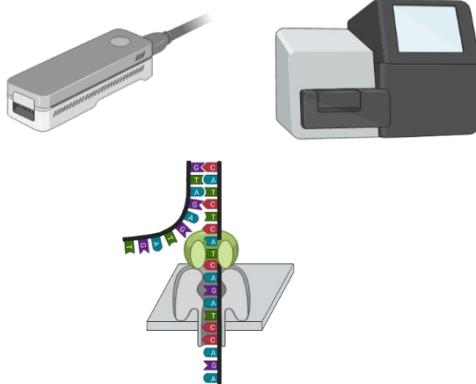
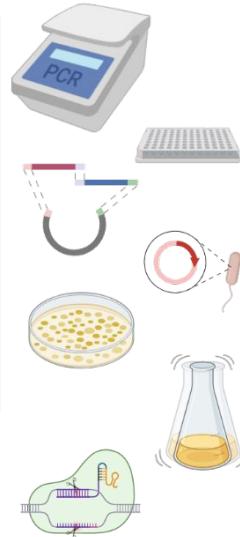
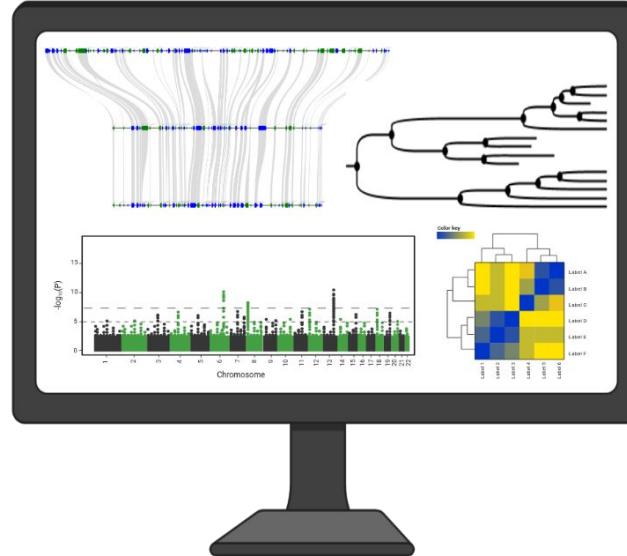




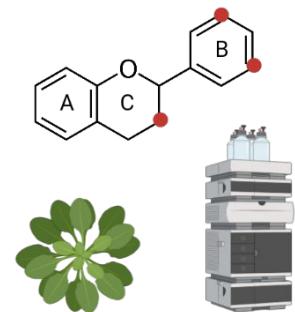
Technische
Universität
Braunschweig



Plant Biotechnology
and Bioinformatics



species biosynthesis proteins analysis different conditions
biosynthesis biosynthesis variants H293-MYB
within genes functional variants belarum
duplex site data divergently Col loco variant
sequencer ICGM single reference multiple annotations non-canonical
sites synthesis pre-mRNA annotation level identified
plants plants genes plant structure evolutionary
pigments model genome synthesis pathway
Keyvanda genes key against canonical loc accession
genomes across systems biology long insertion
flavonoid conservation sequencing evolution
gene read transcription synthetic MYB introns residues RNA-Seq



Utilization of RNA-seq data

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

Availability of slides

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

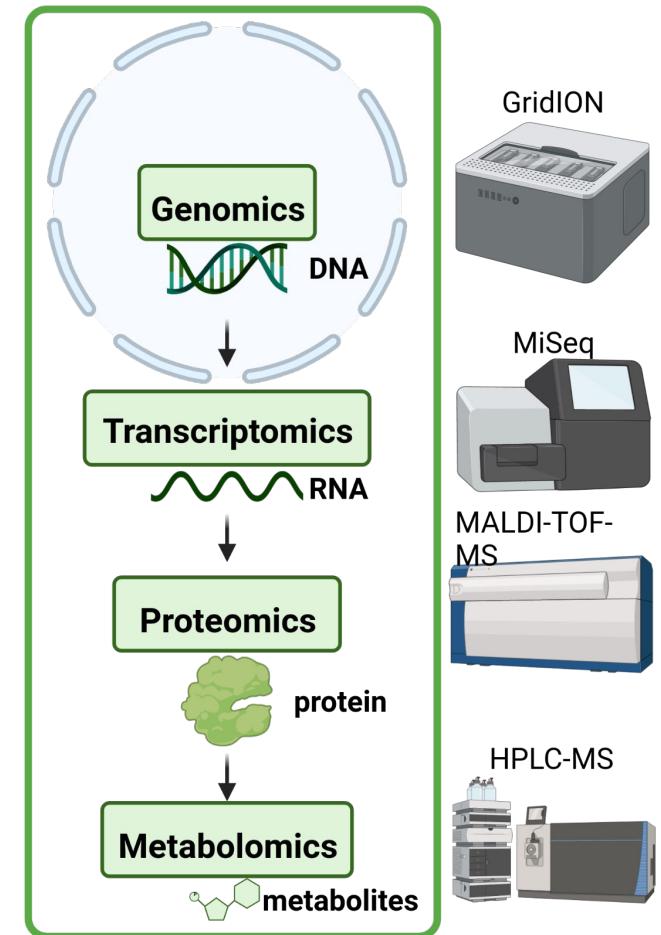
My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

Overview

- Gene prediction
- Coexpression and coexpression networks
- Neo- and subfunctionalization
- Mapping-by-sequencing
- Cross-species transcriptomics

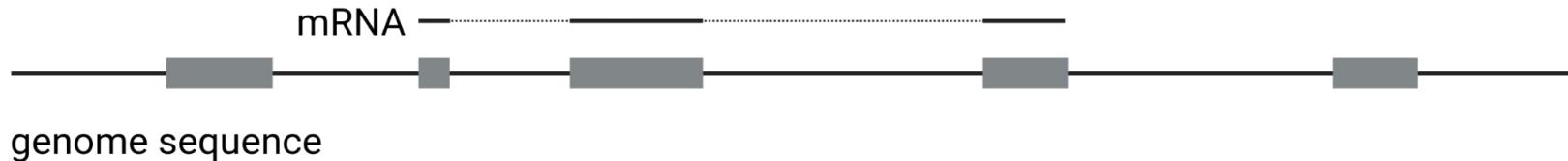
Transcriptomics in context

- Transcriptomics is most powerful in the context of other omics data sets
- Genomics and transcriptomics benefit from molecules with uniform properties
- Proteomics and metabolomics involve more chemically diverse molecules thus comprehensive quantification is not feasible



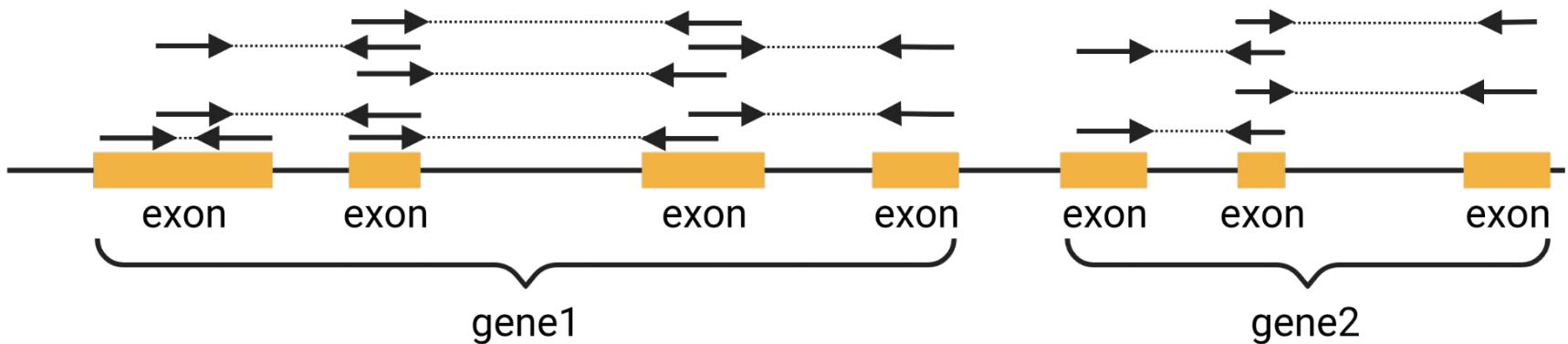
De novo transcriptome assembly - summary

- Cost-effective way to represent the most important parts of a genome
- Computational costs are substantially lower than for a genome sequence assembly
- Reveals information about ‘active’ genes
- Can provide hints for the annotation of a genome sequence



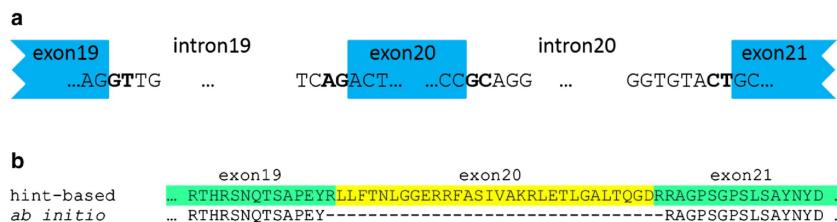
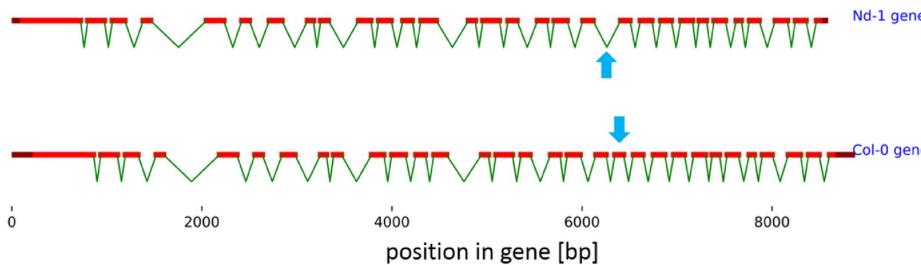
Hints for gene prediction

- RNA-seq reads can inform about positions of transcribed parts in a genome genes
- Positions of introns can be inferred from alignment gaps
- Connection of exons can be inferred from split read alignments
- Stranded RNA-seq data sets are best for gene prediction



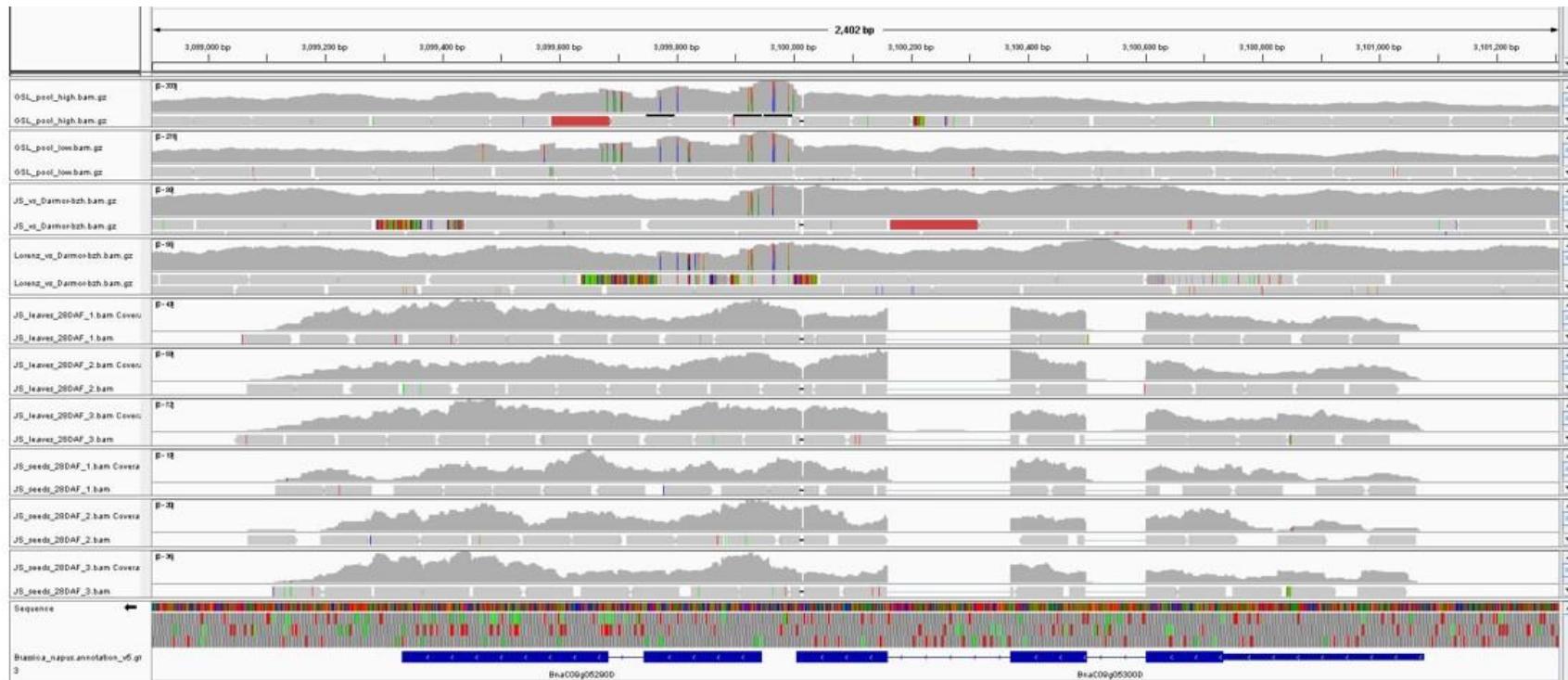
Annotation of non-canonical splice sites requires hints

- Canonical splice sites: GT...AG
- Major non-canonical splice sites: GC...AG, AT...AC
- Minor non-canonical splice sites: NN...NN



<https://doi.org/10.1186/s13104-017-2985-y>

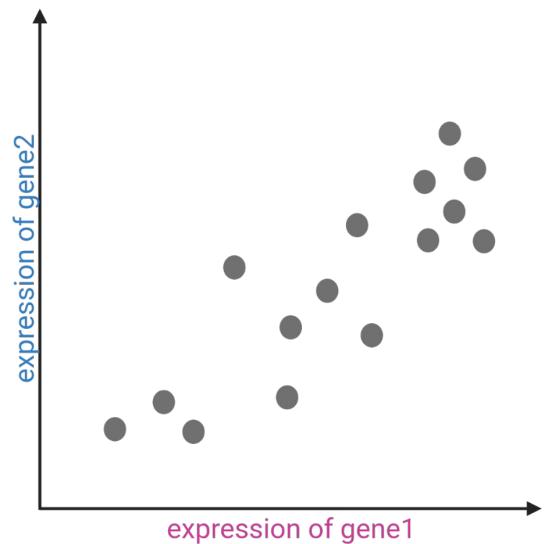
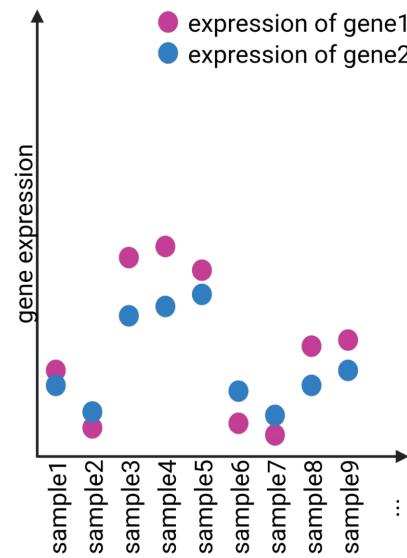
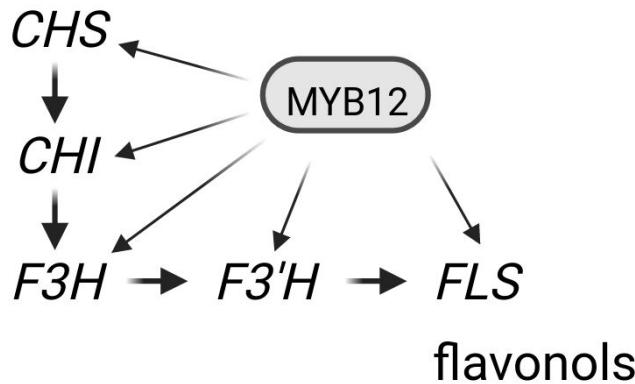
Finding non-functional alleles



<https://doi.org/10.3390/genes13071131>

Co-expression analysis

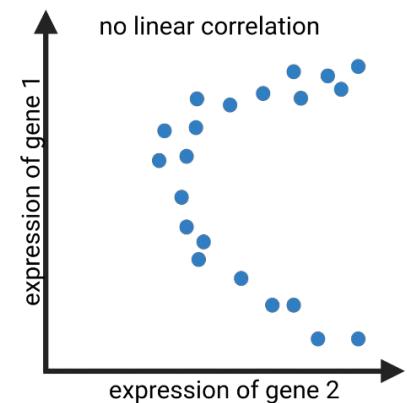
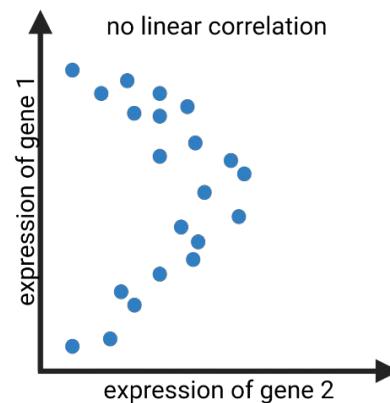
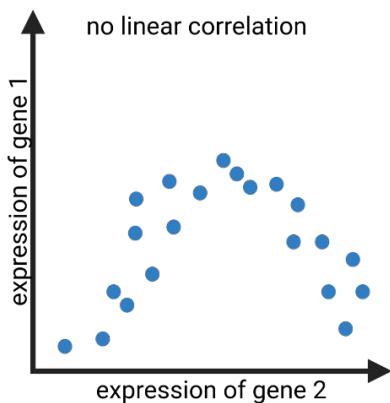
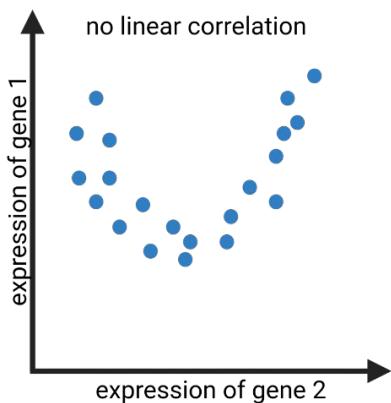
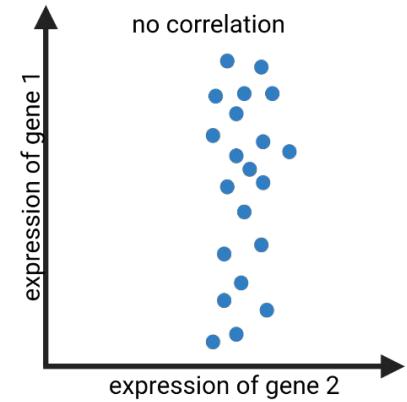
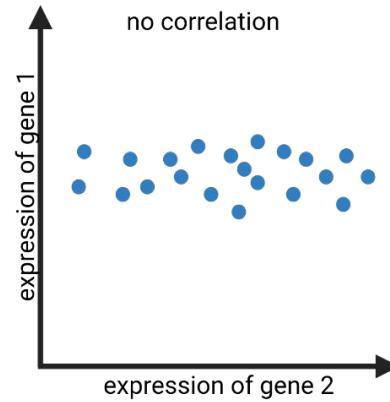
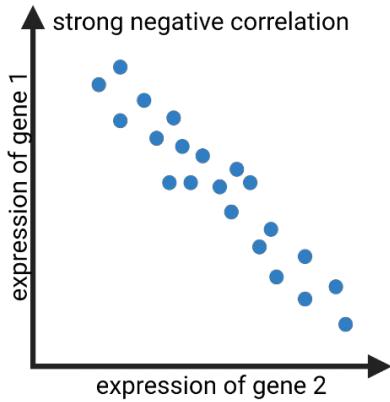
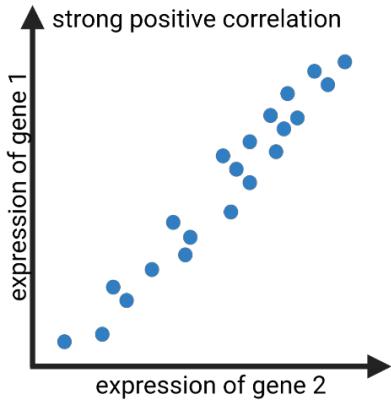
- Genes of the same biosynthesis pathway show similar expression patterns
- Co-expression is a very powerful indication for connection between genes



Correlation vs. causation

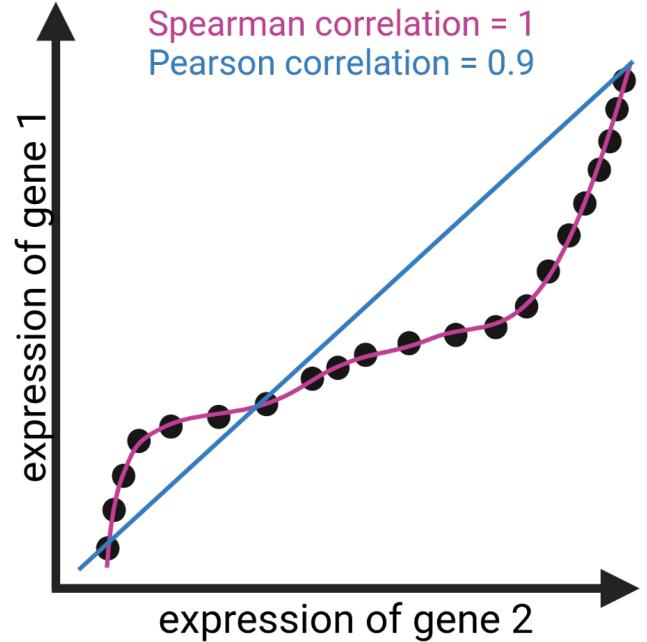
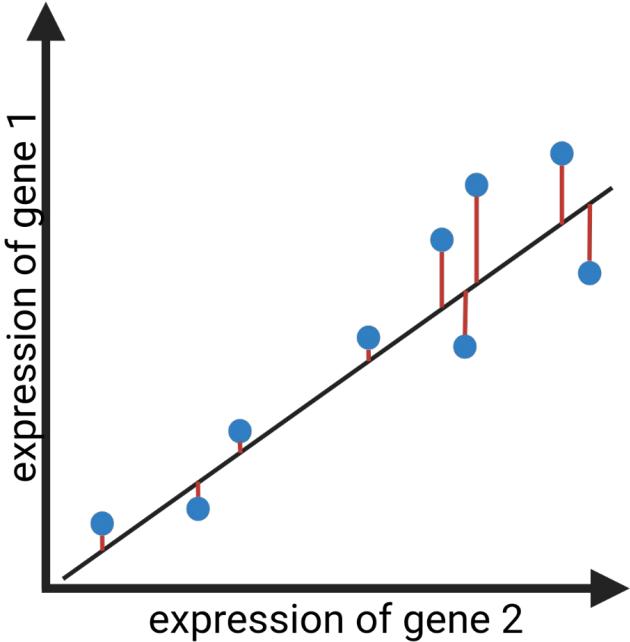
- Correlation does not allow conclusions about the direction i.e. cause and consequence
- Guilt-by-association is frequently used to find genes in a pathway/network
- Correlation can be caused by a direct or indirect connection

Examples



Types of correlation coefficients

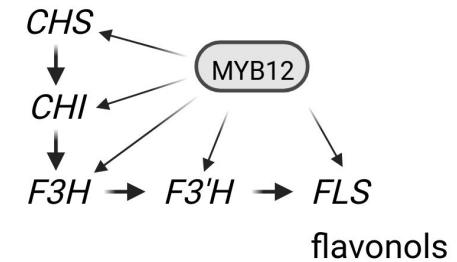
- Pearson is only suitable for linear correlation
- Spearman is more tolerant towards outliers and non-linear correlations



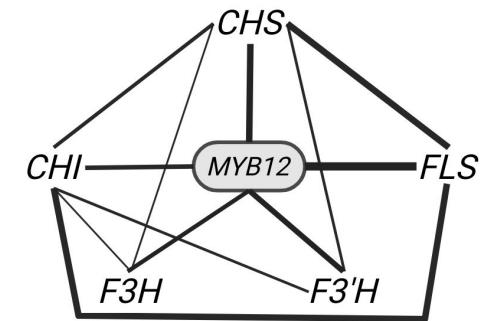
Coexpression network connection

- Calculation of pairwise correlation coefficients of gene expression
- Strength of network edges based on correlation coefficient
- Correlation coefficient is visualized by line width

MYB12	
genes	correlation
<i>CHS</i>	0.7
<i>CHI</i>	0.5
<i>F3H</i>	0.45
<i>F3'H</i>	0.6
<i>FLS</i>	0.9



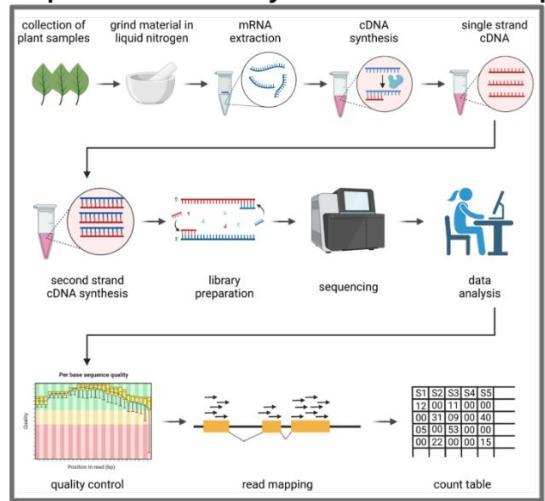
CHS	
genes	correlation
<i>CHI</i>	0.4
<i>F3H</i>	0.2
<i>F3'H</i>	0.3
<i>FLS</i>	0.7



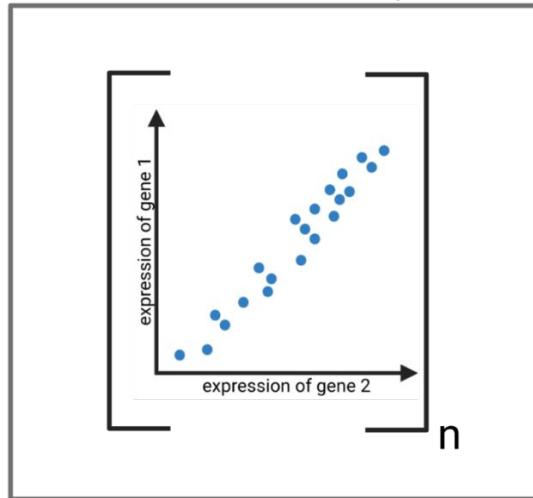
CHI	
genes	correlation
<i>F3H</i>	0.2
<i>F3'H</i>	0.3
<i>FLS</i>	0.7

Summary: co-expression workflow

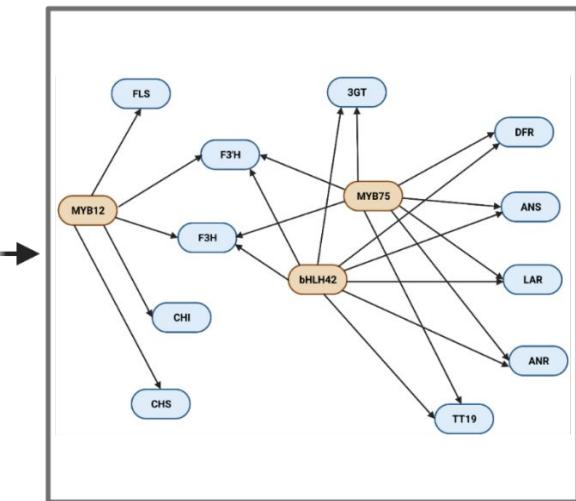
Expression analysis via RNA-Seq



Coexpression analysis

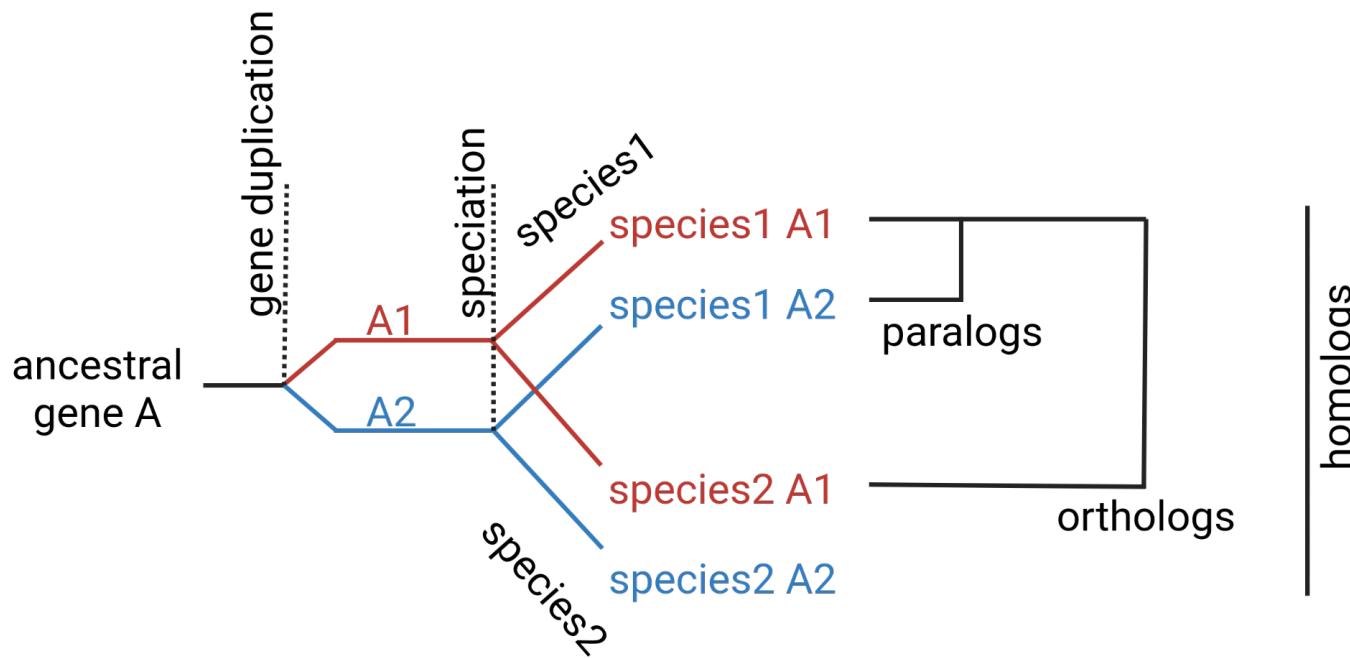


Network construction



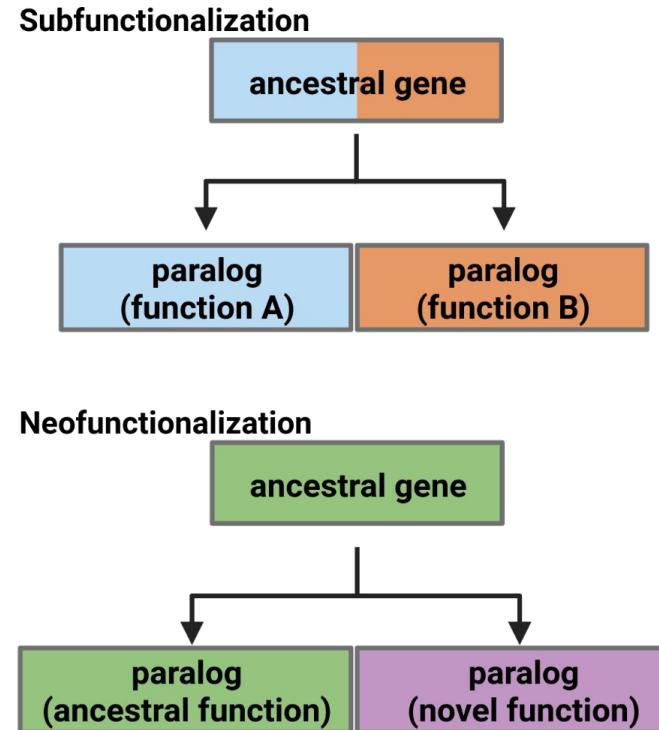
Phylogenetic relationships of genes

- Orthologs are same genes in different species
- Paralogs are gene copies in one species

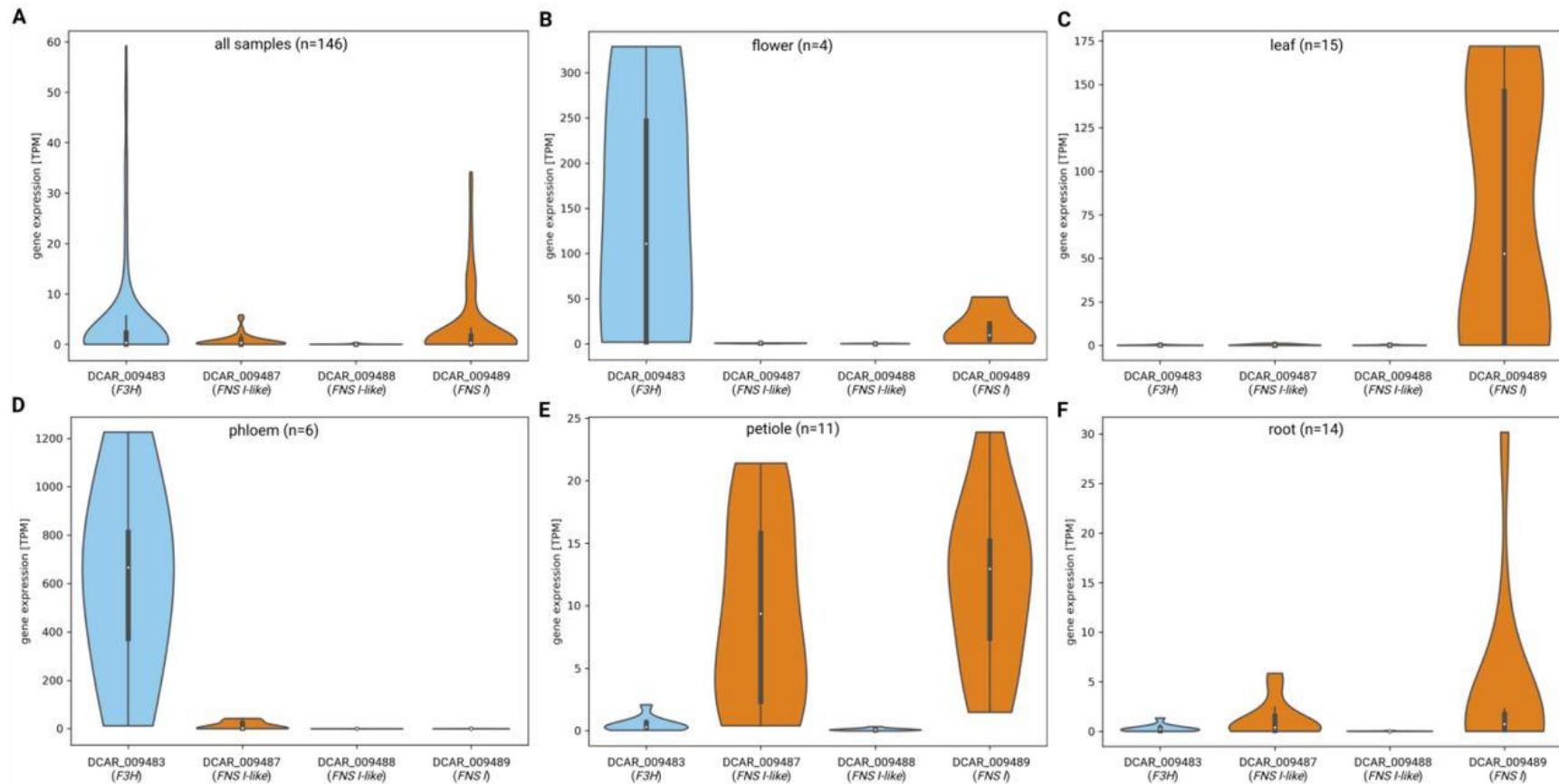


Sub/Neofunctionalization

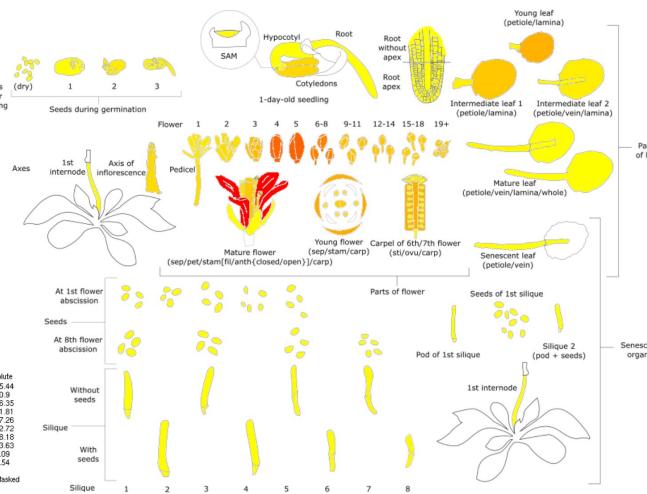
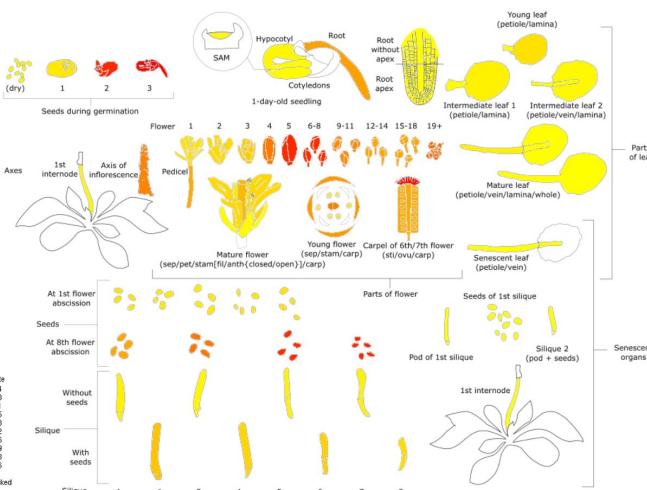
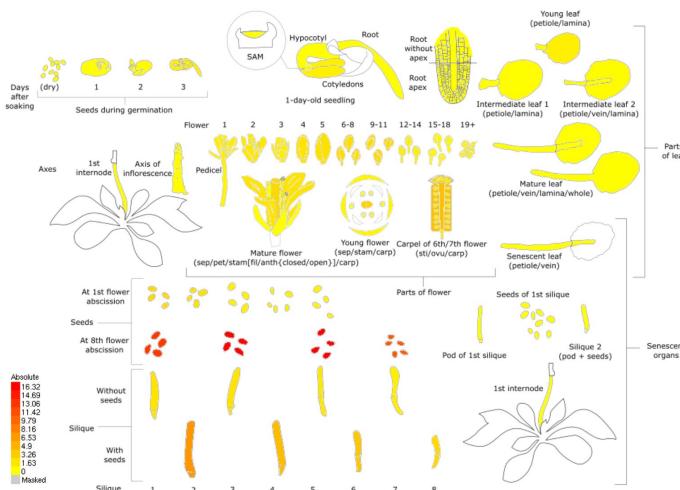
- Sub/neofunctionalization can be achieved through changes in gene expression following the duplication
- Gene copies can be restricted to specific tissues/conditions



Example: *FNS I* sub/neofunctionalization through expression

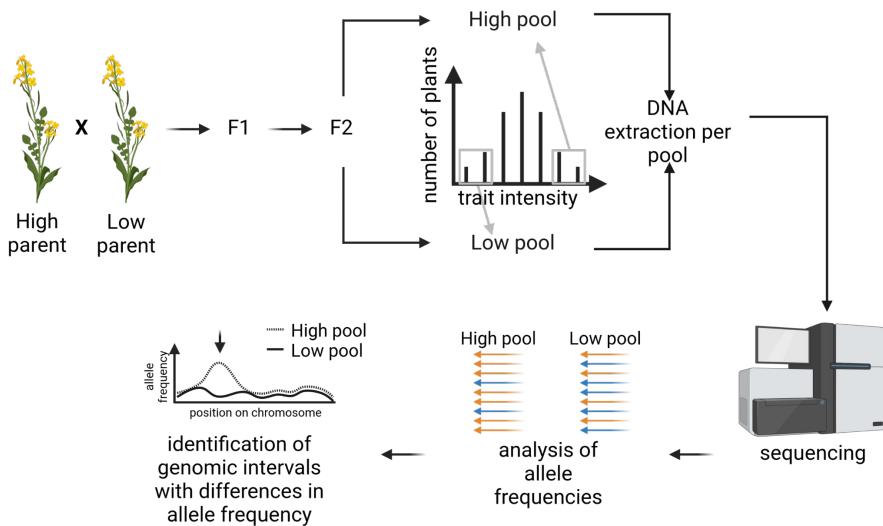
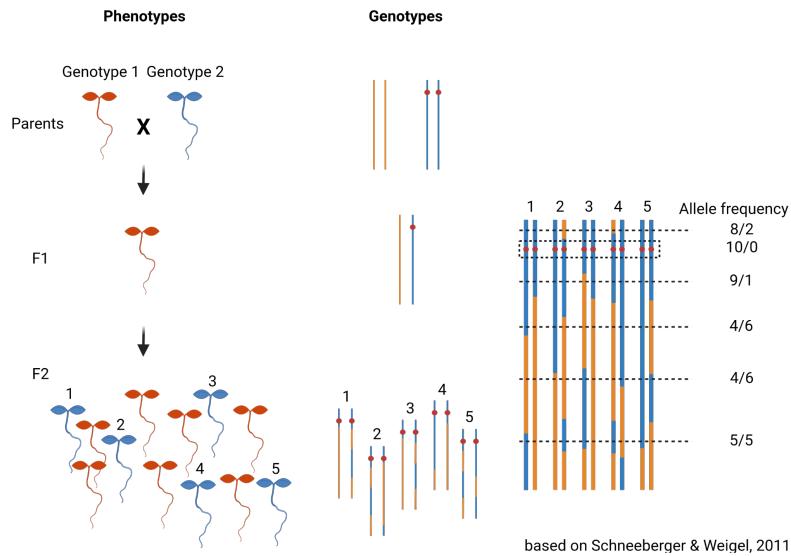


Example: MYB11/12/111



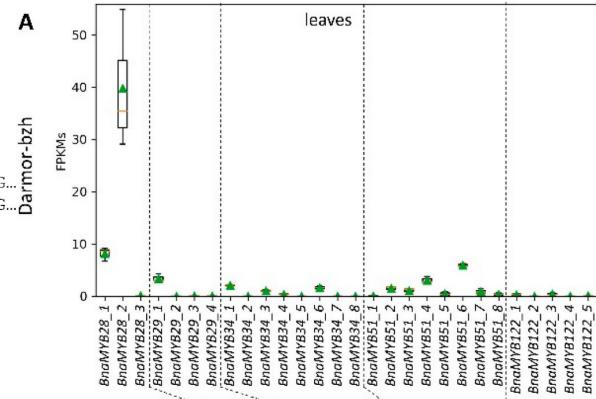
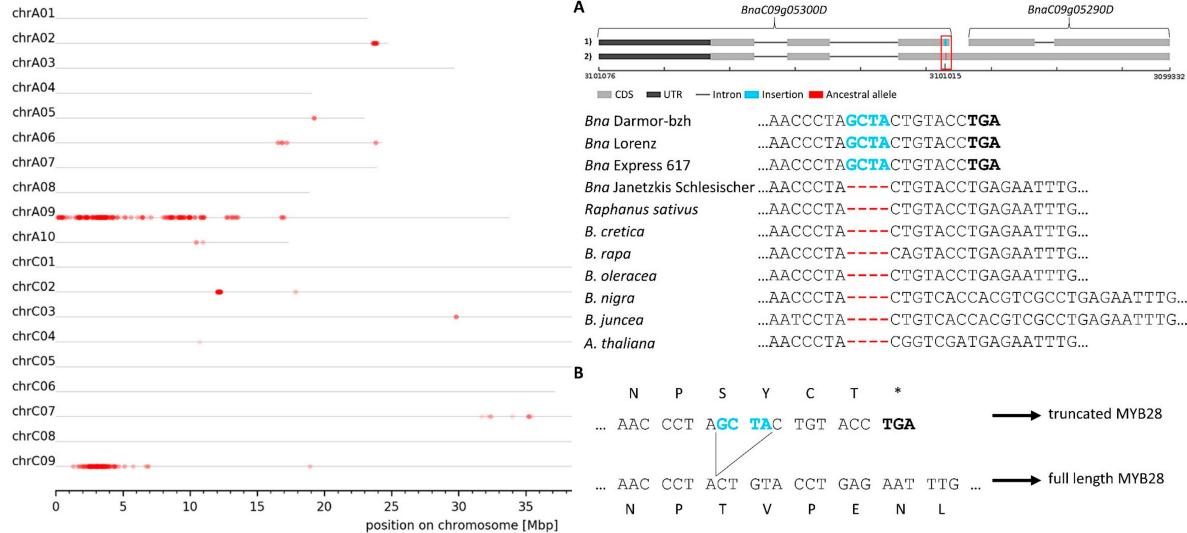
<https://bar.utoronto.ca/thalemine>

Mapping-by-sequencing



MBS support for candidate gene search

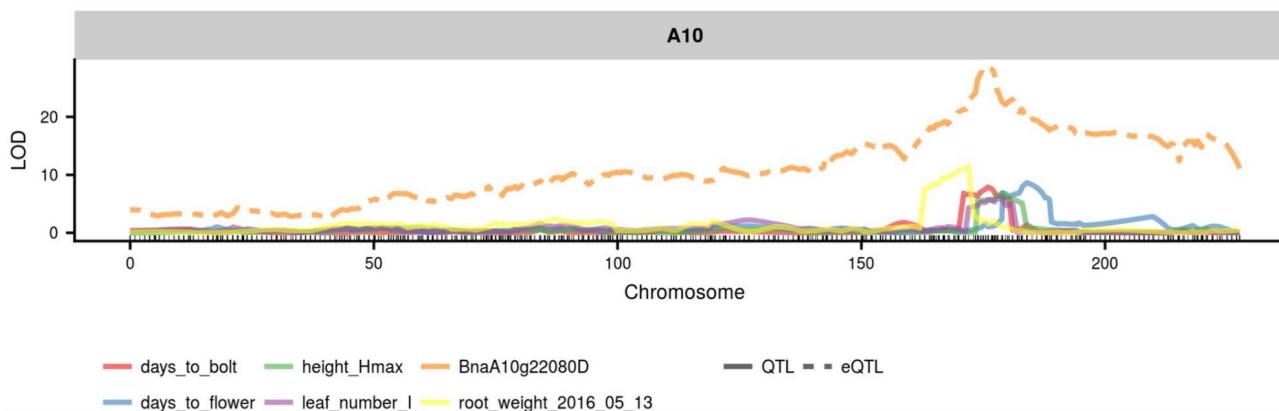
- Mapping-by-sequencing (MBS) reveals genomic region responsible for a trait
- Sequence variant effects can reveal candidate genes
- Identification of causal gene among a few hundred candidates benefits from expression



Schilbert & Pucker et al., 2022: <https://doi.org/10.3390/genes13071131>

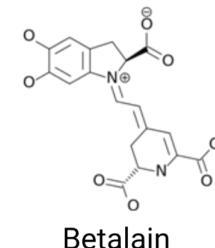
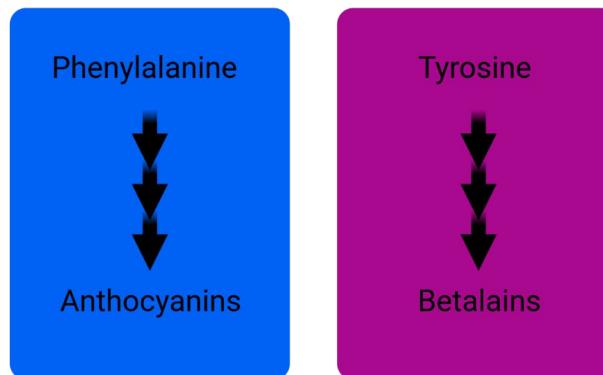
eQTL

- eQTL = expression quantitative trait locus
- Identification of a genomic region responsible for a trait based on gene expression



<https://doi.org/10.3389/fpls.2018.01632>

Functional redundancy of anthocyanins and betalains



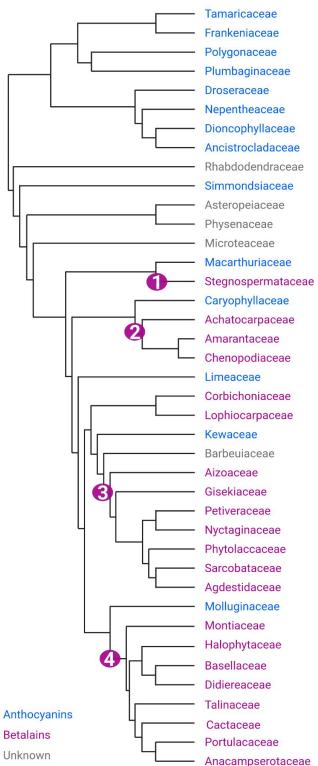
Anthocyanin color range



Betalain color range

Brockington et al., 2011

Complex pigment evolution in Caryophyllales



- At least four independent origins of betalain biosynthesis
- Mutual exclusion: anthocyanins and betalains were never observed in same (natural) plants
- Functional redundancy of both pigments

Timoneda *et al.*, 2019

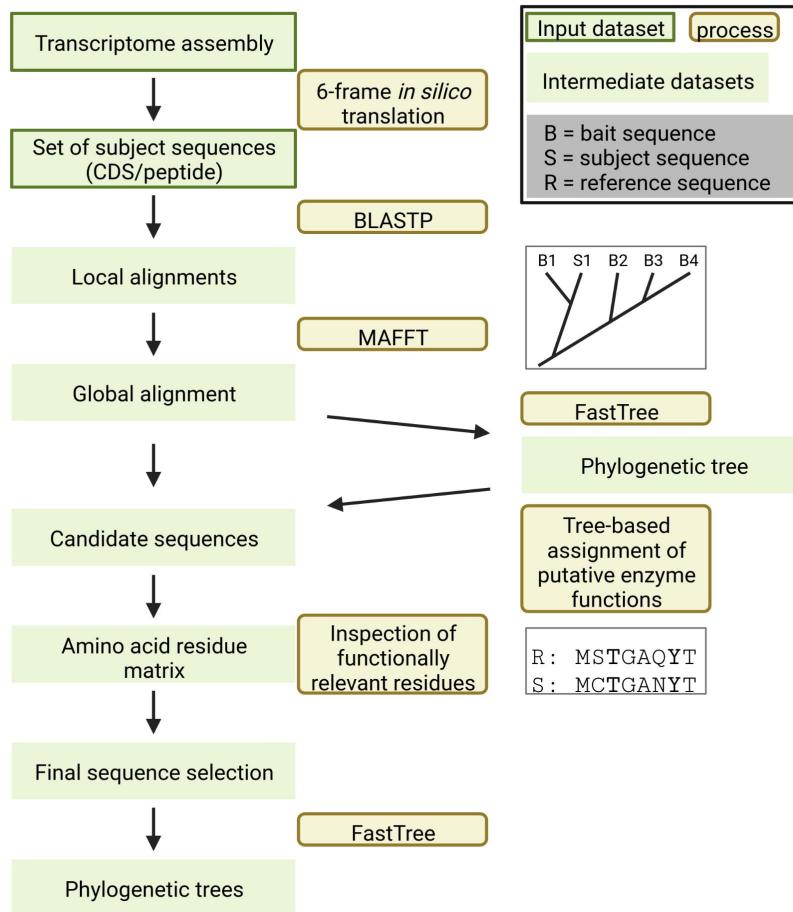
Sheehan *et al.*, 2020

[1] Dick Culbert [2] Stan Shebs [3] Emöke Denes

Hypotheses

- Lack of anthocyanin biosynthesis genes in the Caryophyllales?
- Low/no expression of anthocyanin biosynthesis genes in the Caryophyllales?

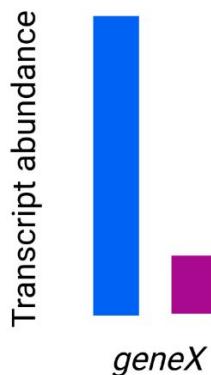
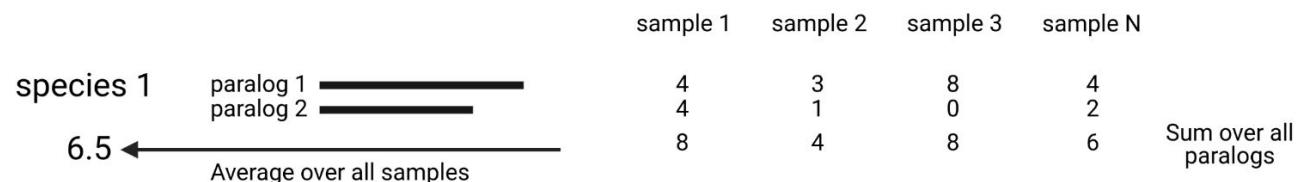
Finding anthocyanin biosynthesis genes



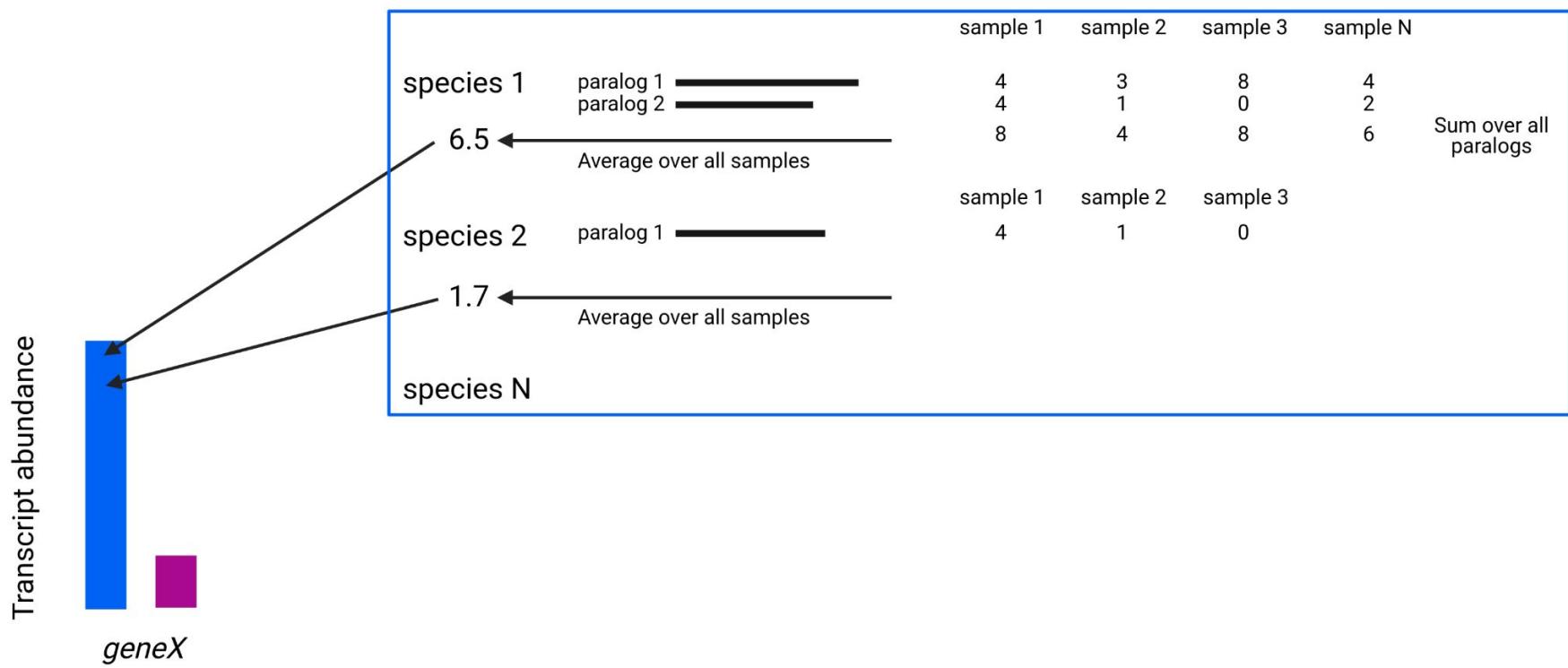
- Identification of sequences across all species in the Caryophyllales
- Knowledge-based Identification of Pathway Enzymes (KIPES) allows automatic identification of anthocyanin pathways genes
- >300 transcriptome assemblies are covering the families of the Caryophyllales
- Additional genome sequences

Pucker *et al.*, 2020
Rempel & Pucker, 2022

Cross-species transcriptomics

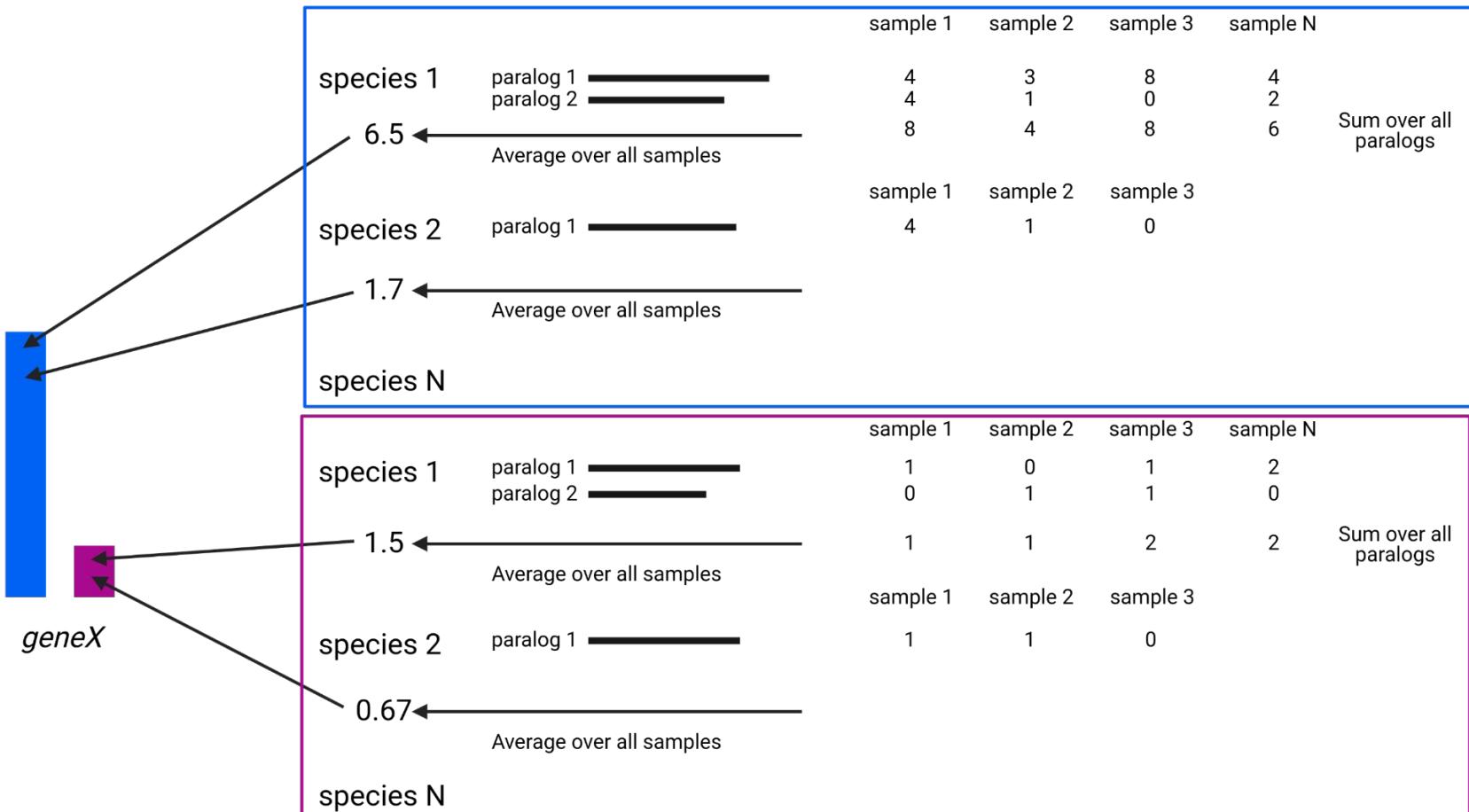


Cross-species transcriptomics

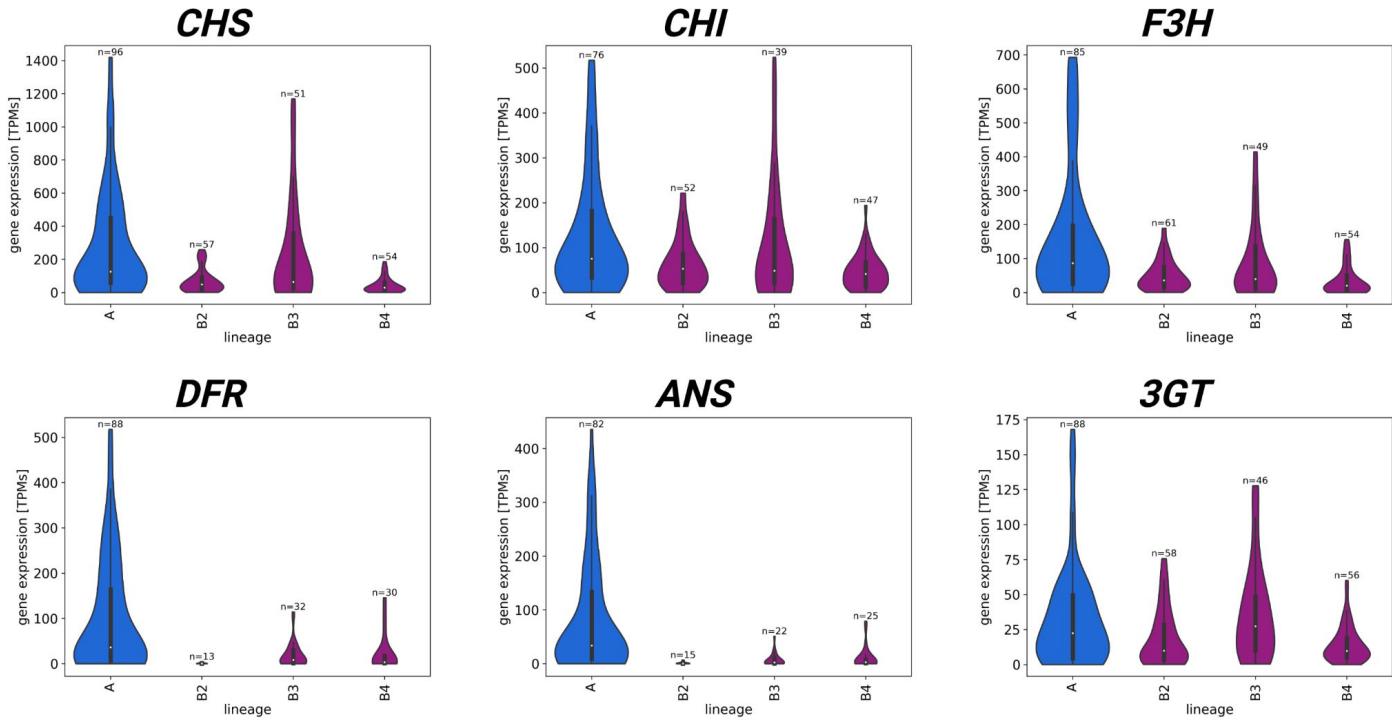


Cross-species transcriptomics

Transcript abundance



Results of cross-species transcriptomics



Time for questions!



Questions

1. How can RNA-seq support the gene prediction process?
2. Which methods can be applied to study pairwise co-expression?
3. What are the two evolutionary events that can follow gene duplication?
4. Which tool allows an inspection of tissue-specific gene expression?
5. What is an eQTL?