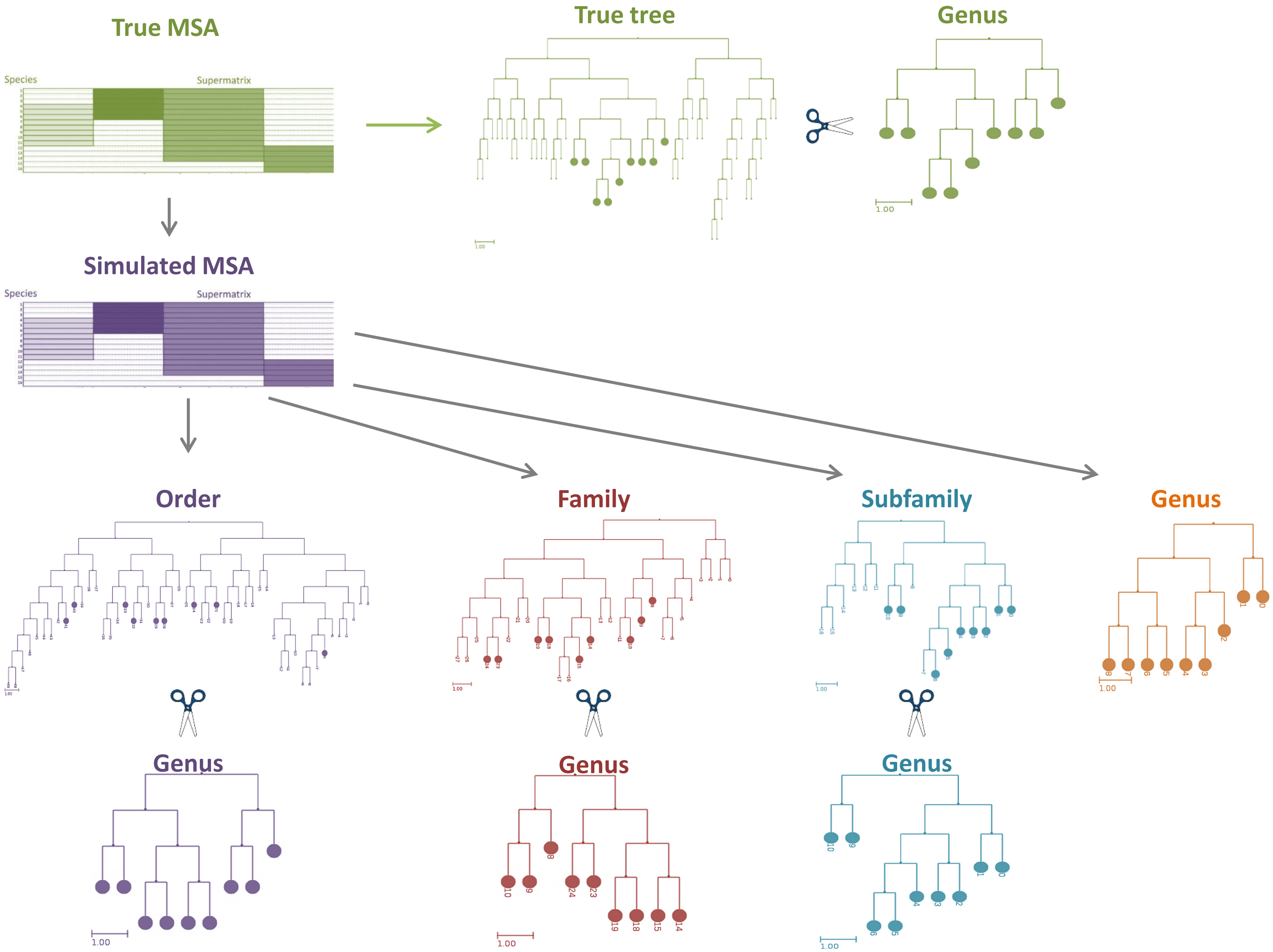
Simulated MSA

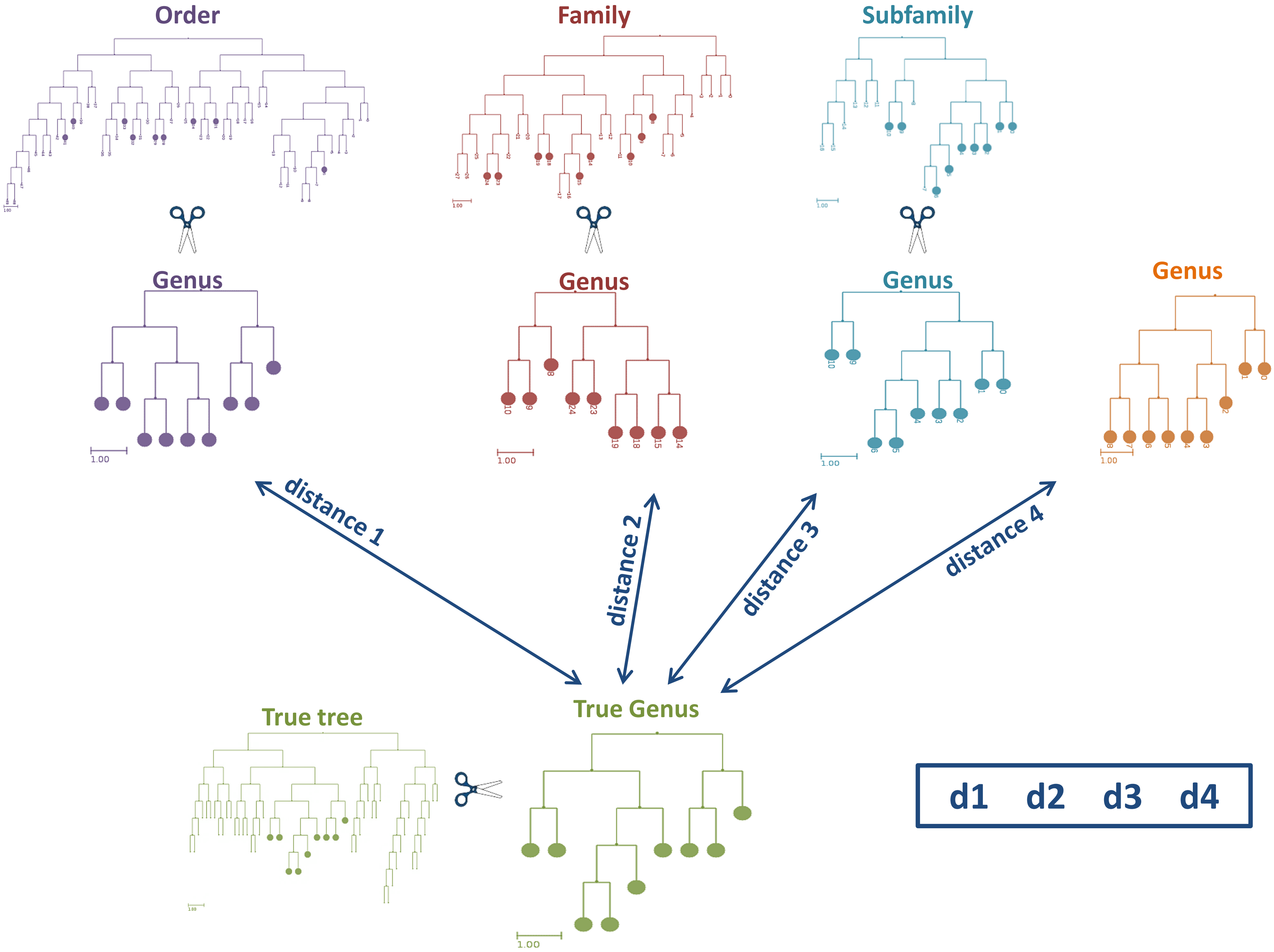
Order

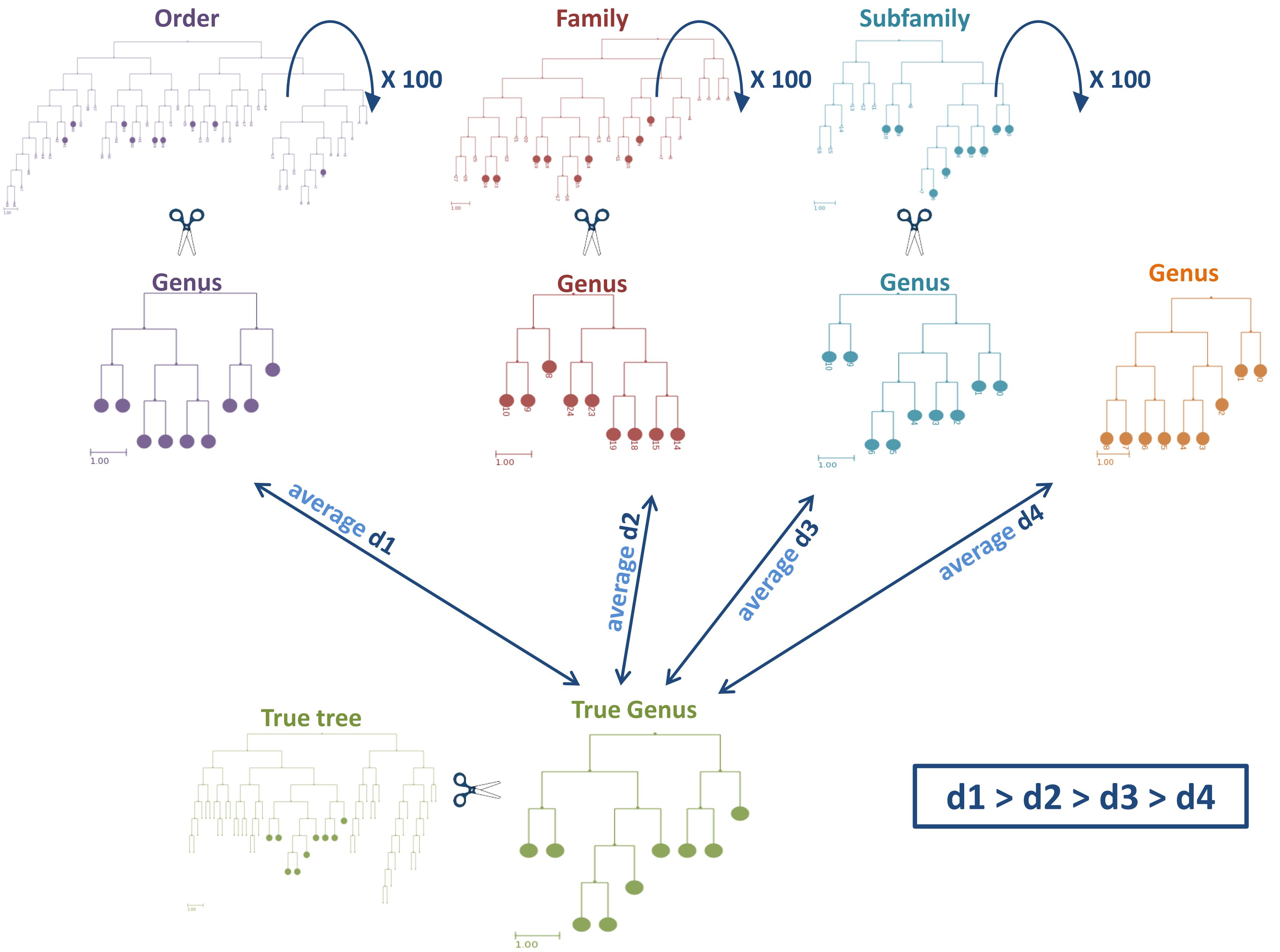
Family

Subfamily

Genus


True MSA
Species
Supermatrix
True tree
Genus
1.00
1.00
Simulated MSA
Species
Supermatrix
Order
Family
Subfamily
Genus
1.00
1.00
1.00
1.00
1
0
2
8
7
6
5
4
3
Genus
1.00
Genus
8
10
9
24
23
19
18
15
14
1.00
Genus
10
9
1
0
4
3
2
6
5
1.00


Order
Family
Subfamily
Genus
Genus
Genus
Genus
1.00
distance 1
distance 2
distance 3
distance 4
True tree
True Genus
1.00
d1 d2 d3 d4


Order
X 100
Family
X 100
Subfamily
X 100
Genus
1.00
Genus
10
9
8
24
23
19
18
15
14
1.00
Genus
10
9
1
0
4
3
2
6
5
1.00
Genus
1
0
2
8
7
6
5
4
3
1.00
average d1
average d2
average d3
average d4
True tree
True Genus
1.00
d1 > d2 > d3 > d4

How can we solve the problem?

Locus 1

```
ATGGCATATCCCATACACTAGGAT--CAA-GT-A
AT---GCTTACCCATTTCAACTTGGCT-ACAAGAT
ATGGCCA-----ACTCCCAACTAGGCTTTCAAGTT
ATGGC-ATCC--CACACAATTAGGATTCCAAGAT
ATGGCCTACCCATTCCAACTTG—TCTAC-AGA--
```

→

Locus 1

```
ATGGCATATCCCATACACTAGGAT--CAA-GT-A
AT---GCTTACCCATTTCAACTTGGCT-ACAAGAT
ATGGCCA-----ACTCCCAACTAGGCTTTCAAGTT
ATGGC-ATCC--CACACAATTAGGATTCCAAGAT
ATGGCCTACCCATTCCAACTTG—TCTAC-AGA-
```

Locus 2

```
TATCCCGGG - - -AAGCTAATT -AATGGAGGAATTTCAAGTATATT
TAT - - - -GGGCA - - - - - - - - - - - - -ATGGA - - AATTTCAA - - -TAT
TATCCCTGG - - -AAAGCTAAT -CAATGGAGGAATTTCAAGTATAT
TAA - - - -GGGCA - - - - - - - - - - - - -ATGGA - -AATTGCAA - - -TAT
TAT - - - - GCGCA - - - - - - - - - - - - -ATGCA - - AATTTCAA - - -TAT
```

→

Locus 2a Locus 2b

```
TATCCCGGGAAGCTAATTAATGGAGGAATTTCAAGTATATT-------------------------------------------
TATCCCTGGAAAGCTAATCAATGGAGGAATTTCAAGTATAT-------------------------------------------
------------------------------------------------------------------------TATGGGCAATGGAAATTTCAATATT
------------------------------------------------------------------------TAAGGGCAATGGAAATTGCAATAT
------------------------------------------------------------------------TATGCGCAATGCAAATTTCAATATT
```

Supermatrix:

```
ATGGCATATCCCATACAACTAGGATTCCAAGTTATCCCGGGAAGCTAATTAATGGAGGAATTTCAAGTATATT-------------------------------------------
ATGGCTTACCCATTTCAACTTGGCTTACAAGATATCCCTGGAAAGCTAATCAATGGAGGAATTTCAAGTATAT-------------------------------------------
ATGGCCAACCACTCCCAACTAGGCTTTCAAGT------------------------------------------------------------------------TATGGGCAATGGAAATTTCAATATT
ATGGCACATCCCACACAATTAGGATTCCAAGA------------------------------------------------------------------------TAAGGGCAATGGAAATTGCAATAT
ATGGCCTACCCATTCCAACTTGGTCTACAAGA------------------------------------------------------------------------TATGCGCAATGCAAATTTCAATATT
```

ACTGTGCAHTGATGNGTAGCKGS