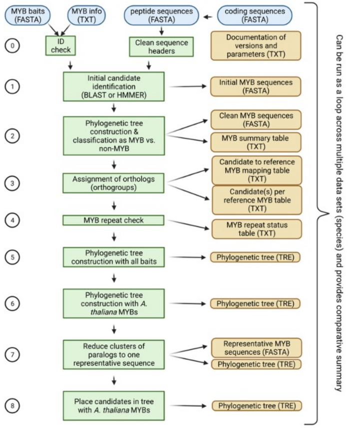
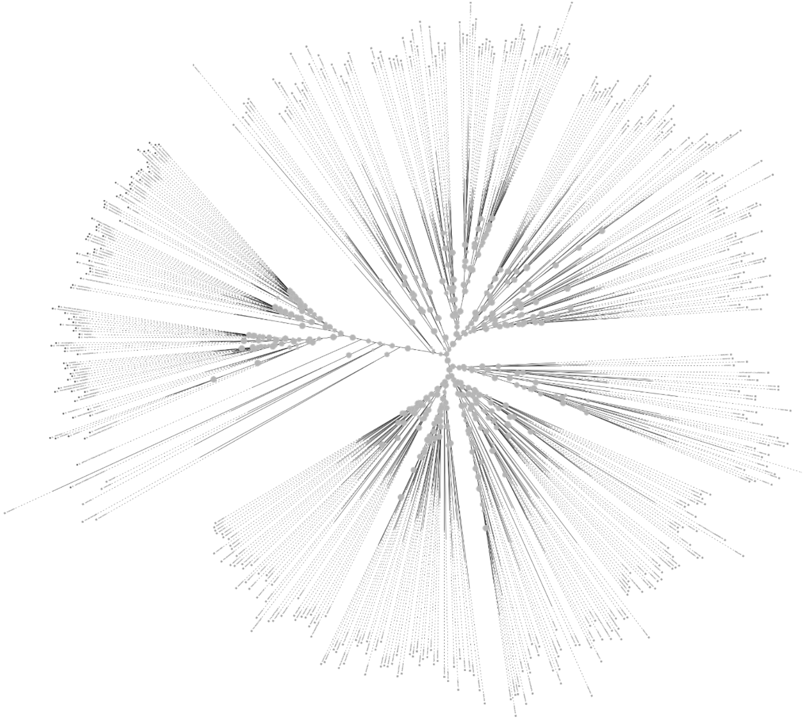
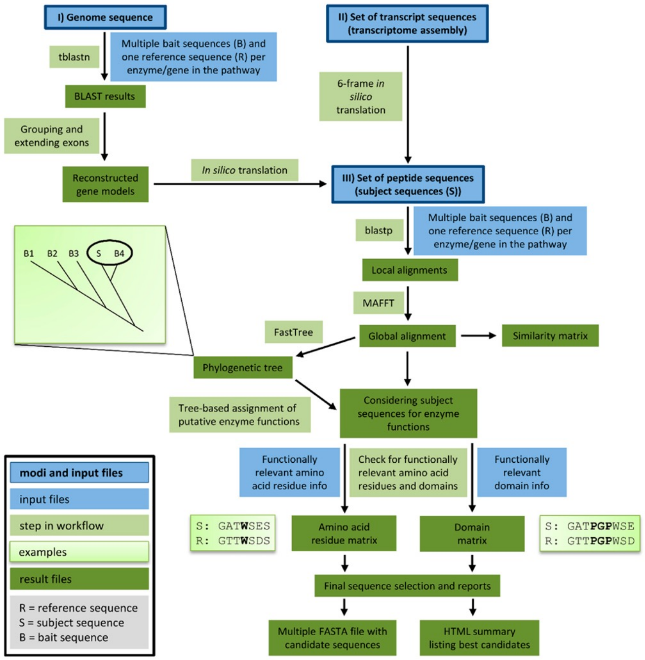
Technische
Universität
Braunschweig





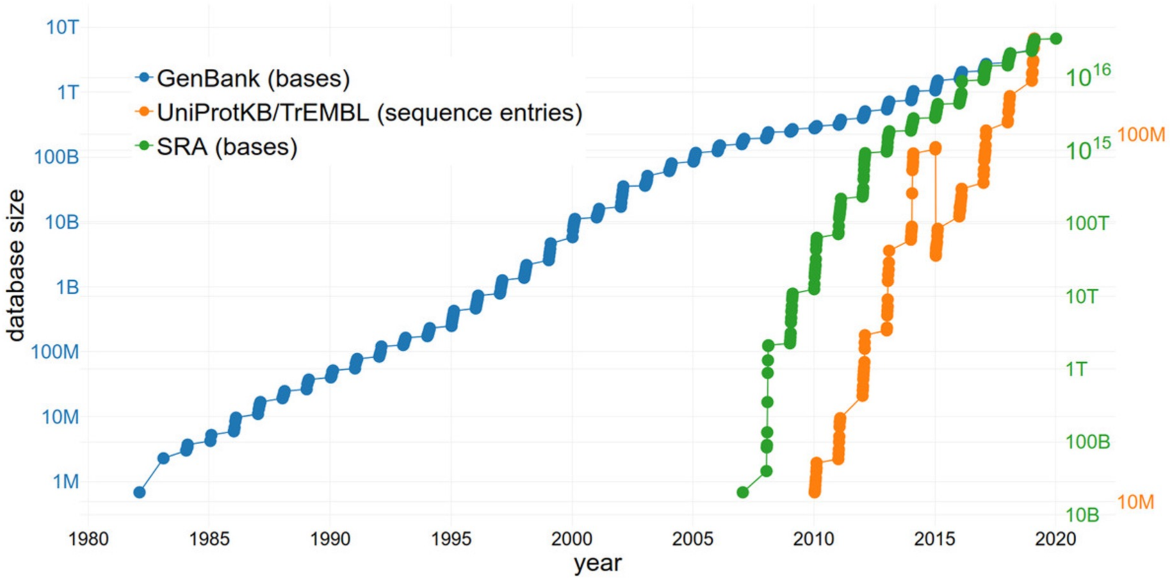


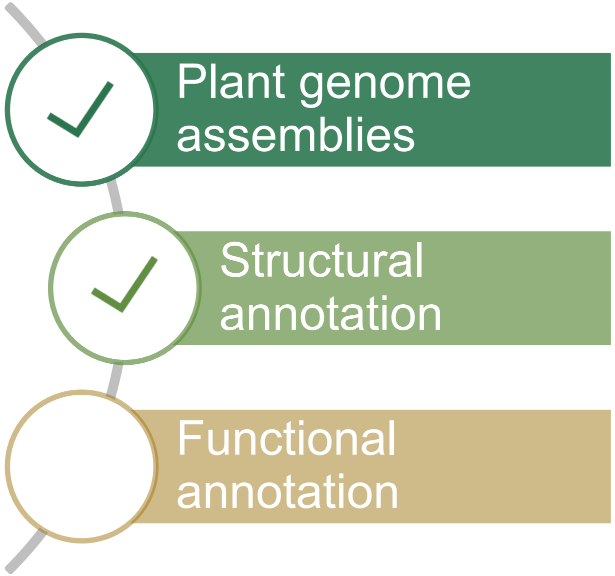
Automatic approaches for cross-species knowledge transfer based on orthology

Corinna Thoben (Institute of Plant Biology & BRICS, TU Braunschweig, Germany)

@PuckerLab https://genomic.social/@PuckerLab

Availability of plant genome sequences



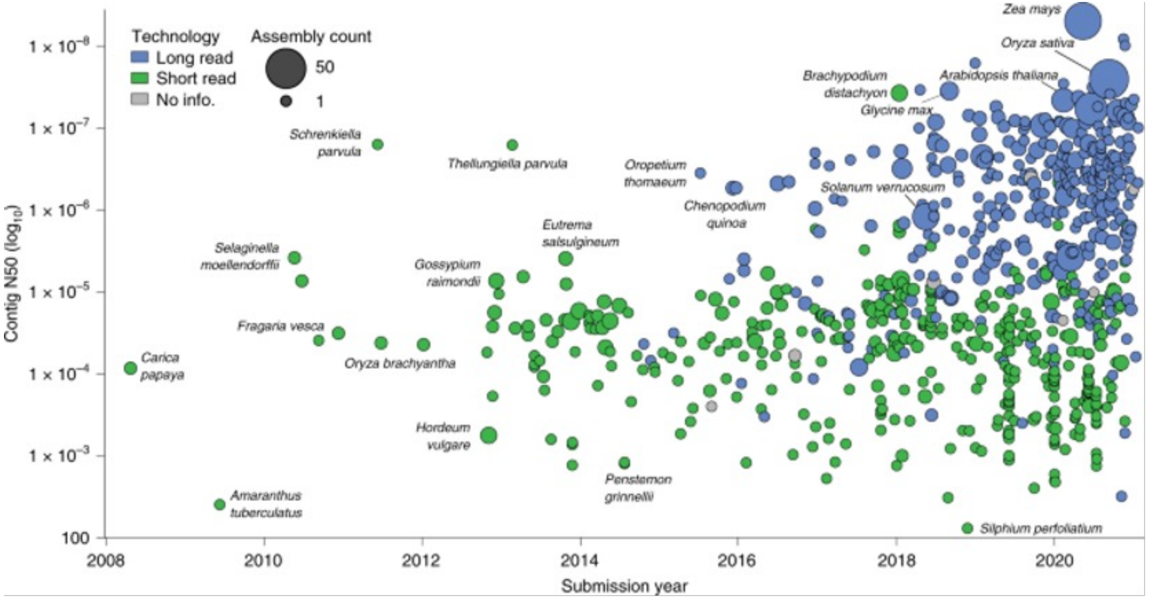

✓ Plant genome assemblies

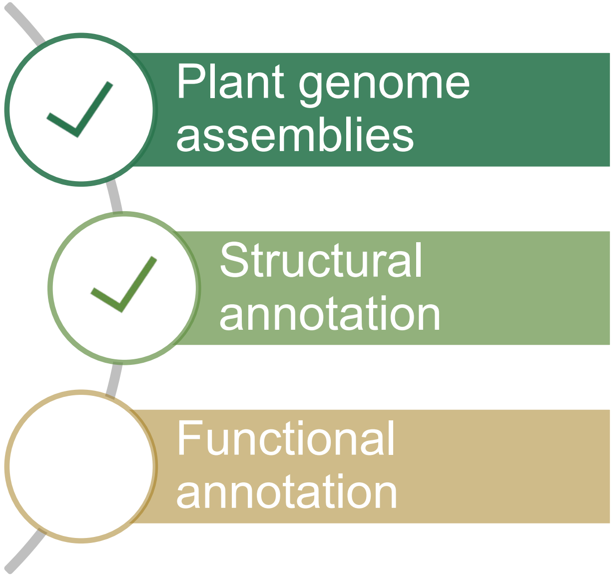
✓ Structural annotation

Functional annotation

Sielemann *et al.*, 2020: 10.7717/peerj.9954

Availability of plant genome sequences





✓ Plant genome assemblies

✓ Structural annotation

Functional annotation

Marks *et al.*, 2021: 10.1038/s41477-021-01031-8

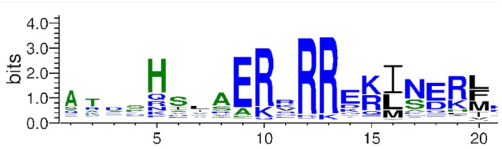
Functional annotation approaches

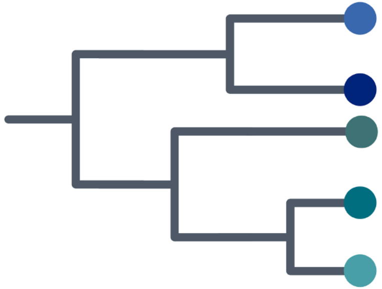
Three general methods:

Sequence similarity

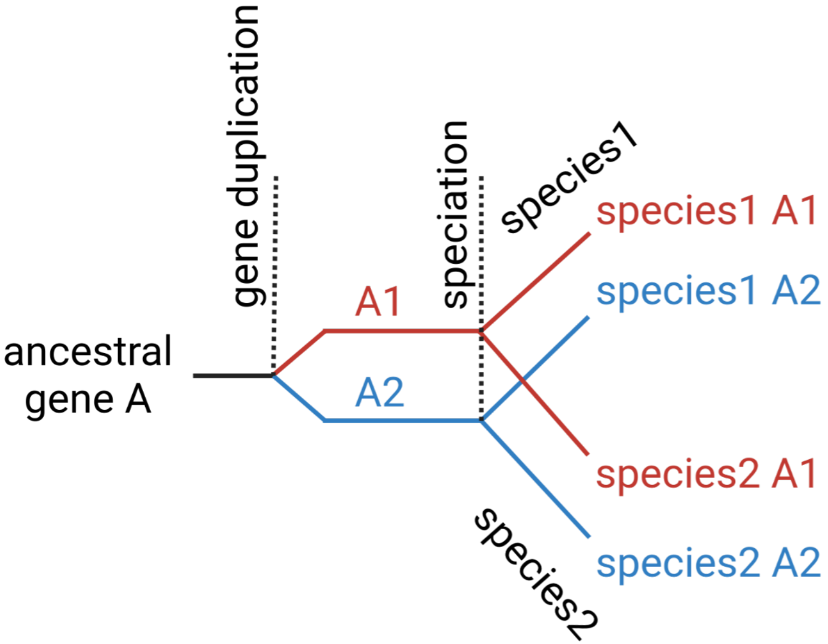
HMMs - Hidden Markov Models

Phylogenetic approaches

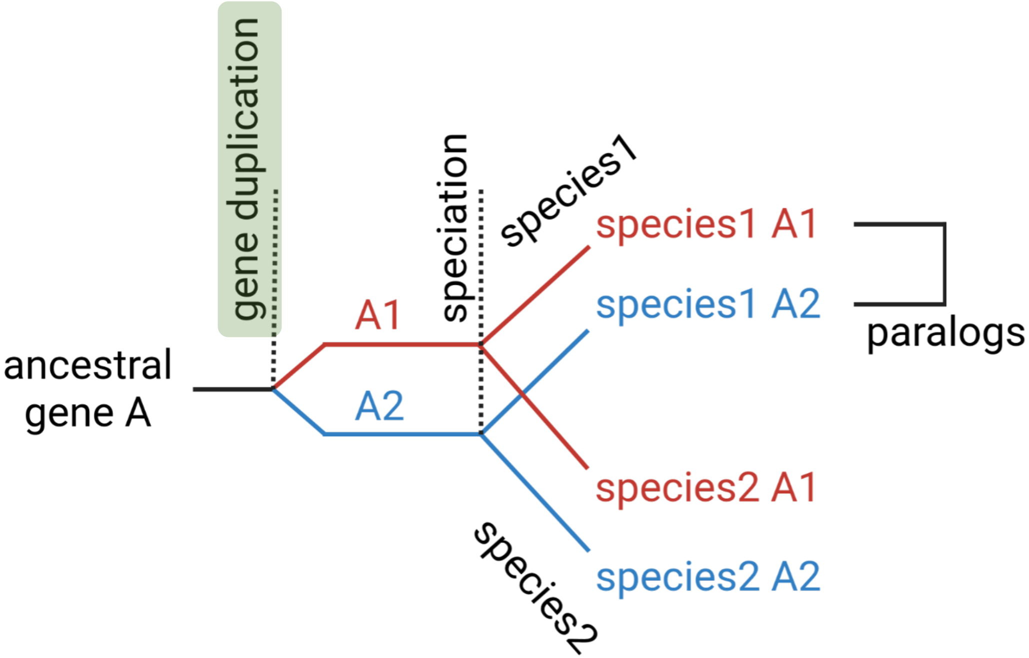




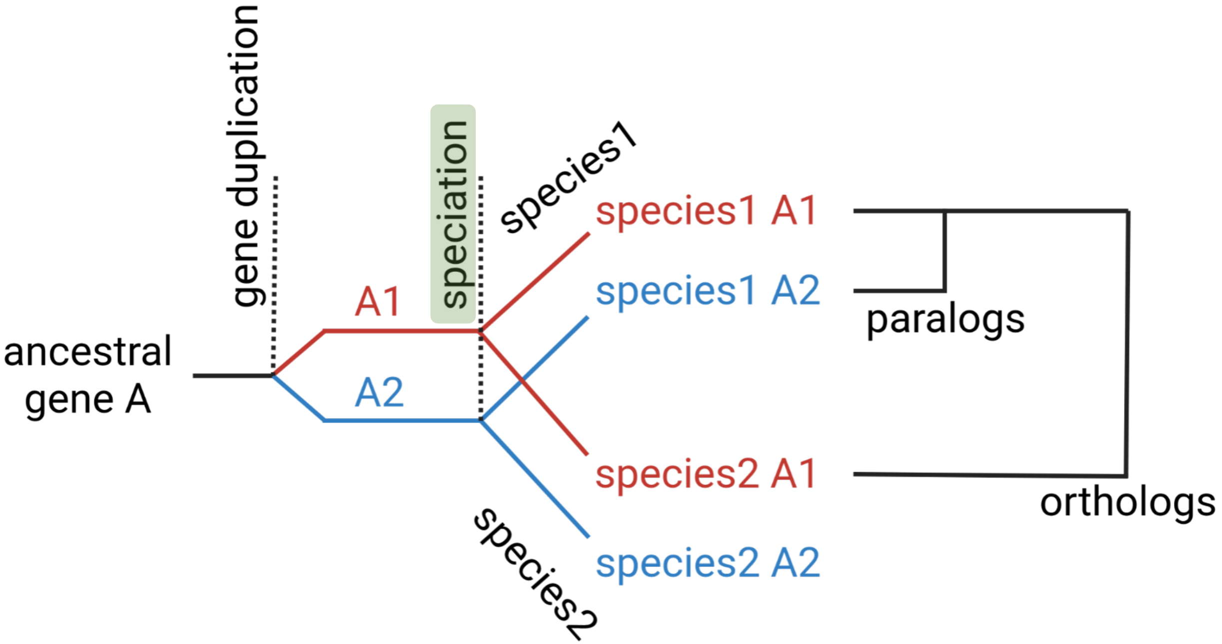
A phylogenetic approach:
Considering the evolutionary perspective



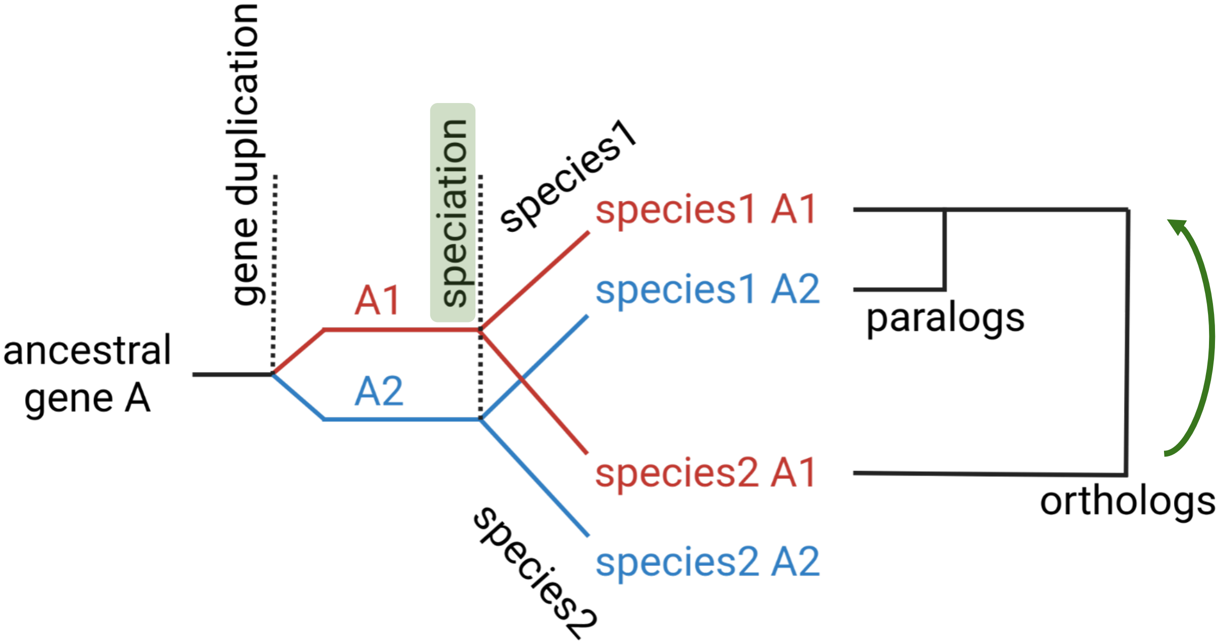
A phylogenetic approach:
Considering the evolutionary perspective



A phylogenetic approach:
Considering the evolutionary perspective



A phylogenetic approach:
Considering the evolutionary perspective



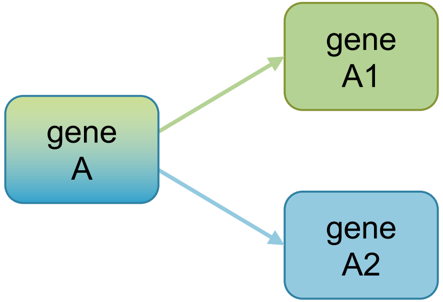
Cross-species knowledge transfer

A phylogenetic approach:
Considering the evolutionary perspective

Paralogs:

A phylogenetic approach:
Considering the evolutionary perspective

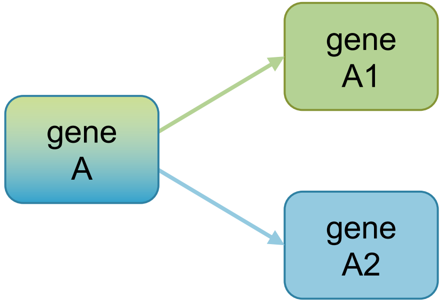
Paralogs:

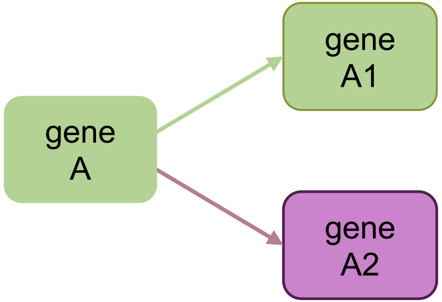


Subfunctionalization

A phylogenetic approach:
Considering the evolutionary perspective

Paralogs:

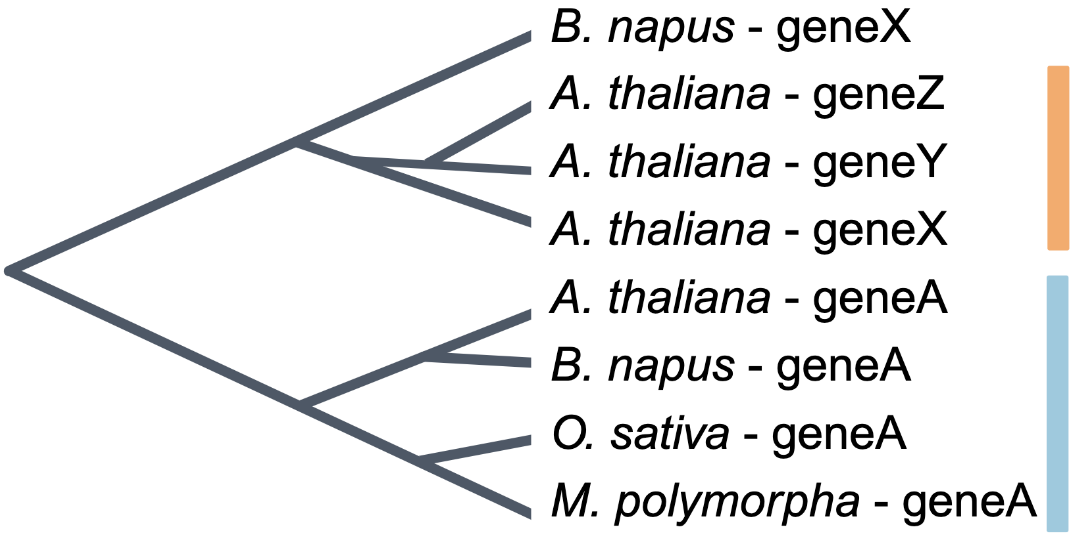




Subfunctionalization

Neofunctionalization

A phylogenetic approach:
Functional annotation transfer between orthologs



B. napus - geneX

A. thaliana - geneZ

A. thaliana - geneY

A. thaliana - geneX

A. thaliana - geneA

B. napus - geneA

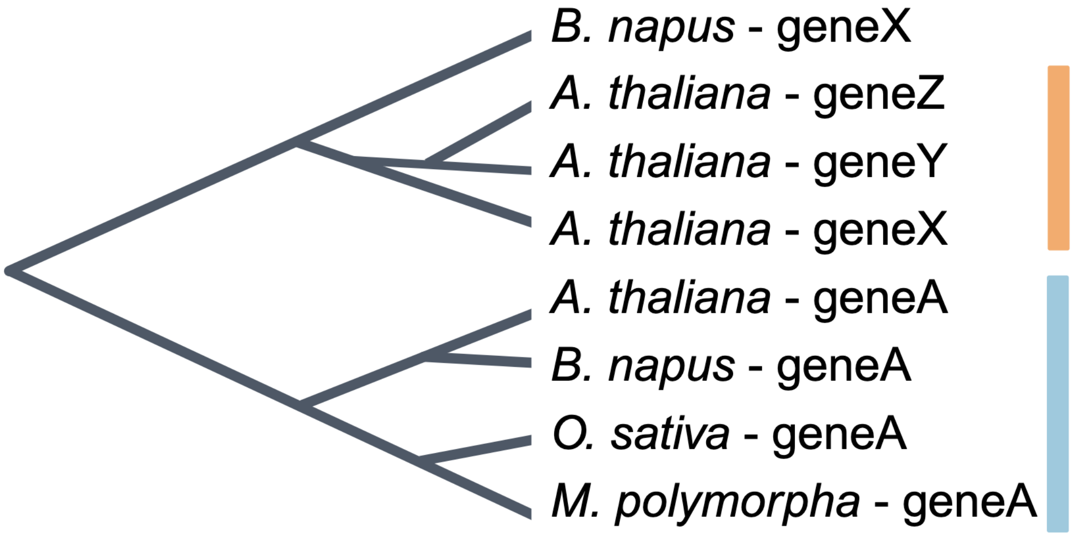
O. sativa - geneA

M. polymorpha - geneA

sub-/neofunctionalization of paralogs

same function among orthologs expected

A phylogenetic approach:
Functional annotation transfer between orthologs



B. napus - geneX
A. thaliana - geneZ
A. thaliana - geneY
A. thaliana - geneX
A. thaliana - geneA
B. napus - geneA
O. sativa - geneA
M. polymorpha - geneA

sub-/neofunctionalization of paralogs

same function among orthologs expected

geneA?

geneX?

How can we automate this?

- **Aim**: Functionally annotate geneA, geneX and geneY in *Vitis vinifera*

- Candidate identification via *bona fide* sequences as ‚baits‘

  → Bait collection

- Transfer knowledge through orthologs of well characterized sequences

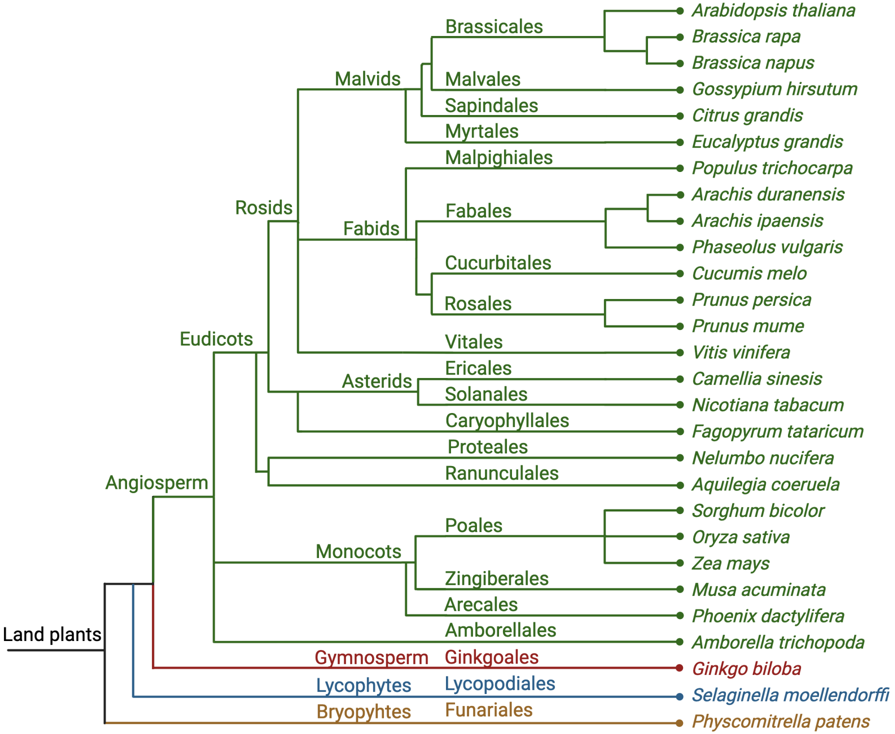
  → Reference sequences

How can we automate this?

Bait collection

→ Described in literature

→ Broad phylogenetic coverage



Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

How can we automate this?

geneA?

geneX?

Annotation steps

① Identification of candidates → BLAST search

How can we automate this?

geneA?

geneX?

Annotation steps

① Identification of candidates

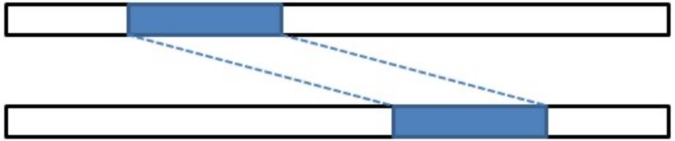
② Global alignment

How can we automate this?

Annotation steps

① Identification of candidates

② Global alignment


Local Alignment


Global Alignment

Issa *et al*., 2018: 10.1007/978-3-319-74690-6_11

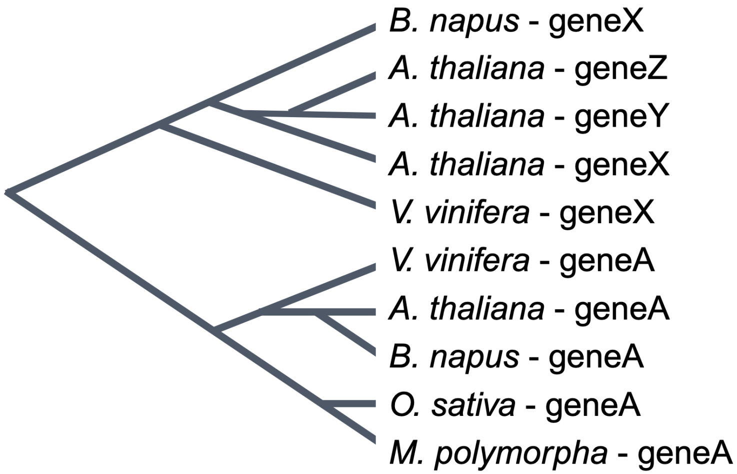
How can we automate this?

geneA?

geneX?

Annotation steps

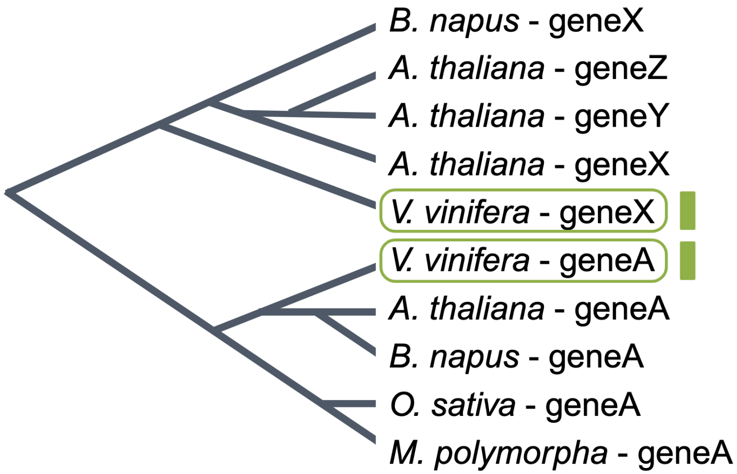
1. Identification of candidates
2. Global alignment
3. Phylogenetic tree construction



How can we automate this?

Annotation steps

1. Identification of candidates
2. Global alignment
3. Phylogenetic tree construction
4. Ortholog assignment

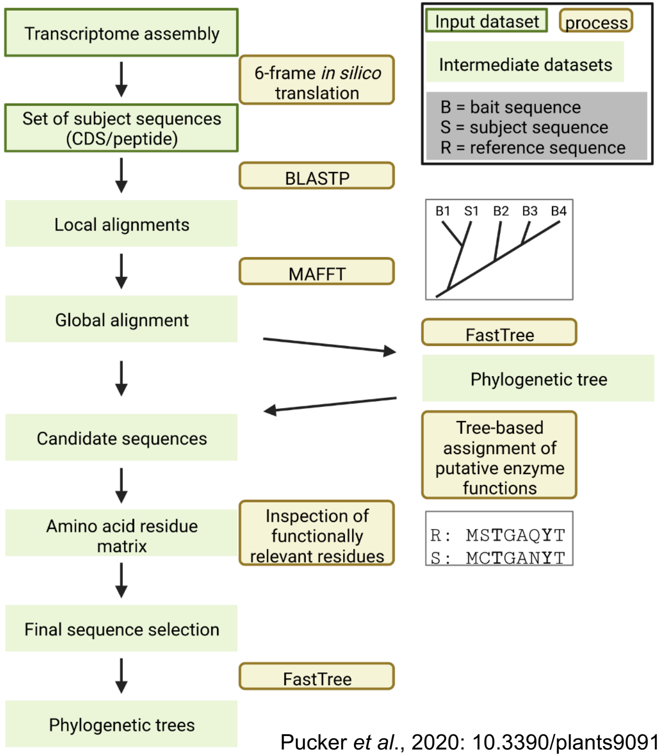


Automatic knowledge transfer

KIPEs	Knowledge-based Identification of Pathway Enzymes
MYB_annotator	Identification and annotation of MYB transcription factors
bHLH_annotator	Identification and annotation of bHLH transcription factors

KIPEs: Knowledge-based Identification of Pathway Enzymes

- Annotation of the flavonoid biosynthesis
- Bait sequences for each enzyme of the pathway
- Reference sequences conserving functional relevant amino acid positions



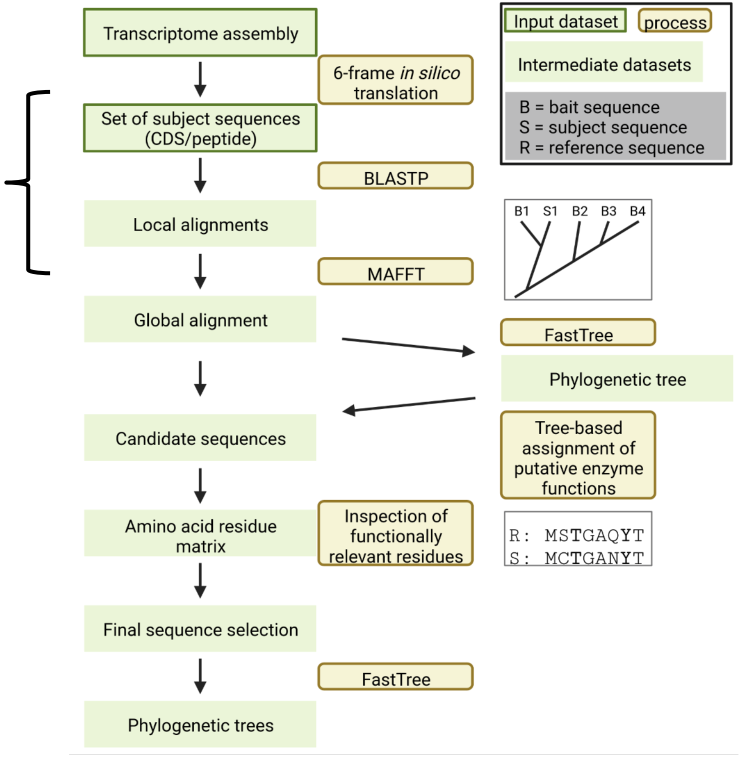
Input dataset
process
Intermediate datasets
B = bait sequence
S = subject sequence
R = reference sequence

Transcriptome assembly
6-frame *in silico* translation
Set of subject sequences (CDS/peptide)
BLASTP
Local alignments
MAFFT
Global alignment
FastTree
Phylogenetic tree
Tree-based assignment of putative enzyme functions
Candidate sequences
Amino acid residue matrix
Inspection of functionally relevant residues
Final sequence selection
FastTree
Phylogenetic trees

Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

KIPEs: Knowledge-based Identification of Pathway Enzymes

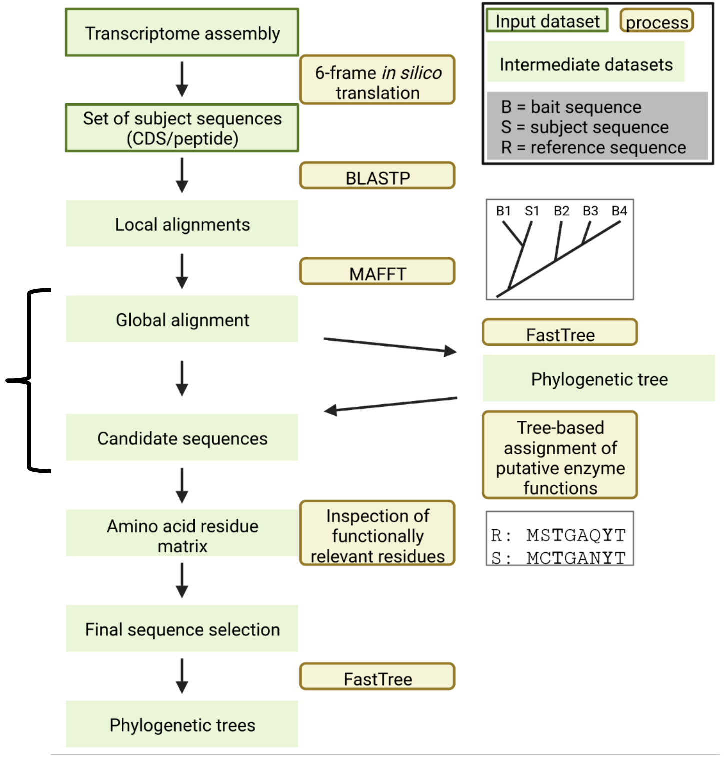
Identification of initial candidates



Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

KIPEs: Knowledge-based Identification of Pathway Enzymes

Best ortholog
candidate


Transcriptome assembly

6-frame *in silico* translation

Set of subject sequences (CDS/peptide)

BLASTP

Local alignments

MAFFT

Global alignment

FastTree

Phylogenetic tree

Tree-based assignment of putative enzyme functions

Candidate sequences

Amino acid residue matrix

Inspection of functionally relevant residues

Final sequence selection

FastTree

Phylogenetic trees

Input dataset
process
Intermediate datasets

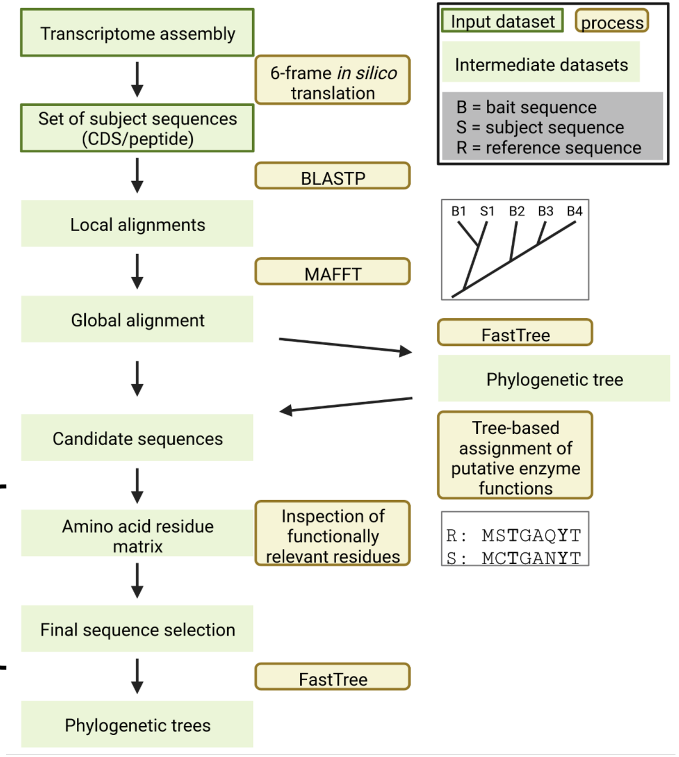
B = bait sequence
S = subject sequence
R = reference sequence

R: MSTGAQYT
S: MCTGANYT

Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

KIPEs: Knowledge-based Identification of Pathway Enzymes

Prediction
of functionality

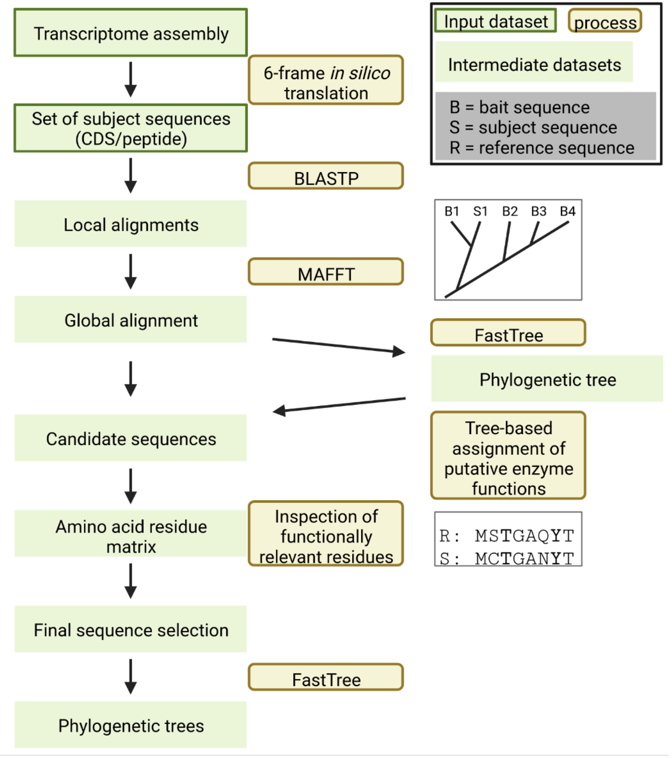


Input dataset
process
Intermediate datasets
B = bait sequence
S = subject sequence
R = reference sequence

Transcriptome assembly
6-frame *in silico* translation
Set of subject sequences (CDS/peptide)
BLASTP
Local alignments
MAFFT
Global alignment
FastTree
Phylogenetic tree
Tree-based assignment of putative enzyme functions
Candidate sequences
Amino acid residue matrix
Inspection of functionally relevant residues
Final sequence selection
FastTree
Phylogenetic trees

Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

KIPEs: Knowledge-based Identification of Pathway Enzymes



Input dataset
process
Intermediate datasets
B = bait sequence
S = subject sequence
R = reference sequence

Transcriptome assembly
6-frame *in silico* translation
Set of subject sequences (CDS/peptide)
BLASTP
Local alignments
MAFFT
Global alignment
FastTree
Phylogenetic tree
Tree-based assignment of putative enzyme functions
Candidate sequences
Amino acid residue matrix
Inspection of functionally relevant residues
Final sequence selection
FastTree
Phylogenetic trees

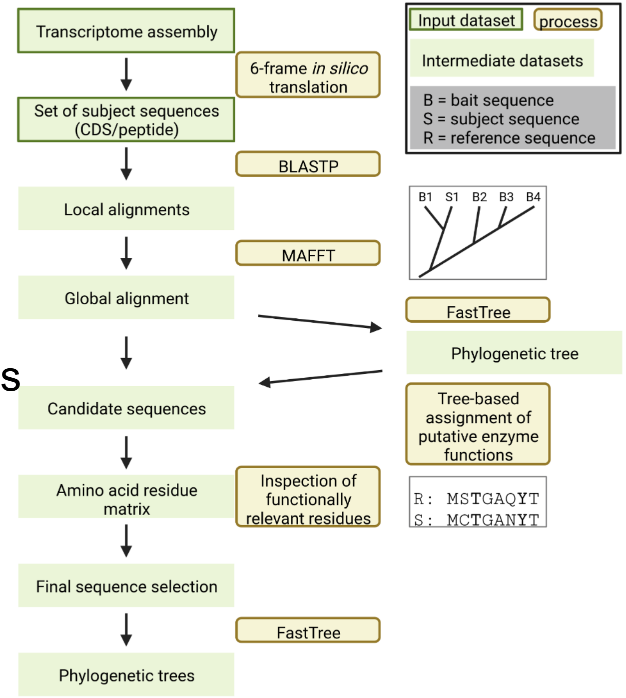
```
R:  MSTGAQYT
S:  MCTGANYT
```

Final phylogenetic tree

Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

KIPEs: Knowledge-based Identification of Pathway Enzymes

- Identification of flavonoid pathway enzymes
- Prediction of functionality by existing knowledge
- Effective and automatic screening of large data sets possible
- Systematic comparisons of flavonoid biosynthesis across species



Input dataset

process

Intermediate datasets

B = bait sequence
S = subject sequence
R = reference sequence

Transcriptome assembly

6-frame *in silico* translation

Set of subject sequences (CDS/peptide)

BLASTP

Local alignments

MAFFT

Global alignment

FastTree

Phylogenetic tree

Tree-based assignment of putative enzyme functions

Candidate sequences

Amino acid residue matrix

Inspection of functionally relevant residues

R: MSTGAQYT
S: MCTGANYT

Final sequence selection

FastTree

Phylogenetic trees

Pucker *et al.*, 2020: 10.3390/plants9091103
Rempel *et al.*, 2023: 10.1101/2022.06.30.498365

MYB_annotator & bHLH_annotator

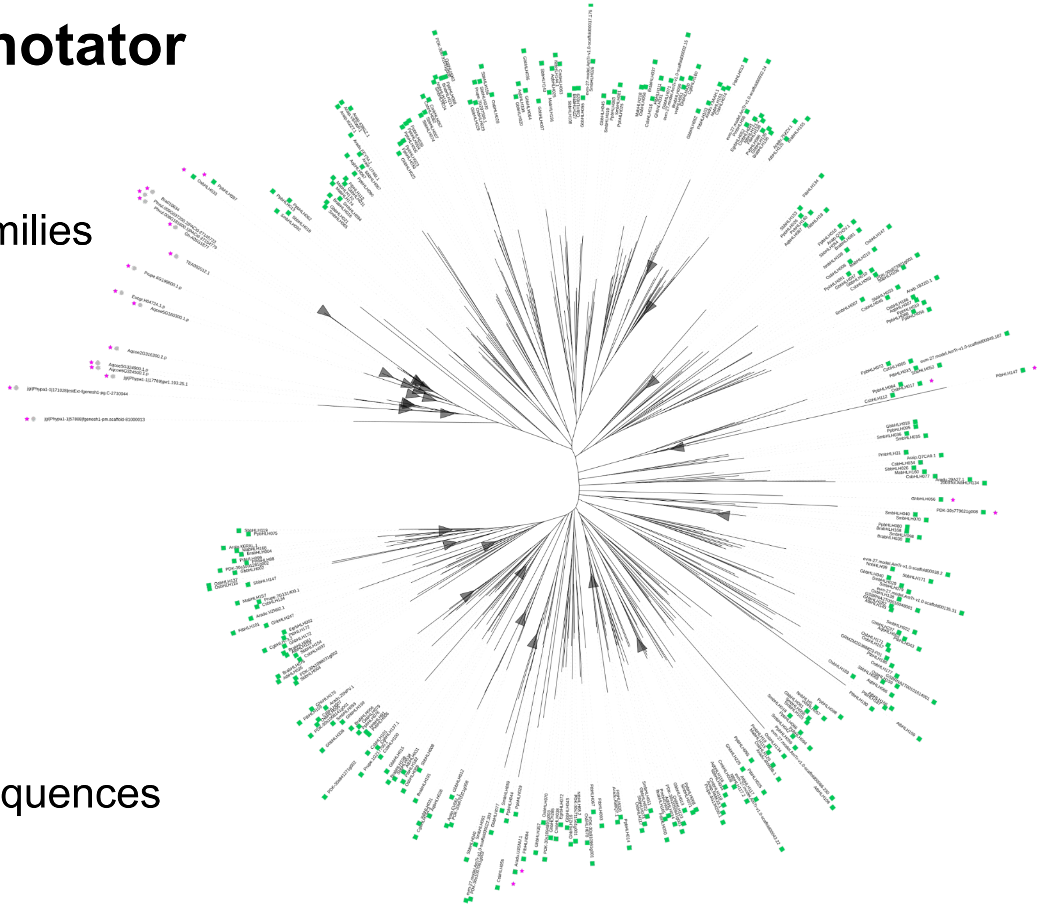
- Annotation of transcription factor families

- Phylogenetic distinct outgroup

  → BLAST: false positives

  → similar non-homologs vs.
  homolog family members

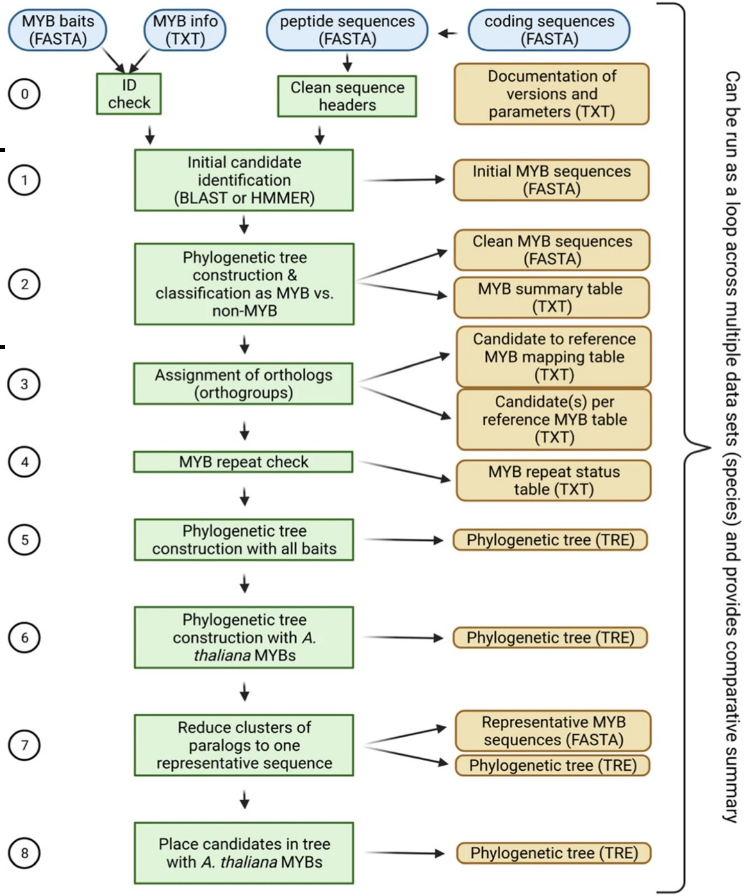
- Functionally annotated reference sequences



Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

MYB_annotator

Identification of candidates


MYB baits (FASTA)

MYB info (TXT)

peptide sequences (FASTA)

coding sequences (FASTA)

0 ID check

Clean sequence headers

Documentation of versions and parameters (TXT)

1 Initial candidate identification (BLAST or HMMER) → Initial MYB sequences (FASTA)

2 Phylogenetic tree construction & classification as MYB vs. non-MYB → Clean MYB sequences (FASTA); MYB summary table (TXT)

3 Assignment of orthologs (orthogroups) → Candidate to reference MYB mapping table (TXT); Candidate(s) per reference MYB table (TXT)

4 MYB repeat check → MYB repeat status table (TXT)

5 Phylogenetic tree construction with all baits → Phylogenetic tree (TRE)

6 Phylogenetic tree construction with *A. thaliana* MYBs → Phylogenetic tree (TRE)

7 Reduce clusters of paralogs to one representative sequence → Representative MYB sequences (FASTA); Phylogenetic tree (TRE)

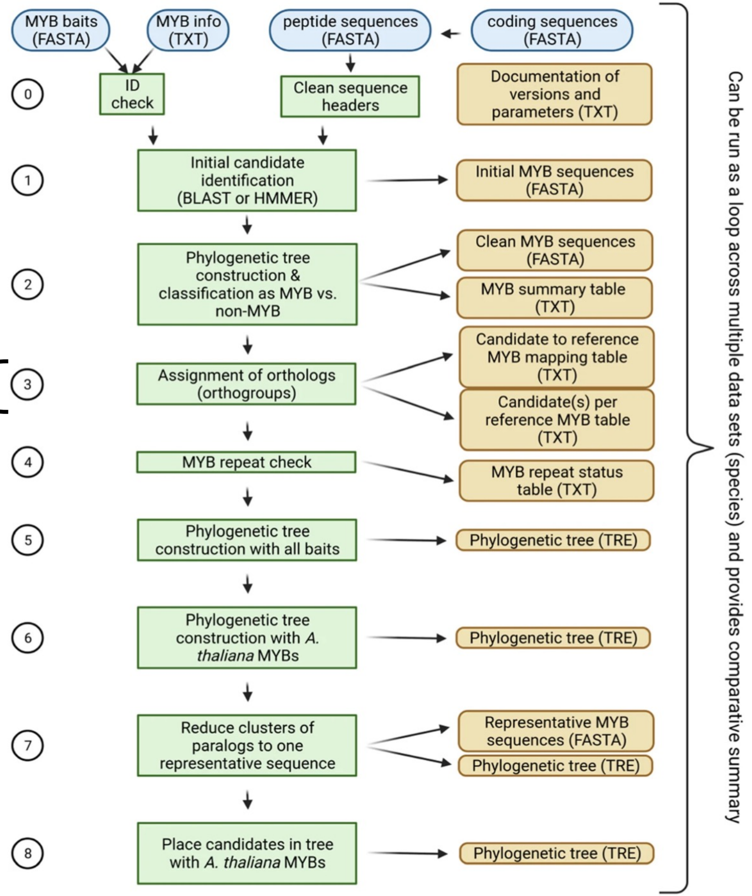
8 Place candidates in tree with *A. thaliana* MYBs → Phylogenetic tree (TRE)

Can be run as a loop across multiple data sets (species) and provides comparative summary

Pucker, 2022: 10.1186/s12864-022-08452-5

MYB_annotator

Functional annotation by reference sequences



MYB baits (FASTA)
MYB info (TXT)
peptide sequences (FASTA)
coding sequences (FASTA)

0 ID check
Clean sequence headers
Documentation of versions and parameters (TXT)

1 Initial candidate identification (BLAST or HMMER)
Initial MYB sequences (FASTA)

2 Phylogenetic tree construction & classification as MYB vs. non-MYB
Clean MYB sequences (FASTA)
MYB summary table (TXT)

3 Assignment of orthologs (orthogroups)
Candidate to reference MYB mapping table (TXT)
Candidate(s) per reference MYB table (TXT)

4 MYB repeat check
MYB repeat status table (TXT)

5 Phylogenetic tree construction with all baits
Phylogenetic tree (TRE)

6 Phylogenetic tree construction with *A. thaliana* MYBs
Phylogenetic tree (TRE)

7 Reduce clusters of paralogs to one representative sequence
Representative MYB sequences (FASTA)
Phylogenetic tree (TRE)

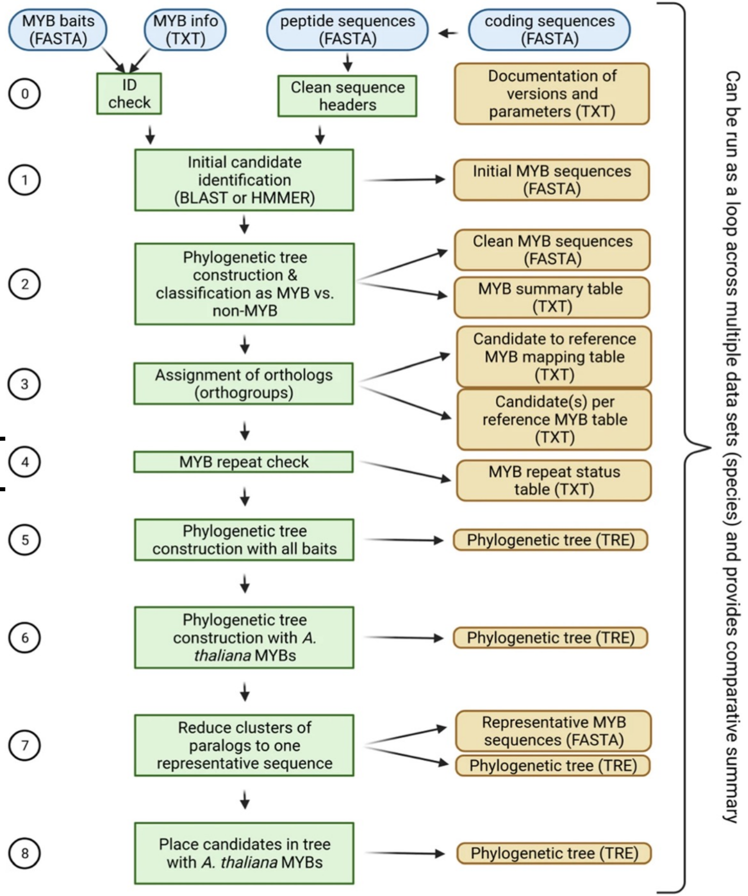
8 Place candidates in tree with *A. thaliana* MYBs
Phylogenetic tree (TRE)

Can be run as a loop across multiple data sets (species) and provides comparative summary

Pucker, 2022: 10.1186/s12864-022-08452-5

MYB_annotator

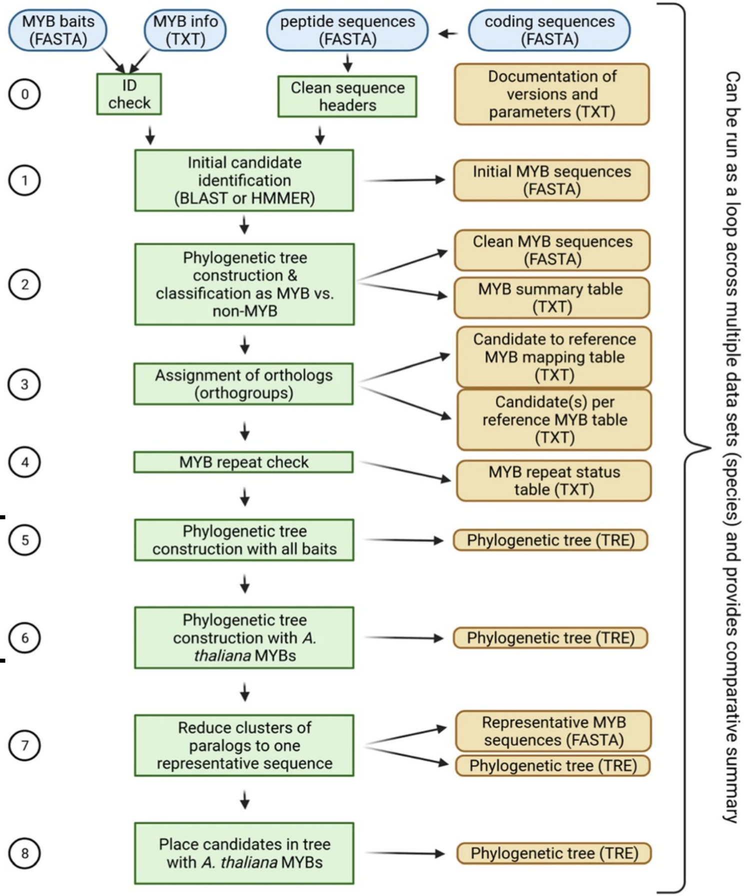
MYB repeat check



Pucker, 2022: 10.1186/s12864-022-08452-5

MYB_annotator

Phylogenetic tree construction



MYB baits (FASTA)

MYB info (TXT)

peptide sequences (FASTA)

coding sequences (FASTA)

0 ID check

Clean sequence headers

Documentation of versions and parameters (TXT)

1 Initial candidate identification (BLAST or HMMER) → Initial MYB sequences (FASTA)

2 Phylogenetic tree construction & classification as MYB vs. non-MYB → Clean MYB sequences (FASTA); MYB summary table (TXT)

3 Assignment of orthologs (orthogroups) → Candidate to reference MYB mapping table (TXT); Candidate(s) per reference MYB table (TXT)

4 MYB repeat check → MYB repeat status table (TXT)

5 Phylogenetic tree construction with all baits → Phylogenetic tree (TRE)

6 Phylogenetic tree construction with *A. thaliana* MYBs → Phylogenetic tree (TRE)

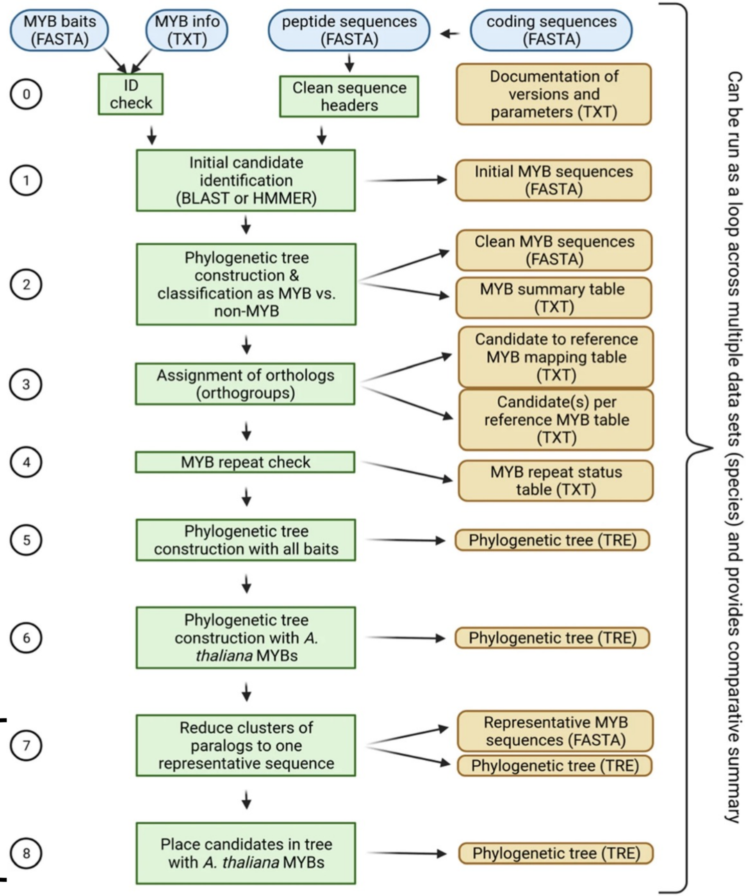
7 Reduce clusters of paralogs to one representative sequence → Representative MYB sequences (FASTA); Phylogenetic tree (TRE)

8 Place candidates in tree with *A. thaliana* MYBs → Phylogenetic tree (TRE)

Can be run as a loop across multiple data sets (species) and provides comparative summary

Pucker, 2022: 10.1186/s12864-022-08452-5

MYB_annotator

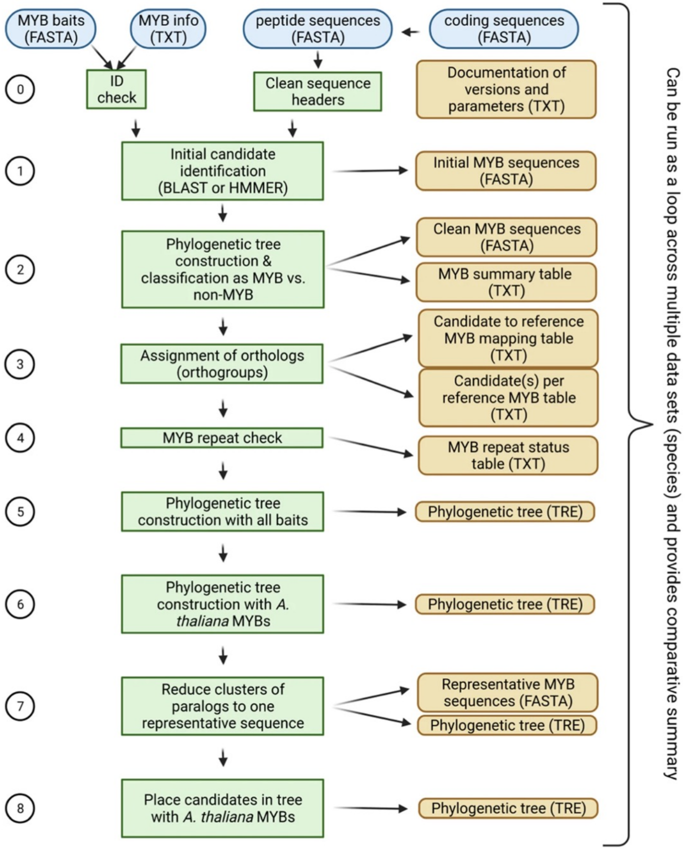


Large data sets
Identification of representatives

Pucker, 2022: 10.1186/s12864-022-08452-5

MYB_annotator

- Identification and functional annotation of MYB transcription factor family
- Family-specific characteristics: MYB repeat check
- Identification of candidates with lost domain
  → Loss of function



MYB baits (FASTA)
MYB info (TXT)
peptide sequences (FASTA)
coding sequences (FASTA)

0
ID check
Clean sequence headers
Documentation of versions and parameters (TXT)

1
Initial candidate identification (BLAST or HMMER)
Initial MYB sequences (FASTA)

2
Phylogenetic tree construction & classification as MYB vs. non-MYB
Clean MYB sequences (FASTA)
MYB summary table (TXT)

3
Assignment of orthologs (orthogroups)
Candidate to reference MYB mapping table (TXT)
Candidate(s) per reference MYB table (TXT)

4
MYB repeat check
MYB repeat status table (TXT)

5
Phylogenetic tree construction with all baits
Phylogenetic tree (TRE)

6
Phylogenetic tree construction with *A. thaliana* MYBs
Phylogenetic tree (TRE)

7
Reduce clusters of paralogs to one representative sequence
Representative MYB sequences (FASTA)
Phylogenetic tree (TRE)

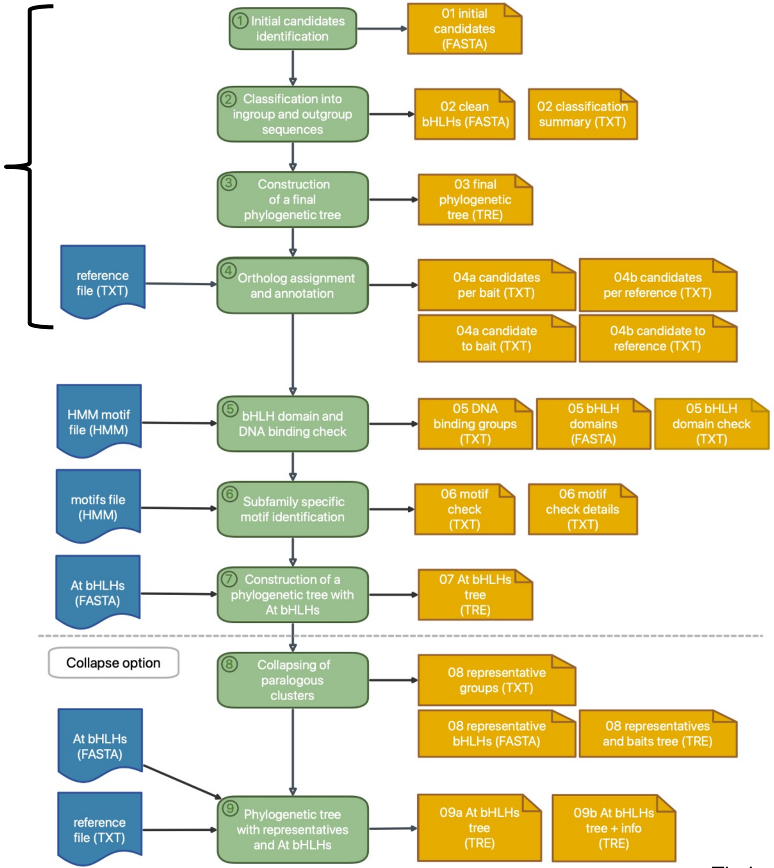
8
Place candidates in tree with *A. thaliana* MYBs
Phylogenetic tree (TRE)

Can be run as a loop across multiple data sets (species) and provides comparative summary

Pucker, 2022: 10.1186/s12864-022-08452-5

bHLH_annotator

Identification of ortholog candidates



① Initial candidates identification

01 initial candidates (FASTA)

② Classification into ingroup and outgroup sequences

02 clean bHLHs (FASTA)

02 classification summary (TXT)

③ Construction of a final phylogenetic tree

03 final phylogenetic tree (TRE)

reference file (TXT)

④ Ortholog assignment and annotation

04a candidates per bait (TXT)

04b candidates per reference (TXT)

04a candidate to bait (TXT)

04b candidate to reference (TXT)

HMM motif file (HMM)

⑤ bHLH domain and DNA binding check

05 DNA binding groups (TXT)

05 bHLH domains (FASTA)

05 bHLH domain check (TXT)

motifs file (HMM)

⑥ Subfamily specific motif identification

06 motif check (TXT)

06 motif check details (TXT)

At bHLHs (FASTA)

⑦ Construction of a phylogenetic tree with At bHLHs

07 At bHLHs tree (TRE)

Collapse option

⑧ Collapsing of paralogous clusters

08 representative groups (TXT)

08 representative bHLHs (FASTA)

08 representatives and baits tree (TRE)

At bHLHs (FASTA)

reference file (TXT)

⑨ Phylogenetic tree with representatives and At bHLHs

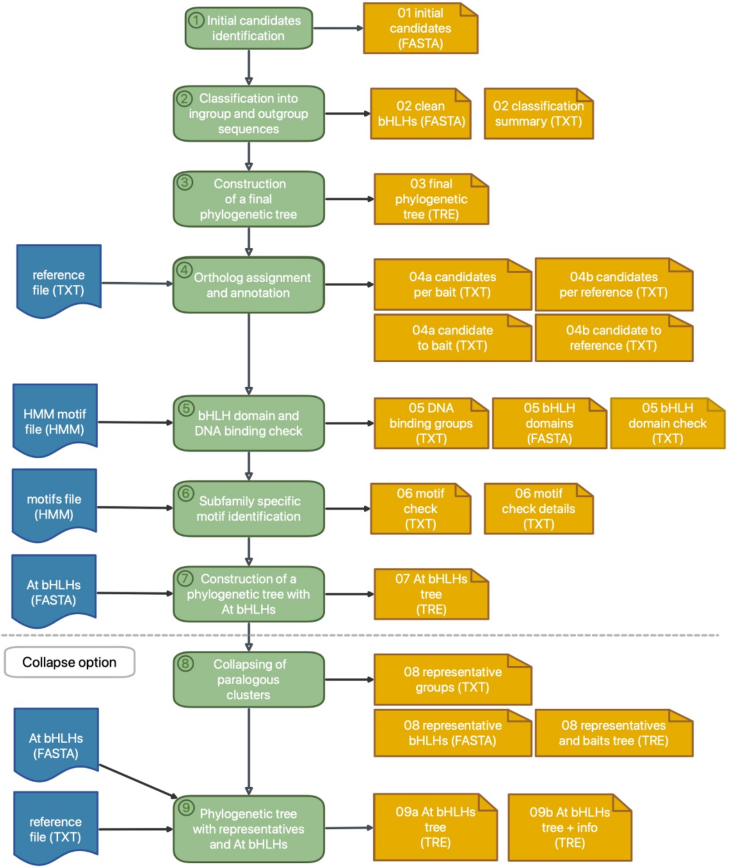
09a At bHLHs tree (TRE)

09b At bHLHs tree + info (TRE)

Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

bHLH_annotator

Subfamily specific motifs & DNA binding check



① Initial candidates identification → 01 initial candidates (FASTA)

② Classification into ingroup and outgroup sequences → 02 clean bHLHs (FASTA), 02 classification summary (TXT)

③ Construction of a final phylogenetic tree → 03 final phylogenetic tree (TRE)

reference file (TXT) → ④ Ortholog assignment and annotation → 04a candidates per bait (TXT), 04b candidates per reference (TXT), 04a candidate to bait (TXT), 04b candidate to reference (TXT)

HMM motif file (HMM) → ⑤ bHLH domain and DNA binding check → 05 DNA binding groups (TXT), 05 bHLH domains (FASTA), 05 bHLH domain check (TXT)

motifs file (HMM) → ⑥ Subfamily specific motif identification → 06 motif check (TXT), 06 motif check details (TXT)

At bHLHs (FASTA) → ⑦ Construction of a phylogenetic tree with At bHLHs → 07 At bHLHs tree (TRE)

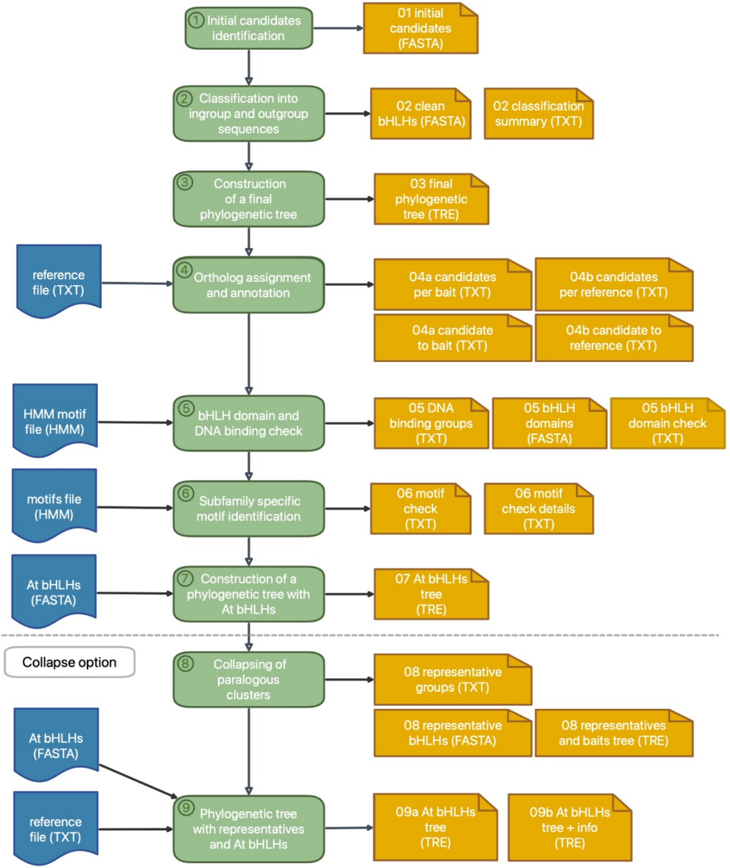
Collapse option

⑧ Collapsing of paralogous clusters → 08 representative groups (TXT), 08 representative bHLHs (FASTA), 08 representatives and baits tree (TRE)

At bHLHs (FASTA), reference file (TXT) → ⑨ Phylogenetic tree with representatives and At bHLHs → 09a At bHLHs tree (TRE), 09b At bHLHs tree + info (TRE)

Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

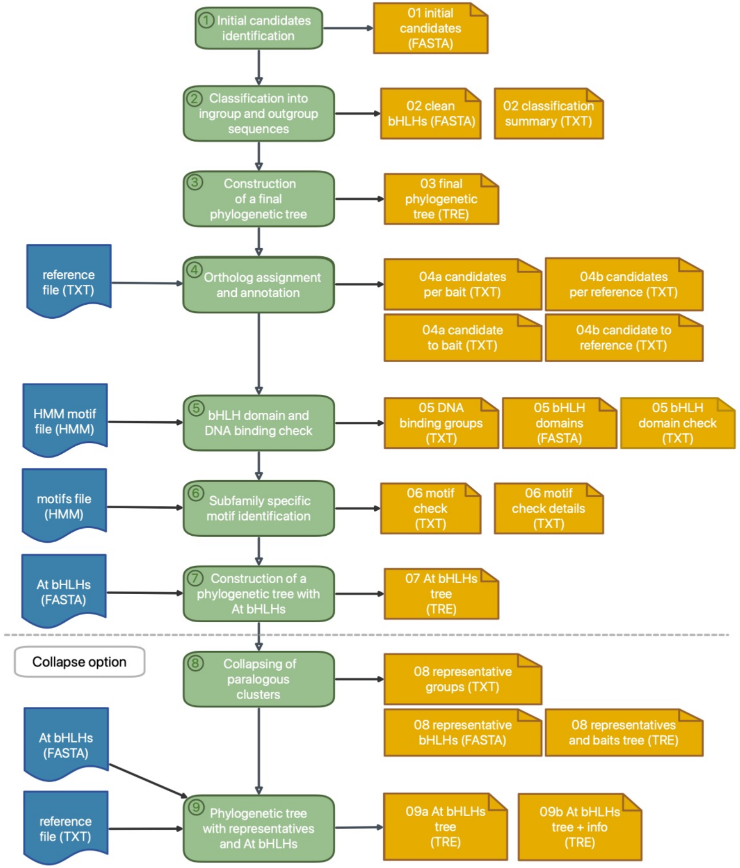
bHLH_annotator



Phylogenetic tree construction

Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

bHLH_annotator



Large data sets
Identification of representatives

Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

bHLH_annotator

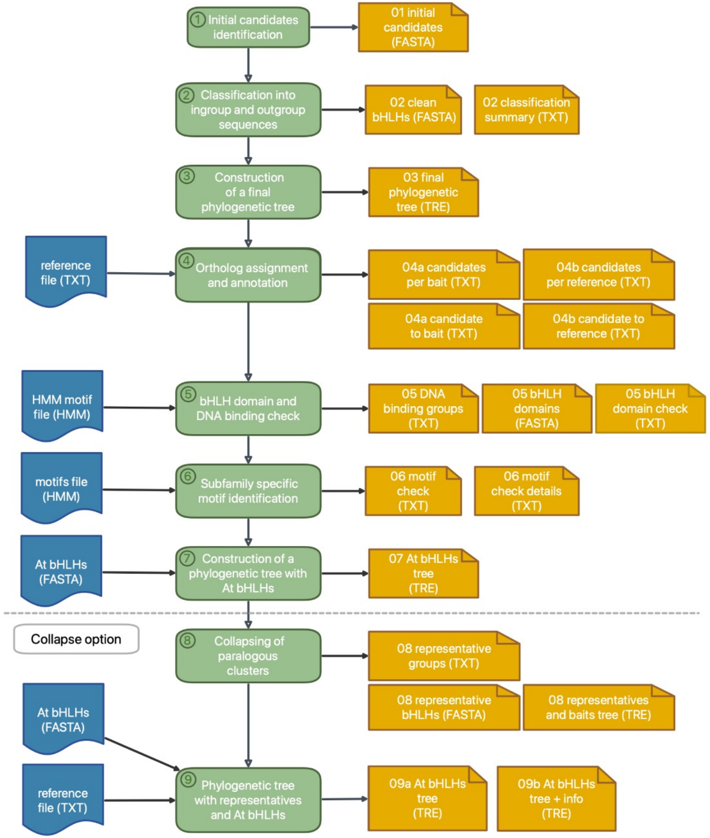
- bHLH transcription factor gene family

- Family-specific characteristics:

  → Subfamily motifs

  → Prediction of DNA binding group

- Generalized concept: Annotation of other (transcription factor) gene families



① Initial candidates identification
01 initial candidates (FASTA)

② Classification into ingroup and outgroup sequences
02 clean bHLHs (FASTA)
02 classification summary (TXT)

③ Construction of a final phylogenetic tree
03 final phylogenetic tree (TRE)

reference file (TXT)
④ Ortholog assignment and annotation
04a candidates per bait (TXT)
04b candidates per reference (TXT)
04a candidate to bait (TXT)
04b candidate to reference (TXT)

HMM motif file (HMM)
⑤ bHLH domain and DNA binding check
05 DNA binding groups (TXT)
05 bHLH domains (FASTA)
05 bHLH domain check (TXT)

motifs file (HMM)
⑥ Subfamily specific motif identification
06 motif check (TXT)
06 motif check details (TXT)

At bHLHs (FASTA)
⑦ Construction of a phylogenetic tree with At bHLHs
07 At bHLHs tree (TRE)

Collapse option
⑧ Collapsing of paralogous clusters
08 representative groups (TXT)
08 representative bHLHs (FASTA)
08 representatives and baits tree (TRE)

At bHLHs (FASTA)
reference file (TXT)
⑨ Phylogenetic tree with representatives and At bHLHs
09a At bHLHs tree (TRE)
09b At bHLHs tree + info (TRE)

Thoben & Pucker, 2023: 10.1101/2023.05.02.539087

Automatic knowledge transfer based on orthology

KIPEs

MYB_annotator

bHLH_annotator

→ Adaption for own research purposes

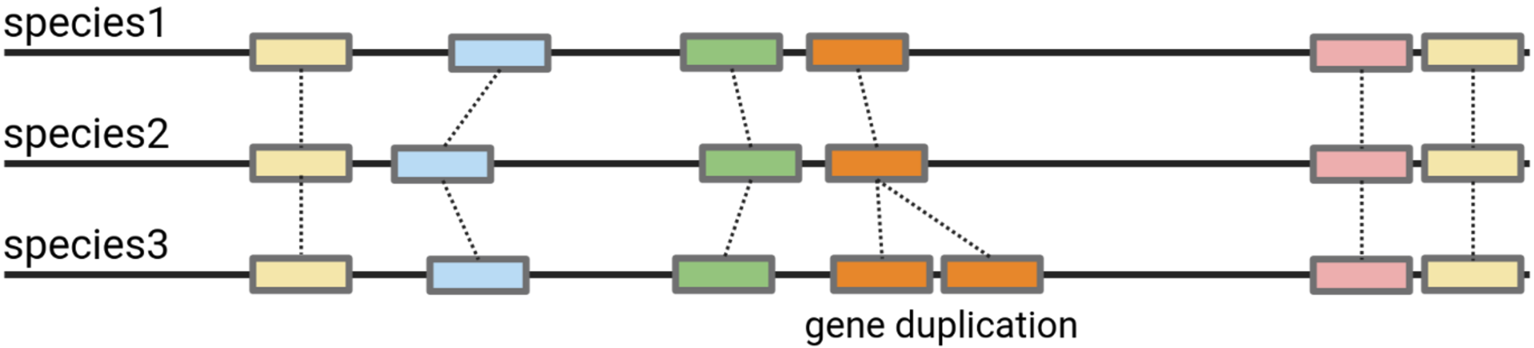
Bait collection

Reference sequences

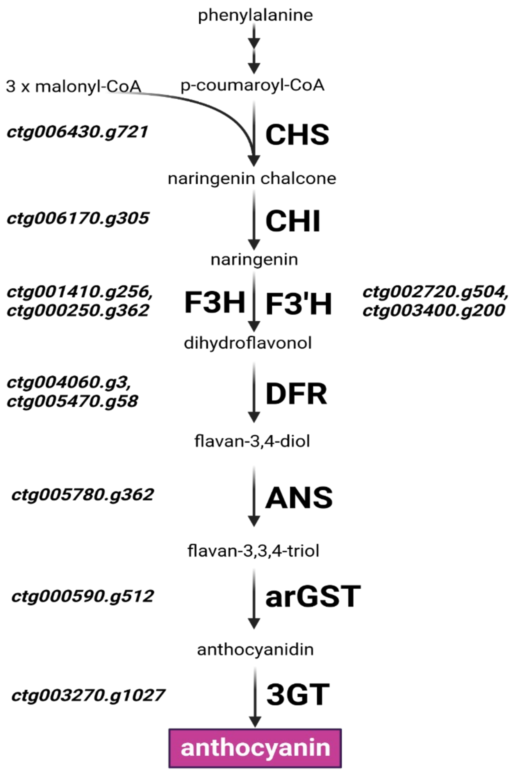
Additional analysis

→ Evaluation of findings

Outlook: Integration of synteny analysis



Application: Anthocyanin biosynthesis in *Digitalis purpurea*



phenylalanine

3 x malonyl-CoA

p-coumaroyl-CoA

ctg006430.g721 **CHS**

naringenin chalcone

ctg006170.g305 **CHI**

naringenin

ctg001410.g256, ctg000250.g362 **F3H** | **F3'H** *ctg002720.g504, ctg003400.g200*

dihydroflavonol

ctg004060.g3, ctg005470.g58 **DFR**

flavan-3,4-diol

ctg005780.g362 **ANS**

flavan-3,3,4-triol

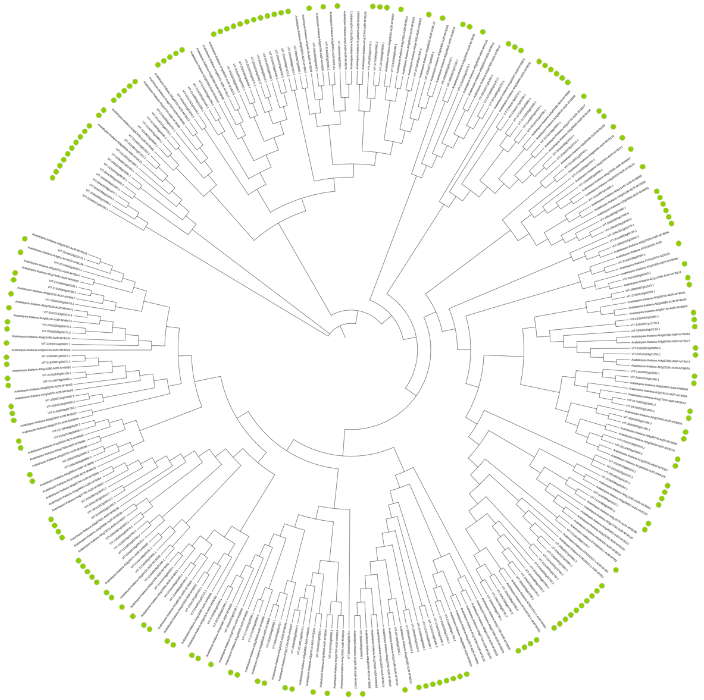
ctg000590.g512 **arGST**

anthocyaninidin

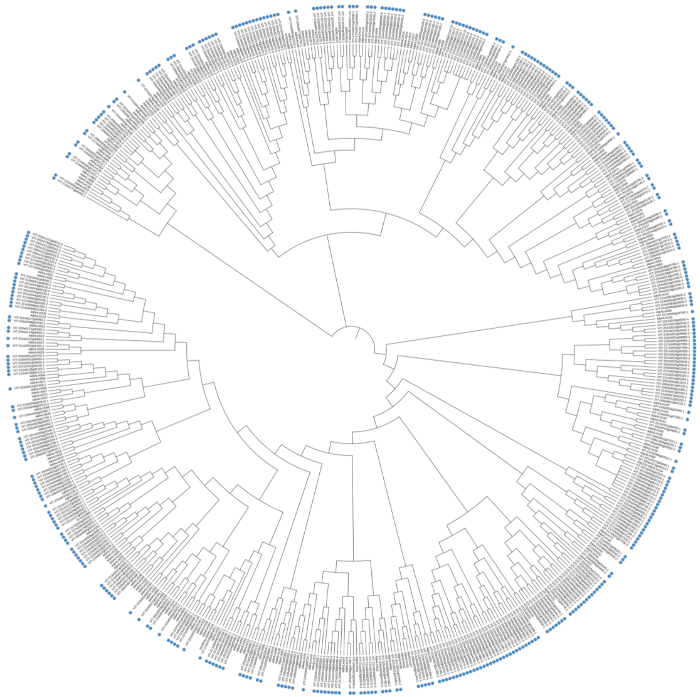
ctg003270.g1027 **3GT**

anthocyanin

Application: Identification of MYBs and bHLHs in *Vitis vinifera*



MYBs



bHLHs

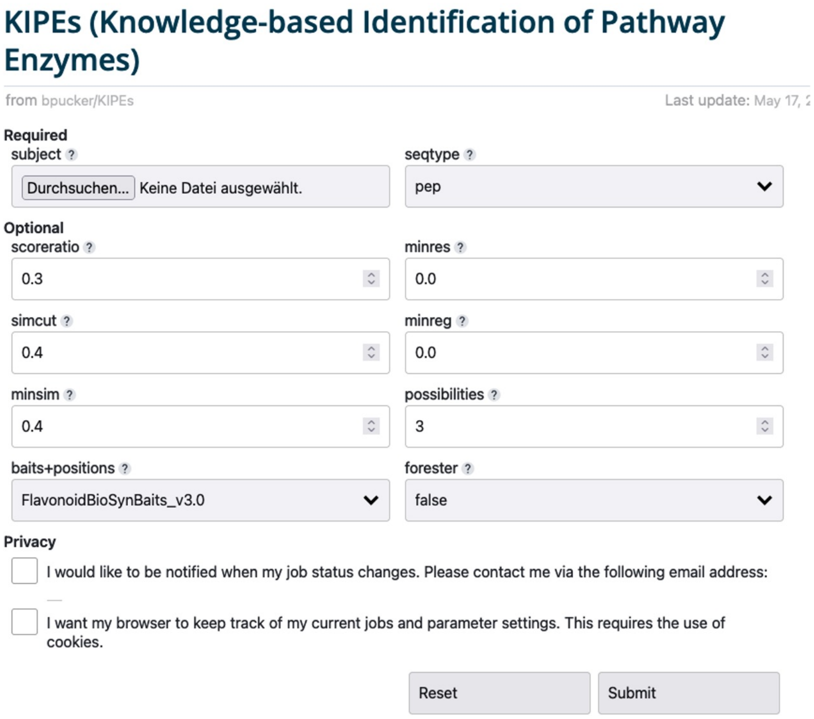
Availability

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

BioInfToolServer

GitHub

https://github.com/bpucker/KIPEs
https://github.com/bpucker/MYB_annotator
https://github.com/bpucker/bHLH_annotator



KIPEs (Knowledge-based Identification of Pathway Enzymes)

from bpucker/KIPEs

Last update: May 17,

Required

subject ?	seqtype ?
Durchsuchen... Keine Datei ausgewählt.	pep

Optional

scoreratio ?	minres ?
0.3	0.0
simcut ?	minreg ?
0.4	0.0
minsim ?	possibilities ?
0.4	3
baits+positions ?	forester ?
FlavonoidBioSynBaits_v3.0	false

Privacy

☐ I would like to be notified when my job status changes. Please contact me via the following email address:

☐ I want my browser to keep track of my current jobs and parameter settings. This requires the use of cookies.

https://www.tu-braunschweig.de/en/ifp/pbb/tools

Acknowledgements

- Plant Biotechnology and Bioinformatics (TU Braunschweig):
  Boas Pucker
  Nancy Choudhary
- Genetics and Genomics of Plants (Bielefeld University):
  Hanna Schilbert
- Genome Informatics (Bielefeld University):
  Andreas Rempel


Plant Biotechnology and Bioinformatics (July 2023)


Hanna Marie Schilbert


Andreas Rempel


Nancy Choudhary


Boas Pucker

Some figures created with BioRender.com