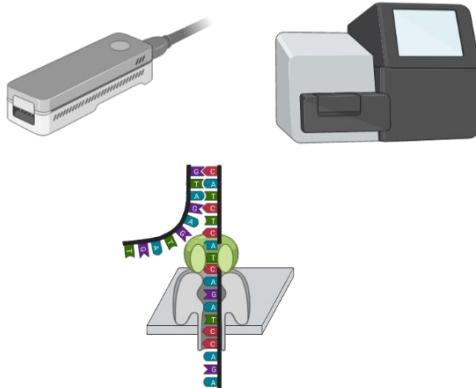
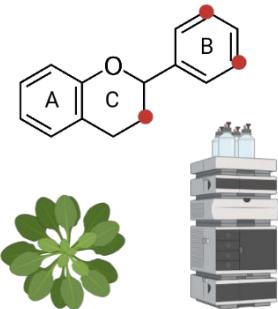
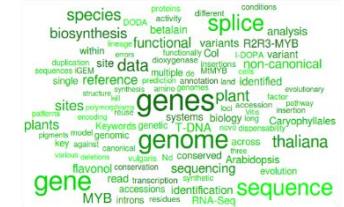
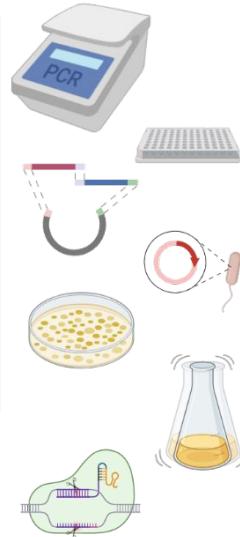
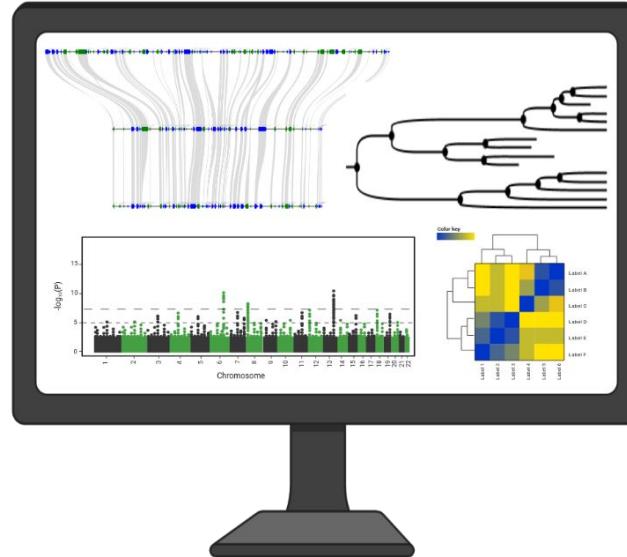




Technische  
Universität  
Braunschweig



# Plant Biotechnology and Bioinformatics



# Phylogenetic Analyses

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

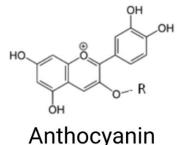
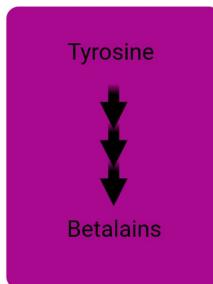
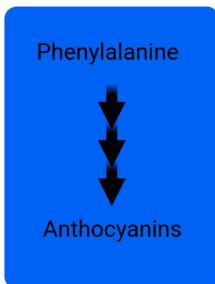
# Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: LMChemBSc12
  - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: b.pucker[a]tu-bs.de

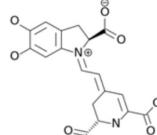
My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

# Theodosius Dobzhansky (1900-1975)

“Nothing in biology makes sense except in the light of evolution”



*Antirrhinum majus*



*Mirabilis jalapa*

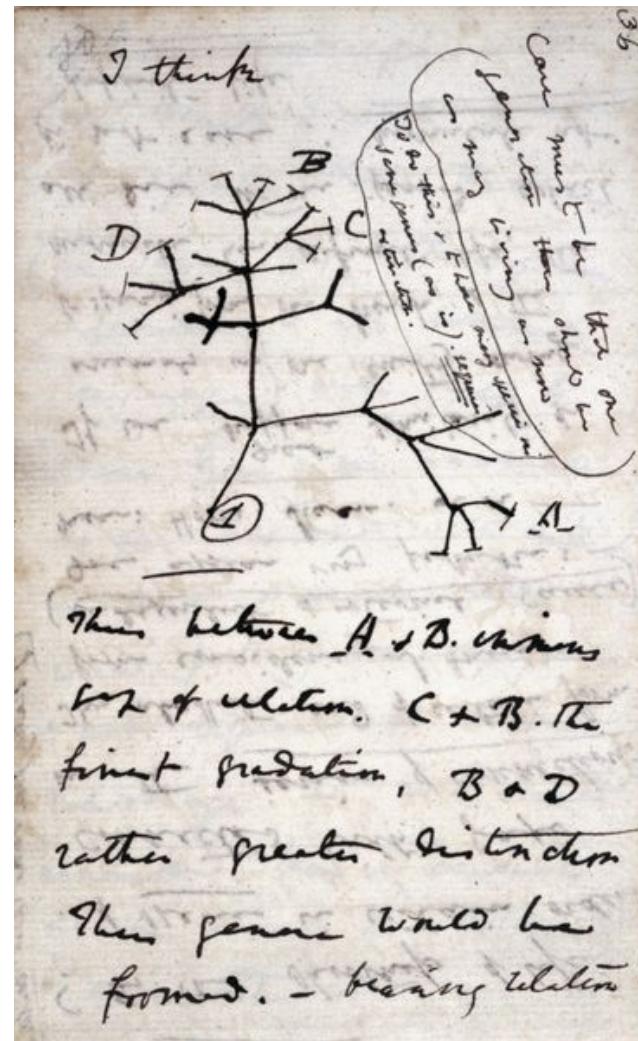
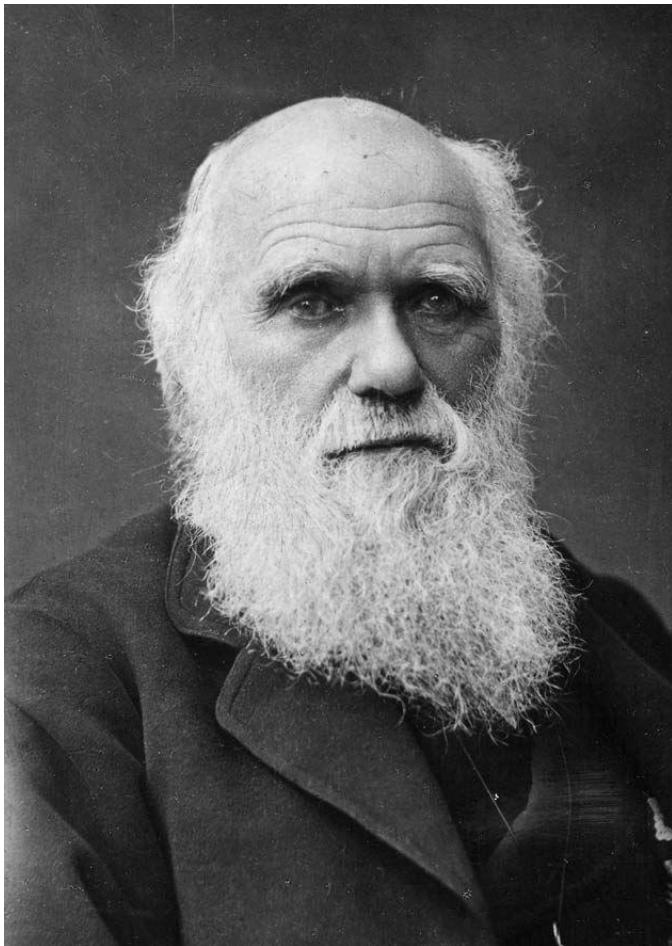


Anthocyanin color range



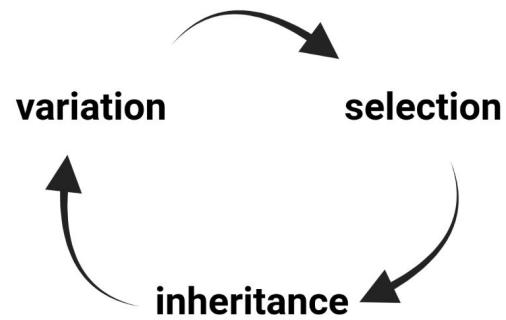
Betalain color range

# Charles Darwin (1809-1882)

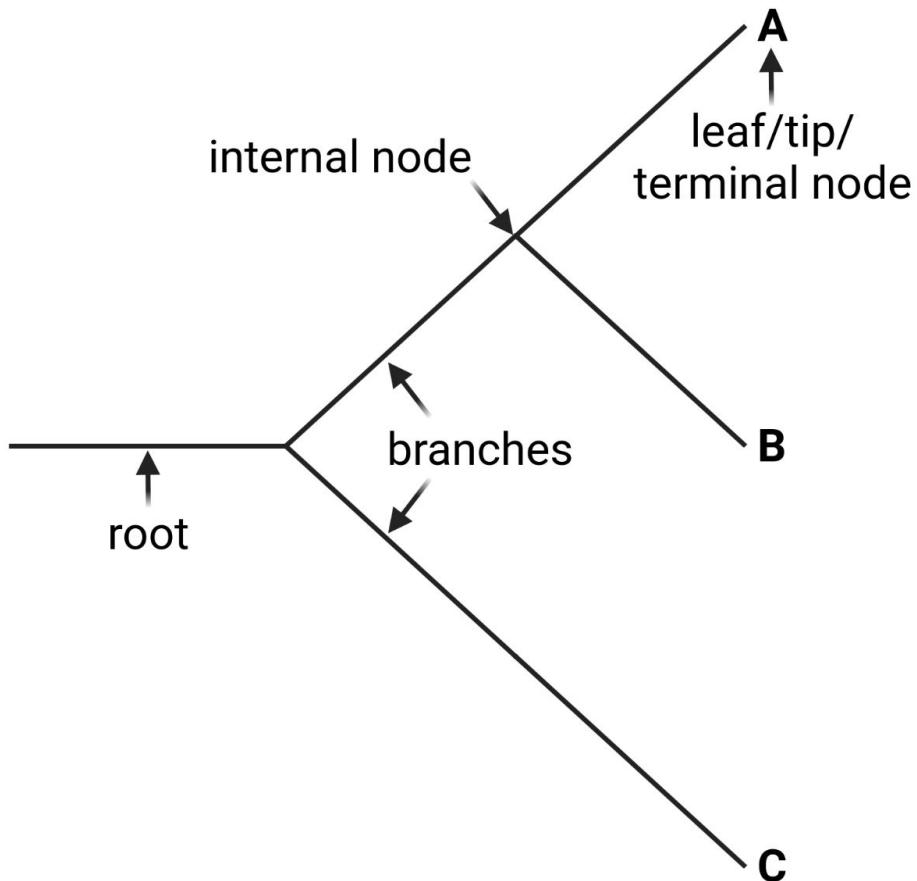


# Basic evolution concept

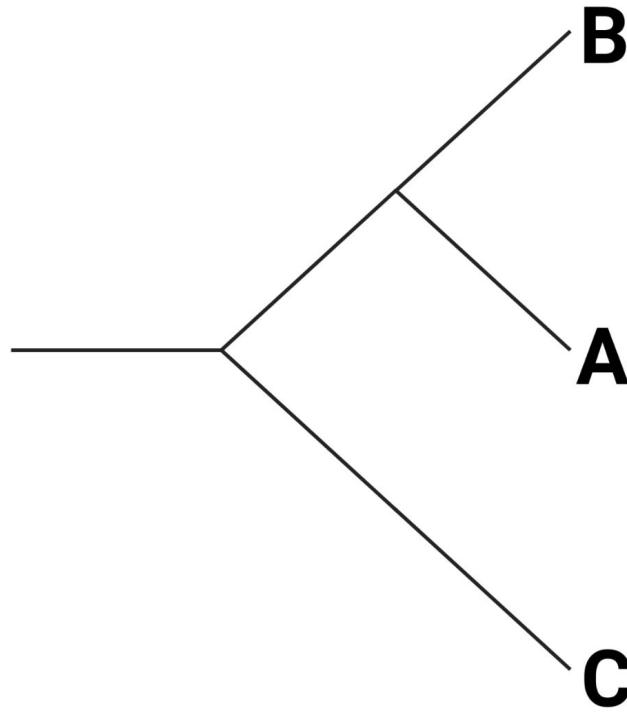
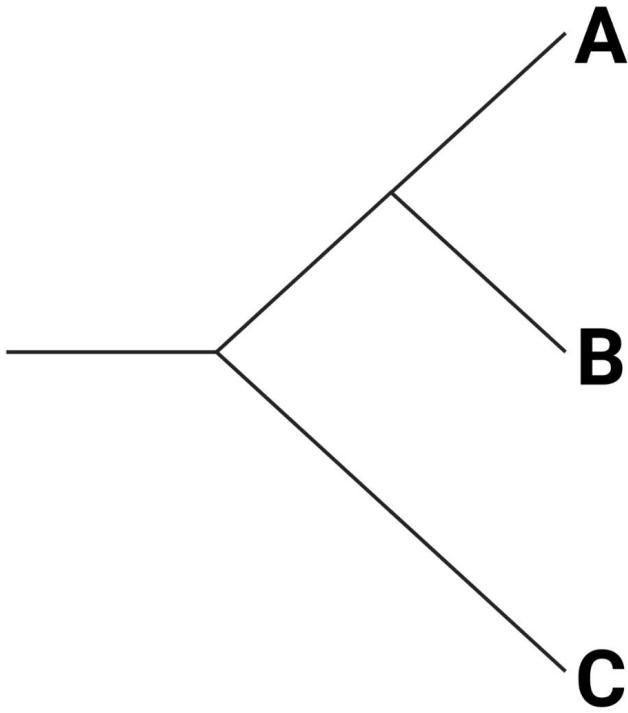
- Selection is performed by nature ('theory of natural selection')
- Selection requires (1) variation to act upon and (2) more offspring than sustainable
- Inheritance (genotype) determines the properties of individuals (phenotype)



# Elements of a phylogenetic tree

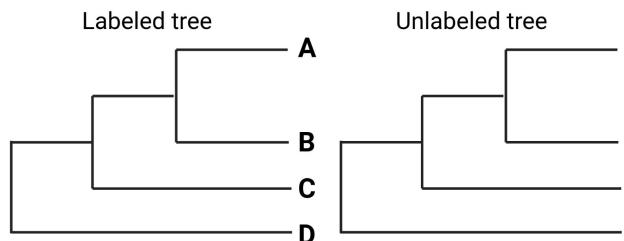
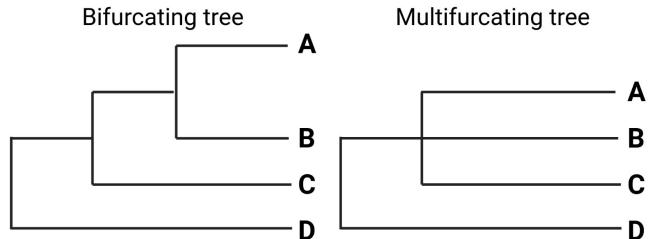
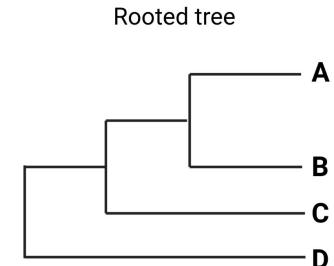
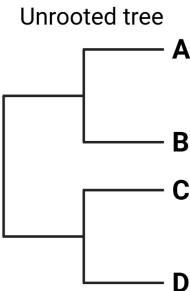


# Tree topology and rotation



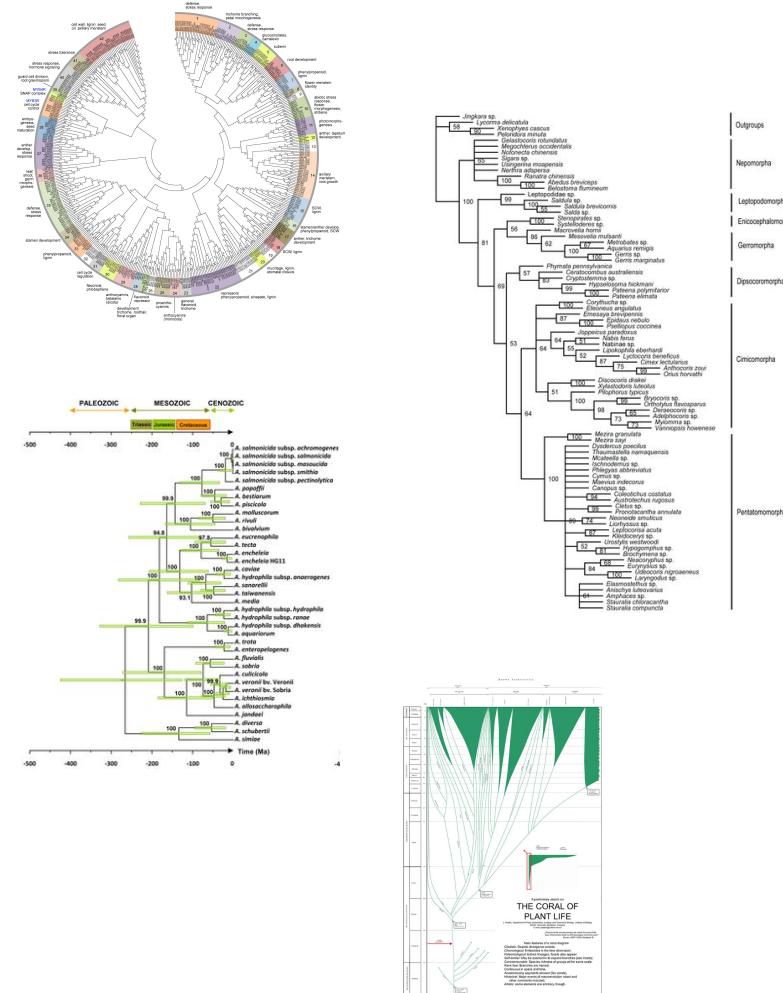
# Types of trees

- Rooted vs. unrooted tree
- Bifurcating vs. multifurcating tree
- Labeled vs. unlabeled trees



# Types of special trees

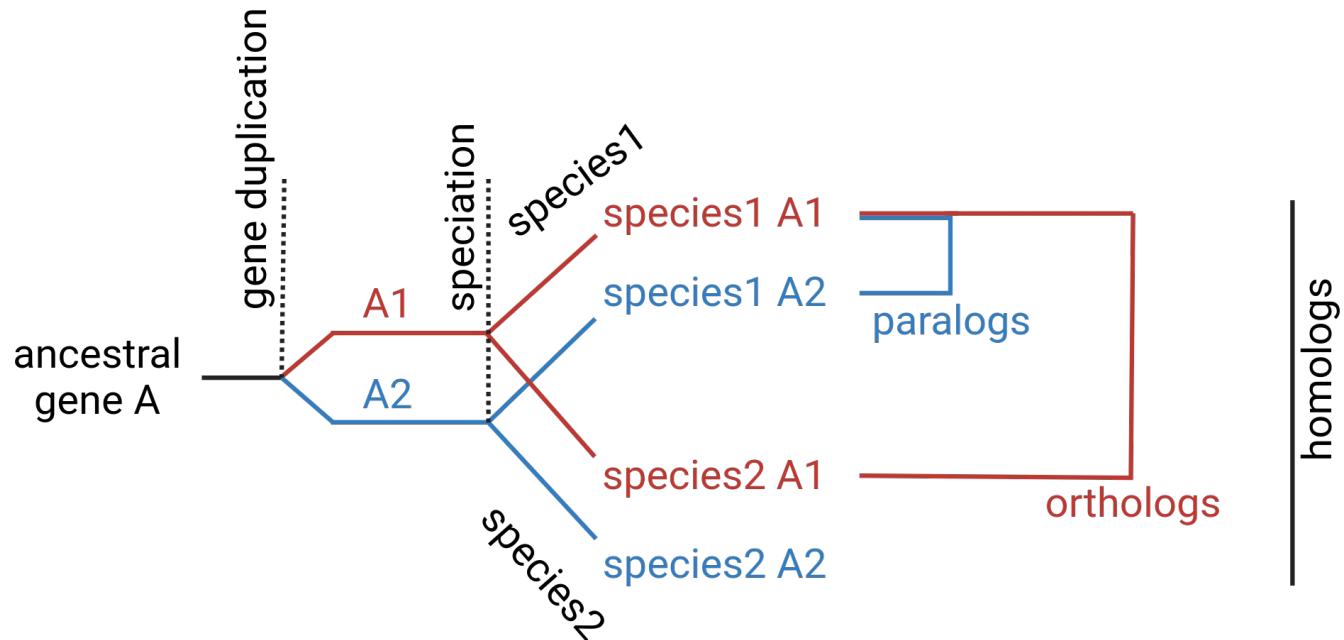
- Cladogram: only represent the branching pattern without branch lengths
- Phylogram: branch lengths indicate number of changes
- Chronogram: branch lengths indicate time
- Spindle diagram: shows taxonomic diversity over time



Pucker et al., 2020: 10.1371/journal.pone.0239275  
Li et al., 2012: 10.1371/journal.pone.0032152  
Loren et al., 2014: 10.1371/journal.pone.0088805  
Podani et al., 2020: 10.5586/asbp.8937

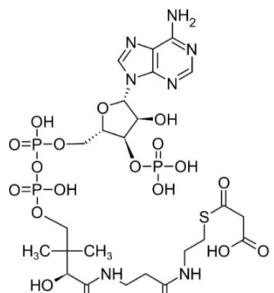
# Homologs, orthologs, and paralogs

- Homolog = genes with common ancestors
- Ortholog = same gene in different species
- Paralog = gene copies in one species

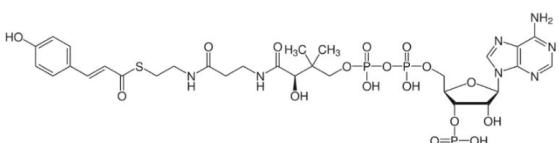


# Example: orthologs

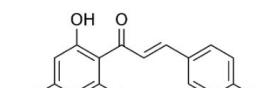
- CHS in different species has the same function
- Reaction is initial step of the flavonoid biosynthesis
- Flavonoid biosynthesis is conserved across plant species



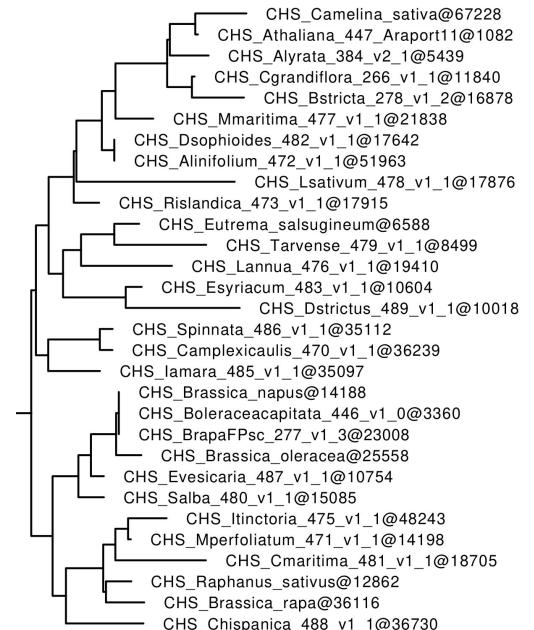
3x malonyl-CoA



4-coumaroyl-CoA

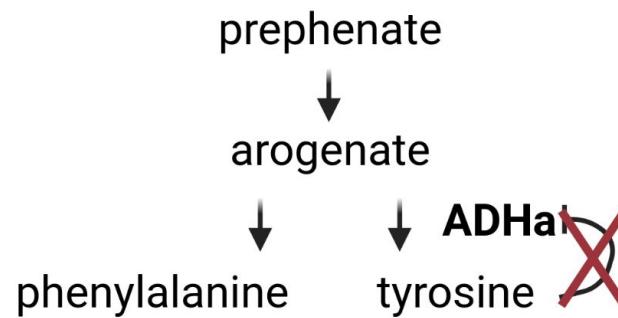
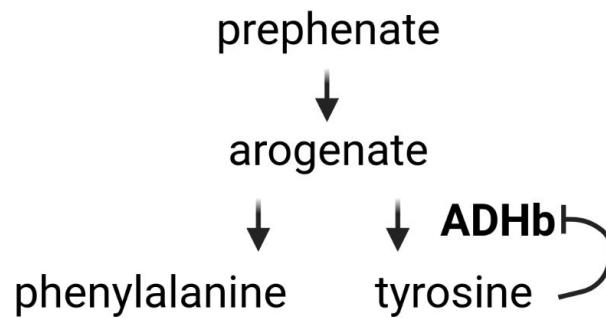


naringenin chalcone  
(4x CoA + 3x CO<sub>2</sub>)



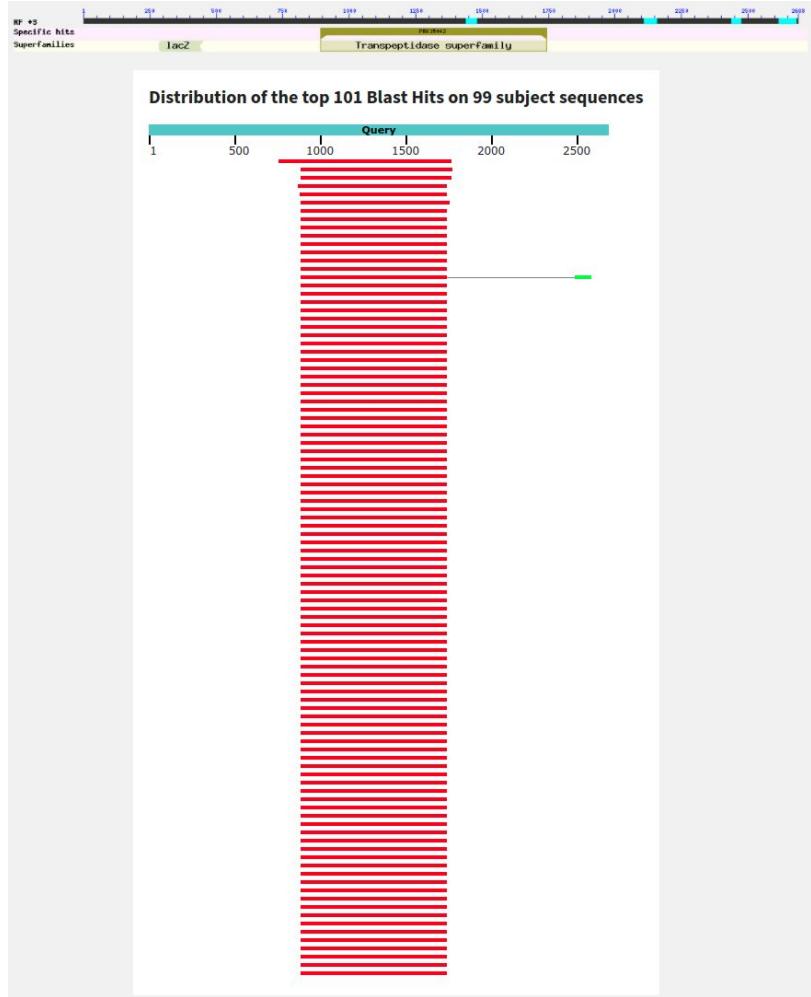
# Example: paralogs

- ADH was duplicated in Caryophyllales
  - ADHa = feedback-resistance enzyme
  - ADHb = feedback-inhibited enzyme



# BLAST - simple sequence searches

- BLAST = Basic Local Alignment Search Tool
- Identification of short stretches of sequence similarity
- Quick search against a large database
- Sequence similarity suggest homology i.e. common ancestors and similar function



# NCBI website vs. local BLAST

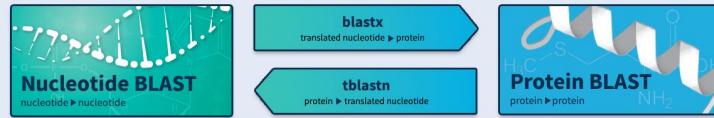
**Basic Local Alignment Search Tool**  
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.  
[Learn more](#)

**NEWS**

BLAST+ 2.13.0 is here!  
Starting with this release, we are including the blastn\_vdb and tblastn\_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

**Web BLAST**



**BLAST Genomes**  
Enter organism common name, scientific name, or tax id  **Search**  
Human Mouse Rat Microbes

```
USAGE
blastn [-h] [-help] [-import_search_strategy filename] [-task task_name] [-db database_name]
[-export_search_strategy filename] [-taxid taxid] [-taxidlist filename]
[-negative_glist filename] [-negative_taxidlist filename]
[-taxids taxids] [-negative_taxids taxids] [-taxidlist filename]
[-negative_taxidlist filename] [-entrez_query entrez_query]
[-ob soft_mask filtering_algorithm] [-db hard_mask filtering_algorithm]
[-subject subject input_file] [-subject loc range] [-query input_file]
[-out output_file] [-evalvalue evalvalue] [-word_size int_value]
[-gapopen open_penalty] [-gapexten extend_penalty]
[-perc_identity float_value] [-qcov_hsp_perc float_value]
[-max_hsps int_value] [-xdrop_ungap float_value] [-xdrop_gap float_value]
[-xdrop_stop float_value] [-xdrop_end float_value] [-xdrop_start float_value]
[-reward reward] [-no_greedy] [-min_raw_gapped_score int_value]
[-template_type type] [-template_length int_value] [-dust DUST_options]
[-filtering_db filtering_database]
[-window masker taxid window masker taxid]
[-window masker db window masker db] [-soft_masking soft_masking]
[-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
[-best_hit_score_edge float_value] [-subject besthit]
[-window_size int_value] [-off_diagonal_range int_value]
[-use_swapping] [-use_swapping_threshold float_value]
[-use_swapping_min_length int_value] [-use_swapping_max_length int_value]
[-use_swapping_min_percent float_value]
[-use_swapping_max_percent float_value]
[-use_swapping_min_align_length int_value]
[-use_swapping_max_align_length int_value]
[-show_gis] [-num_descriptions int_value] [-num_alignments int_value]
[-line_length line_length] [-html] [-sorthits sort_hits]
[-sortthsp sort_hsps] [-max_target_seqs num_sequences]
[-num_threads int_value] [-remote] [-version]
```

DESCRIPTION  
Nucleotide-Nucleotide BLAST 2.9.0+

NCBI website	Local BLAST
Convenient to use	Requires some command line knowledge
Requires no local computational resources	Computational resources needed
Transfer of query sequence	Query kept secret
Large databases available	Download of databases required

# BLAST alternatives

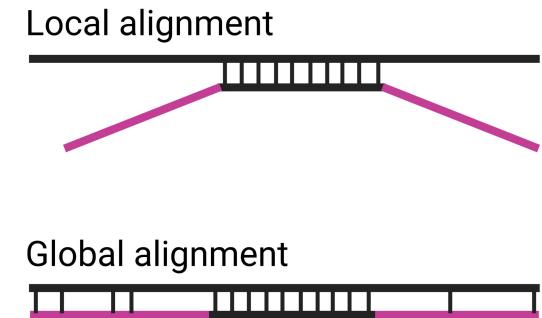
- HMMER: <http://hmmer.org/>
  - Dedicated to the identification of specific domains
- DIAMOND: <https://github.com/bbuchfink/diamond>
  - Much faster, but higher memory requirements

<http://hmmer.org/>  
Mistry, 2013: 10.1093/nar/gkt263  
<https://github.com/bbuchfink/diamond>  
Buchfink et al., 2021: 10.1038/s41592-021-01101-x



# Local vs. global alignment

- Local alignment highlights small stretches with high similarity
  - random hit possible
  - better for strong sequence length differences
- Global alignment only works by overall high similarity
  - Full length similarity supports similar function



# MAFFT: online vs. local

## Multiple Sequence Alignment

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.

We have recently changed the default parameter settings for MAFFT. Alignments should run much more quickly and larger DNA alignments can be carried out by default. Please click the 'More options' button to review the defaults and change them if required.

**Important note:** This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of  
AUTOMATIC

sequences in any supported format:

Or upload a file:  No file selected.      [Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your Parameters

OUTPUT FORMAT  
Pearson/FASTA

The default settings will fulfill the needs of most users.  
[More options...](#) (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

```
MAFFT v7.453 (2019/Nov/8)
https://mafft.cbrc.jp/alignment/software/
MBE 30:772-780 (2013), NAR 30:3059-3066 (2002)

-----
```

**High speed:**

```
% mafft in > out
% mafft --retree 1 in > out (fast)
```

**High accuracy (for <~200 sequences x <~2,000 aa/nt):**

```
% mafft --maxiterate 1000 --localpair in > out (% linsi in > out is also ok)
% mafft --maxiterate 1000 --genafpair in > out (% einsi in > out)
% mafft --maxiterate 1000 --globalpair in > out (% ginsi in > out)
```

**If unsure which option to use:**

```
% mafft --auto in > out
```

--op # : Gap opening penalty, default: 1.53  
--ep # : Offset (works like gap extension penalty), default: 0.0  
--maxiterate # : Maximum number of iterative refinement, default: 0  
--clustalout : Output: clustal format, default: fasta  
--reorder : Outorder: aligned, default: input order  
--quiet : Do not report progress  
--thread # : Number of threads (if unsure, --thread -1)  
--dash : Add structural information (Rozewicki et al, submitted)

<https://www.ebi.ac.uk/Tools/msa/mafft/>

# Alignment construction

- Letters representing amino acid residues are arranged to highlight similarities
- Gaps (-) are inserted to show differences

Input FASTA file with sequences

```
>seq1  
MDTEKYMEKWIDQGHALFPEEDQ  
>seq2  
MDTDIKYSEKWIQGHSLFPPDDQ  
>seq3  
METEKYMEKWIQGHSIFPEDQ  
>seq4  
MDTEIKYMEKWIQGHALFPDDQP
```

Alignment of sequences

```
seq1    MDTE-KYMEKWIDQGHALFPEEDQ-  
seq2    MDTD-IKYSEKWI-QGHSLFPD-DQ-  
seq3    METE-KYMEKWI-QGHSIFPE-DQ-  
seq4    MDTEIKYMEKWI-QGHALFPD-DQP
```



# Alignment formats

## CLUSTAL

AthCHS	MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLOAEYPDYYFRITNSEHMTDLKEK
BvuCHS	M---ATPSVQEIRDAQRSNGPATILAIGTANPANE MYQAEYPDFYFRVTKEHMSELKQK
	* .:.*:***:***:*** * **** * .: ****:***:***:***:***
AthCHS	FKRMCDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLDTRQDIVVEVPKLGKEAAVKAIK
BvuCHS	FKRMCDKSMIKKRYMHVTQELLEENPHMCDYNASSLNTRQDILATEVPKLGKEAAVKAIK
	***** * :***:***:***:***:*** * * .***:***:***:***:***
AthCHS	EWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLRPSVKRLMMYQQGCFAGGTVLRIAK
BvuCHS	EWGQPRSKITHVIFCTTSGVDMPGADYQLTKLLRPSVKRFMLYQQGCYAGGTVLRLAK
	*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****
AthCHS	DIAENNRRGARVLVVCEITAVTFRGPDSDTHLDLSLVGQALFSDGAAALIVGSDPDTSVGK
BvuCHS	DIAENNRRGARVLVVCAEITVICFRGPTETHLDMSIGQALFGDGAGAIVVGADLDEI-ER
	:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****
AthCHS	PIFEMVSAAQTILPDSGAI DGHREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISD
BvuCHS	PIFQLWAAQTILPDSGAI DGHREVGLAFHLLKDVPGLISKNIKEVALVEAFKPIGIDD
	***:***:*****:*****:*****:*****:*****:*****:*****:*****:*****
AthCHS	WNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHLVSEYGNMSSACVLFILDEMRRKSA
BvuCHS	WNSIFWAAHPGGPAILDQVESKLGLQDKLSTTRHVLSEFGNMSSACVLFILDEMRRKRSM
	***:***:*****:*****:*****:*****:*****:*****:*****:*****:*****
AthCHS	KDGVATTGEGLEWGVLFGFPGPLTVETVVLHSVPL--
BvuCHS	KEGMATTGEGLEWGVLFGFPGPLTVETVMLHSVPIAN
	*****:*****:*****:*****:*****:*****:*****:*****:*****:*****

## FASTA

```
>AthCHS
MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLOAEYPDYYFRITNSEHMTDLKEK
FKRMCDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLDTRQDIVVEVPKLGKEAAVKAIK
EWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLRPSVKRLMMYQQGCFAGGTVLRIAK
DIAENNRRGARVLVVCEITAVTFRGPDSDTHLDLSLVGQALFSDGAAALIVGSDPDTSVGK
PIFEMVSAAQTILPDSGAI DGHREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISD
WNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHLVSEYGNMSSACVLFILDEMRRKSA
KGDVATTGEGLEWGVLFGFPGPLTVETVVLHSVPL--
>BvuCHS
M---ATPSVQEIRDAQRSNGPATILAIGTANPANE MYQAEYPDFYFRVTKEHMSELKQK
FKRMCDKSMIKKRYMHVTQELLEENPHMCDYNASSLNTRQDILATEVPKLGKEAAVKAIK
EWGQPRSKITHVIFCTTSGVDMPGADYQLTKLLRPSVKRFMLYQQGCYAGGTVLRLAK
DIAENNRRGARVLVVCAEITVICFRGPTETHLDMSIGQALFGDGAGAIVVGADLDEI-ER
PIFQLWAAQTILPDSGAI DGHREVGLAFHLLKDVPGLISKNIKEVALVEAFKPIGIDD
WNSIFWAAHPGGPAILDQVESKLGLQDKLSTTRHVLSEFGNMSSACVLFILDEMRRKSM
KEGMATTGEGLEWGVLFGFPGPLTVETVMLHSVPIAN
```

# Alignment trimming

- Removal of alignment columns with low abundance
- Columns with low abundance are not informative
- Reduced alignment lengths makes tree construction easier

Alignment of sequences

seq1	MDTE - KYMEKWI DQGHALFP EEDQ -
seq2	MDTDI KYSEKWI - QGHSLFPD - DQ -
seq3	METE - KYMEKWI - QGHSIFPE - DQ -
seq4	MDTEIKYMEKWI - QGHALFPD - DQP



Trimmed alignment of sequences

seq1	MDTE - KYMEKWI QGHALFP EDQ
seq2	MDTDI KYSEKWI QGHSLFP DDQ
seq3	METE - KYMEKWI QGHSIFP EDQ
seq4	MDTEIKYMEKWI QGHALFP DDQ

# Construction of a phylogenetic tree

- Construction of phylogenetic trees is based on sequence alignments
- Similar sequences (recent shared ancestors) are grouped

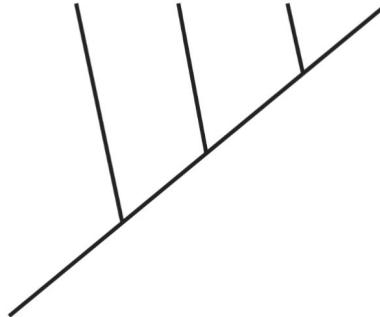
Trimmed alignment of sequences

seq1	MDTE - KYMEKWIQGHALFPEDQ
seq2	MDTDIKYSEKWIQGHSLFPDDQ
seq3	METE - KYMEKWIQGHSIFPEDQ
seq4	MDTEIKYMEKWIQGHALFPDDQ

Distance calculation

	seq1	seq2	seq3	seq4
seq1	.	5	3	2
seq2	.	.	6	3
seq3	.	.	5	.
seq4	.	.	.	.

seq2 seq3 seq1 seq4



# Substituting amino acids with codons

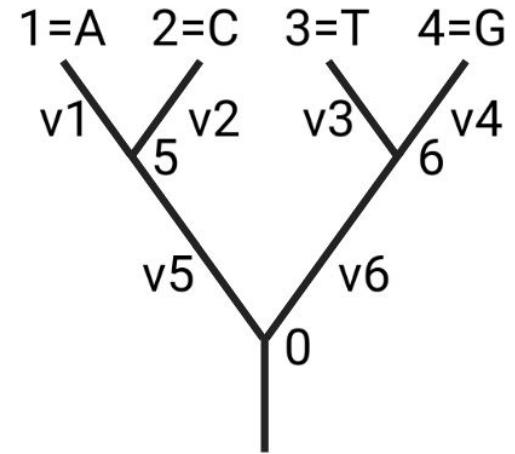
- Phylogenetic signal is boosted by substitution of amino acids by codons
- More positions that can show variation
- Alignment can still be based on amino acid residues (avoids splitting codons)

ATG GAT ACT GAA ...

seq1 MDTE - KYMEKWI IDQGHALFPEEDQ -  
seq2 MDTD IKY SEKWI - QGHSLFPD - DQ -  
seq3 METE - KYMEKWI - QGHSIFPE - DQ -  
seq4 MDTEIKYMEKWI - QGHALFPD - DQP

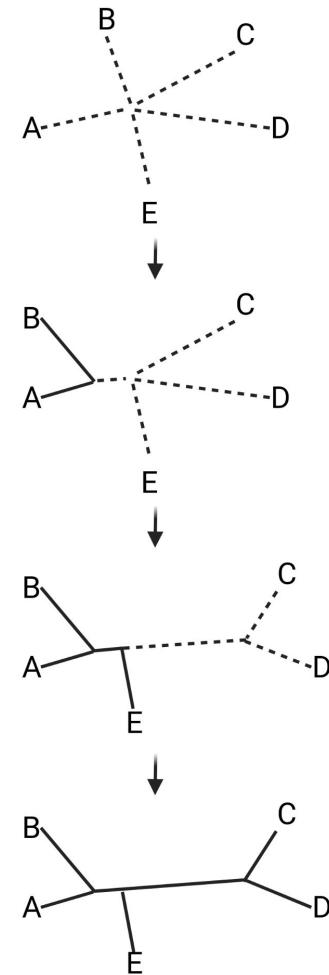
# Maximum Likelihood (ML) trees

- Finding the tree with highest likelihood to explain the observed sequences
- Number of possible trees depends on number of sequences
- Often impossible to explore all possible trees (very computationally demanding)
- Tool: RAxML-NG



# Neighbor Joining (NJ) tree

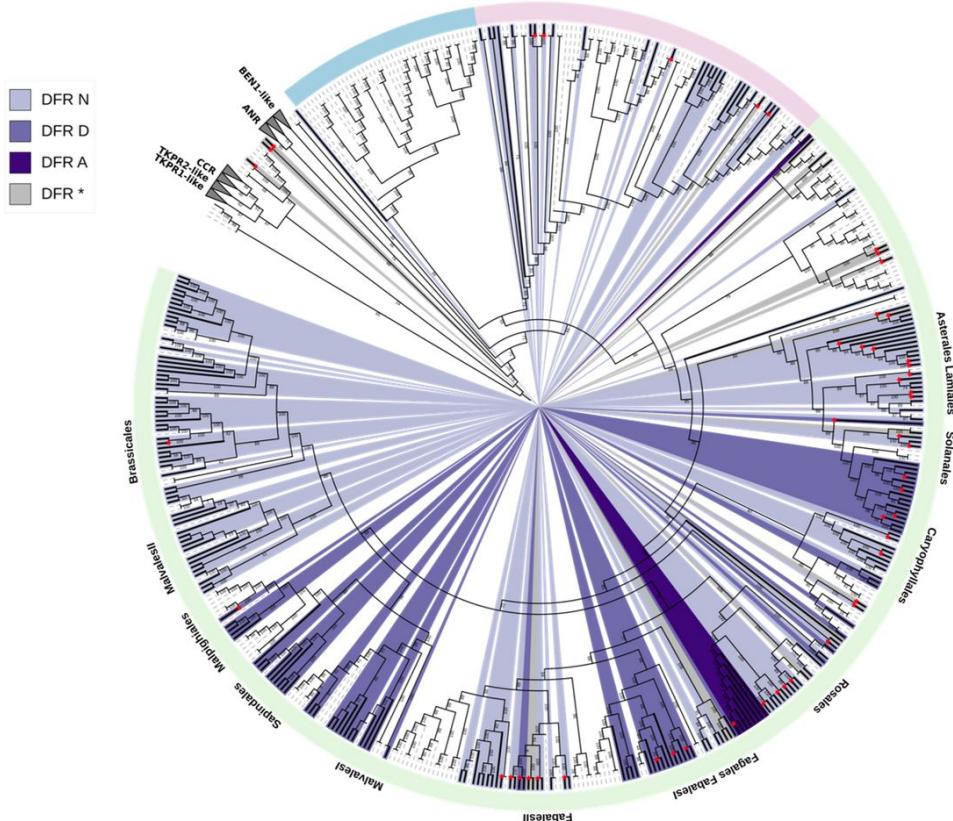
- Combination of sequences (groups of sequences) with smallest distance
- Quick and computationally cheap approach
- Less reliable than maximum likelihood (ML) trees
- Tool: MEGA



Kumar et al., 2016: 10.1093/molbev/msw054

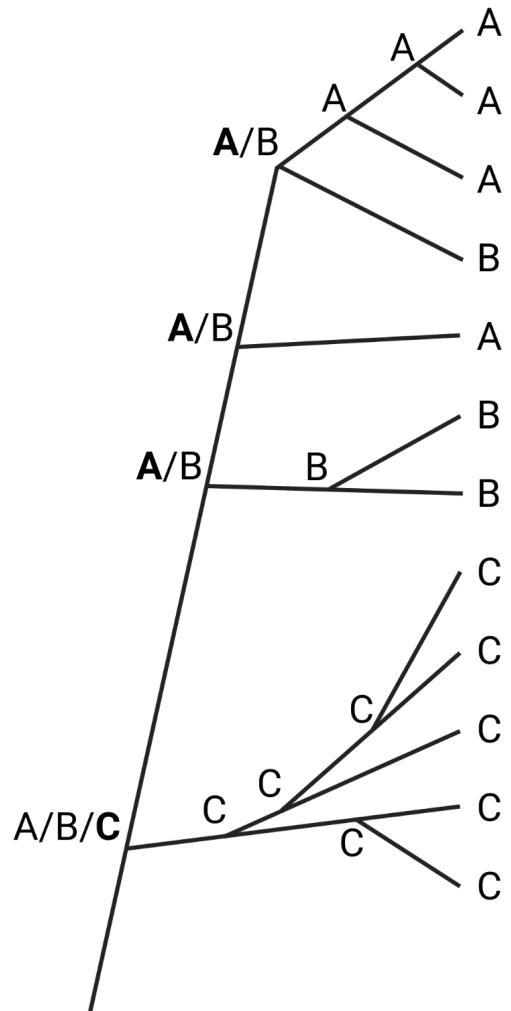
# **EXAMPLE: gene tree of *DFR***

- Evolution of *DFR* in land plants
  - Multiple clades visible
  - Sequences of closely related species cluster in lineages



# Ancestral state reconstruction

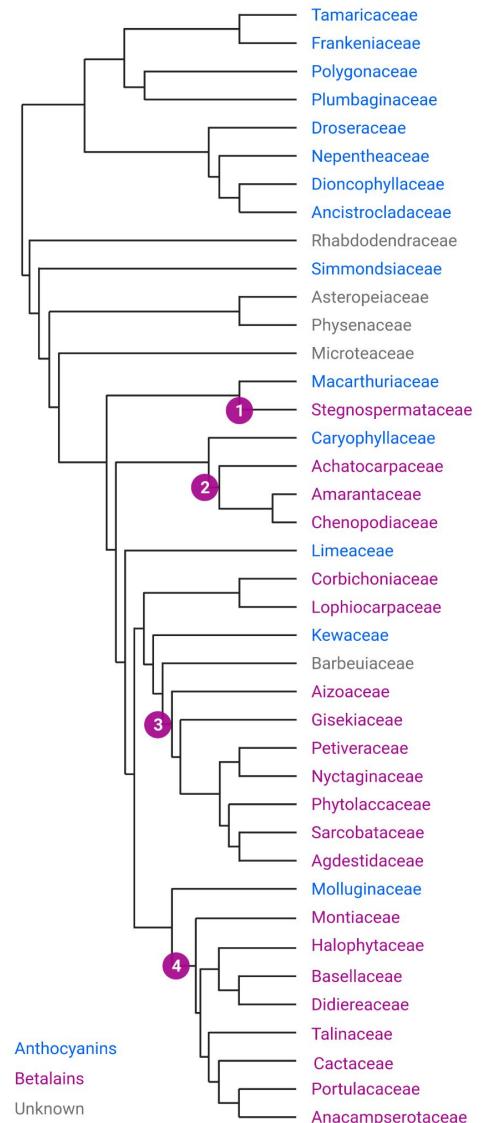
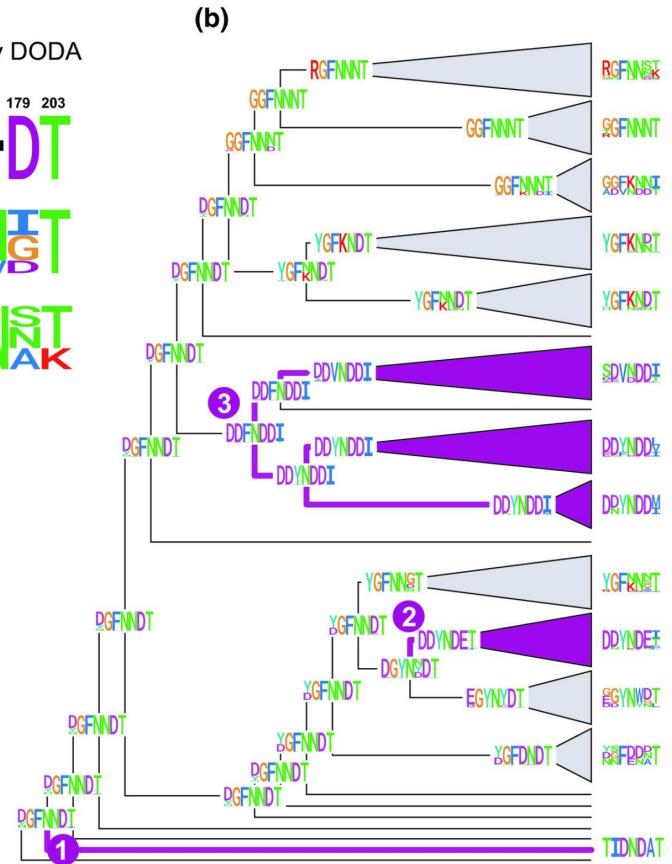
- Identify the state of a certain sequence/trait in a last common ancestor
- Multiple states are possible in some cases
- Ancestral states are often predicted with a probability



# Example: Ancestral state reconstruction

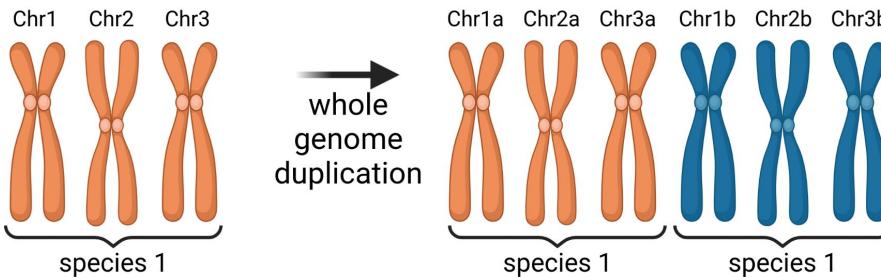
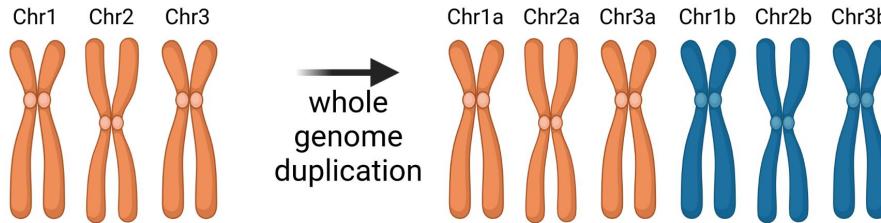
(a)

High activity DODA	No/Low activity DODA
74 75 76 152 178 179 203	74 75 76 152 178 179 203
<b>1</b>	
<b>2</b>	
<b>3</b>	

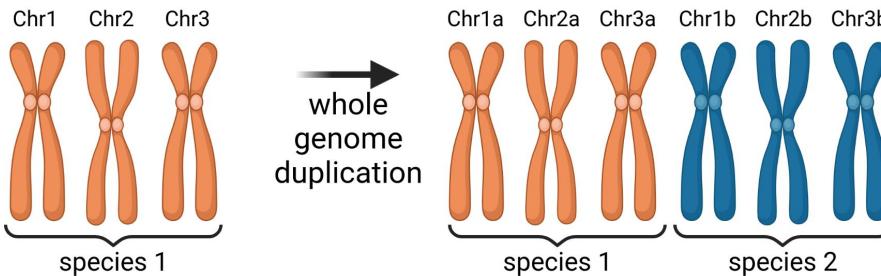


Sheehan et al., 2020: 10.1111/nph.16089

# Whole genome duplications



autopolyploidy



allopolyploidy

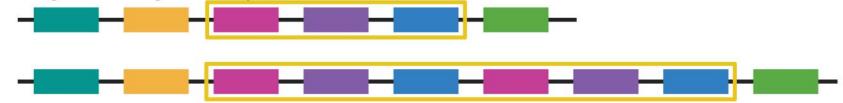
# Classification of duplications: tandem, segment, WGD

- Gene copies can be the results of different mechanisms:
  - Tandem duplication: copies are located next to each other on the same chromosome
  - Segmental duplication: like tandem duplications, but multiple adjacent genes are also duplicated
  - Whole genome duplication (WGD): all genes are duplicated

Tandem gene duplication



Segmental gene duplication

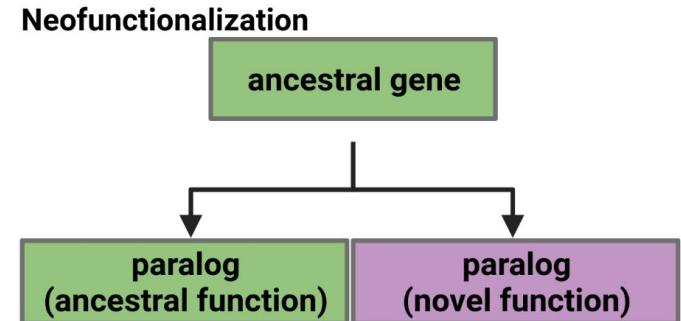
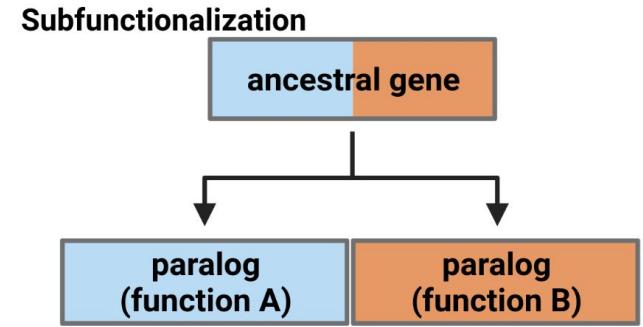


Whole genome duplication

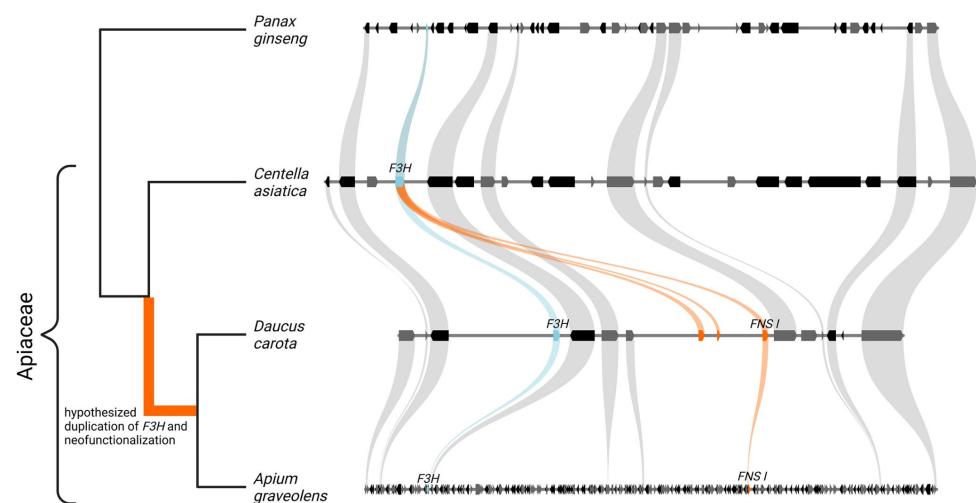
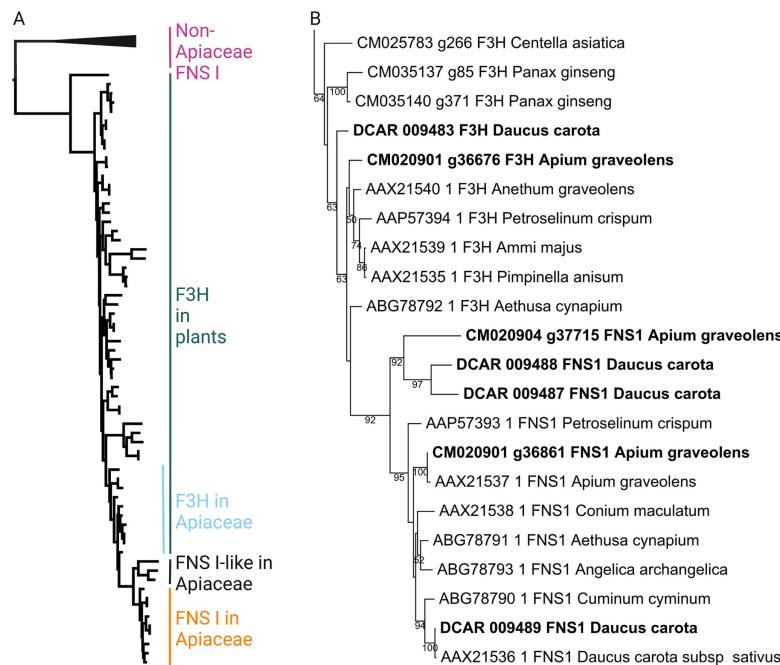


# Subfunctionalization & Neofunctionalization

- Subfunctionalization: gene copies fulfill parts of the ancestral function
- Neofunctionalization: one copy maintains ancestral function, while other copy can acquire a novel function



# EXAMPLE: *F3H/FNS I* neofunctionalization following gene duplication in Apiaceae



<https://doi.org/10.1371/journal.pone.0280155>



# Summary

- Concepts of evolution
- Homology (orthology & paralogy)
- Alignment and phylogenetic tree construction
- Ancestral state reconstruction
- Neo- and subfunctionalization

# Time for questions!



# Questions

1. What is the difference between homolog, paralog, and ortholog?
2. What are the advantages/disadvantages of running BLAST locally vs. at the NCBI website?
3. What are suitable BLAST alternatives?
4. What is the difference between a local and a global alignment?
5. What are the different methods for the construction of a phylogenetic tree?
6. What is ancestral state reconstruction?
7. What are possible mechanisms underlying gene duplications?
8. What is neofunctionalization/subfunctionalization?
9. What is the relationship of A and B? (paralog/ortholog)
10. What is the relationship of D and E (paralog/ortholog)

