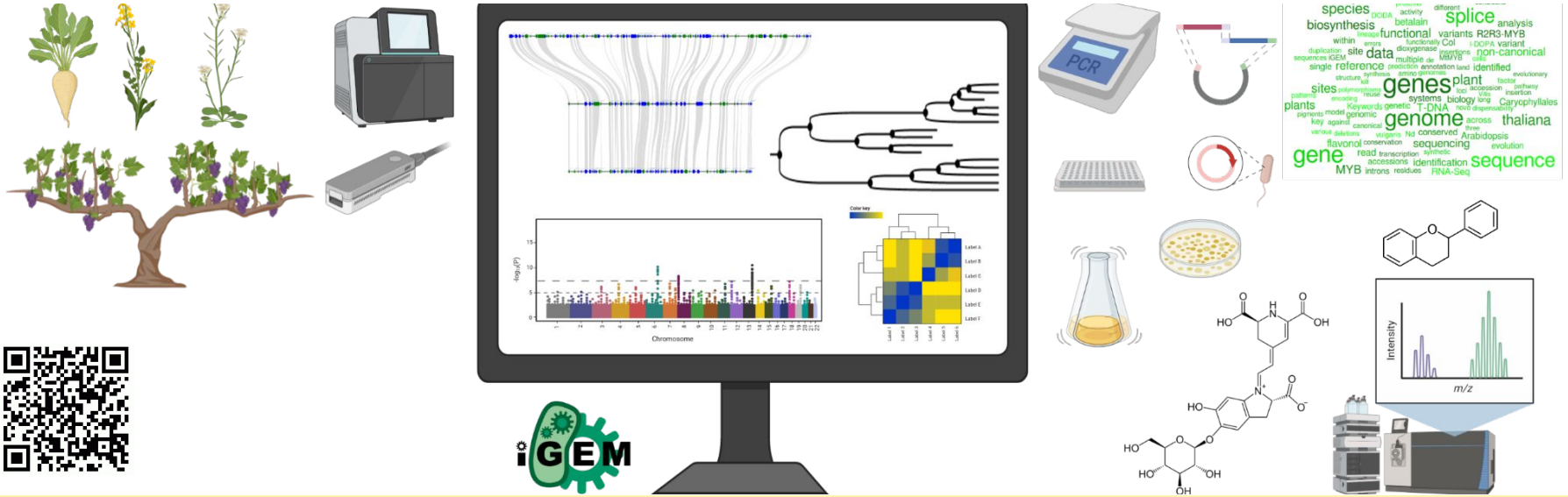
Technische Universität Braunschweig




# RNA-seq data reuse

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

# Data sources

- Gene Expression Omnibus (GEO)

## Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Gene Expression Omnibus

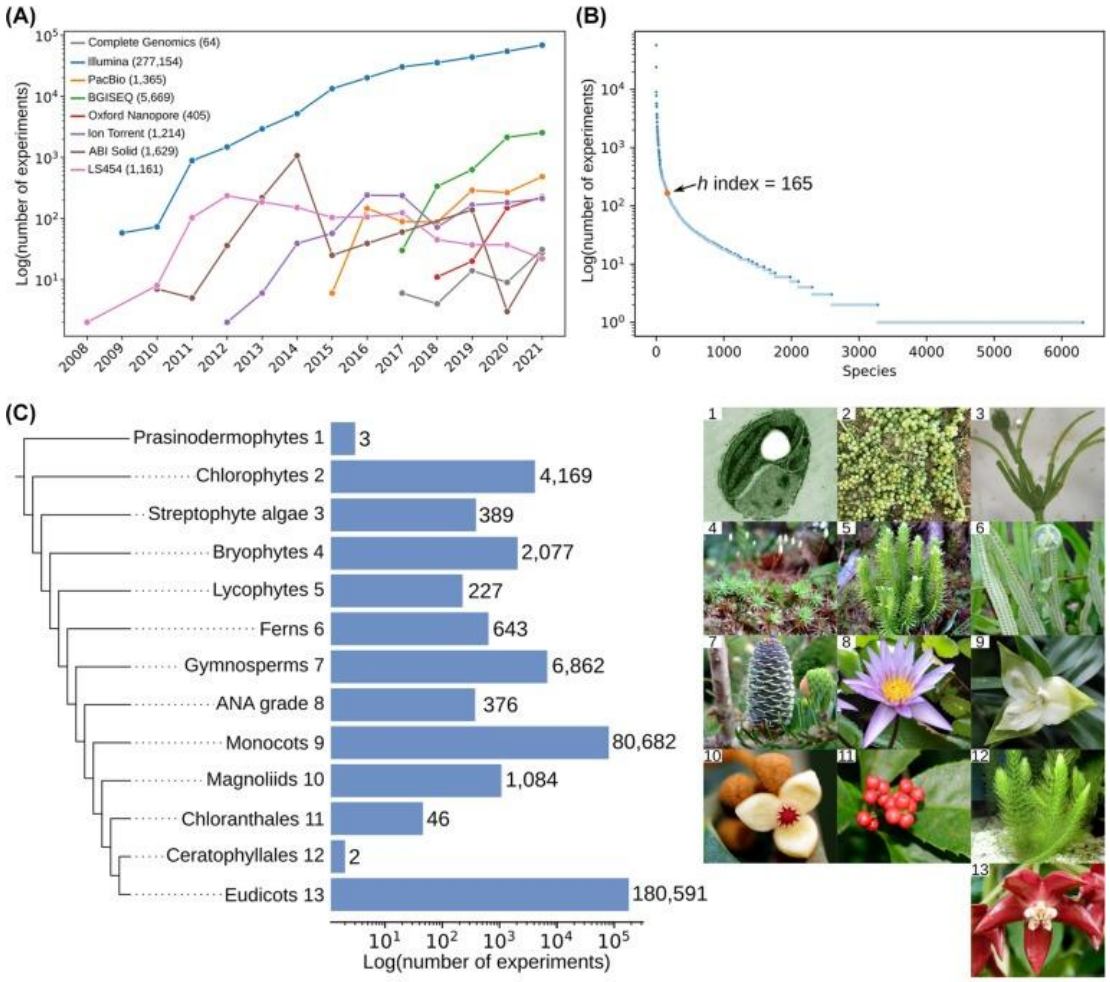
Keyword or GEO Accession Search

- Sequence Read Archive (SRA)

## SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

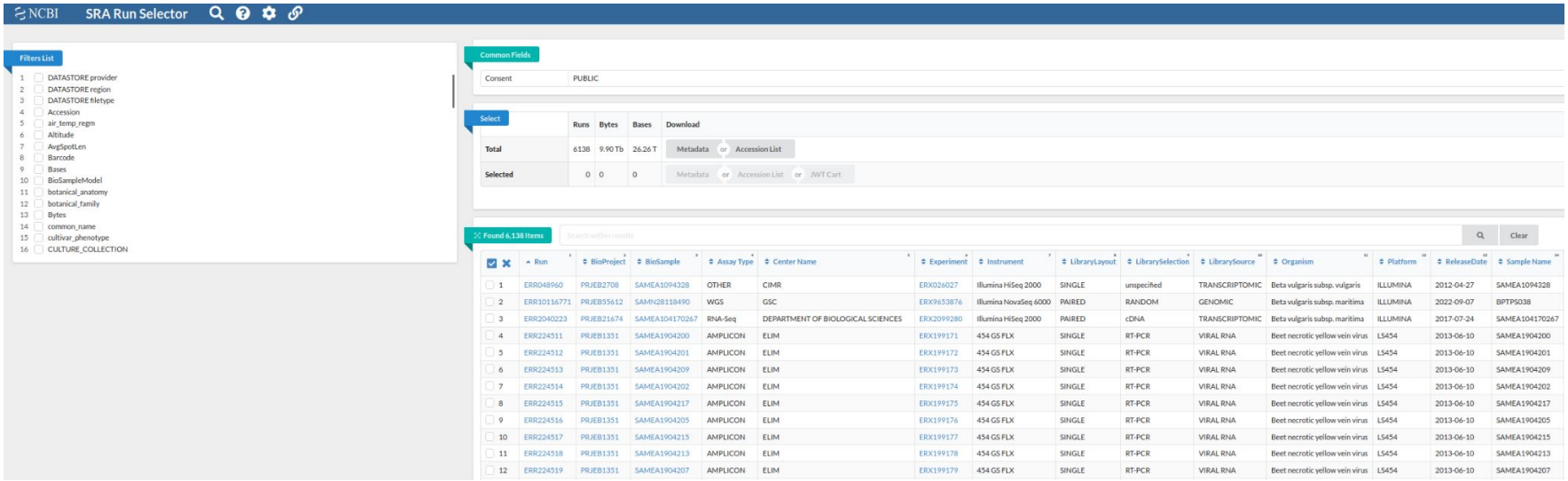
# Available plant RNA-seq data sets




Julca et al., 2022: 10.1016/j.tplants.2022.09.007

# SRA Run Selector

- Search based on various parameters and filtering
- Download of data set IDs and metadata



# Retrieving expression data

- Fastq-dump: automatic download of data set (FASTQ) based on ID
- Part of SRA tools; Faster alternatives are available
- GEO: download of count tables
- Latest development: bring analysis to data i.e. run analyses in cloud

# Metadata

- ID of run, experiment, and study
- Name and taxon ID of species
- Technical details about library preparation and sequencing
- Additional data about the samples are optional:
  - Sampled tissue
  - Date of sampling
  - Growth conditions/treatments

# Checking/filtering RNA-seq data sets

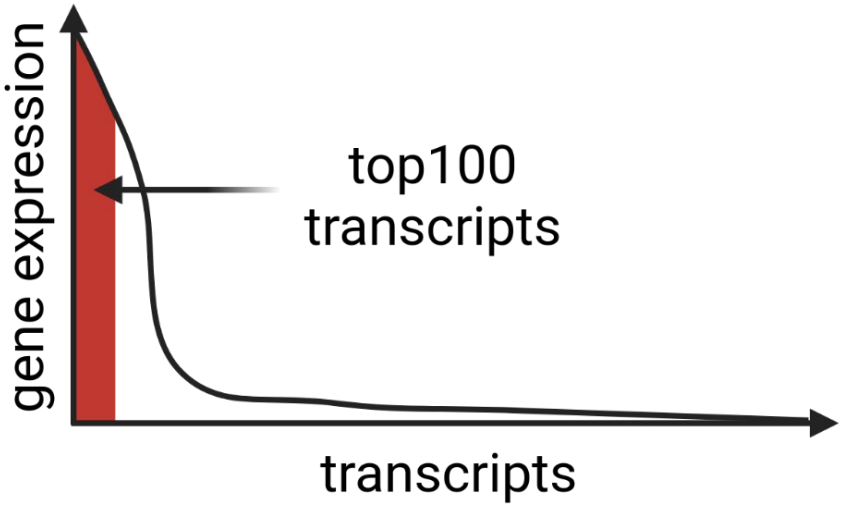
- Size of data set: count number of reads or read pairs
  - >5 million recommended

- Check species identity
  - Exclude mislabeled samples e.g. symbiosis

- Check tissue identity
  - Photosynthesis genes should be expressed in green tissue
  - No photosynthesis gene expression in roots
  - …

# Plausibility checks

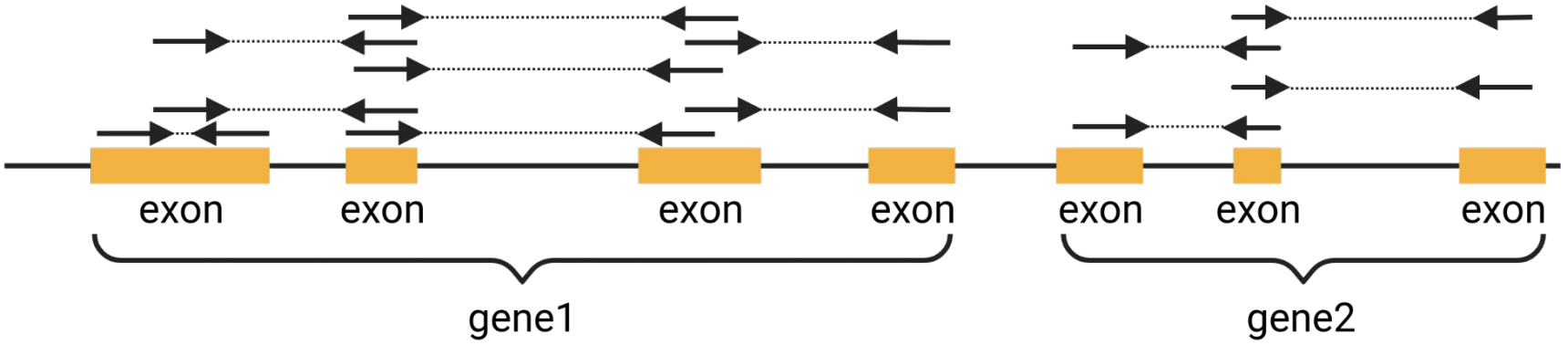
- Check for RNA character of samples (validate metadata)
- Exclude mislabeled DNA data sets
- Top100 transcripts should account for half of all transcripts (reads)



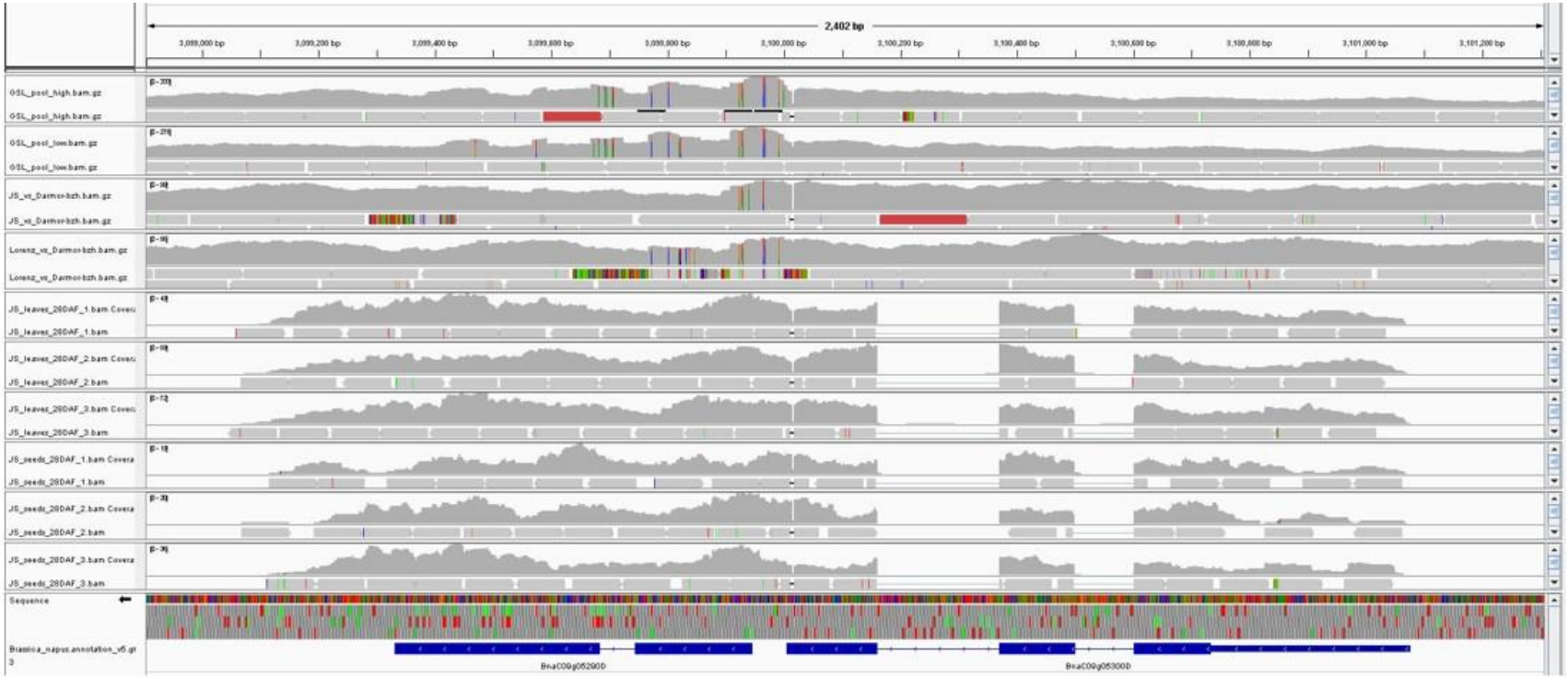

https://github.com/bpucker/CoExp

# Hints for gene prediction

- Alignments (mappings) of RNA-seq reads against a genome sequence
- RNA-seq reads indicate exon positions
- Splitted RNA-seq reads indicate introns and show connection of exons

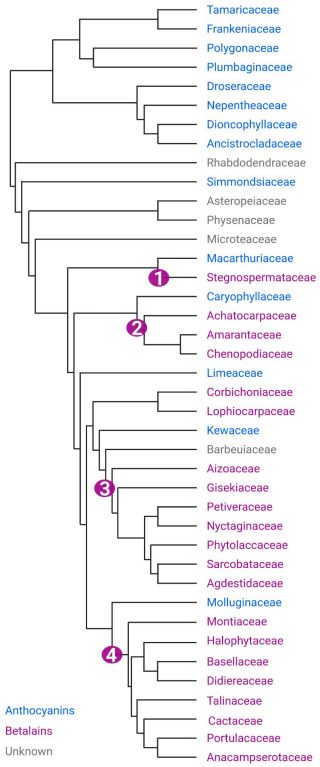


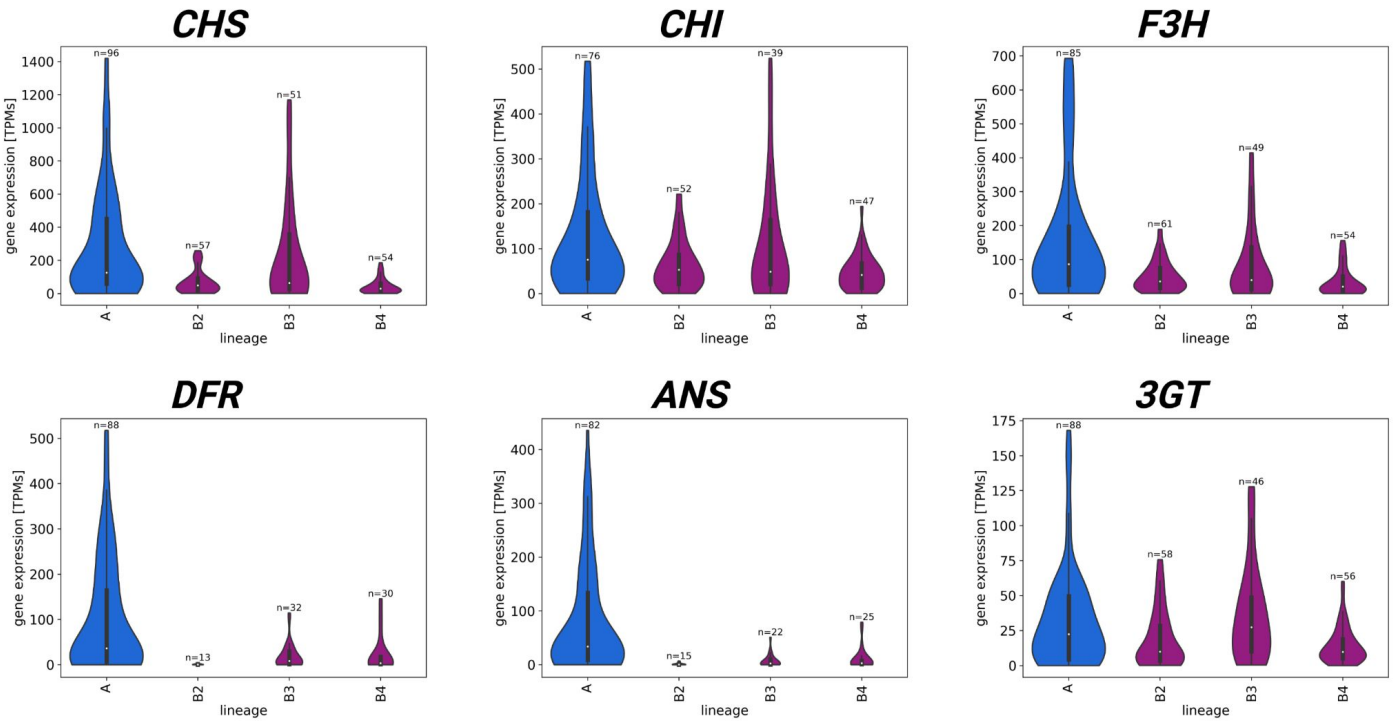
# Annotation of *Bna*MYB28




Schilbert & Pucker *et al*., 2022: 10.3390/genes13071131

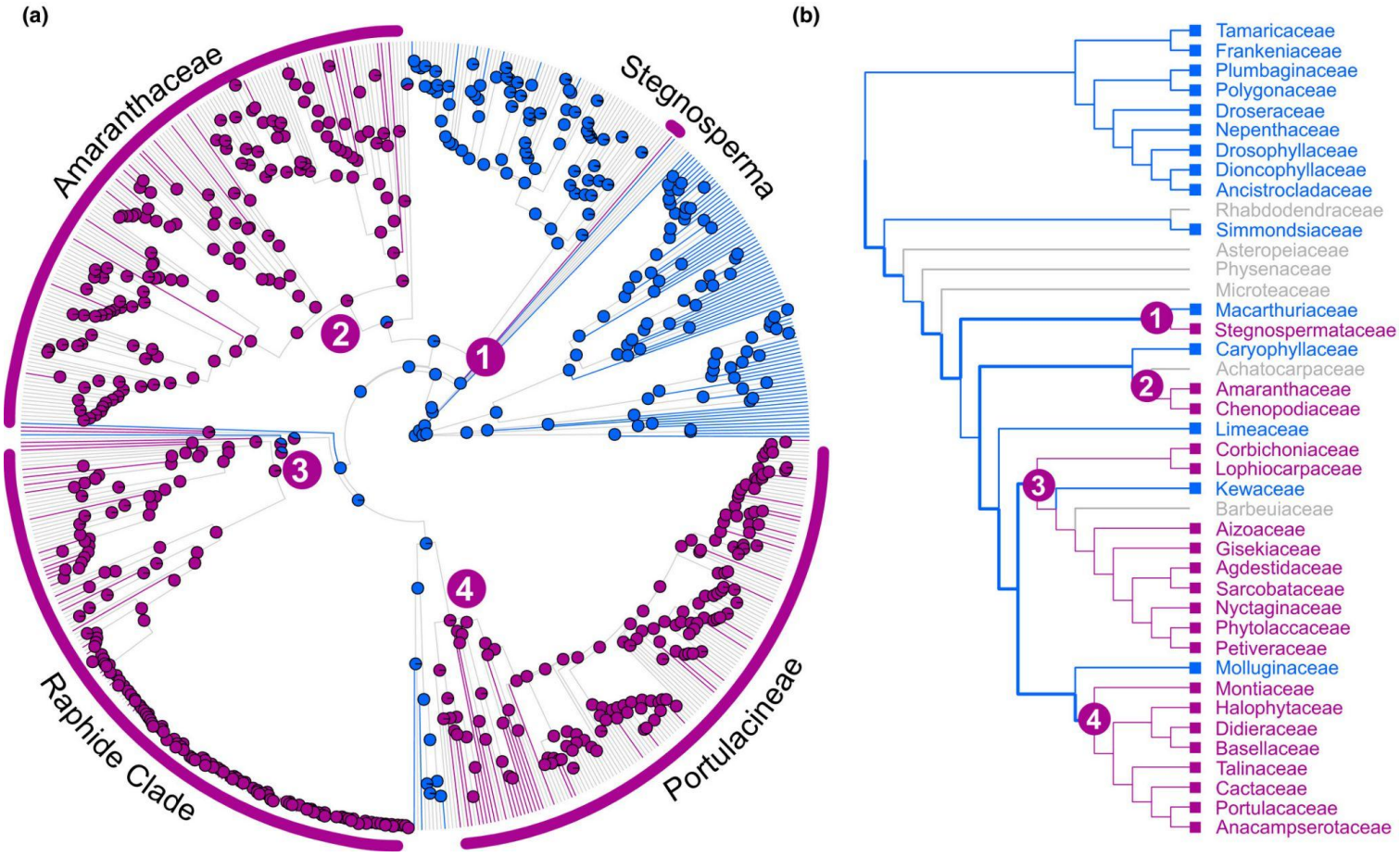
# Cross-species transcriptomics




Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

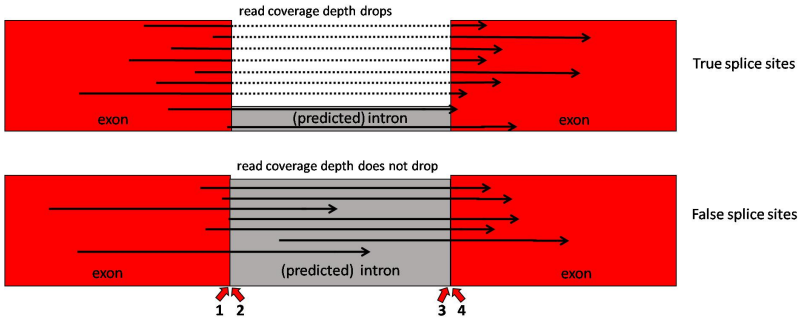
# Analysis of DODA evolution in the Caryophyllales

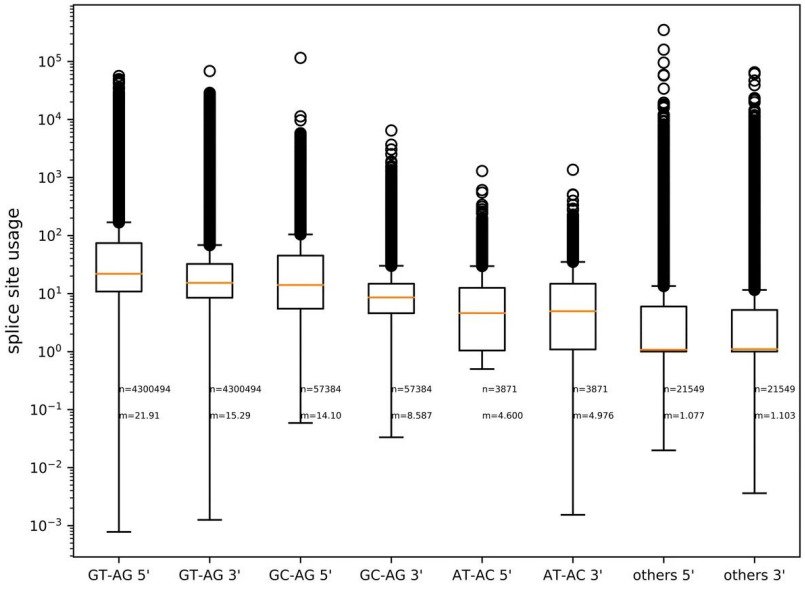



Sheehan *et al.*, 2020: 10.1111/nph.16089

# Non-canonical splice sites

- RNA-seq reads can be used to study the usage of splice sites
- Splitting of reads in the alignments is crucial to investigate introns
- STAR/HISAT2 are suitable read mappers to generate RNA-seq read alignments
- Splice site usage inferred from difference between terminal exon and terminal intron coverage

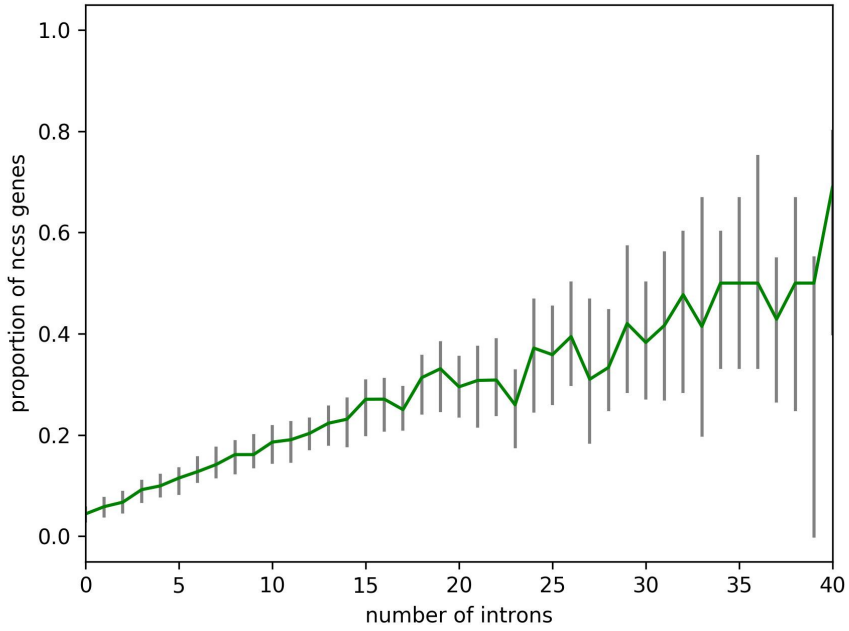


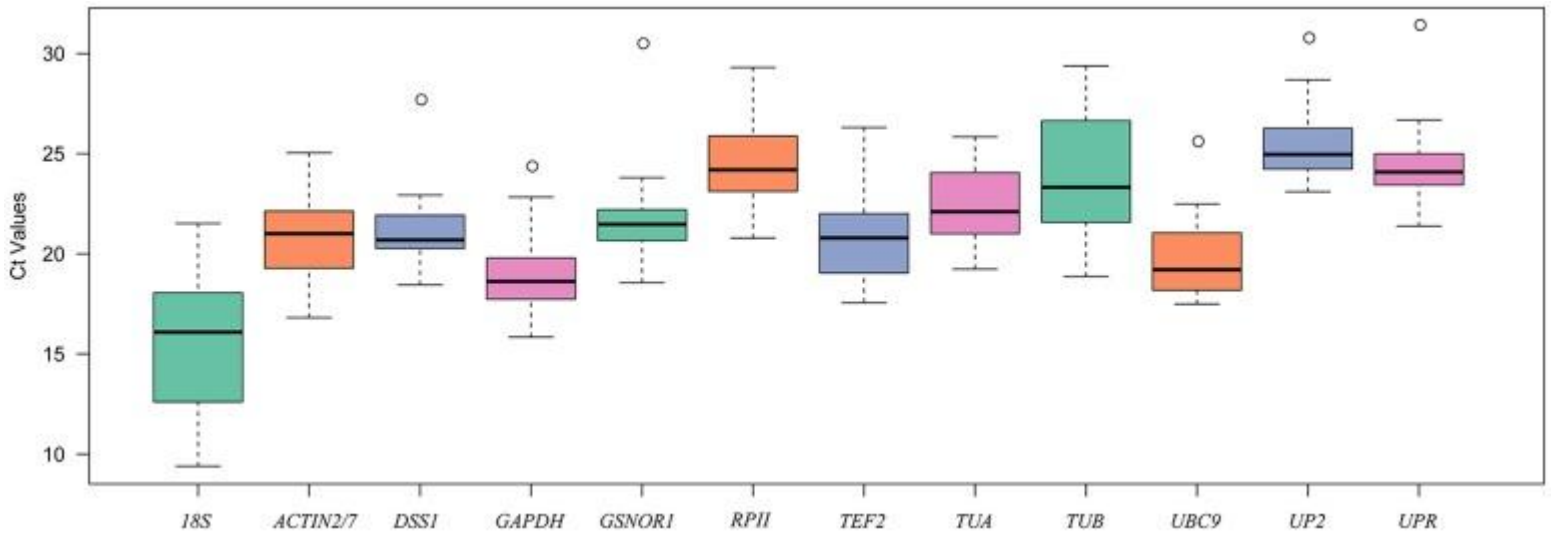

Pucker & Brockington, 2019: 10.1186/s12864-018-5360-z

# Relevance of non-canonical splice sites

- Non-canonical splice sites are affecting a substantial proportion of genes
- Percentage of affected genes depends on the number of introns (possibilities)
- About 5-10% of multi-exon genes affected by non-canonical splice sites

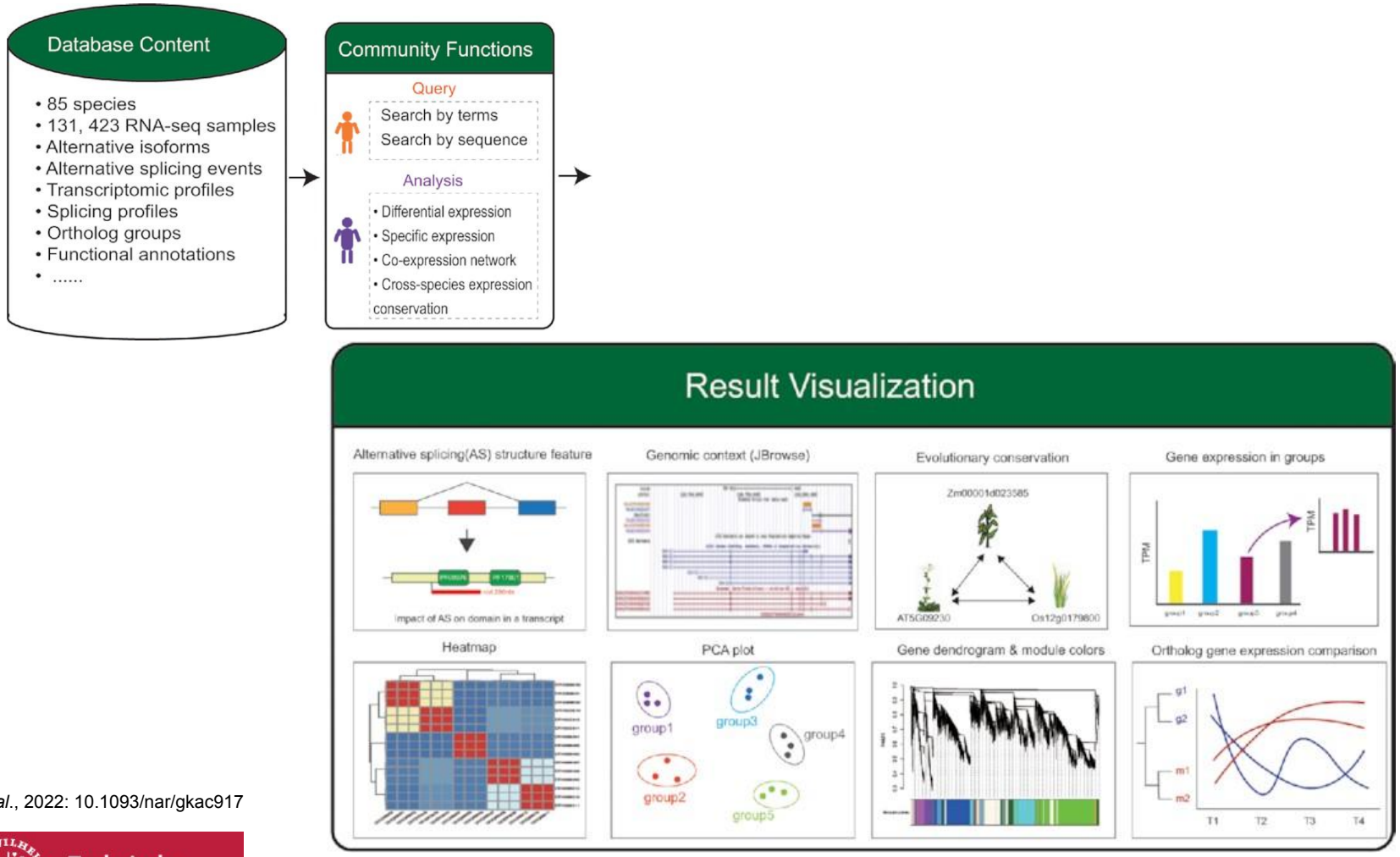


# Identification of qPCR reference genes



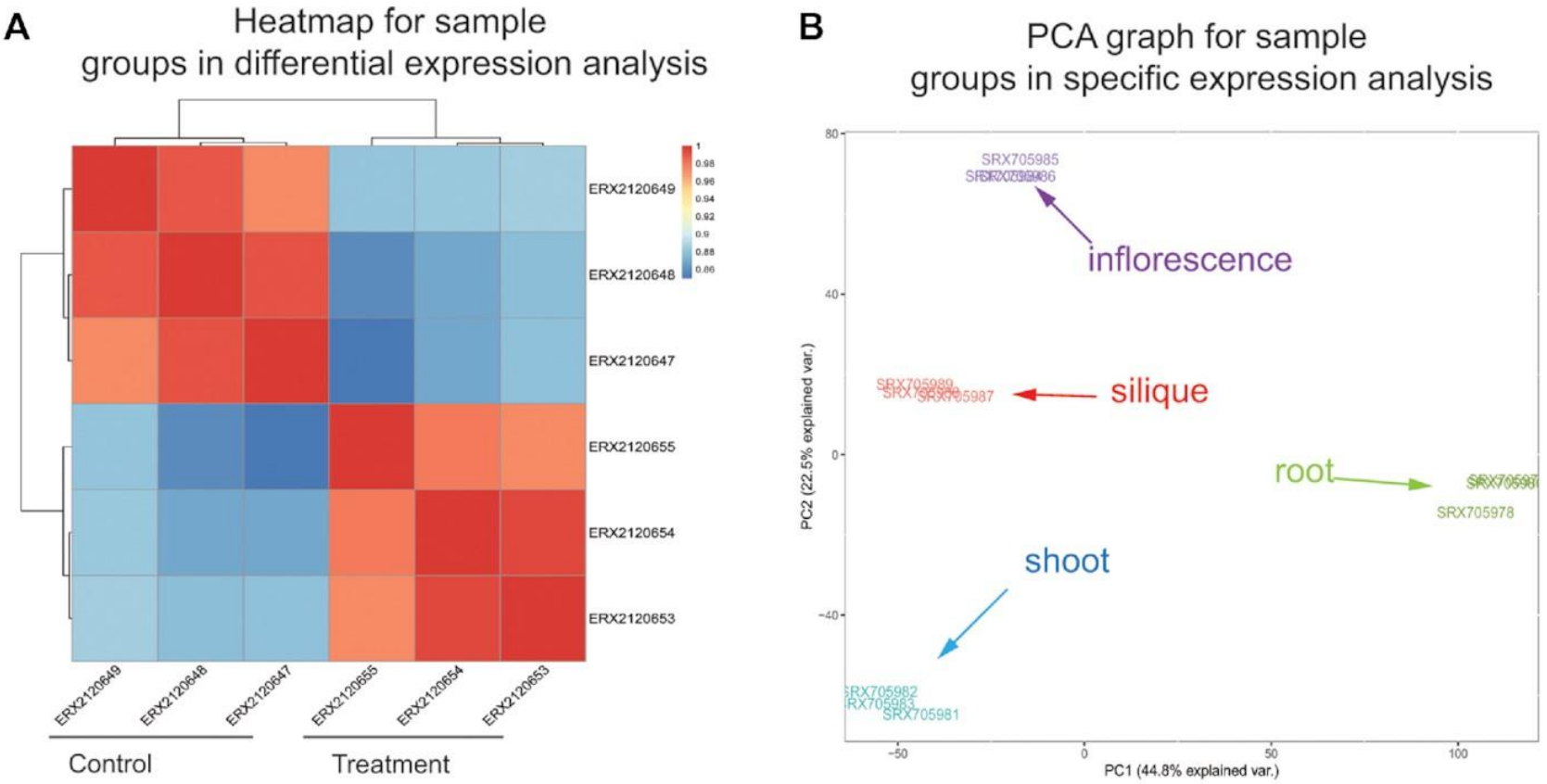

Duan *et al.*, 2017: 10.3389/fpls.2017.01605

# PlantExp: gene expression & alternative splicing



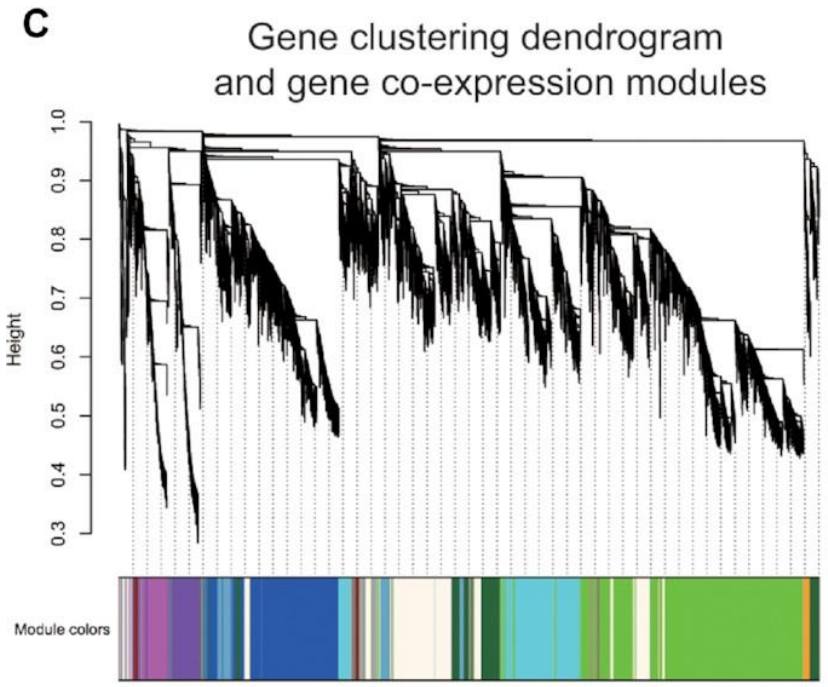

Liu *et al*., 2022: 10.1093/nar/gkac917

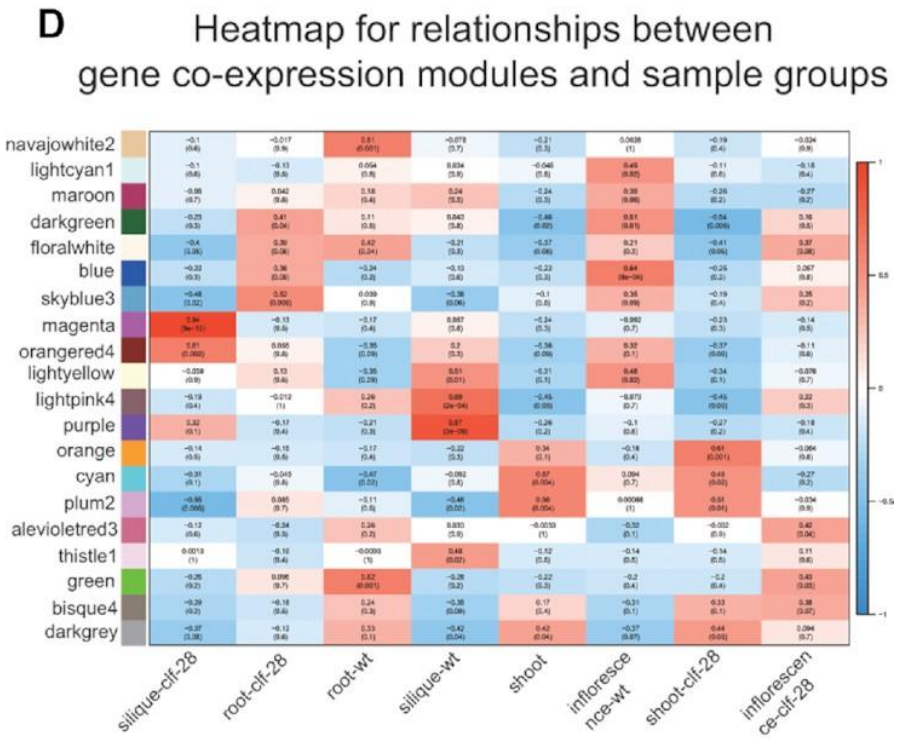
# PlantExp: example plots (1)




Liu *et al.*, 2022: 10.1093/nar/gkac917

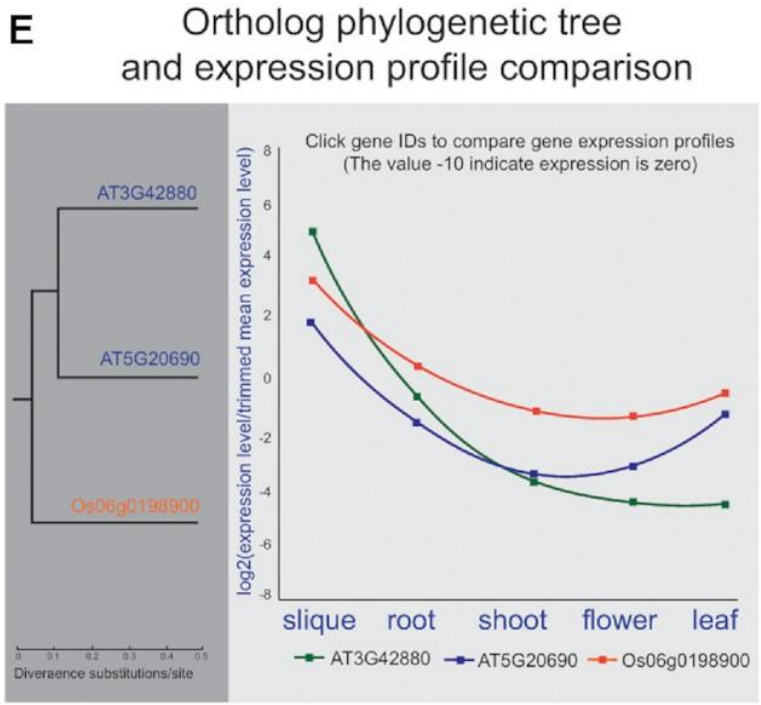
# PlantExp: example plots (2)

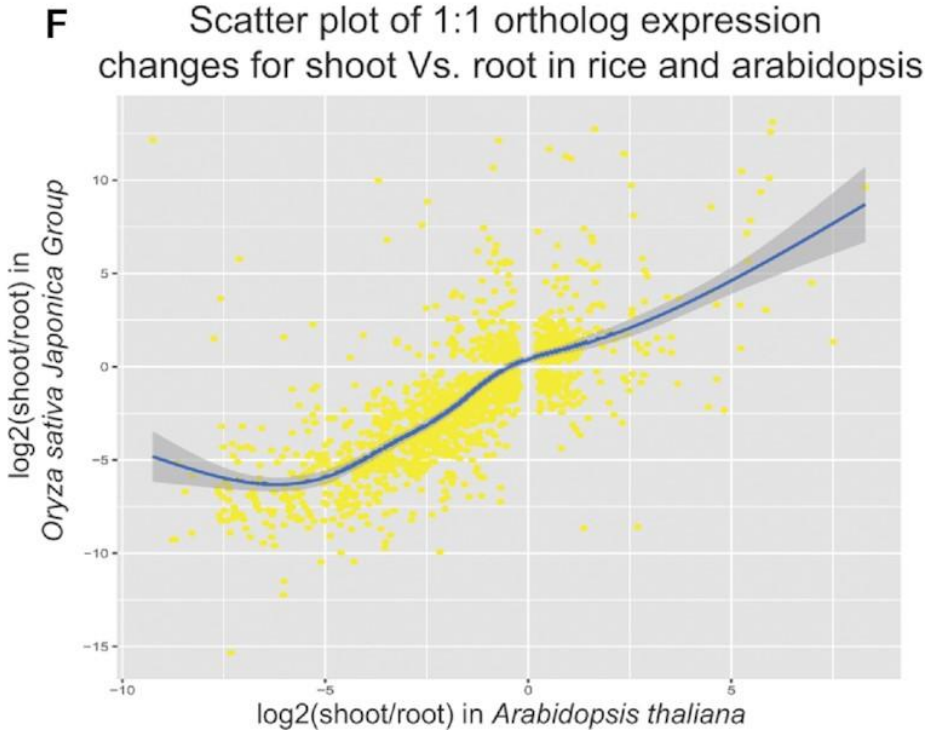



Liu *et al.*, 2022: 10.1093/nar/gkac917

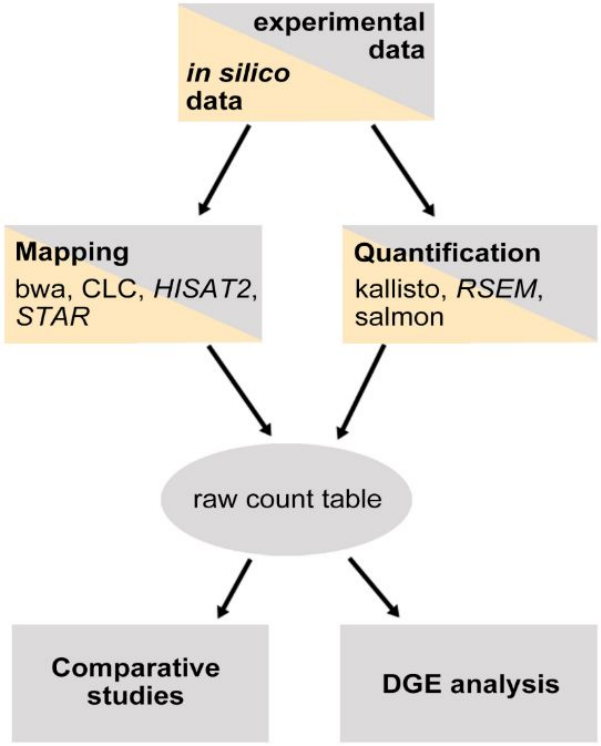
# PlantExp: example plots (3)

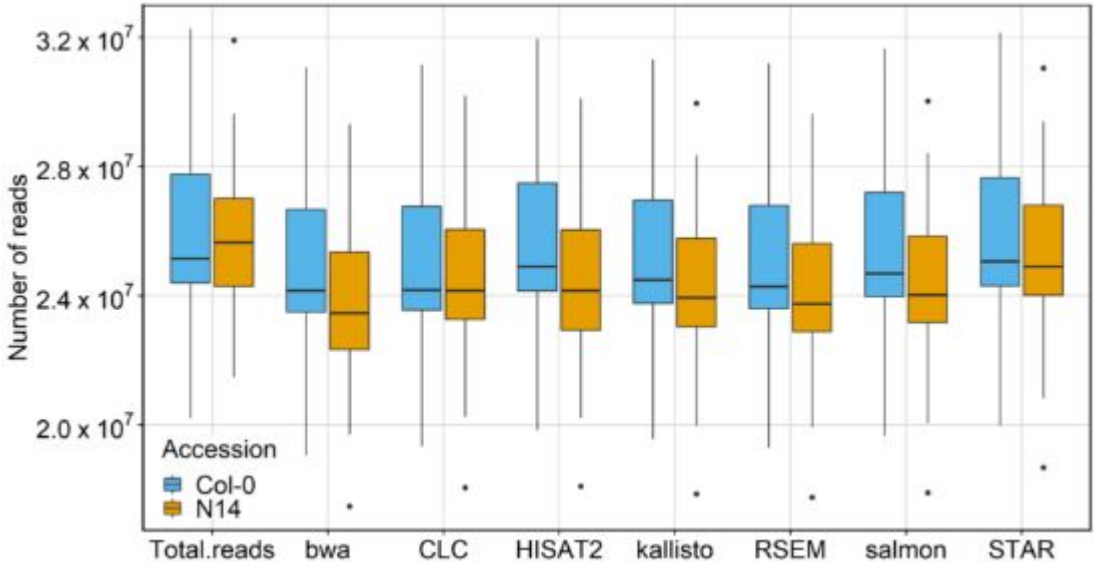



Liu *et al*., 2022: 10.1093/nar/gkac917

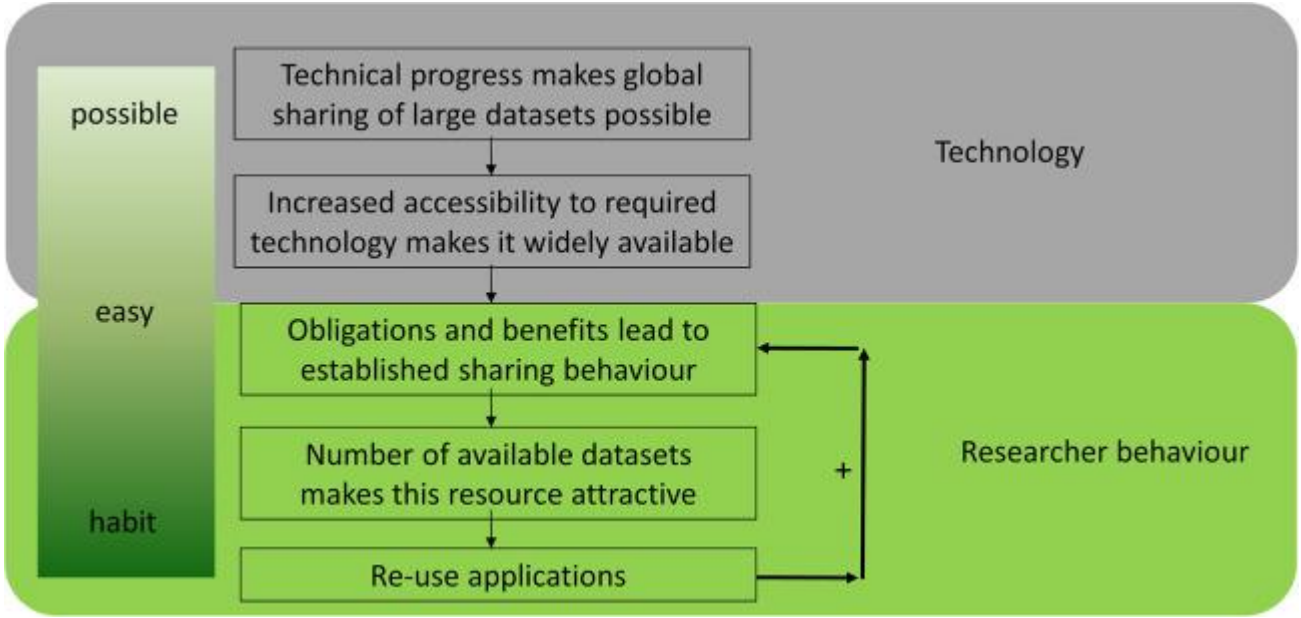
# RNA-seq read mapper benchmarking




Schaarschmidt et al., 2020: 10.3390/ijms21051720

# How to facilitate data reuse?




Sielemann *et al*., 2020

# Time for questions!

# Questions

1. Where can you find RNA-seq data sets for reuse?
2. What are metadata?
3. How to filter retrieved RNA-seq data sets?
4. What are examples of RNA-seq reuse applications?