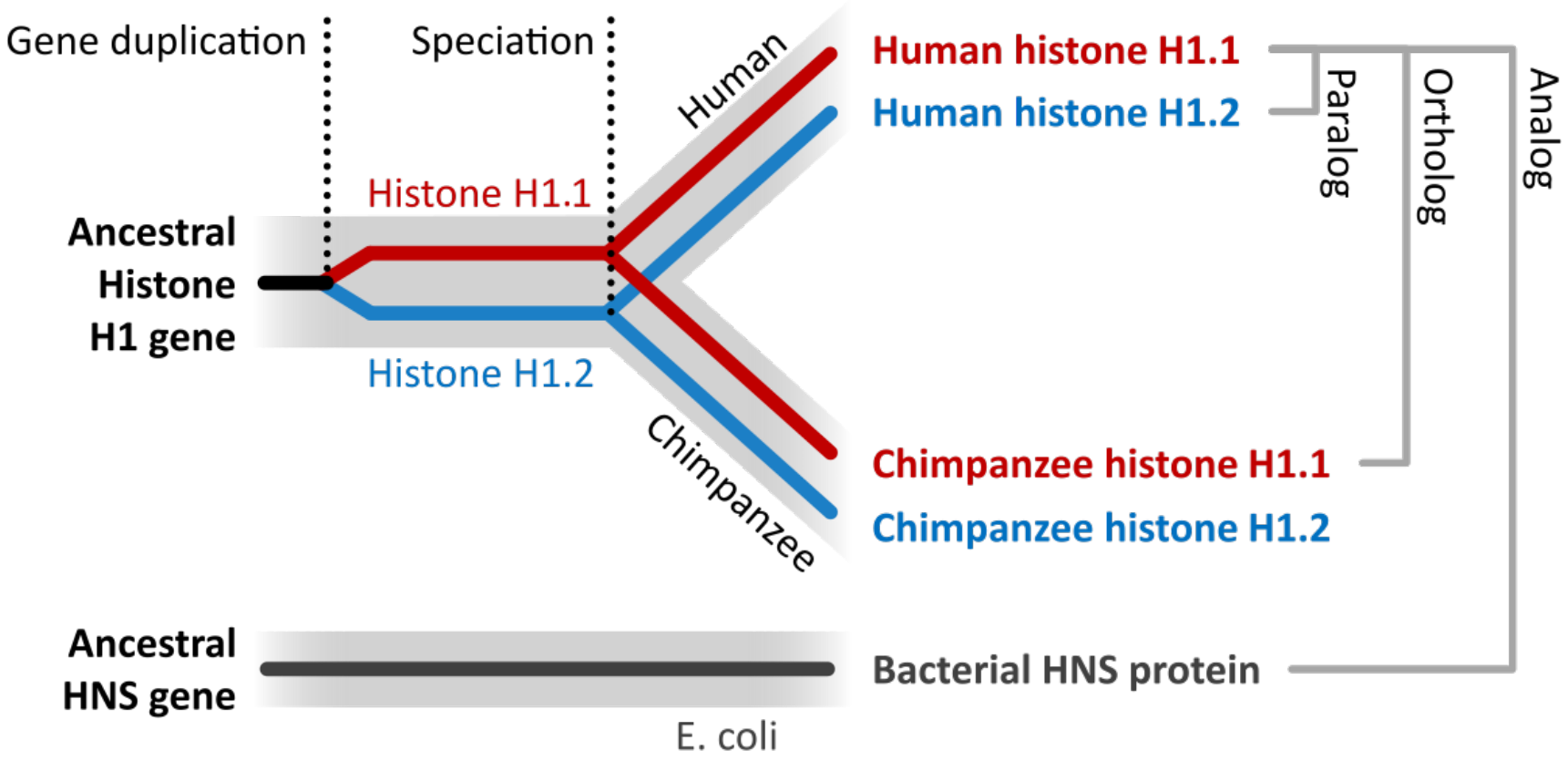
Sequence Clustering & Phylogenetic Analysis

Qi Sun, Cheng Zou

Bioinformatics Facility

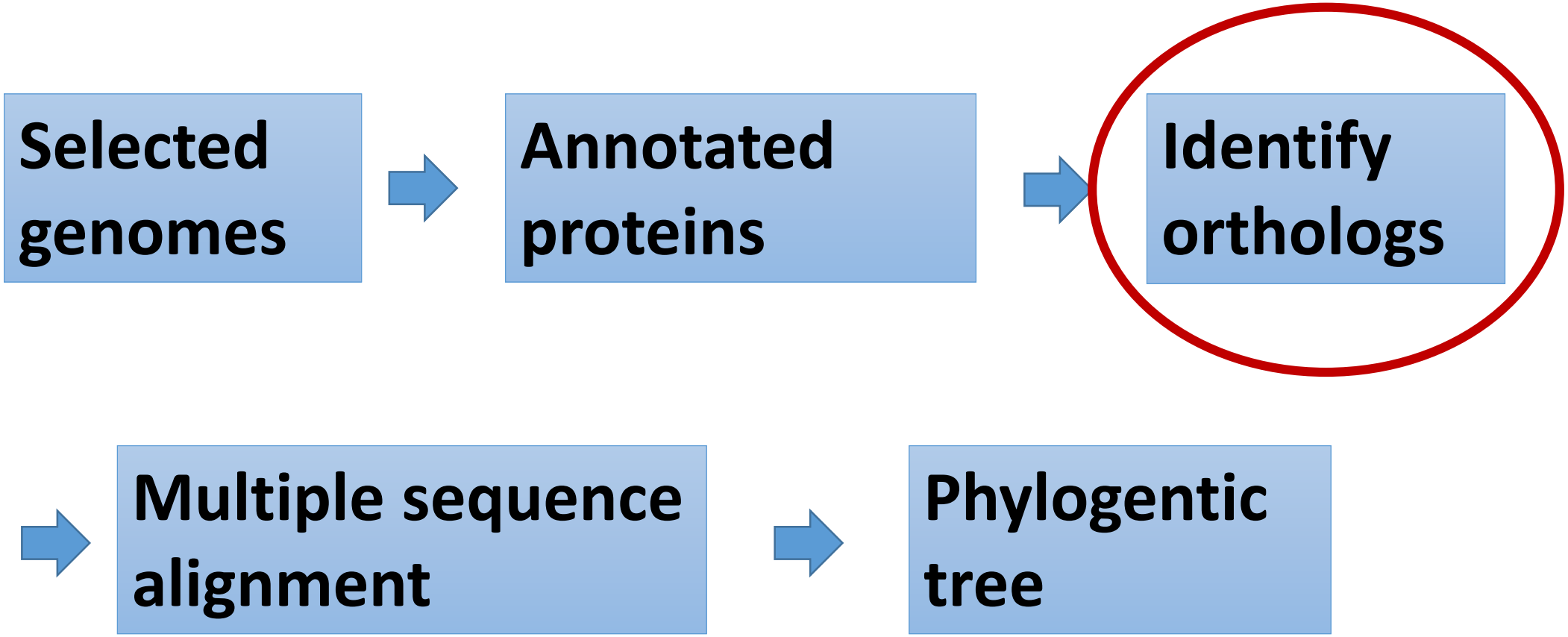
Cornell University

Evolution of genes: Ortholog vs Paralog



https://en.wikipedia.org/wiki/Sequence_homology

A pipeline for phylogenetic analysis



Cluster analysis

FASTA file

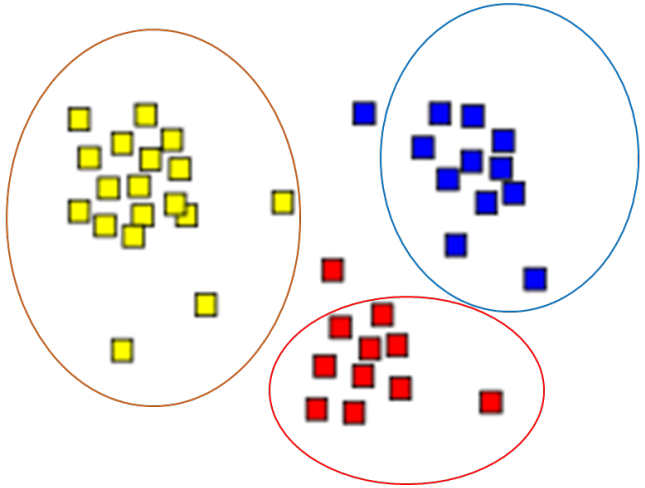
```
>NP_001014992.2 inositol 1,4,5-triphosphate kinase [Apis mellifera]
MSRSINMDQEKKNNVENLKSGGSTTPASPTLSTPPTLNLMEQILLAKIEKQNLHESDDLHESDGRVGGKRRNILLRRTDS
MDSQNSASTYNSFLSSDSASSGNVYCKCDDCLLGIVDDYQRNPSVVGRKKSSGWRKLRNIVHWTPFFQTYKKQRYPWVQL
AGHQGNFRAGPTPGTILKKLCPQEEACFRLLMNDILRPYVPEFKGVLDVKDVEEGNVEETNSEETHQKDGSSDSVIKRTV
VSSYLQLQDLLGDFEHPCVMDCKVGVRTYLESELAKAKERPKLRKDMYEKMVQVDPTAPNAEERRVQGVTKPRYMVWRET
ISSTATLGFRVEGIKLAHGGSSKDFKTTRTREQVTEALRRFVEGYPHAVPKYIQRLKAIRATLKASPFFASHEVVGSSLL
FVHDTKNAGIWMIDFAKTLPLPQHLPRIHHDAEWKVGNHEDGYLIGVNNLIDIFQDIRNSEET
>NP_001014993.1 elongation factor 1-alpha [Apis mellifera]
MGKEKIHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAQEMGKGSFKYAWVLDKLKAERERGITIDIALWKF
ETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGTGEFEAGISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEP
PYSETRFEEIKKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEVSSKMPWFKGWTVERKEGKVEGKCLIEALDAILPPTR
PTDKALRLPLQDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPAGLTTEVKSVEM
>NP_001014994.1 glycerol-3-phosphate dehydrogenase [Apis mellifera]
MAEKLRICIVGSGNWGSTIAKIIGINAANFSNFEDRVTMYVYEEIINGKKLTEIINETHENVKYLPGHKLPPNIIAIPDV
VEAAKDADILTFVVPHQFIKRICSALFGKIKPTAIGLSLIKGFDKKQGGGIELISHIISKQLHIPVSVLMGANLASEVAN
EMFCETTIGCKDKNMAPILKDLMETSYFKVVVVEDVDSVECCGALKNIVACGAGFIDGLGLGDNTKAAVMRLGLMEIIKF
VNIFFPGGKKTTFFESCGVADLIATCYGGRNRKICEAFVKTGKKISELEKEMLNGQKLQGPFTAEEVNYMLKAKNMENRF
PLFTTVHRICIGETMPMELIENLRNHPEYIDETRNYQECKCSI
>NP_001019868.1 major royal jelly protein 9 precursor [Apis mellifera]
MSFNIWWLILYFSIVCQAKAHYSLRDFKANIFQVKYQWKYFDYNFGSDEKRQAAIQSGEYNYKNNVPIDVDRWNGKTFVT
ILRNDGVPSSLNVISNKIGNGGPLLEPYPNWSWAKNQNCSGITSVYRIAIDEWDRLWVLDNGISGETSVCPSQIVVFDLK
NSKLLKQVKIPHDIAINSTTGKRNVVTPIVQSFDYNNTWVYIADVEGYALIIYNNADDSFQRLTSSTFVYDPRYTKYTIN
DESFSLQDGILGMALSHKTQNLYYSAMSSHNLNYVNTKQFTQGKFQANDIQYQGASDILWTQASAKAISETGALFFGLVS
DTALGCWNENRPLKRRNIEIVAKNNDTLQFISGIKIIKQISSNIYERQNNEYIWIVSNKYQKIANGDLNFNEVNFRILNA
PVNQLIRYTRCENPKTNFFSIFL
>NP_001027532.1 follistatin-like 5 [Apis mellifera]
MRCMLEIAARSFLLLSIASTYVVSVAGYKHSRRHRDFTVAESYDASSSNSDSLSMTIPPSIDRSSIHEESYLAESSRSID
PCASKYCGIGKECELSPNSTIAVCVCMRKCPRRHRPVCASNGKIYANHCELHRAACHSGSSLTKSRLMRCLHHDIENAHI
RRTLHMNRTSLKTSKIVSYPKSRSRKKGGLKDNLIPDKNDPDSKECSNQEYEIMKDNLLLYNHARLMSQDNHSKEYLVSI
MFSHYDRNNNGNLEREELEQFAENEDLEELCRGCNLGHMISYDDTDGDGKLNVNEFYMAFSKLYSVSVVSLDKSLEVNHI
SARVGDNVEIKCDVTGTPPPPLVWRRNGADLETLNEPEIRVFNDGSLYLTKVQLIHAGNYTCHAVRNQDVVQTHVLTIHT
IPEVKVTPRFQAKRLKEEANIRCHVAGEPLPRVQWLKNDEALNHDQPDKYDLIGNGTKLIIKNVDYADTGAYMCQASSIG
GITRDISSLVVQEQPTPTTESEERRFFSFHQWGILVYEPSACRPRHEIRSTDVIPGTQEHVCGVKGIPCSWGRAINVANR
IGGLQHPGAVVWFTVSLH
>NP_001032395.1 putative tyramine receptor [Apis mellifera]
MANQTANYYGDVYQWNHTVSSGERDTRTEYYLPNWTDLVLAGLFTMLIIVTIVGNTLVIAAVITTRRLRSVTNCFVSSLA
AADLLVGLAVMPPAVLLQLTGGTWELGPMLCDSWVSLDILLCTASILSLCAISIDRYLAVTQPLIYSRRRRSKRLAGLMI
VAVWVLAGAITSPPLLGCFPRATNRDIKKCSYNMDSSYVIFSAMGSFFLPMLVMLYVYGRISCVIASRHRNLEATESENV
RPRRNVLIERAKSIRARRTECVTNSVTCDRPSDEAEPSSTSKKSGIVRSHQQSCINRVARETKTAGTLAVVVGGFVACWL
PFFILYLATPFVPVEPPDILMPALTWLGWINSAINPFIYAFYSADFRLAFWRLTCRKCFKSRTNLDPSNRKLPAPANWKK
DTTRT
```

...

Similarity matrix

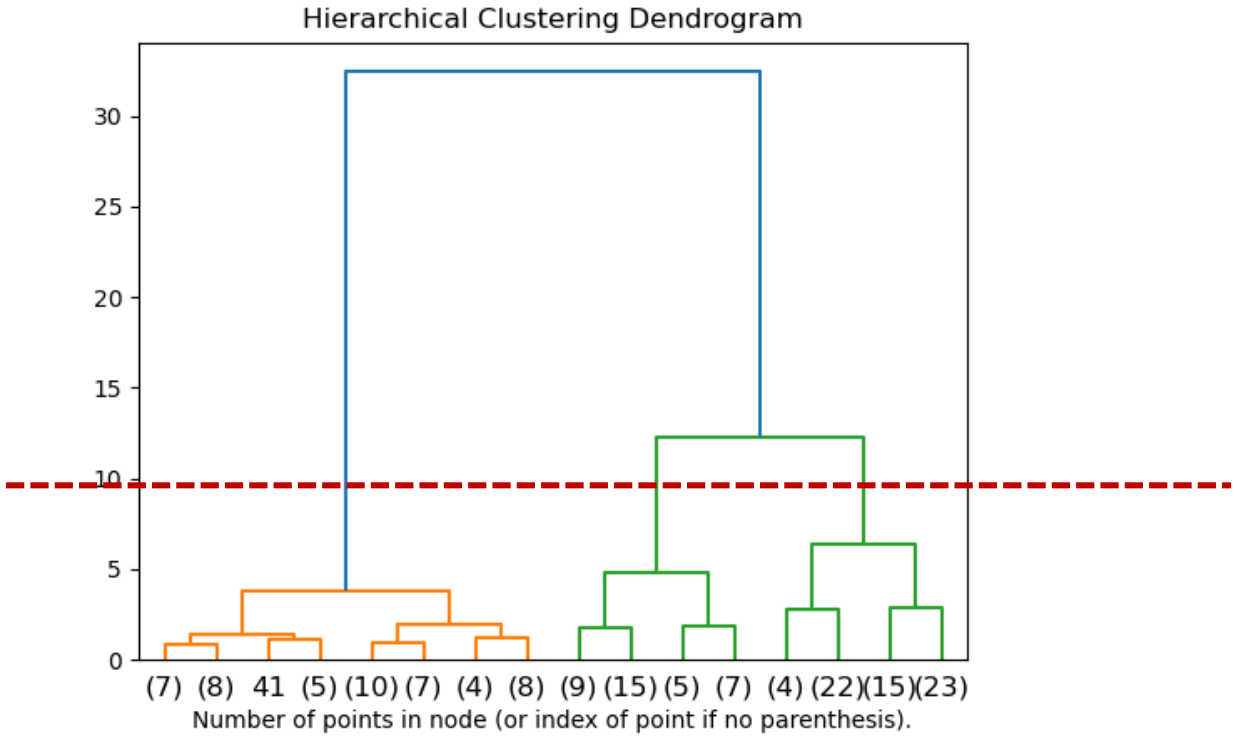
	seq1	seq2	seq3	seq4	seq5
seq1	1				
seq2	0.9	1			
seq3	0.7	0.6	1		
seq4	0.8	0.8	0.6	1	
seq5	0.9	0.5	0.8	0.7	1

Clustering

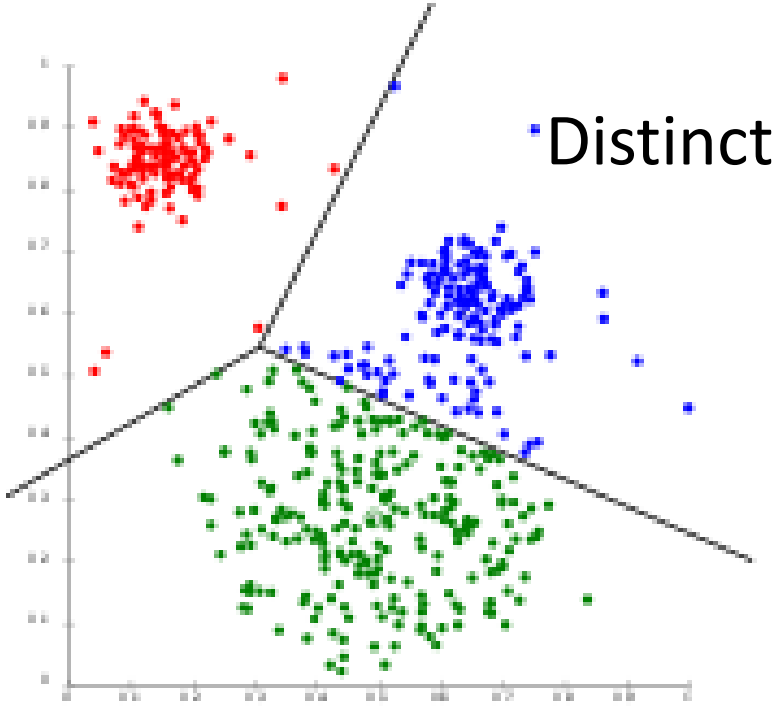


K-means vs Hierarchical Clustering

Hierarchical



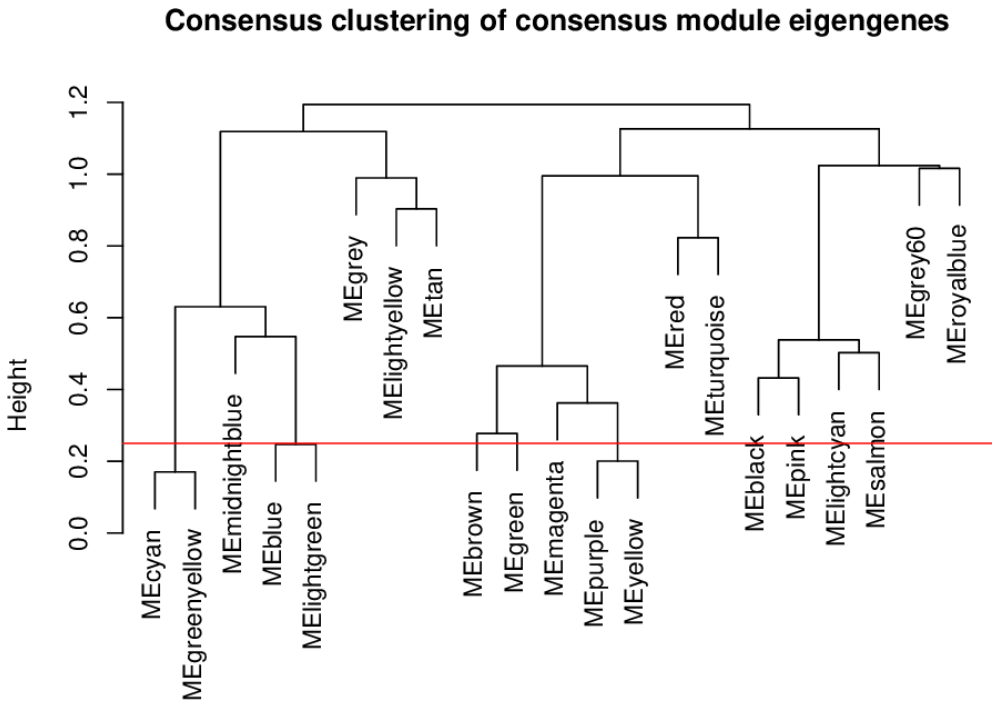
K-means

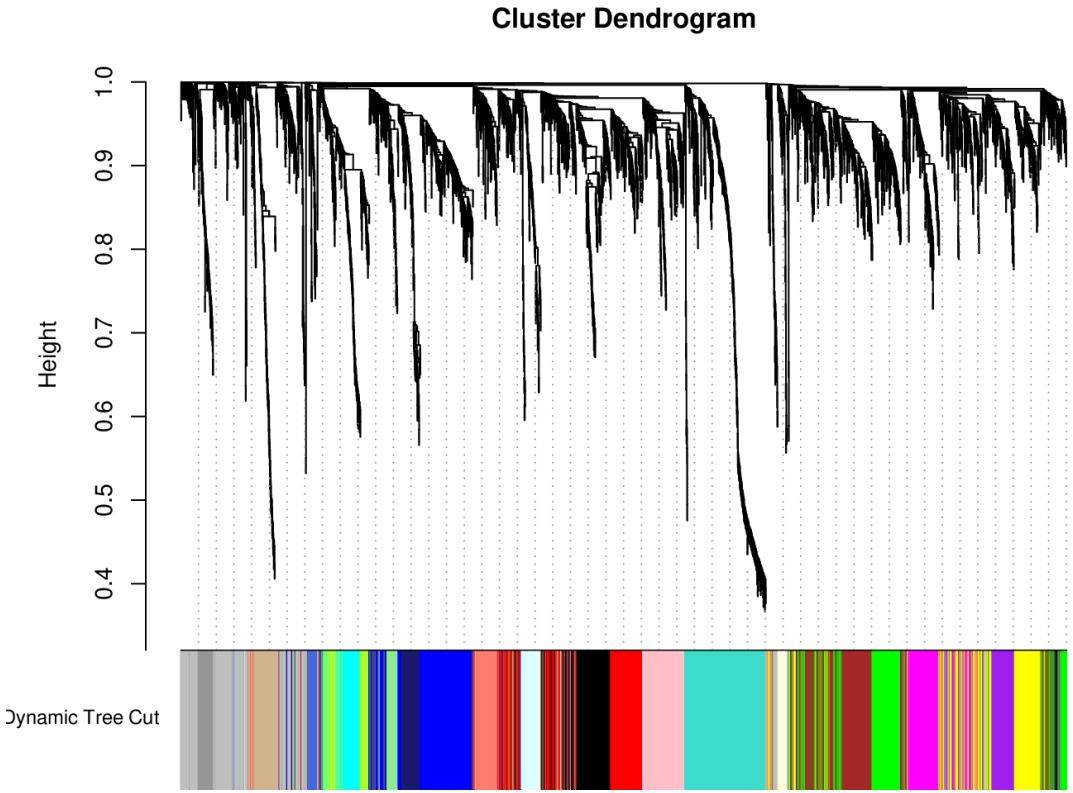


Distinct groups

- pre-specified K (number of groups);
- prefer clusters of similar size

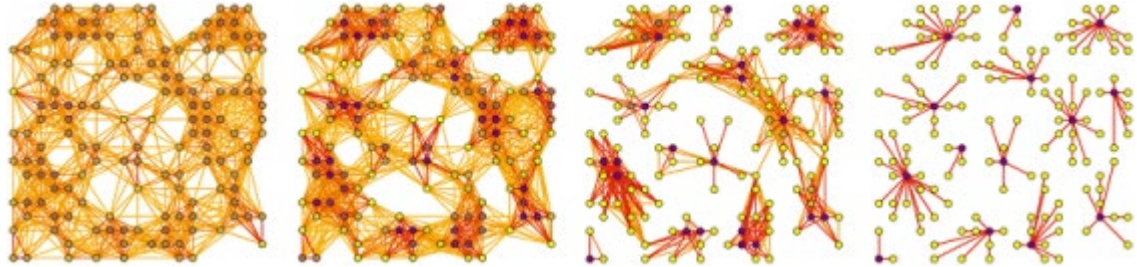
WGCNA transformation





WGCNA tutorial by P. Langfelder, S. Horvath

MCL
- a cluster algorithm for graphs



- Single Parameter: -I (Inflation, range from 1.1 to 10.0)

- Increasing the value of **-I** will increase cluster granularity

MCL on BLAST results (https://micans.org/mcl/)

```
cut -f 1,2,11 blastOutfmt6.txt > seq.abc

mcxload -abc seq.abc --stream-mirror --stream-neg-log10 -stream-tf 'ceil(200)' -o seq.mci -write-tab seq.tab

mcl seq.mci -I 5.0 -use-tab seq.tab
```

#cut the blast output to 3 columns: querID, targetID, evalue

#transform evalues

#run mcl.

Output text file: each line is a cluster, with gene names in the same cluster separated by “tab”

Gene1	Gene234	Gene56
Gene3		
Gene43	Gene12	
Gene653	Gene877	
……		

MMseqs2: an ultra fast protein/DNA sequence clustering tools

Command:

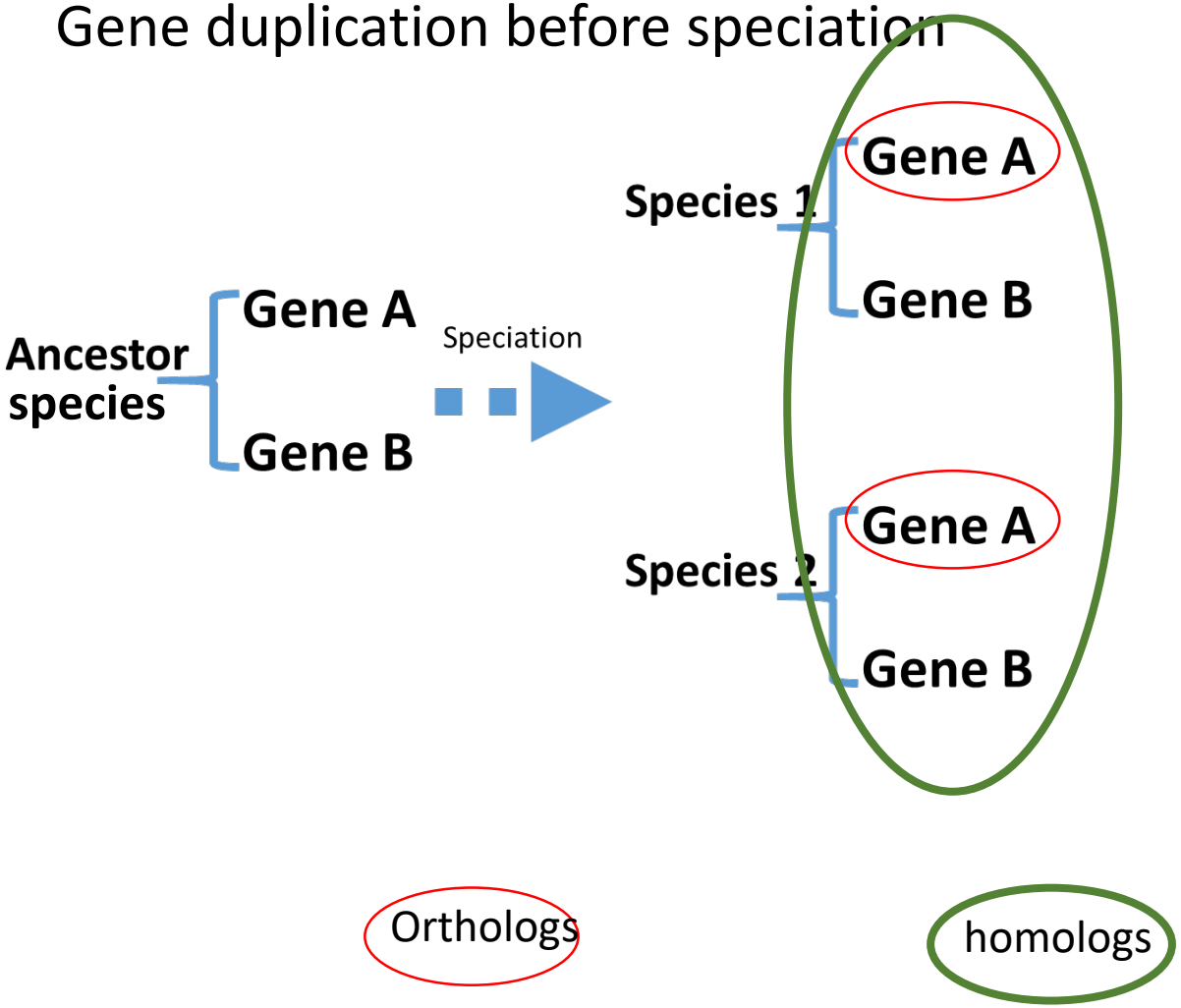
```
mmseqs easy-linclust input.fasta clusterResult tmp
```

- easy-linclust is a tool in the package that scales linearly with number of sequences;
- Fast speed due to pre-filtering through k-mer matching

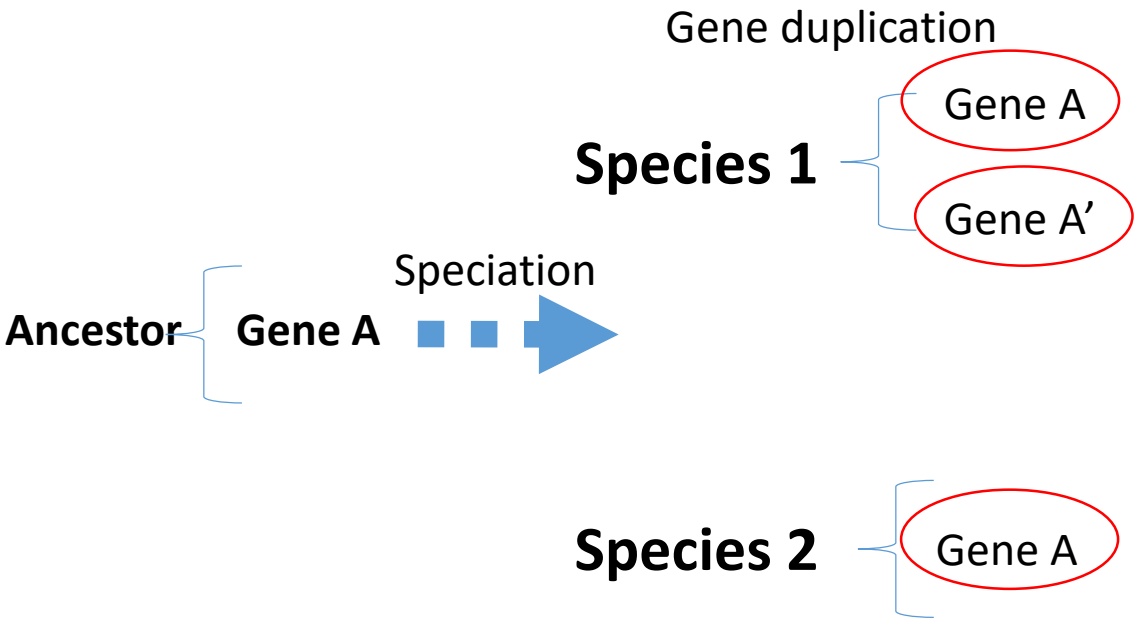
Sequence clustering alone cannot solve orthologous relationship

Gene duplication and speciation, which happens first?

Gene duplication before speciation



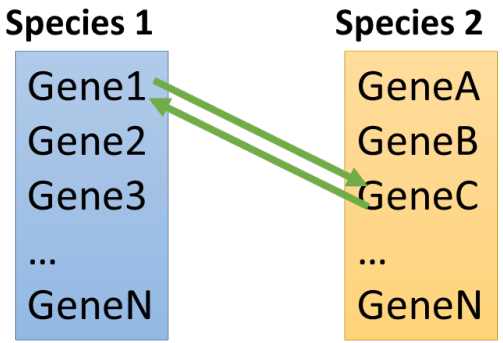
Gene duplication after speciation



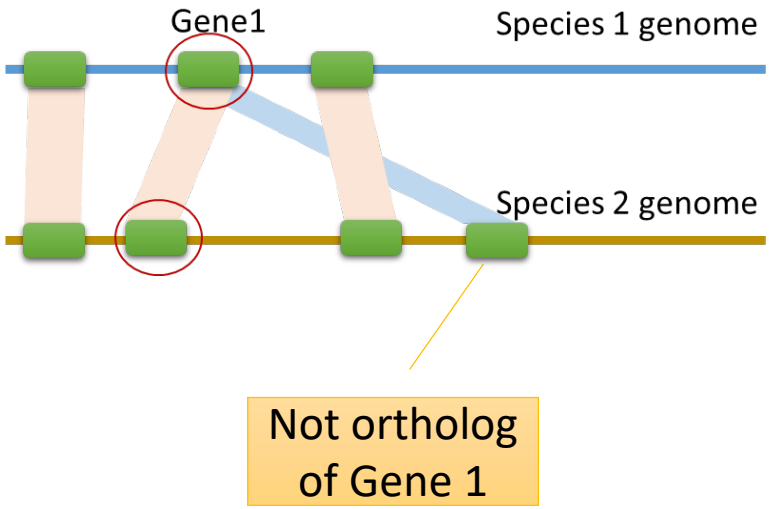
How to define orthologs?

Without direct proof, here are the commonly accepted practices:

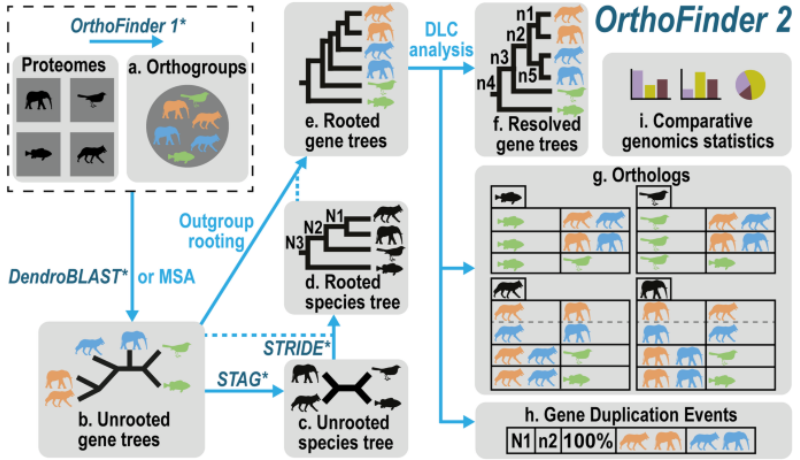
Reciprocal best BLAST hit



Synteny



Inferred from phylogeny



Orthofinder

Software for identifying orthologous genes *

Orthofinder (MCL + reciprocal best blast hits + phylogeny)

Input: A set of fasta files, with each fasta file represents all proteins from a single genome.

Output: Genes in orthologous groups. Genes duplicated after speciation would be in the same group.

MCScanX (synteny)

Input: BLAST (gene sequence similarity) and GFF (gene position on the genome)

Output: Collinear orthologs.

Phylogenetic analysis pipelines:

Concatenated orthologous genes

Orthologous gene identification

MSA for each ortholog group

Build phylogenetic tree using MSA

SNP based

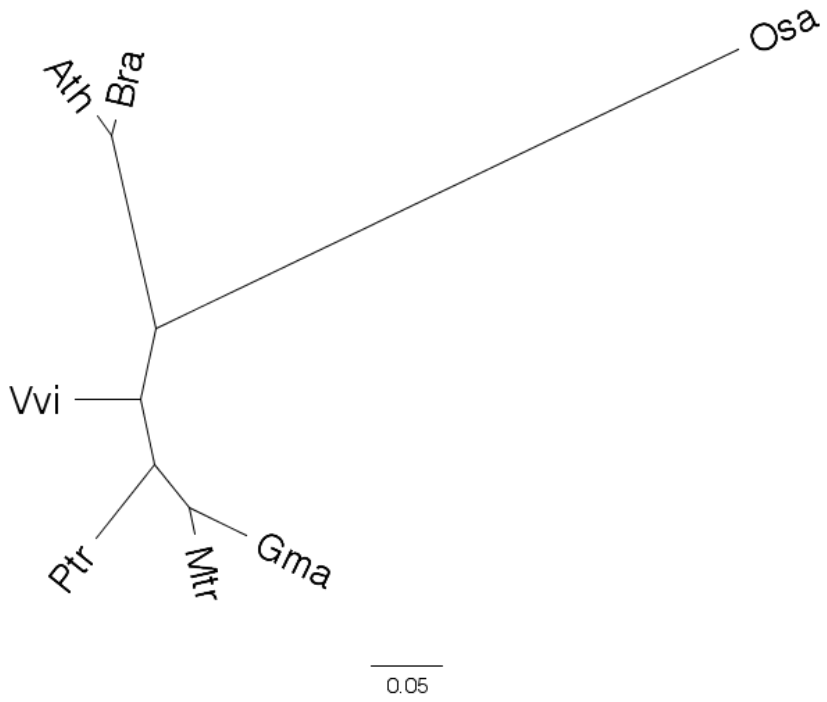
SNP calling on a reference genome

Build phylogenetic tree from vcf file

Phylogeny

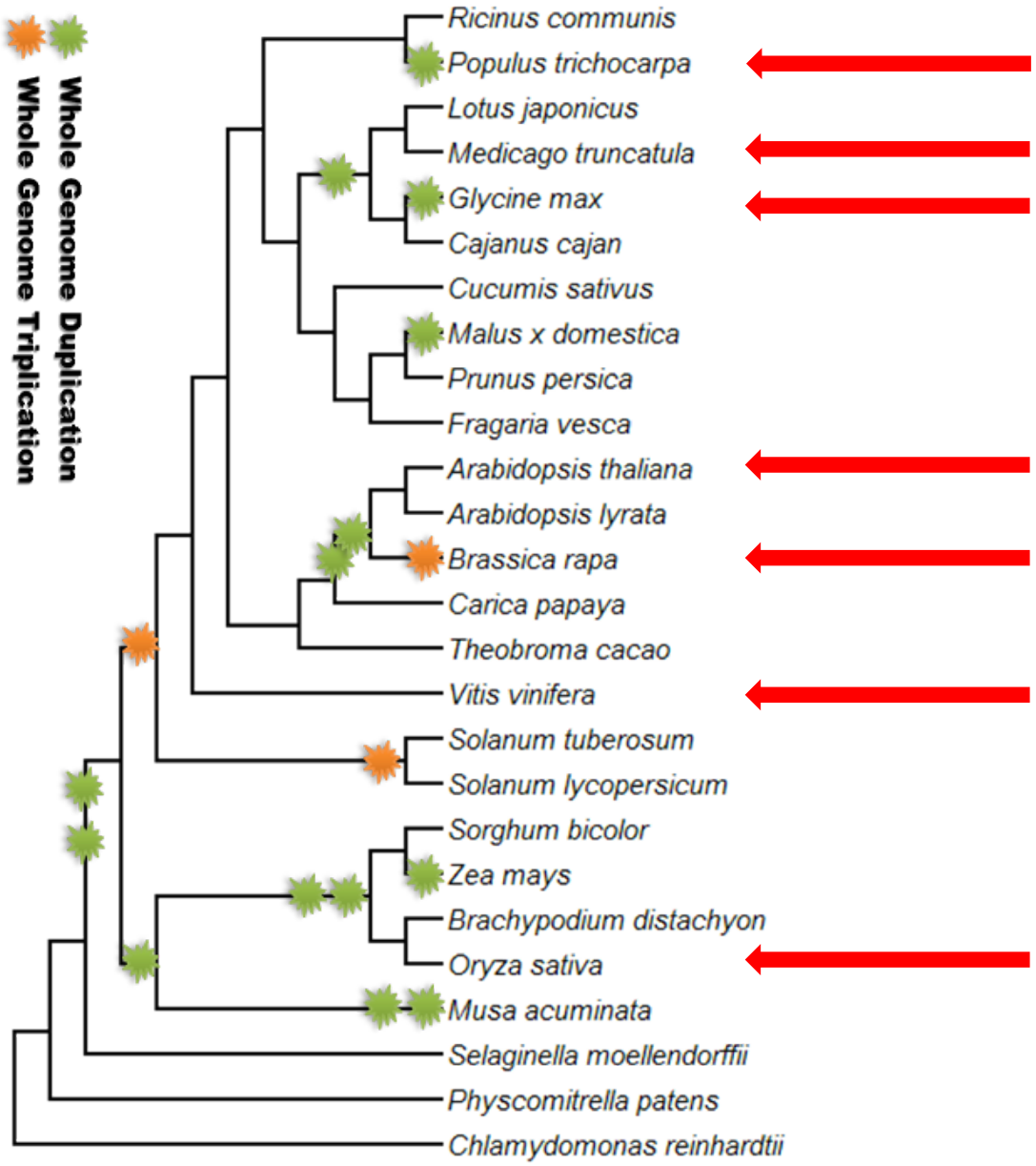
What is phylogeny?

- A phylogeny, also known as a tree, is an explanation of how things evolved, their **evolutionary relationships** between “taxa” (entities such as genes, populations, species, etc.)
- In this introduction, we focus on phylogeny reconstruction by **sequences**



Outline

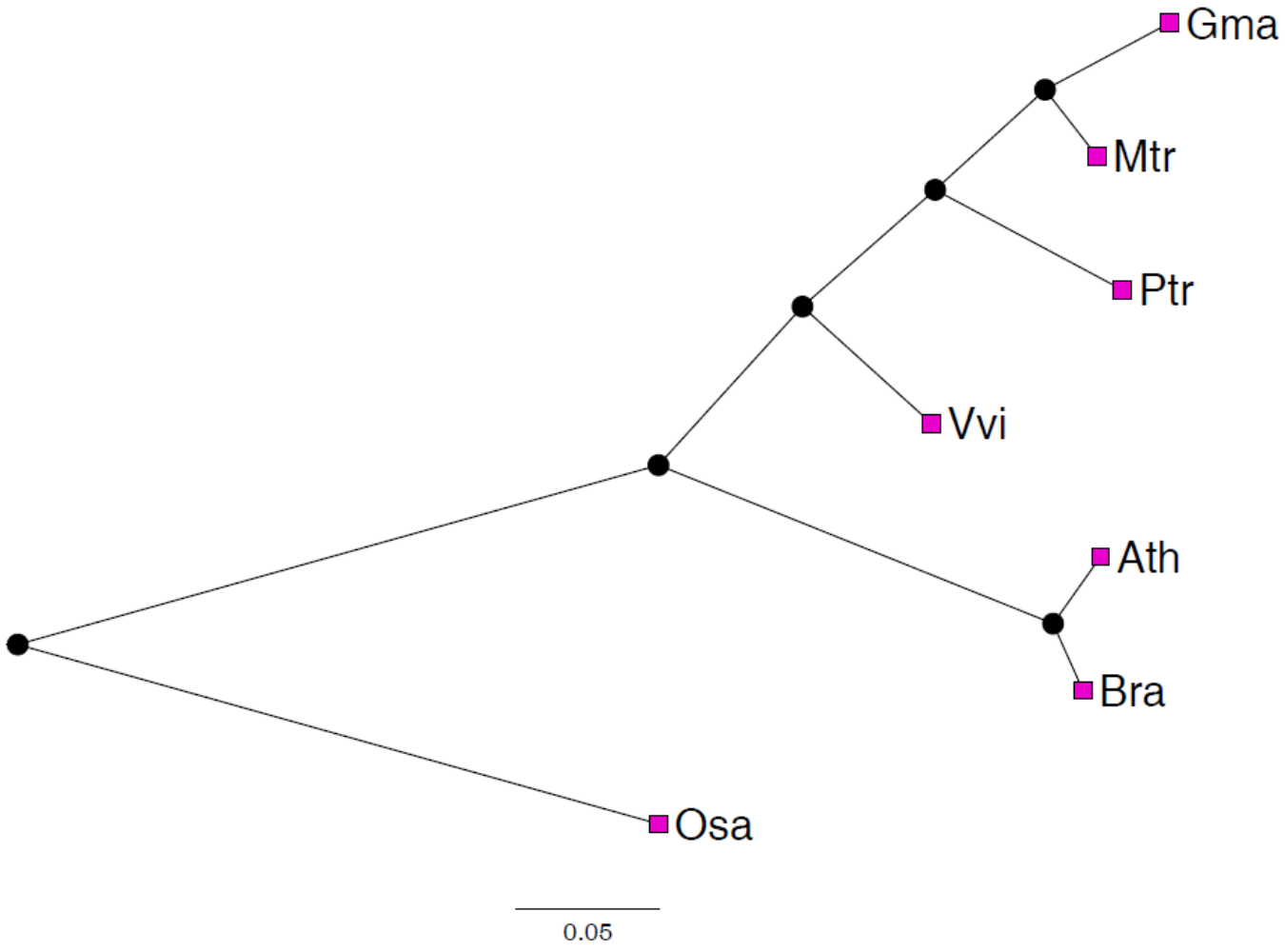
- Elements in the phylogeny
- How to construct a phylogeny
- Phylogeny evaluation and illustration
- Hands on practice



LEAFY Controls Floral Meristem Identity in Arabidopsis

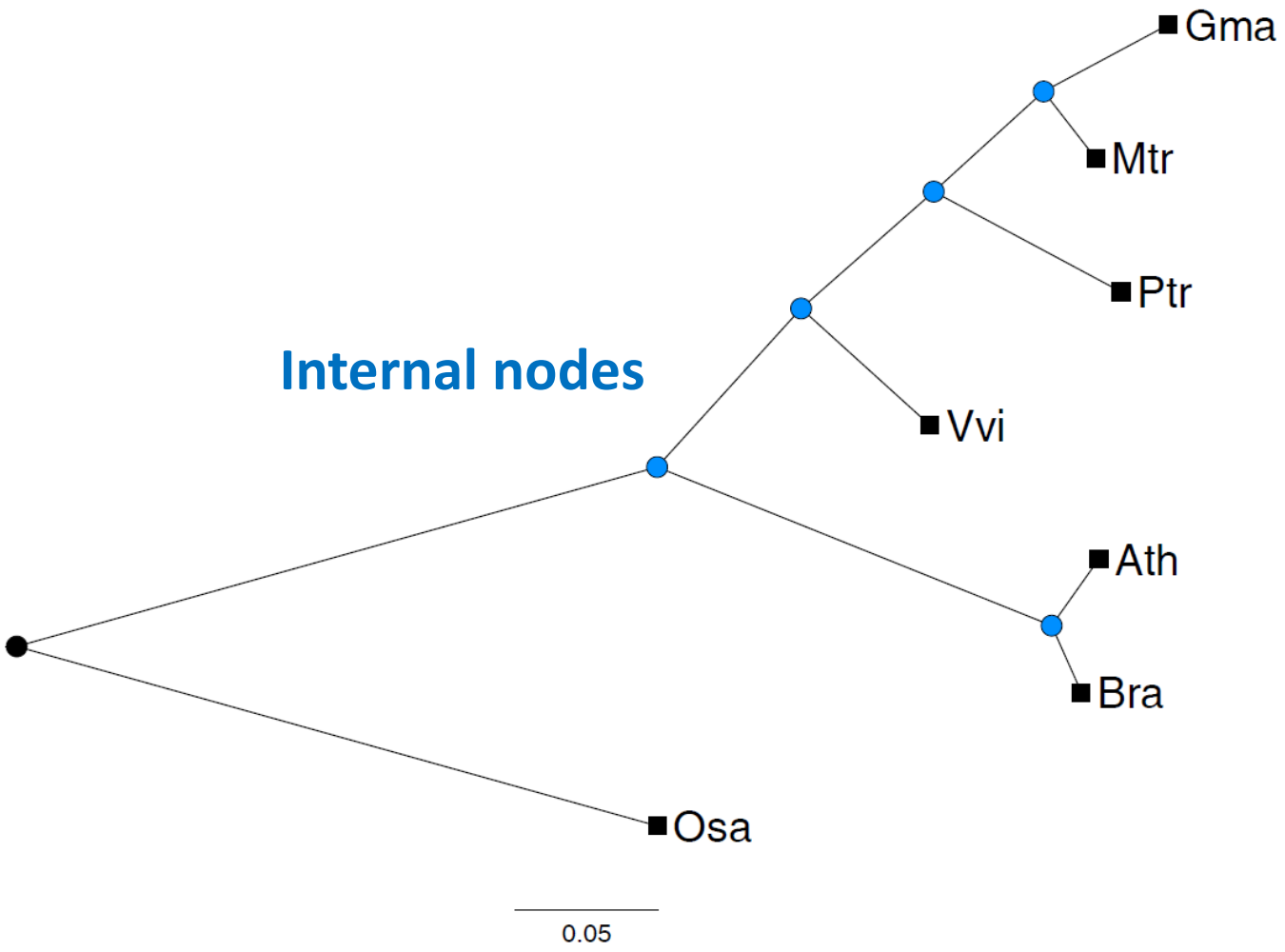
Weigel D, Alvarez J, Smyth D R, et al. LEAFY controls floral meristem identity in Arabidopsis[J]. Cell, 1992, 69(5): 843-859.

Elements of a phylogeny tree

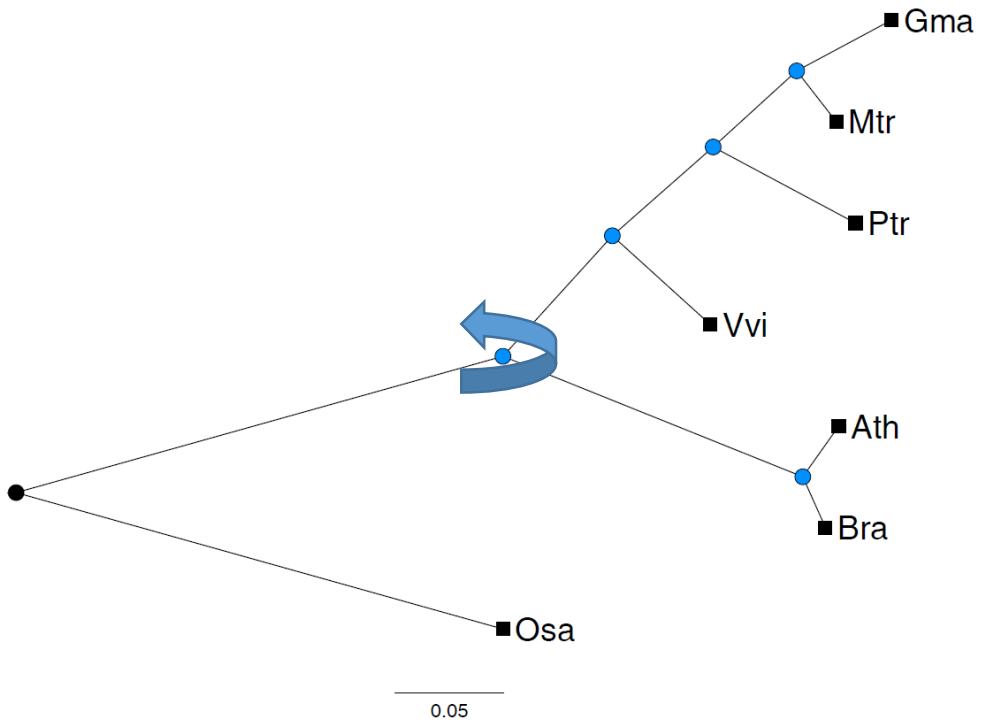


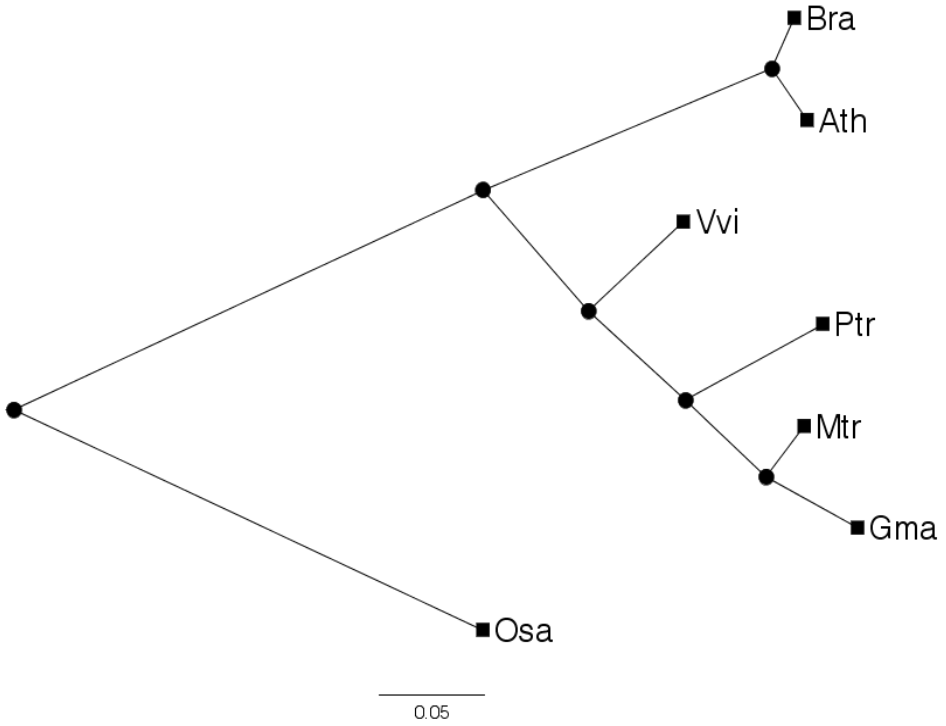
Taxa / Terminal Node/ tips

Elements of a phylogeny tree

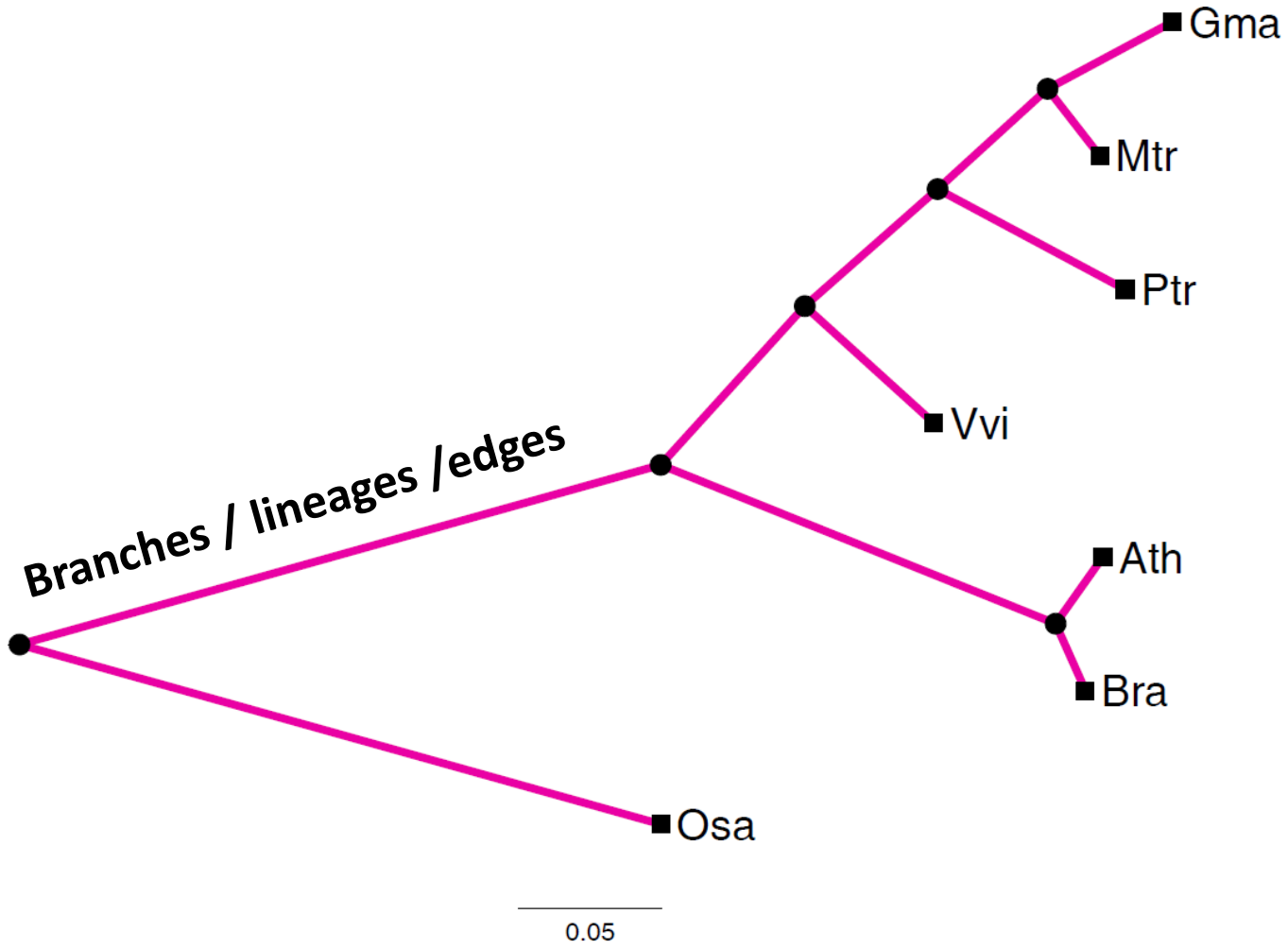


Two branches at any internal note can be flipped without changing the phylogeny

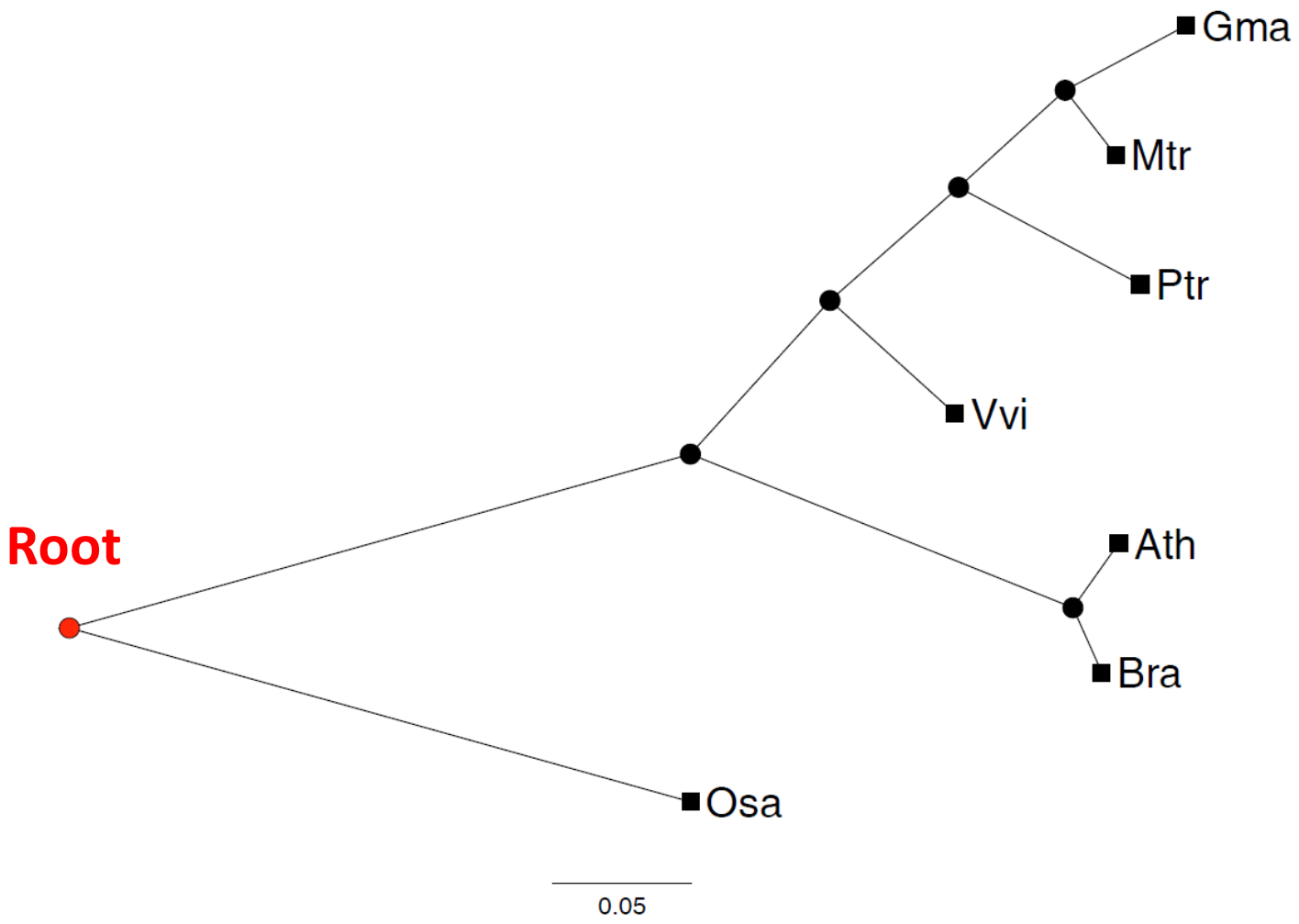




Elements of a phylogeny tree

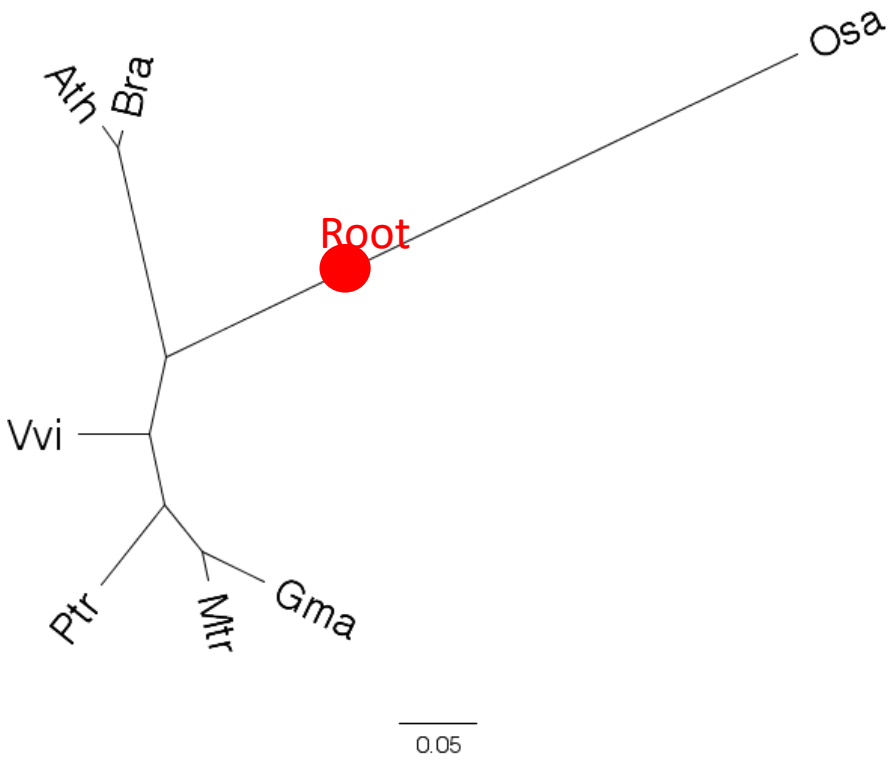


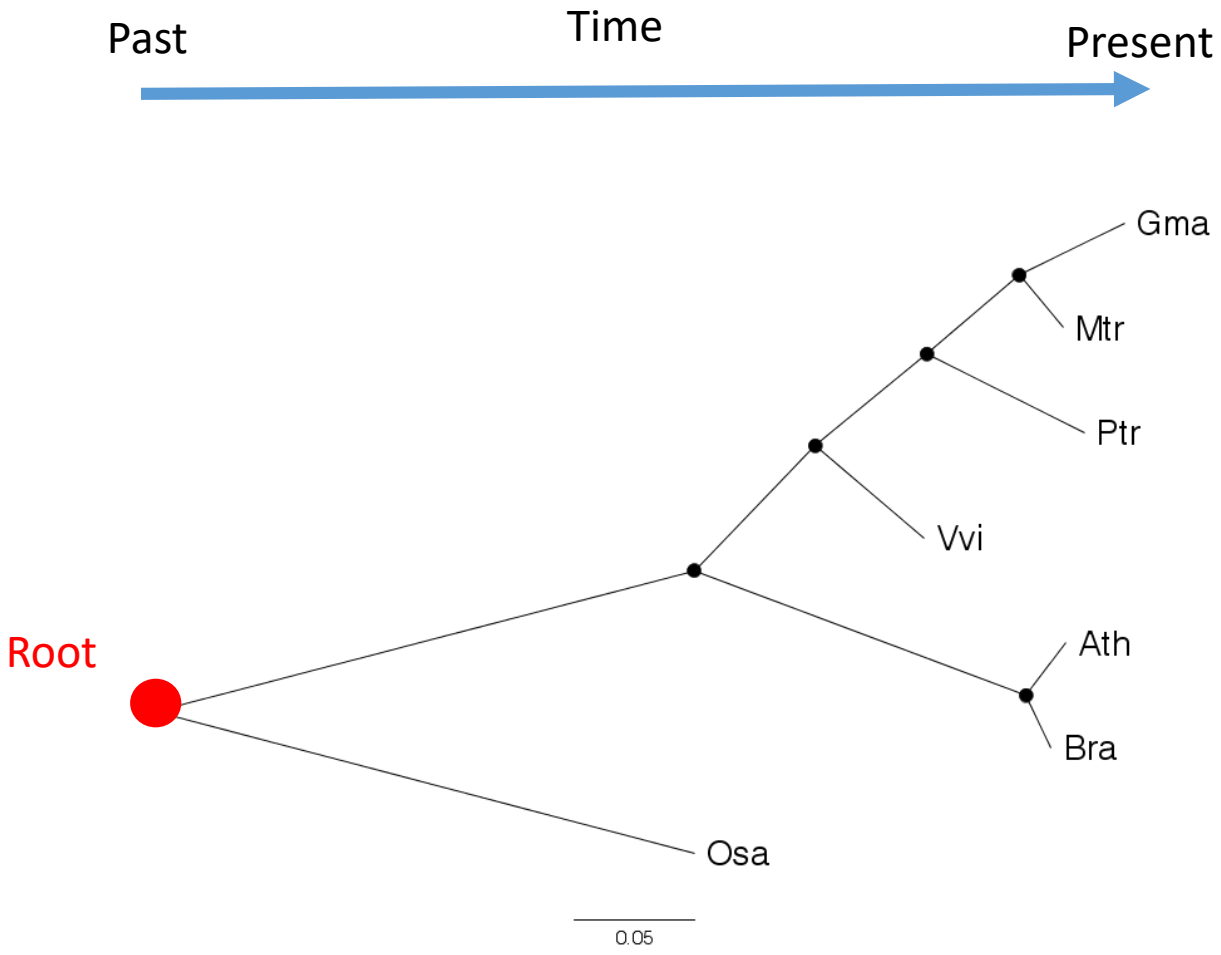
Elements of a phylogeny tree



The root is a very important internal node representing **the most recent common ancestor** of all sequences in the phylogeny.

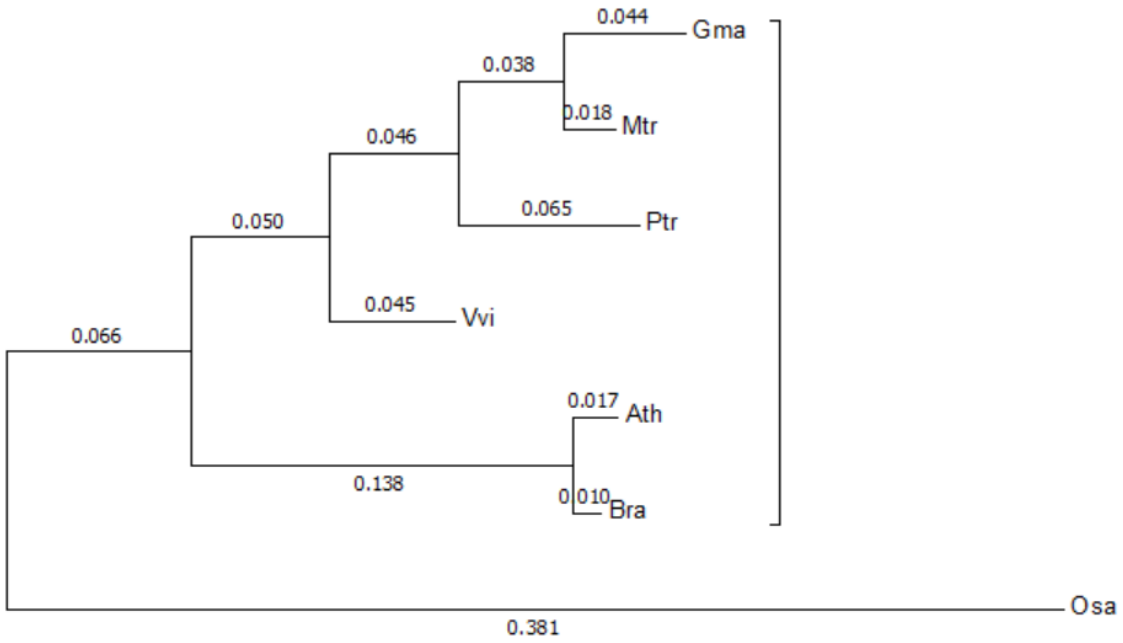
Place a root on the tree



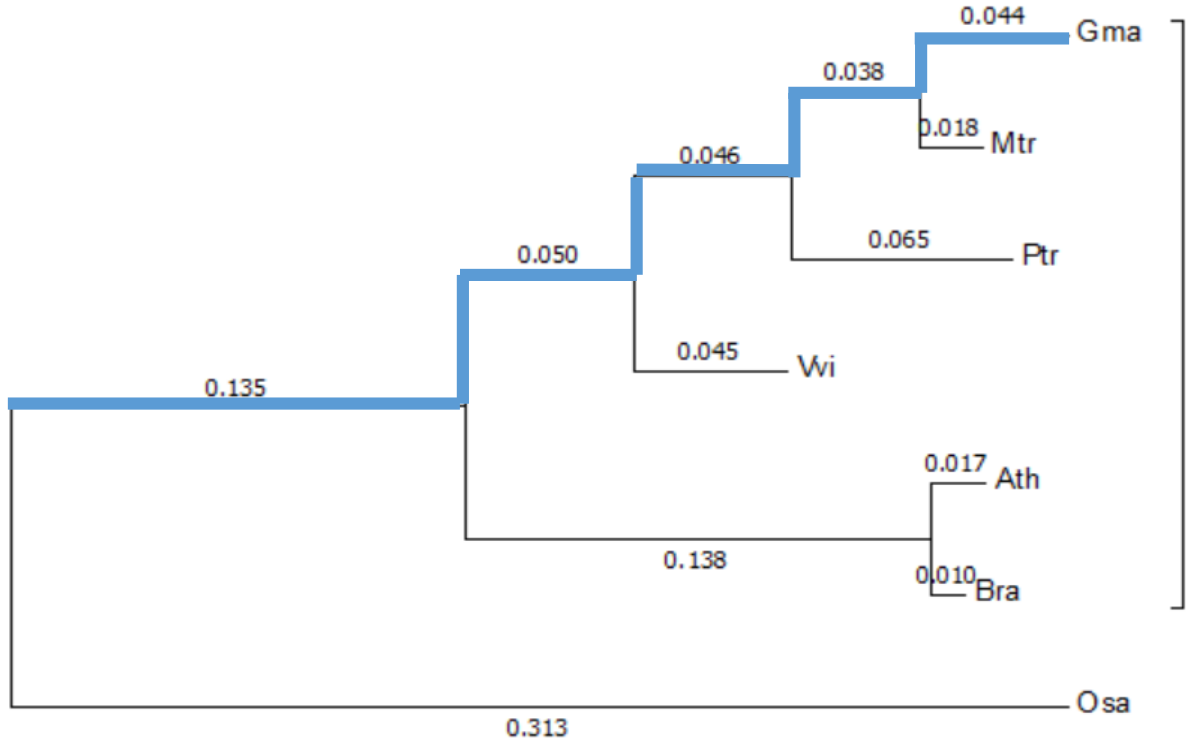


two main approaches to place a root

- Outgroup rooting



Mid-point rooting



the root is positioned at the midpoint between the two longest branches

How to construct a phylogeny

1. distance data

	1	2	3	4	5	6	7
1. Ath							
2. Bra	0.027						
3. Gma	0.239	0.244					
4. Vvi	0.216	0.205	0.164				
5. Osa	0.452	0.444	0.492	0.399			
6. Mtr	0.227	0.227	0.063	0.134	0.475		
7. Ptr	0.239	0.250	0.139	0.154	0.475	0.116	

- **UPGMA**
- **Neighbor-joining**

2.discrete characters

Species/Abbrv △	Gr		*	*			*				*	*	*		*	*		*	*	*		*
1. Ath		E	M	M	N	S	L	S	H	I	F	R	W	E	L	L	V	G	E	R	Y	G
2. Bra		D	M	M	N	S	L	S	H	I	F	R	W	E	L	L	V	G	E	R	Y	G
3. Gma		D	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
4. Mtr		D	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
5. Osa		D	M	M	A	A	L	A	G	L	F	R	W	D	L	L	L	G	E	R	F	G
6. Ptr		E	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
7. Vvi		D	M	M	N	S	L	C	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G

- **Parsimony**
- **Maximum Likelihood**
- **Bayesian Methods**

Distance method

Normal ...GCTATACGCTAGG...

Base pair substitution ...GCTATTCGCTAGG...

↓

G

Substitution refers to the replacement of one amino acid with another amino acid in a protein or the replacement of one nucleotide with another in DNA or RNA

- Distance calculated based on a specific substitution model (Jukes--Cantor
- Model, Kimura, BLOSUM64, etc.)
- Distances from each sequence to all others are calculated and stored in a matrix
- Tree then calculated from the distance matrix using a specific tree-building algorithm

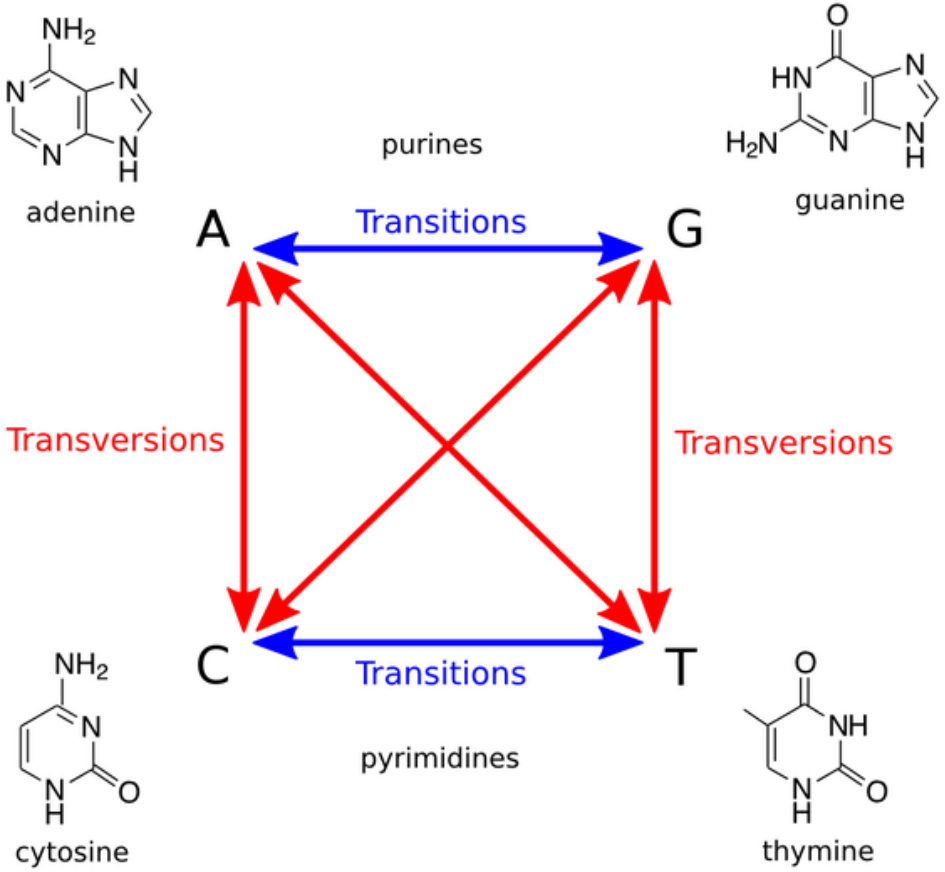
Jukes--Cantor Model

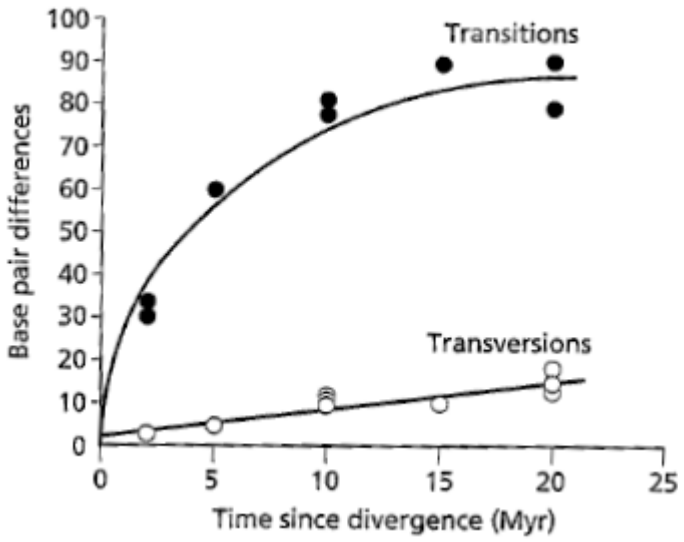
	A	T	C	G
A	1-3α	α	α	α
T	α	1-3α	α	α
C	α	α	1-3α	α
G	α	α	α	1-3α

Jukes-Cantor Model (JC69)

- 1969
- Evolution is described by a single parameter, alpha (α) , the rate of substitution.
- Assumptions:
  - Substitutions among 4 nucleotide types occur with equal probability (rate matrix below)
  - Nucleotides have equal frequency at equilibrium

Kimura's 2--parameter model (K2P)

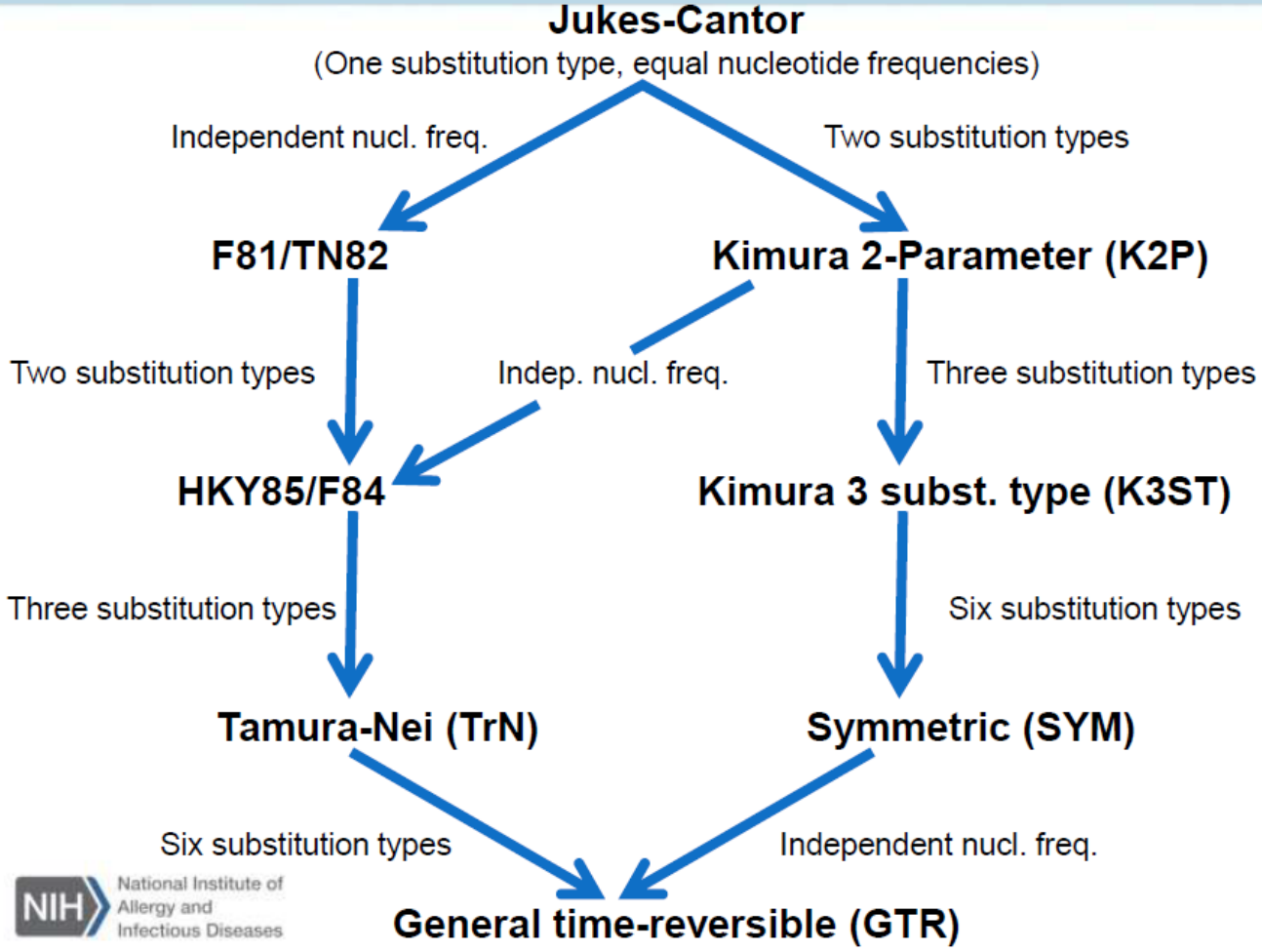




	A	T	C	G
A	1-α-2β	β	β	α
T	β	1-α-2β	α	β
C	β	α	1-α-2β	β
G	α	β	β	1-α-2β

- Models transition and transversion rates separately
- Two parameters
  - **α** for transition rate and **β** for transversion rate.
- Assumption:
  - Nucleotides have equal frequency at equilibrium

Substitution Models



NIH National Institute of Allergy and Infectious Diseases

(revised from NIH lecture)

Protein Substitution Matrices

PAM250: Based on phylogenies where all sequences differ by no more than 15%.

BLOSUM62: Based on clusters of sequences with greater than 62% identical residues

PAM250

	H	E	A	G	A	W	G	H	E	E
P	0	-1	1	0	1	-5	0	0	-1	-1
A	-1	0	2	1	2	-6	1	-1	0	0
W	-3	-7	-6	-7	-6	17	-7	-3	-7	-7
H	6	1	-1	-2	-1	-3	-2	6	1	1
E	1	4	0	0	0	-7	0	1	4	4
A	-1	0	2	1	2	-6	1	-1	0	0
E	1	4	0	0	0	-7	0	1	4	4

BLOSUM62

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	4	0	4	-3	0	-2	-1	-1
W	-2	-3	-3	-2	-3	11	-2	-2	-3	-3
H	8	0	-2	-2	-2	-2	-2	8	0	0
E	0	5	-1	-2	-1	-3	-2	0	5	5
A	-2	-1	4	0	4	-3	0	-2	-1	-1
E	0	5	-1	-2	-1	-3	-2	0	5	5

BLOSUM62 and BLOSUM50 target alignments with 20 – 30% identity

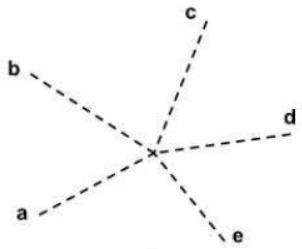
Take home message

1. Transition is more frequent than transversion.

2. Different substitution model lead to different phylogeny

Neighbor-joining algorithm

- Species represented as Points.



- Distance Matrix.

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

Transform distance matrix into a U matrix

$$Q(i,j) = (n-2) * d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k))$$

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

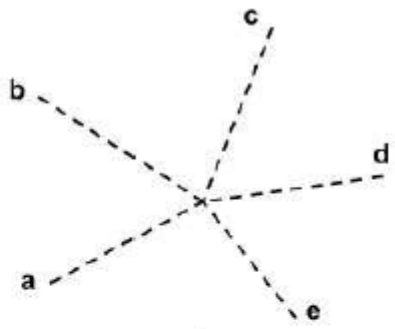
	a	b	c	d	e
a		$\boxed{-50}$	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

$Q_{a,b}$= 4*5 – (5+9+9+8) – (5+10+10+9) = -50

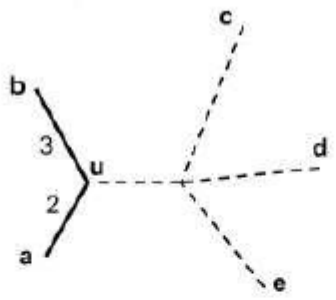
join a and b and repeat the calculation

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

+



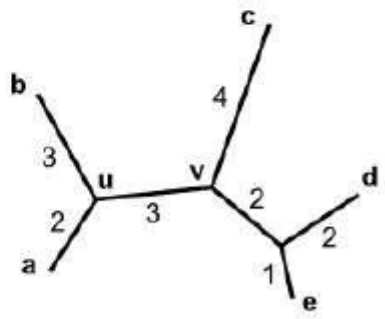
=



Using the formulas mentioned in the previous slide calculate the distances and the new matrix.

	u	c	d	e
u	0	7	7	6
c	7	0	8	7
d	7	8	0	3
e	6	7	3	0

Repeat the steps.



Summary of Neighbor Joining

1. The first step is to build a distance matrix

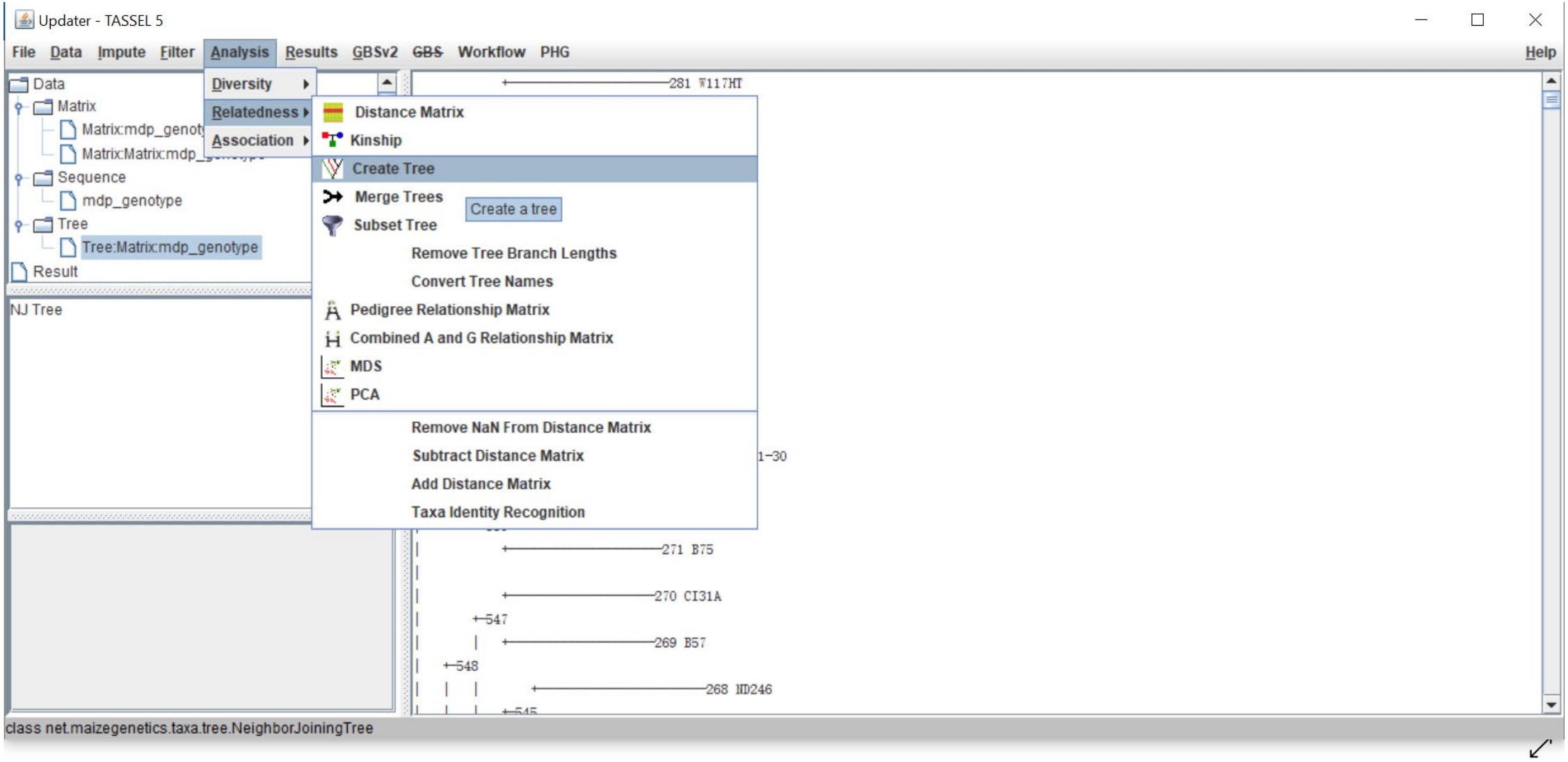
2. Neighbor-joining is a recursive algorithm (step by step). Each step is to cluster the closest branch with the previous step.

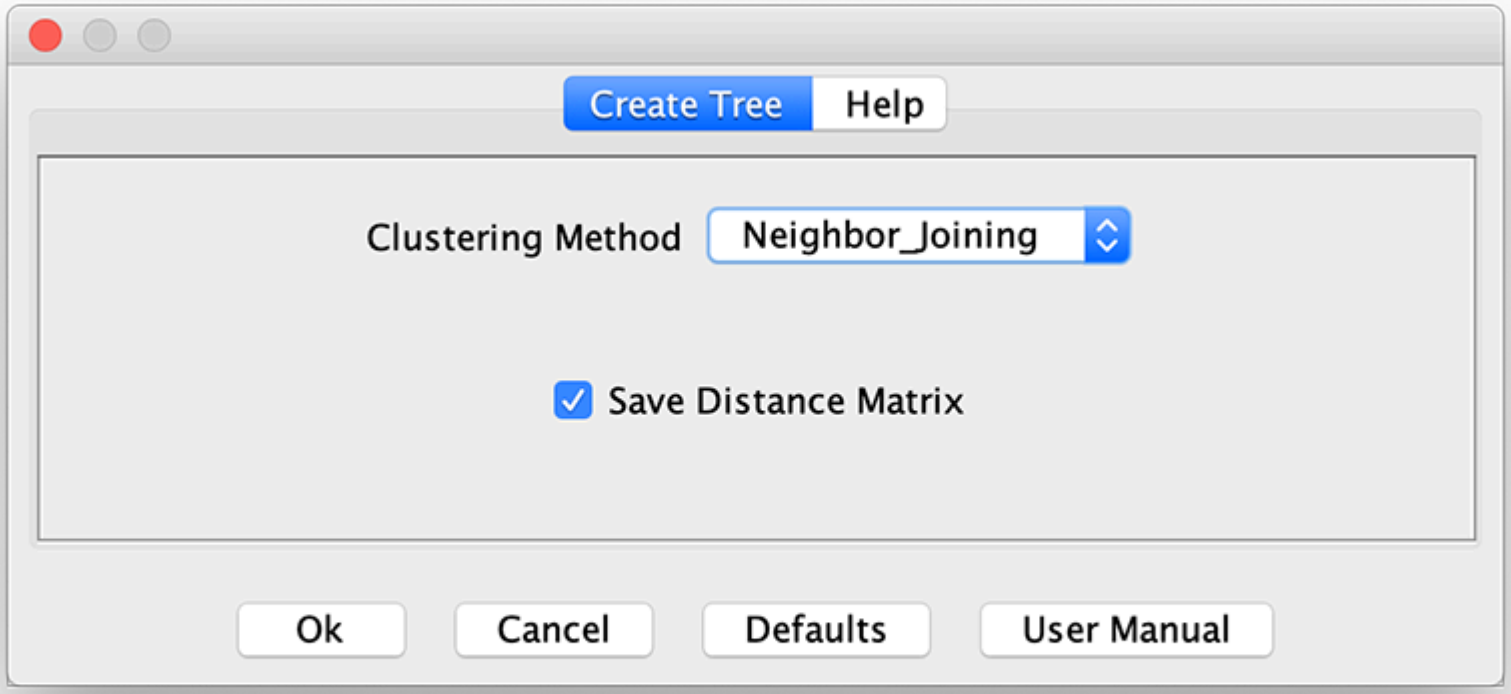
Packages with NJ algorithm

- PAUP* – Phylogenetic Analysis Using Parsimony

  and other methods
- PHYLIP – a suite of phylogenetic programs
- MEGA – An integrated phylogenetic analysis package
- Clustal X, Clustal O , muscle output tree, MAFTT implemented.
- TASSEL -- a Java platform designed for the optimized analysis of crop genomic diversity. TASSEL takes the genotypes as input

TASSEL


Updater - TASSEL 5
File
Data
Impute
Filter
Analysis
Results
GBSv2
GBS
Workflow
PHG
Help
Data
Matrix
Matrix:mdp_genot
Matrix:Matrix:mdp_
Sequence
mdp_genotype
Tree
Tree:Matrix:mdp_genotype
Result
NJ Tree
Diversity
Relatedness
Association
Distance Matrix
Kinship
Create Tree
Merge Trees
Subset Tree
Create a tree
Remove Tree Branch Lengths
Convert Tree Names
Pedigree Relationship Matrix
Combined A and G Relationship Matrix
MDS
PCA
Remove NaN From Distance Matrix
Subtract Distance Matrix
Add Distance Matrix
Taxa Identity Recognition
281 W117HT
271 B75
270 CI31A
547
269 B57
548
268 ND246
class net.maizegenetics.taxa.tree.NeighborJoiningTree

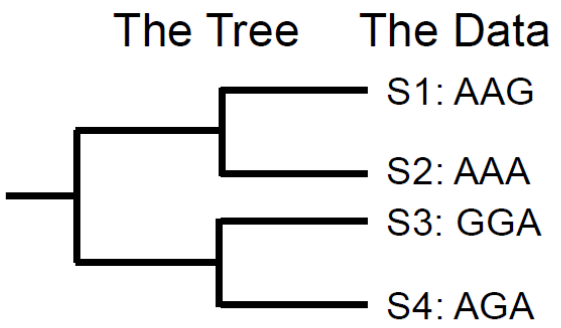

Create Tree
Help
Clustering Method
Neighbor_Joining
Save Distance Matrix
Ok
Cancel
Defaults
User Manual

Maximum likelihood Trees

Most widely used method when accurate trees are required

Complicated models must be maximized through a guided trial-and-error, "hill climbing" algorithm.

1. Set initial parameter values and tree.
2. Calculate likelihood.
3. Propose new parameter value or tree.
4. Calculate likelihood.
5. Decide whether to accept the new value.
6. Repeat steps 3-5 until changes no longer improve likelihood.



$L(\text{Tree}) = \text{Prob}(\text{Data}|\text{Tree}) = \prod_i \text{Prob}(\text{Data}^{(i)}|\text{Tree})$

Log likelihood of different topology

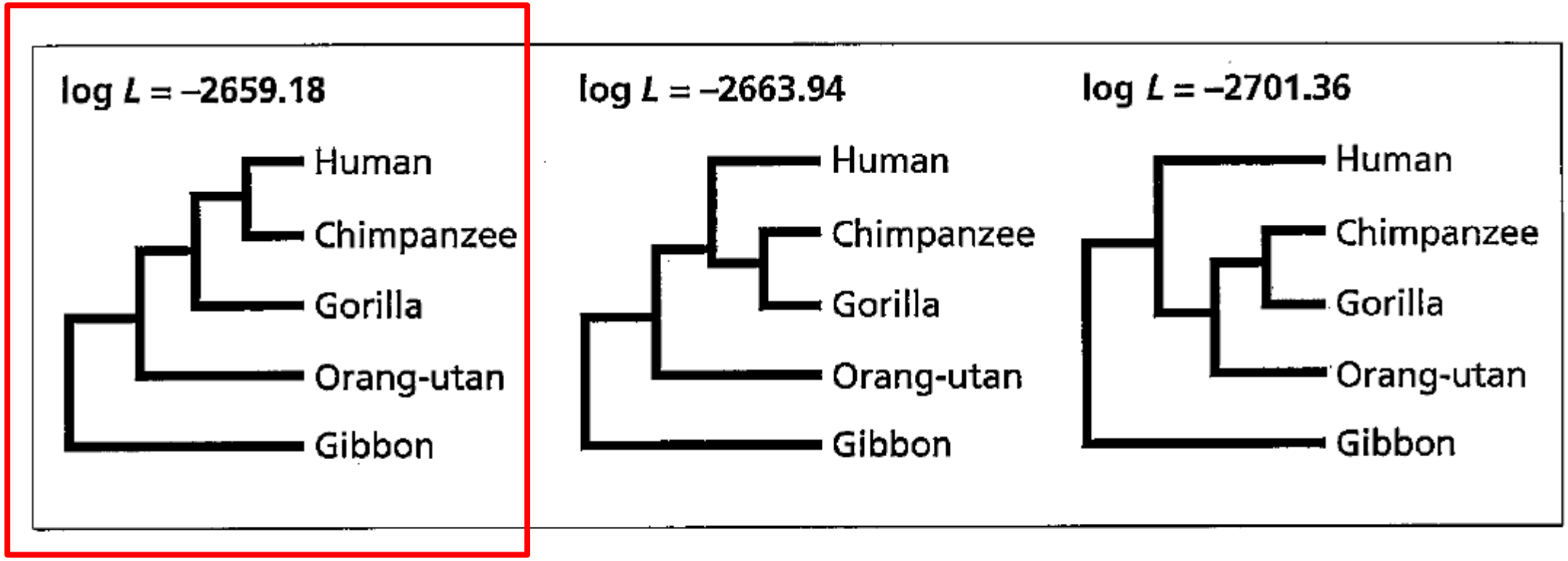


Fig. 6.19 Three different hypotheses of relationship among the hominoids and the likelihoods that each tree has given rise to the observed data.

Packages with ML algorithm

- PAML – tree search only for small datasets
- PHYML – a suite of phylogenetic programs
- MEGA – An integrated phylogenetic analysis package

- RAxML-ng – a fast, scalable and user- friendly tool for maximum likelihood phylogenetic inference, specifically designed for large phylogenomic analyses
- IQ-Tree -- a fast and effective stochastic algorithm for estimating maximum- likelihood phylogenies
- FastTree --Specifically designed for efficiently estimating large phylogenies in terms of number of taxa (up to one million); restricted to a small number of substitution models
- SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data, Reduce SNP redundancy by linkage disequilibrium

Kapli P, Yang Z, Telford M J. Phylogenetic tree building in the genomic age[J]. Nature Reviews Genetics, 2020: 1-17

Bayesian phylogenetic

Calculating the posterior probability of the evolutionary parameters

$$\Pr(\tau, v, \theta | \text{Data}) = \frac{\Pr(D|\tau, v, \theta) \times \Pr(\tau, v, \theta)}{\Pr(D)}$$

where:

τ = tree topology

v = branch lengths

θ = substitution parameters

Packages with Bayesian Inference algorithm

- MrBayes --Bayesian inference of phylogenetic trees. cited 30,000 times
- PhyloBayes-- a Bayesian software package for phylogenetic reconstruction and molecular dating
- P4--Python package for phylogenetic analyses

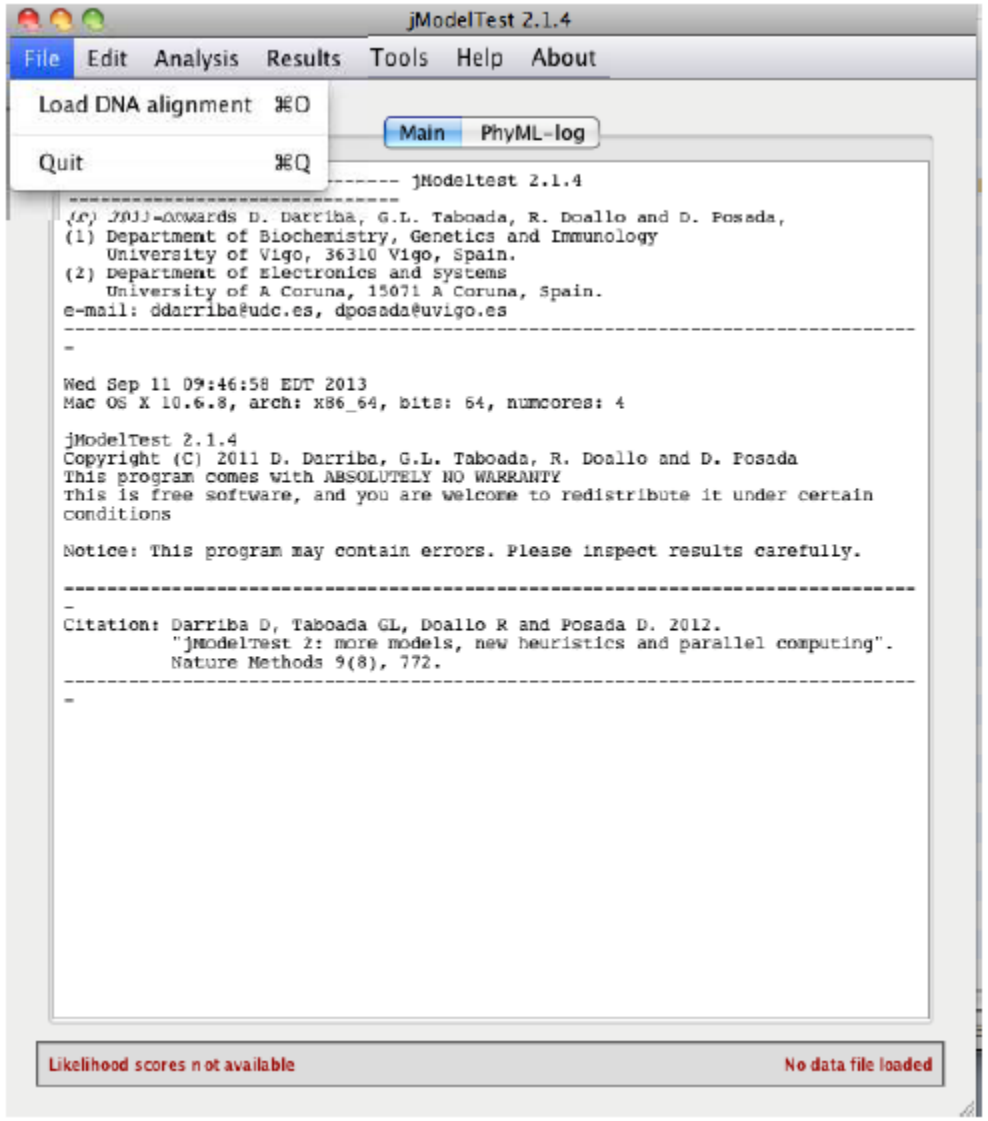
Model selection

- JModelTest
- ProtTest for protein MSAs

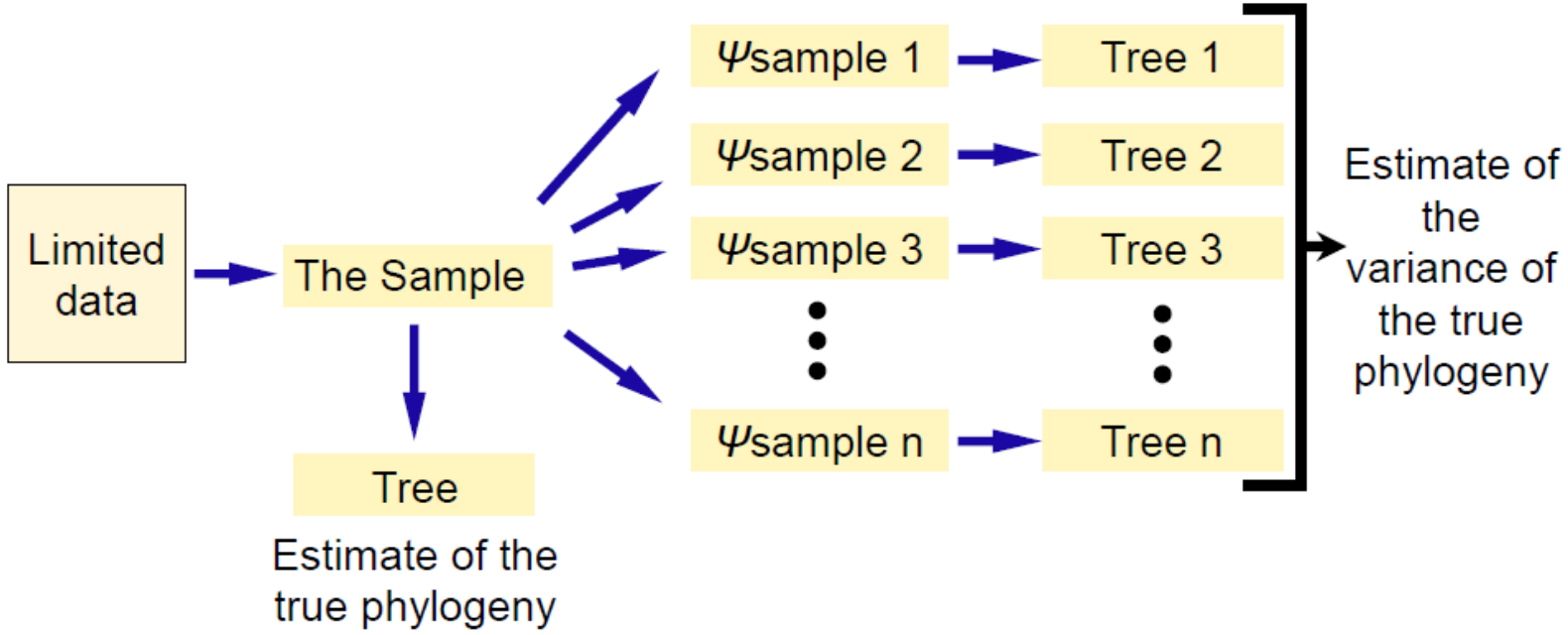
Abadi S, Azouri D, Pupko T, et al. Model selection may not be a mandatory step for phylogeny reconstruction[J]. Nature communications, 2019, 10(1): 1-11.

skipping model selection and using instead the most parameter-rich model, GTR+I+G, leads to similar inferences

general time reversible (GTR, nst=6): variable base frequencies, symmetrical substitution matrix
gamma distribution (G): gamma distributed rate variation among sites
proportion of invariable sites (I): extent of static, unchanging sites in a dataset



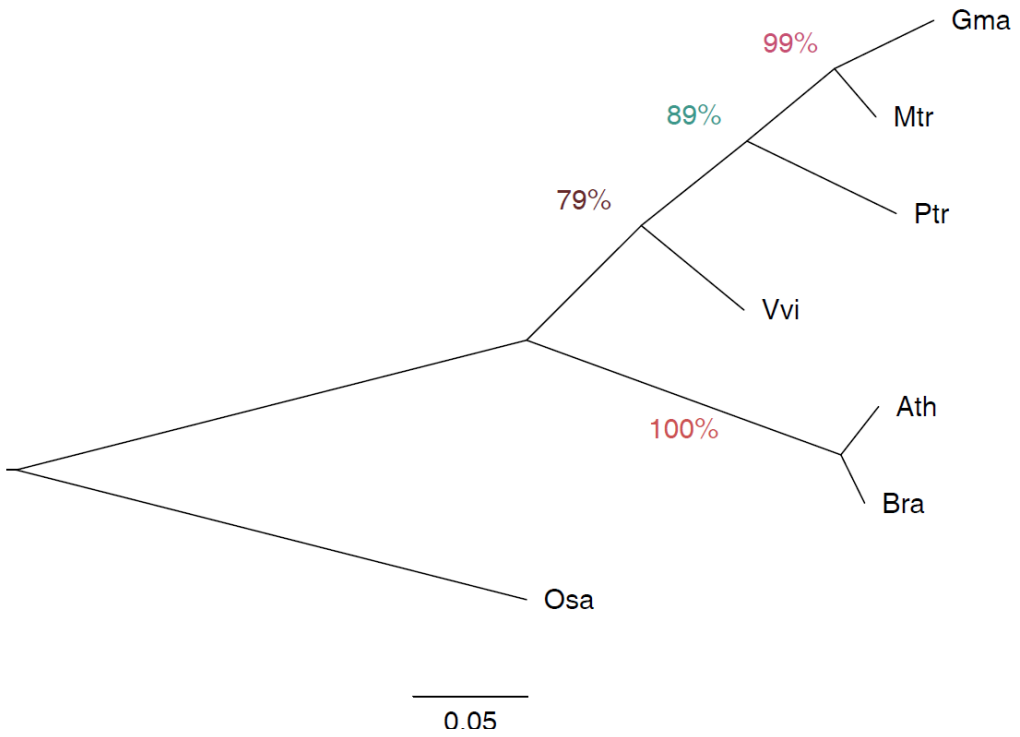
Bootstrapping



Build pseudoreplicates of unlimited data by sampling with replacement from limited data

Bootstrapping

Species/Abbrv △	Gr		*	*			*				*	*	*		*	*		*	*	*		*
1. Ath		E	M	M	N	S	L	S	H	I	F	R	W	E	L	L	V	G	E	R	Y	G
2. Bra		D	M	M	N	S	L	S	H	I	F	R	W	E	L	L	V	G	E	R	Y	G
3. Gma		D	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
4. Mtr		D	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
5. Osa		D	M	M	A	A	L	A	G	L	F	R	W	D	L	L	L	G	E	R	F	G
6. Ptr		E	M	M	N	S	L	S	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G
7. Vvi		D	M	M	N	S	L	C	Q	I	F	R	W	D	L	L	V	G	E	R	Y	G



Hands on practice

- Input format

Fasta format

```
>Ath
MDP-EGFTSGL-FRWNPTRALVQAP-PPVPPPLQQ---QPVTPQTAAFGMR------LGGLEGLFGPYGIRFYTAAKIAELGFTASTLVGMKDEELE
>Bra
MDP-EGFTSGL-FRWNPTRAMVQQPSPPVPPPPQQ--QPPATPQTAAFGMR------LGGLEGLFGPYGVRFYTAAKIAELGFTASTLVGMKDEELE
>Gma
MDP-DAFTASL-FKWDPRTVLPPAPAPPPRPSLLEYAMAPPPVTTAFHPARTAAPRELGGLEELFQAYGIRYYTAAKIAELGFTVSTLVDMKDEELD
>Mtr
MDP-DAFTASL-FKWDPRTVLP--TAPPLRPQLLDYAVTPSTAPSPYYPARL--PRELGGLEELFQAYGIRYYTAAKIAELGFTVSTLVDMKDDELD
>Osa
MDPNDAFSAAHPFRWD-LGPPAPAPVPPPPPP-------PPPPPPANVPRE---------LEELVAGYGVRMSTVARISELGFTASTLLAMTERELD
>Ptr
MDP-EAFTASL-FKWDTRAMV------PHPNRLLEMVPPPQQPPAAAFAVR---PRELCGLEELFQAYGIRYYTAAKIAELGFTVNTLLDMKDEELD
>Vvi
MDP-DAFTASL-FKWDPRGA----VAPPNR--LLE---------------------ALGGLEDLFQEYGVRYYTAAKIAELGFTVSTLLDMKDEELD
```

Phylip format

```
7 451
Ath       MDP-EGFTSG L-FRWNPTRA LVQAP-PPVP PPLQQ---QP VTPQTAAFGM
          R------LGG LEGLFGPYGI RFYTAAKIAE LGFTASTLVG MKDEELEEMM
          NSLSHIFRWE LLVGERYGIK AAVRAERRRL QEEEEEESSR RR-----HLL
          LSAAGDSGTH HALDALSQE- ---GLSEEPV QQQDQTDAAG NNGGGGS---
          GYWDAGQGKM KKQQQQRRRK KPMLTSVETD ED----VNEG EDDDGMDNGN
          GGSGLGTERQ REHPFIVTEP GEVARGKKNG LDYLFHLYEQ CREFLLQVQT
          IAKDRGEKCP TKVTNQVFRY AKKSGASYIN KPKMRHYVHC YALHCLDEEA
          SNALRRAFKE RGENVGSWRQ ACYKPLVNIA CRHGWDIDAV FNAHPRLSIW
          YVPTKLRQLC HLERNNAVAA AAALVG-GIS CTGSSTSGRG GCGGDDLRF*
          -
Bra       MDP-EGFTSG L-FRWNPTRA MVQQPSPPVP PPPQQ--QPP ATPQTAAFGM
          R------LGG LEGLFGPYGV RFYTAAKIAE LGFTASTLVG MKDEELEDMM
          NSLSHIFRWE LLVGERYGVK AAVRAERRRL LEEEEEESSR RR-----HLI
          LSAAGDSGTH HALDALSQED DWTGLSEEPV HQLEHTDAAG NNGGGG----
          GYWDAGQAKM KKPQQ-RRRK KQMVTSVETD DD----MNEG DDDDDGNGGG
          GGGVLGIERQ REHPFIVTEP GEVARGKKNG LDYLFHLYEQ CREFLIQVQT
          IAKDRGEKCP TKVTNQVFRY AKKSGASYIN KPKMRHYVHC YALHCLDEEA
          SNALRRAFKE RGENVGSWRQ ACYKPLVNIA CRHGWDIDAV FNAHPRLSIW
          YVPTKLRQLC HLERSNAVAA ASALVGNGIS CTGSS----- --ASGGLGFN
          *
Gma       MDP-DAFTAS L-FKWDPRTV LPPAPAPPPR PSLLEYAMAP PPVTTAFHPA
          RTAAPRELGG LEELFQAYGI RYYTAAKIAE LGFTVSTLVD MKDEELDDMM
          NSLSQIFRWD LLVGERYGIK AAVRAERRRV ----EDDDIK RRNNNSNNLL
          ---STDT-TT NALDALSQE- ---------- ---------- ----------
          ---------- ---------- ---------- ---------- ----------
          ---------- --------DP GEVARGKKNG LDYLFHLYEQ CREFLMQVQA
          IAKDRGEKCP TKVTNQVFRY AKKAGASYIN KPKMRHYVHC YALHCLDEEV
          SNELRRAFKE RGENVGAWRQ ACYKPLVAIA ARQGWDIDAI FNAHPRLSIW
          YVPTKLRQLC HAERNSVSAS SSVSA----- ---------- --GSAHLPF*
          -
```

1st line: Number of sequences(space)Number of sites
2nd line: Sequence ID (10 characters max) Sequence

Nexus format

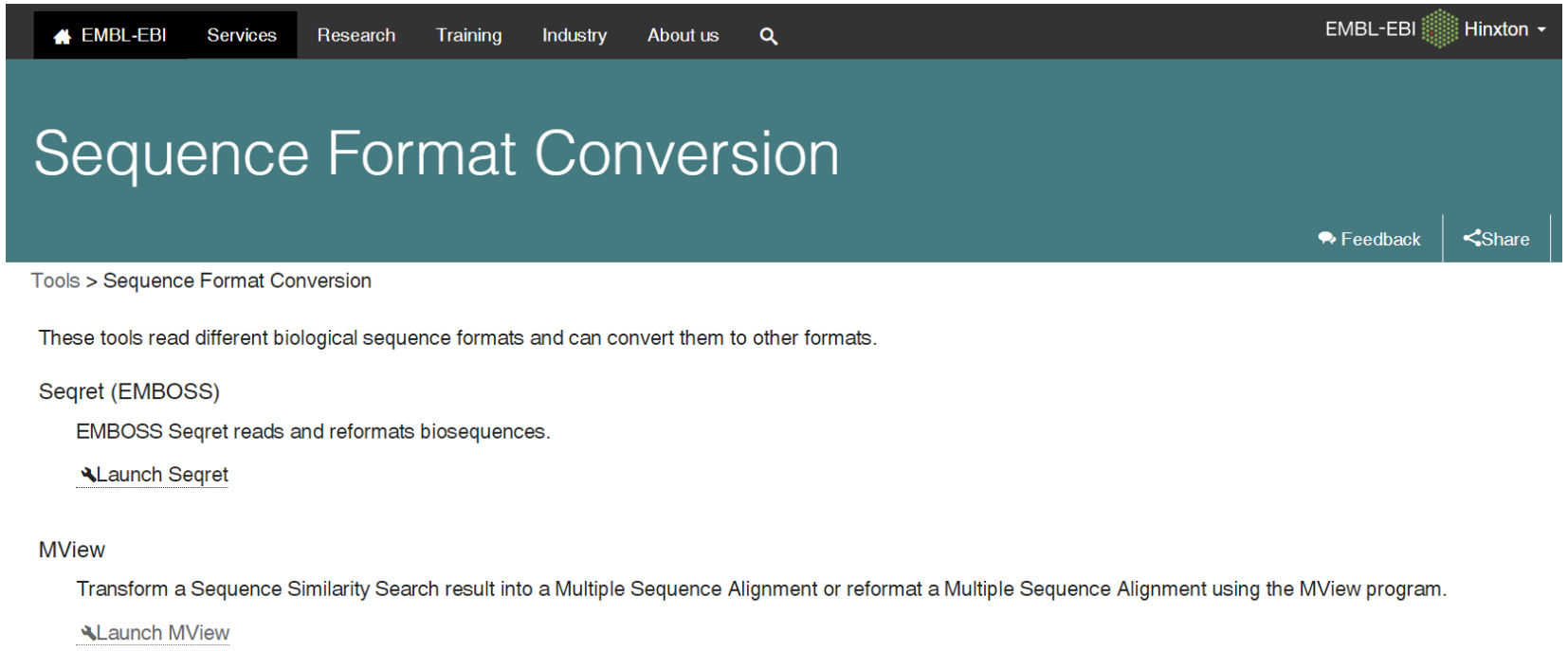
```
#NEXUS
[ Title Phylogenetic Analysis]
begin taxa;
	dimensions ntax=7;
	taxlabels
		Ath
		Bra
		Gma
		Mtr
		Osa
		Ptr
		Vvi
;
end;
begin characters;
	dimensions nchar=451;
	format missing=? gap=- matchchar=. datatype=protein;
	matrix

Ath
MDP-EGFTSGL-FRWNPTRALVQAP-PPVPPPLQQ---QPVTPQTAAFGMR------LGG
LEGLFGPYGIRFYTAAKIAELGFTASTLVGMKDEELEEMMNSLSHIFRWELLVGERYGIK
AAVRAERRRLQEEEEEESSRRR-----HLLLSAAGDSGTHHALDALSQE----GLSEEPV
QQQDQTDAAGNNGGGGS---GYWDAGQGKMKKQQQQRRRKKPMLTSVETDED----VNEG
EDDDGMDNGNGGSGLGTERQREHPFIVTEPGEVARGKKNGLDYLFHLYEQCREFLLQVQT
IAKDRGEKCPTKVTNQVFRYAKKSGASYINKPKMRHYVHCYALHCLDEEASNALRRAFKE
RGENVGSWRQACYKPLVNIACRHGWDIDAVFNAHPRLSIWYVPTKLRQLCHLERNNAVAA
AAALVG-GISCTGSSTSGRGGCGGDDLRF*-

Bra
MDP-EGFTSGL-FRWNPTRAMVQQPSPPVPPPPQQ--QPPATPQTAAFGMR------LGG
LEGLFGPYGVRFYTAAKIAELGFTASTLVGMKDEELEDMMNSLSHIFRWELLVGERYGVK
AAVRAERRRLLEEEEEESSRRR-----HLILSAAGDSGTHHALDALSQEDDWTGLSEEPV
```

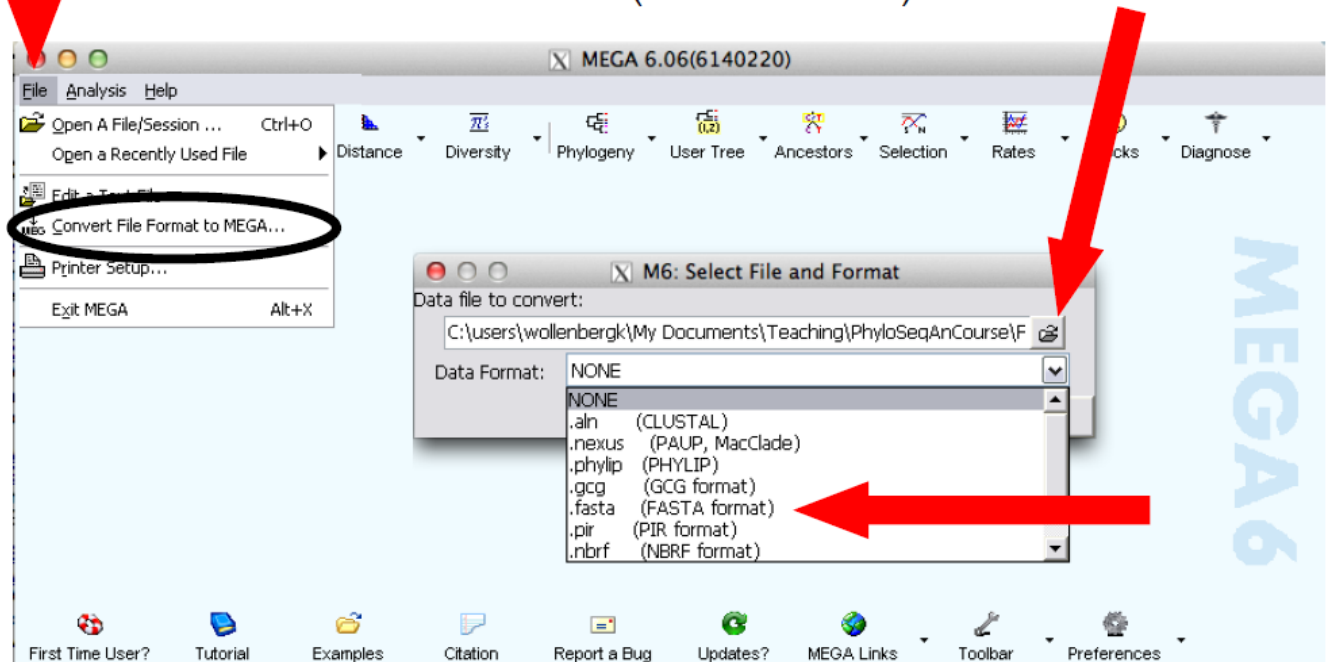
Conversion tools

- https://www.ebi.ac.uk/Tools/sfc/

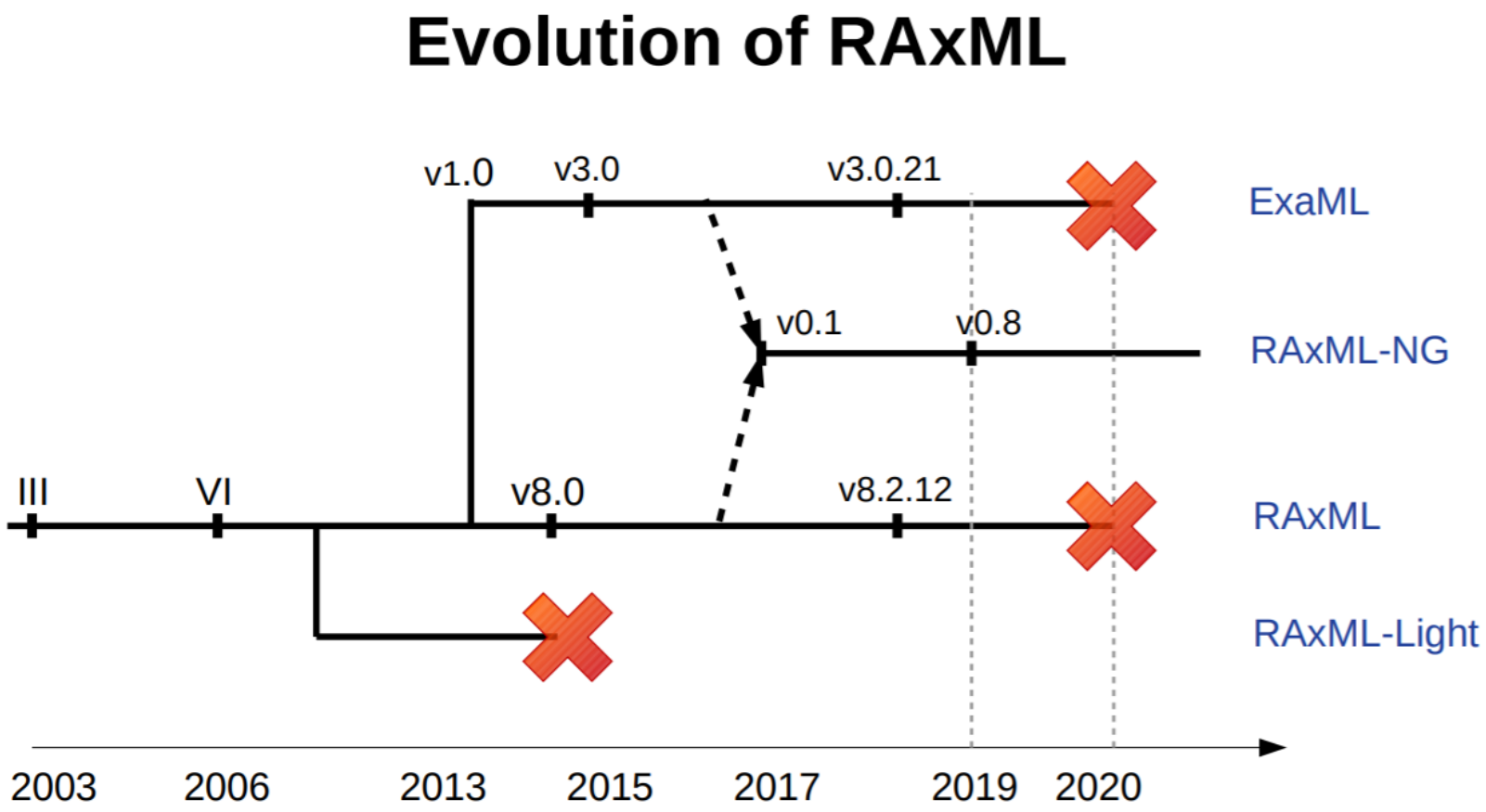


Format conversion in MEGA

Importing a FASTA-formatted alignment
File > Convert to MEGA format
Navigate to FASTA file
Data Format > .fasta (FASTA format)



RAxML replaced RAxML



Adapted from the Exelixis Lab

command line for raxml

- /programs/raxml-ng_v0.9.0/raxml-ng --all --msa leafy_align.fas --model JTT+G4+F --prefix leafy_all

- /programs/raxml-ng_v0.9.0/raxml-ng --all --msa prim.phy --model GTR+G --prefix A1

- The best tree is saved in [prefix].**raxml.bestTree**
- The tree with bootstrap value is saved in [prefix]**.raxml.support**

Illustration of phylogeny

- iTOL – web-based tool
- Mega
- TreeView, FigTree
- R packages ape, ggtree,

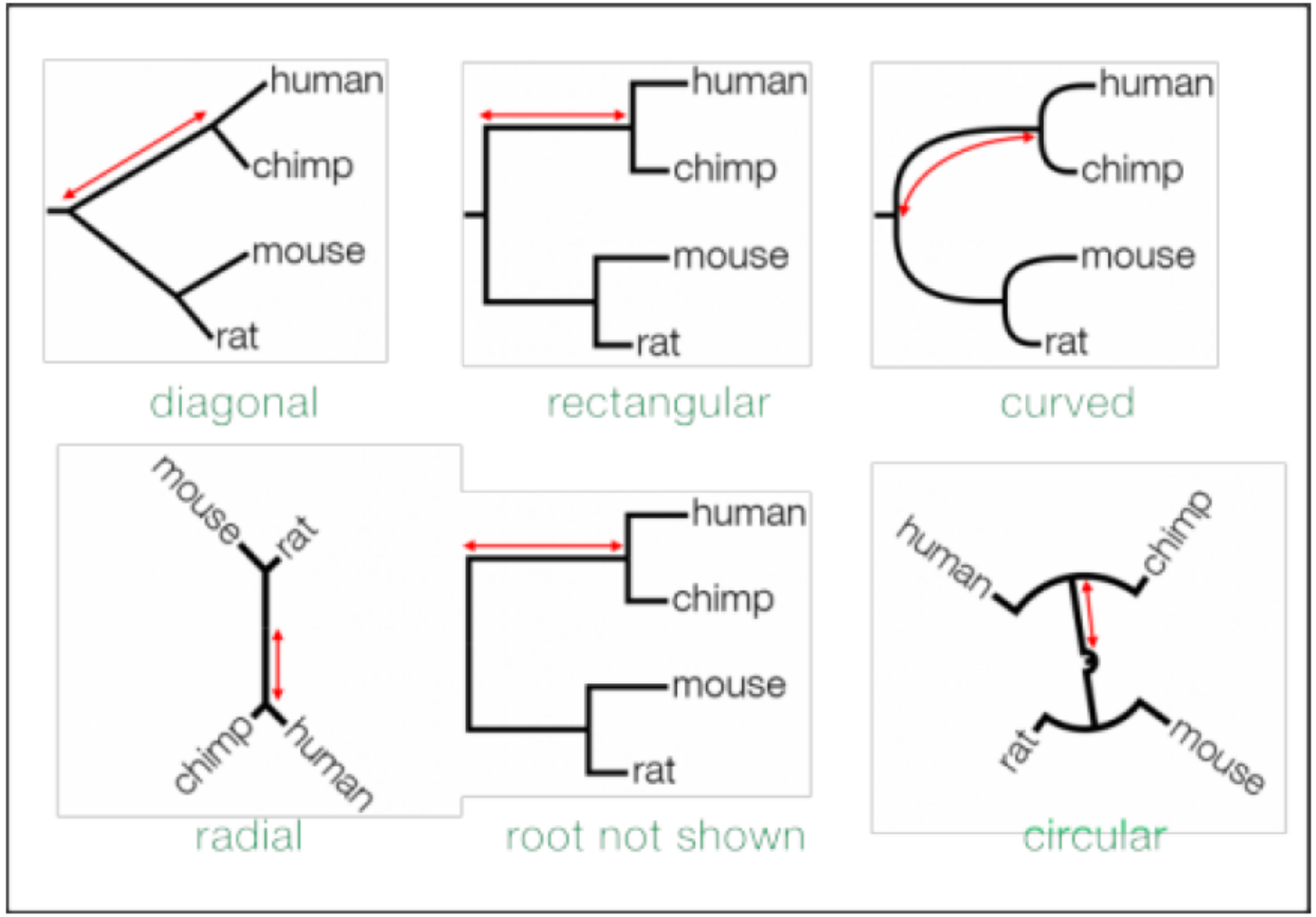


Figure 13 Alternative representations of the same topology. Red lines indicate the same branch in each representation. Trees can be rotated on the page and still depict the same tree. NB: The trees are not drawn to the same scale.