

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](mailto:b.pucker@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.

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.



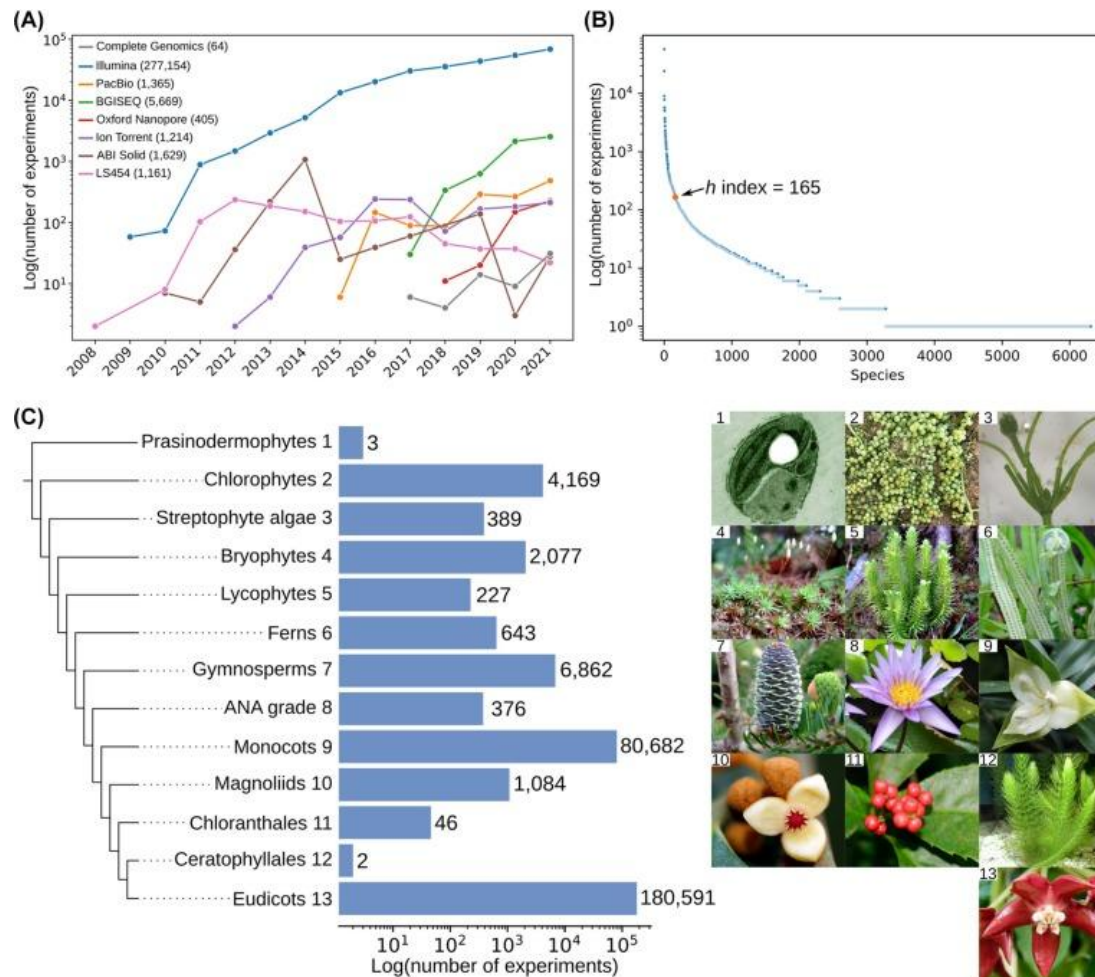
- Sequence Read Archive (SRA)



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

NCBI SRA Run Selector

Filters List

- ☐ DATASTORE provider
- ☐ DATASTORE region
- ☐ DATASTORE filetype
- ☐ Accession
- ☐ air_temp_regm
- ☐ Altitude
- ☐ AvgSpotLen
- ☐ Barcode
- ☐ Bases
- ☐ BioSampleModel
- ☐ botanical_anatomy
- ☐ botanical_family
- ☐ Bytes
- ☐ common_name
- ☐ cultivar_phenotype
- ☐ CULTURE_COLLECTION

Common Fields

Consent PUBLIC

Select

	Runs	Bytes	Bases	Download
Total	6138	9.90 Tb	26.26 T	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List or FMT Cart

Found 6,138 items

	Run	BioProject	BioSample	Assay Type	Center Name	Experiment	Instrument	LibraryLayout	LibrarySelection	LibrarySource	Organism	Platform	ReleaseDate	Sample Name
<input type="checkbox"/>	ERR048960	PRJEB2708	SAMEA1094328	OTHER	CIMR	ERR026027	Illumina HiSeq 2000	SINGLE	unspecified	TRANSCRIPTOMIC	Beta vulgaris subsp. vulgaris	ILLUMINA	2012-04-27	SAMEA1094328
<input type="checkbox"/>	ERR10116771	PRJEB55612	SAMN28118490	WGS	GSC	ERR9653876	Illumina NovaSeq 6000	PAIRED	RANDOM	GENOMIC	Beta vulgaris subsp. maritima	ILLUMINA	2022-09-07	BPTP5038
<input type="checkbox"/>	ERR2040223	PRJEB21674	SAMEA104170267	RNA-Seq	DEPARTMENT OF BIOLOGICAL SCIENCES	ERR2099280	Illumina HiSeq 2000	PAIRED	cDNA	TRANSCRIPTOMIC	Beta vulgaris subsp. maritima	ILLUMINA	2017-07-24	SAMEA104170267
<input type="checkbox"/>	ERR224511	PRJEB1351	SAMEA1904200	AMPLICON	ELIM	ERR199171	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904200
<input type="checkbox"/>	ERR224512	PRJEB1351	SAMEA1904201	AMPLICON	ELIM	ERR199172	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904201
<input type="checkbox"/>	ERR224513	PRJEB1351	SAMEA1904209	AMPLICON	ELIM	ERR199173	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904209
<input type="checkbox"/>	ERR224514	PRJEB1351	SAMEA1904202	AMPLICON	ELIM	ERR199174	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904202
<input type="checkbox"/>	ERR224515	PRJEB1351	SAMEA1904217	AMPLICON	ELIM	ERR199175	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904217
<input type="checkbox"/>	ERR224516	PRJEB1351	SAMEA1904205	AMPLICON	ELIM	ERR199176	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904205
<input type="checkbox"/>	ERR224517	PRJEB1351	SAMEA1904215	AMPLICON	ELIM	ERR199177	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904215
<input type="checkbox"/>	ERR224518	PRJEB1351	SAMEA1904213	AMPLICON	ELIM	ERR199178	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904213
<input type="checkbox"/>	ERR224519	PRJEB1351	SAMEA1904207	AMPLICON	ELIM	ERR199179	454 GS FLX	SINGLE	RT-PCR	VIRAL RNA	Beet necrotic yellow vein virus	LS454	2013-06-10	SAMEA1904207

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

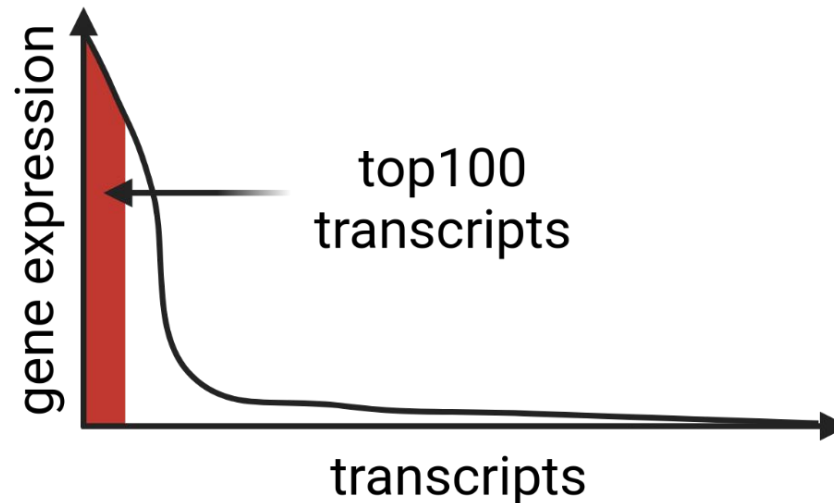
- ID of run, experiment, and study
- Name and taxon ID of species
- Technical details about library preparation and sequencing
- Additional data about the samples are optional:
 - Sampled tissue
 - Date of sampling
 - Growth conditions/treatments

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
 - ...

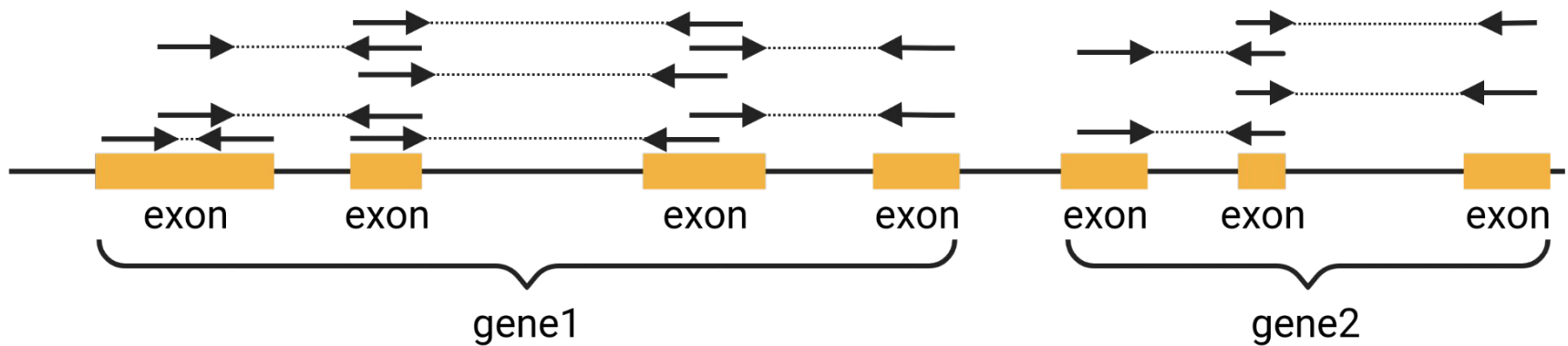
Plausibility checks

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

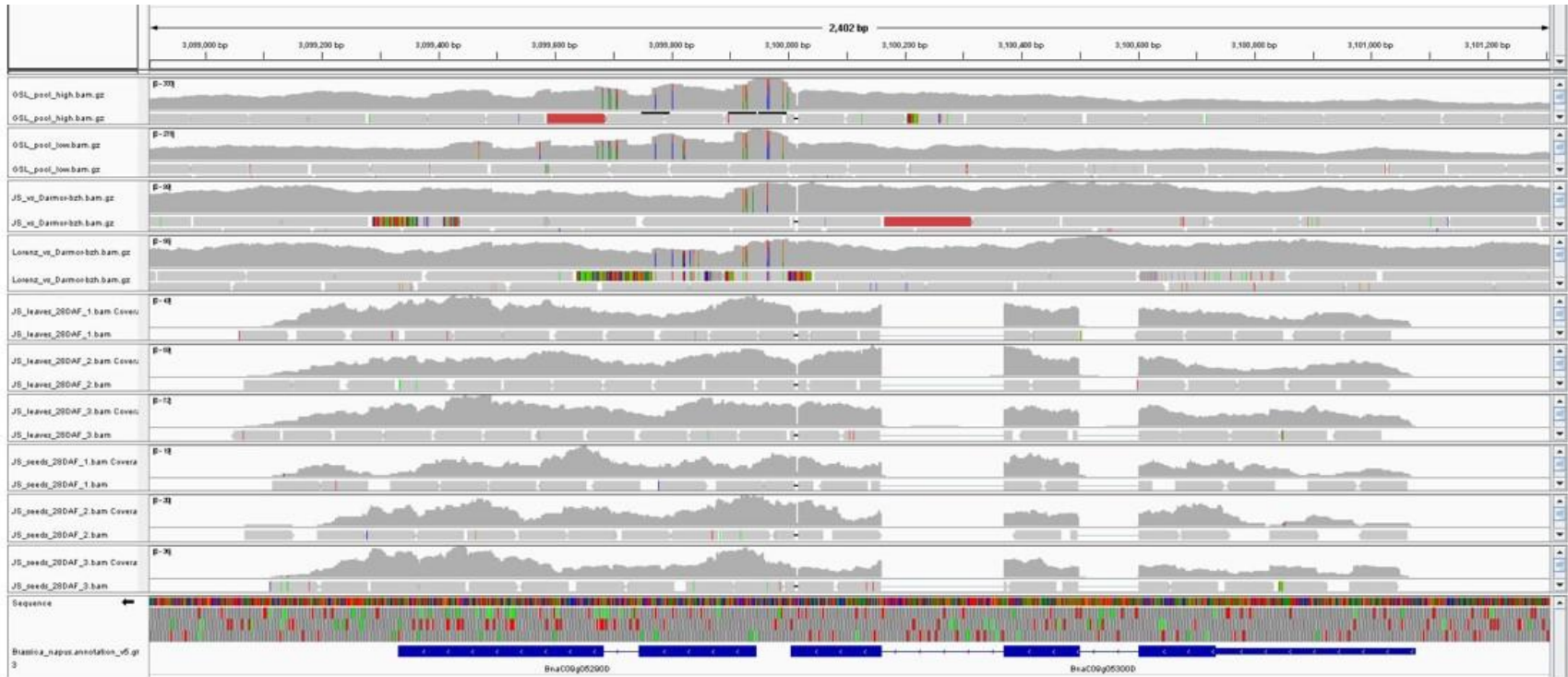


Hints for gene prediction

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

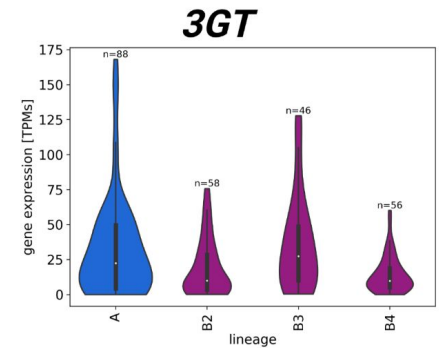
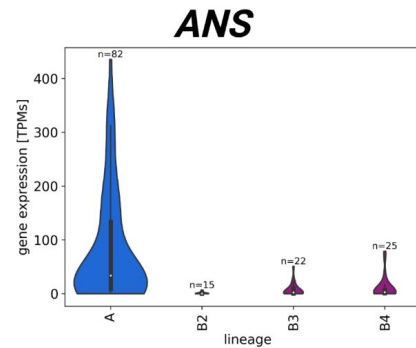
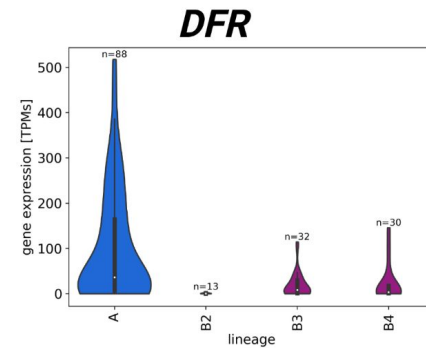
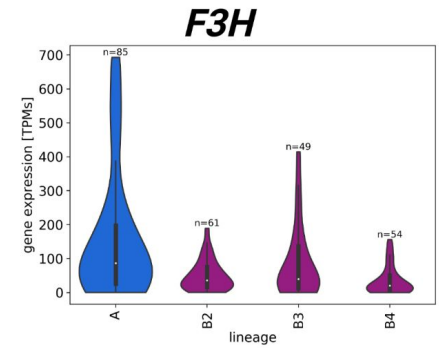
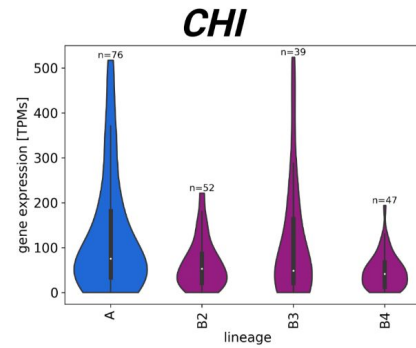
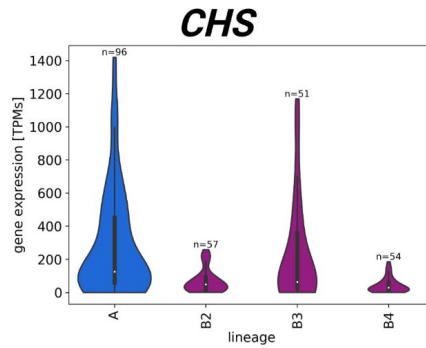
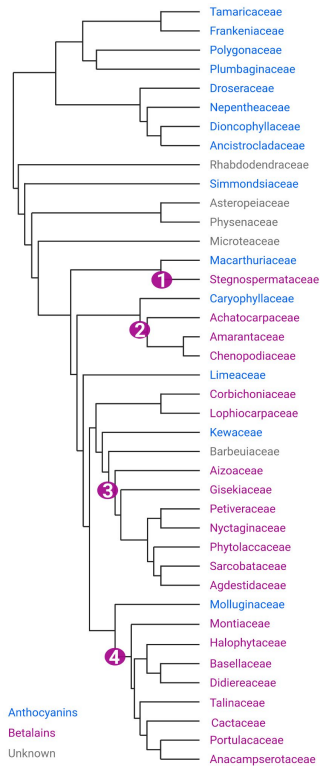


Annotation of *BnaMYB28*



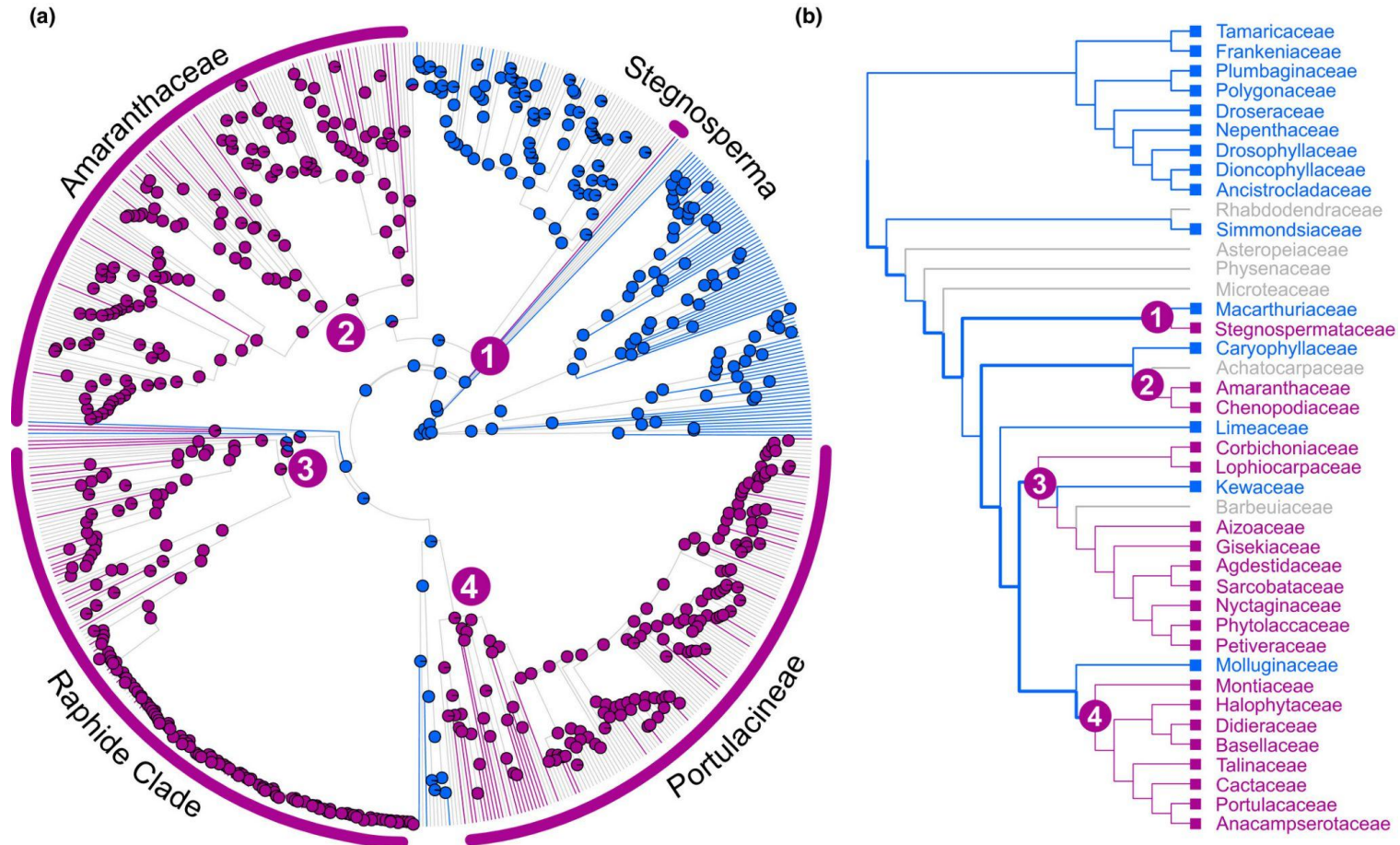
Schilbert & Pucker *et al.*, 2022: 10.3390/genes13071131

Cross-species transcriptomics



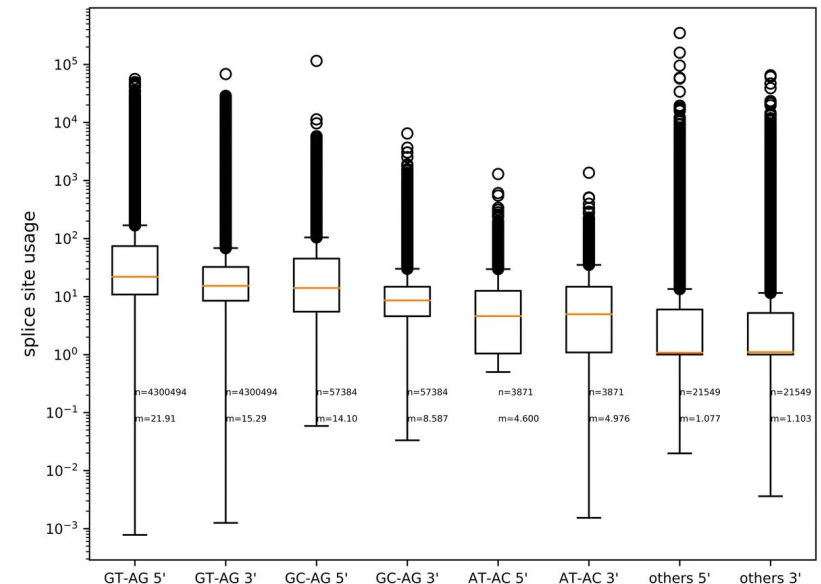
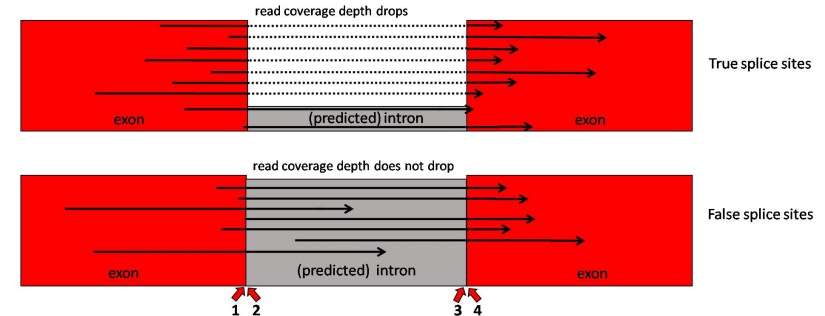
Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

Analysis of DODA evolution in the Caryophyllales



Non-canonical splice sites

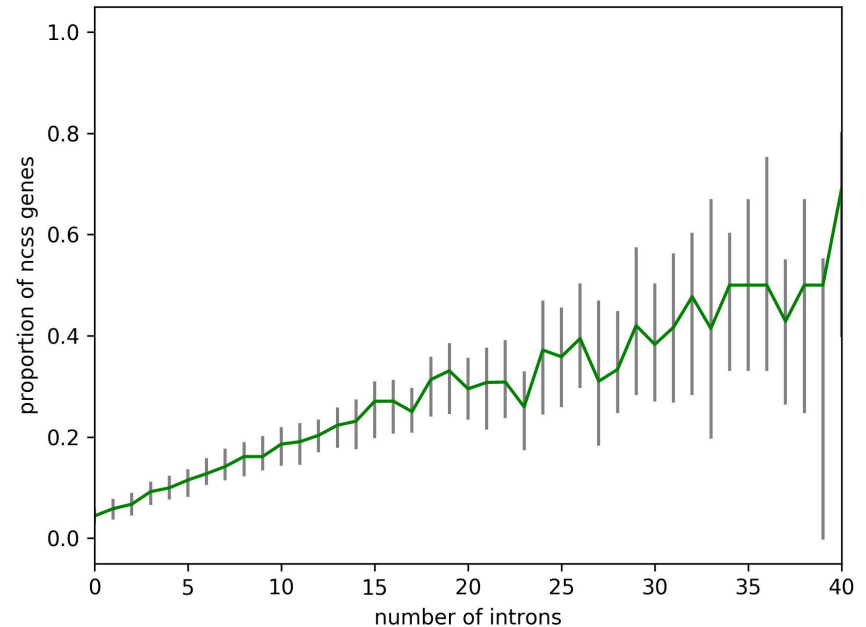
- RNA-seq reads can be used to study the usage of splice sites
- Splitting of reads in the alignments is crucial to investigate introns
- STAR/HISAT2 are suitable read mappers to generate RNA-seq read alignments
- Splice site usage inferred from difference between terminal exon and terminal intron coverage



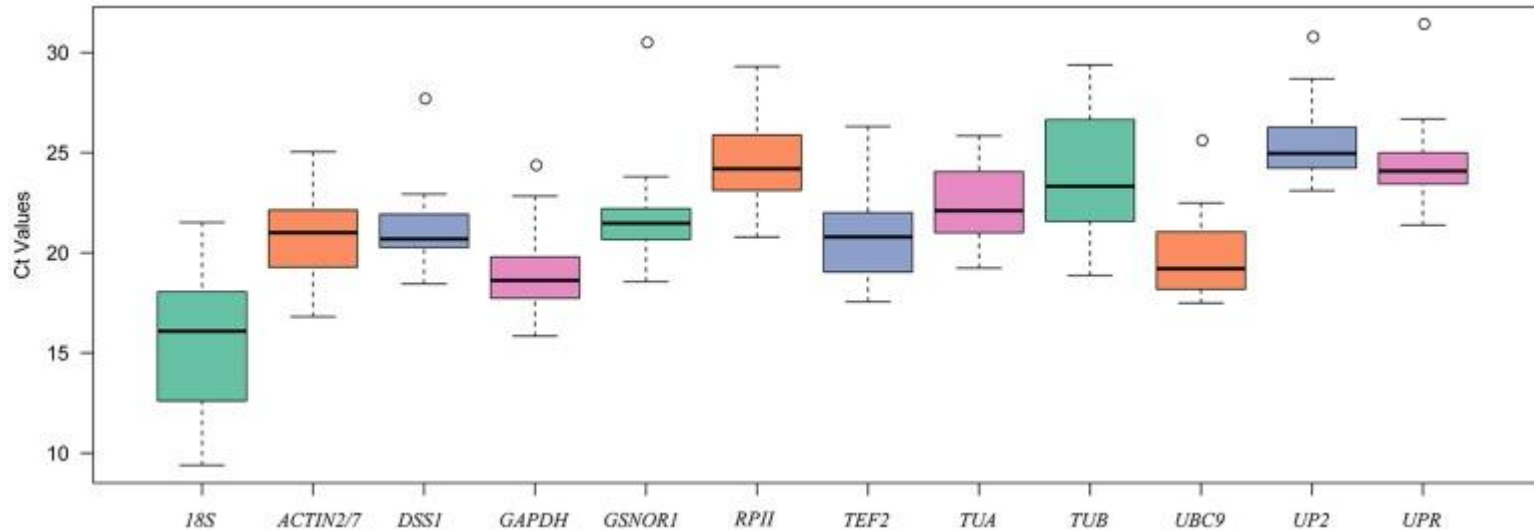
Pucker & Brockington, 2019: 10.1186/s12864-018-5360-z

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

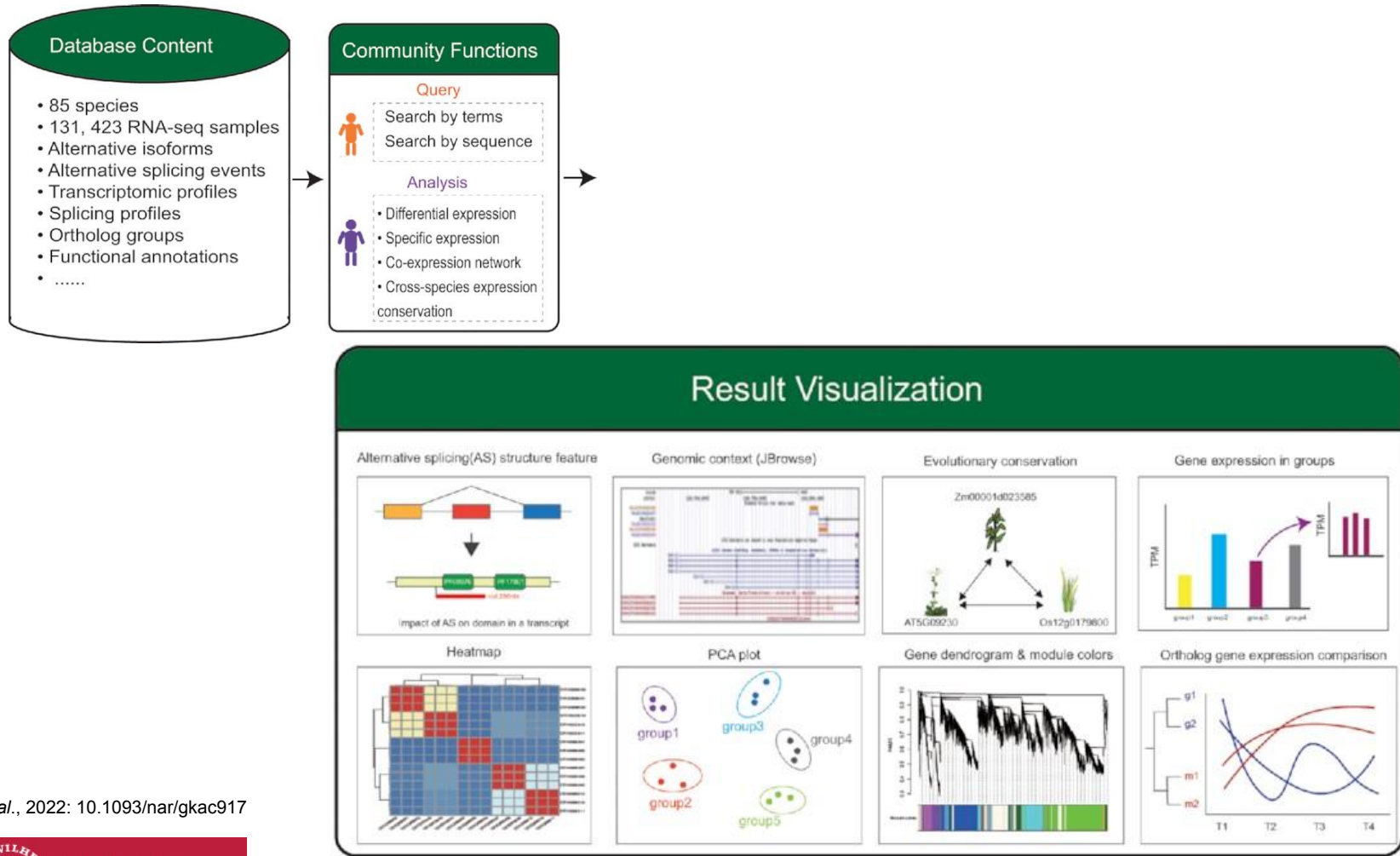


Identification of qPCR reference genes



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

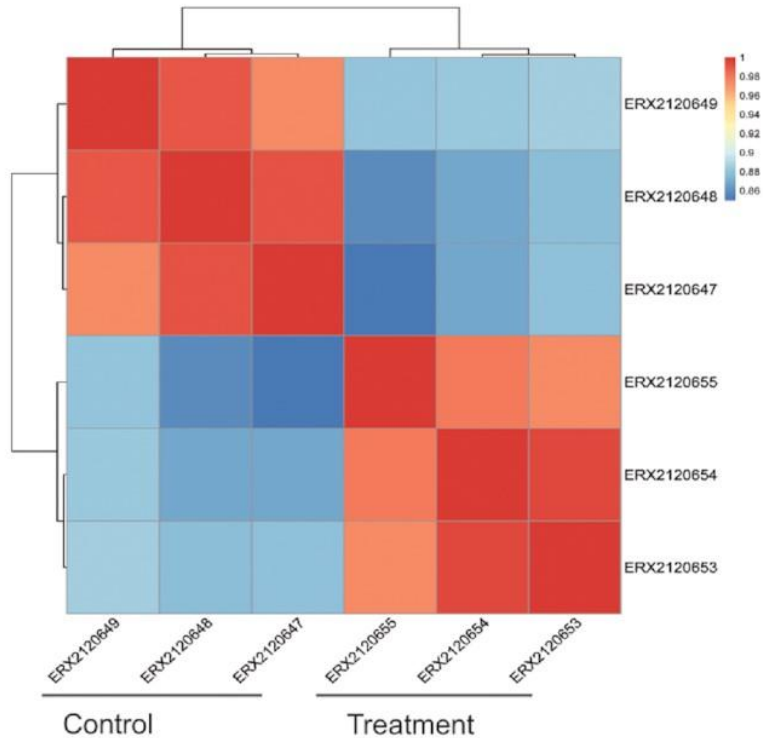
PlantExp: gene expression & alternative splicing



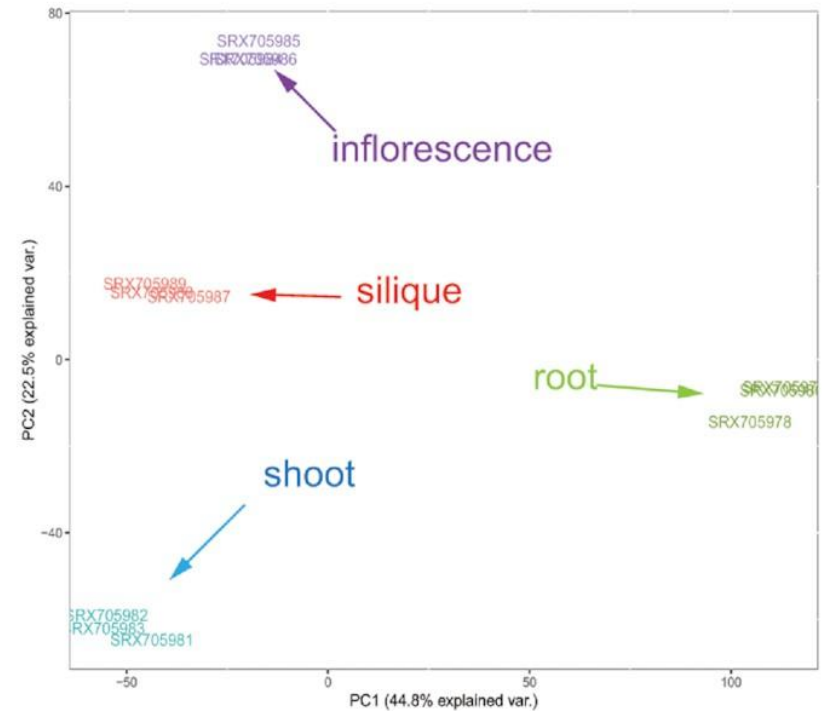
Liu *et al.*, 2022: 10.1093/nar/gkac917

PlantExp: example plots (1)

A Heatmap for sample groups in differential expression analysis

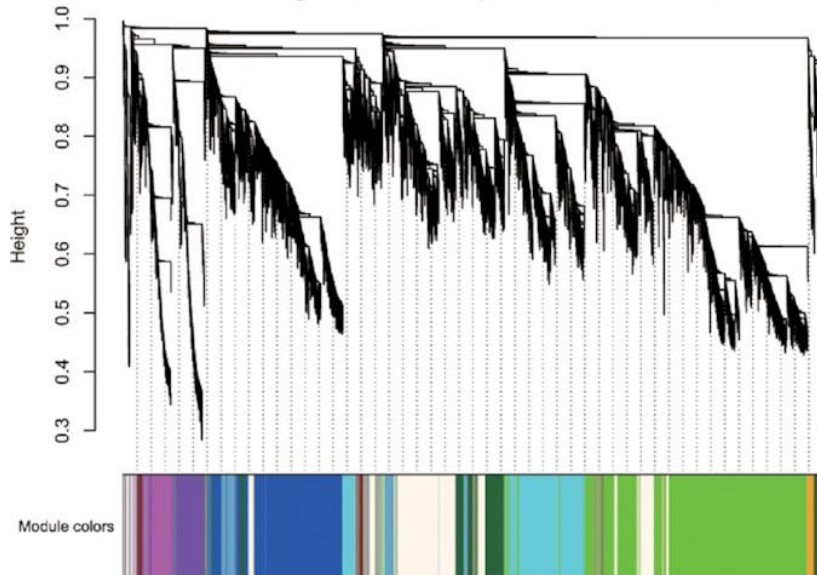


B PCA graph for sample groups in specific expression analysis

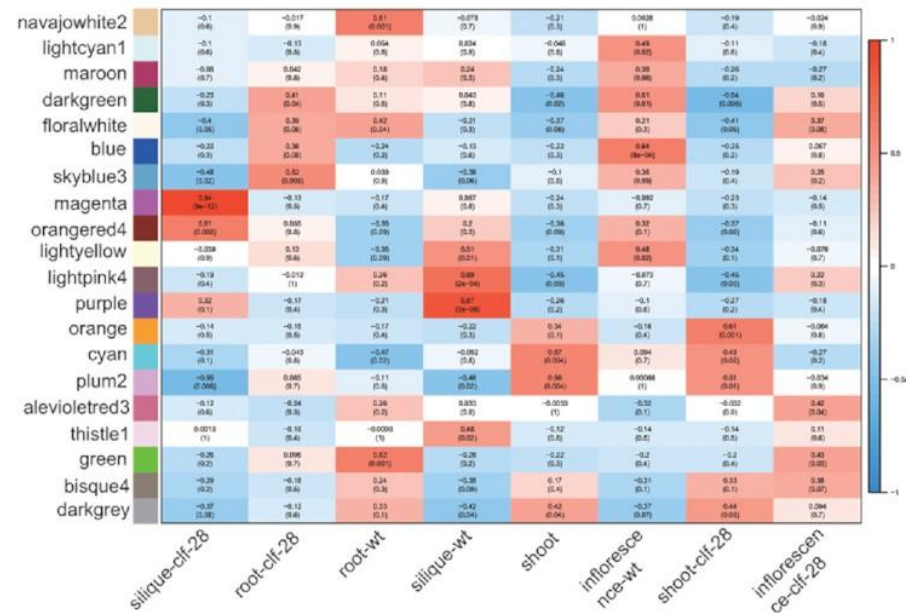


PlantExp: example plots (2)

C Gene clustering dendrogram and gene co-expression modules

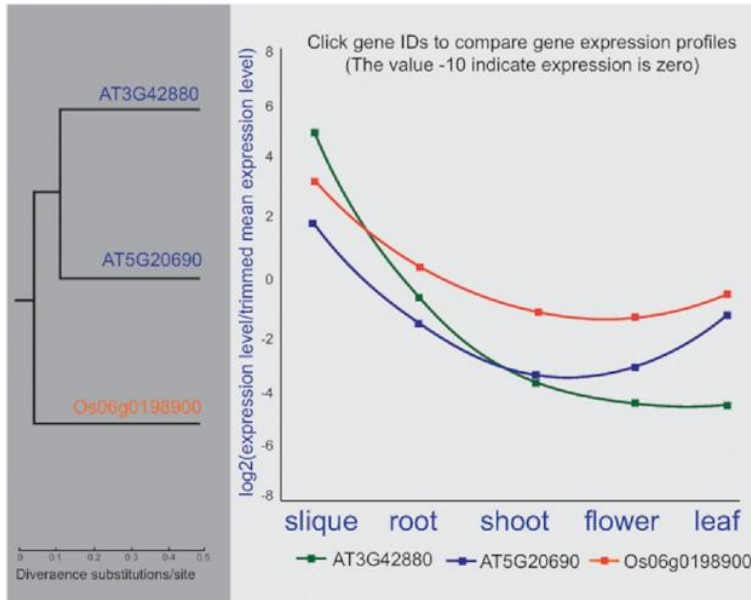


D Heatmap for relationships between gene co-expression modules and sample groups

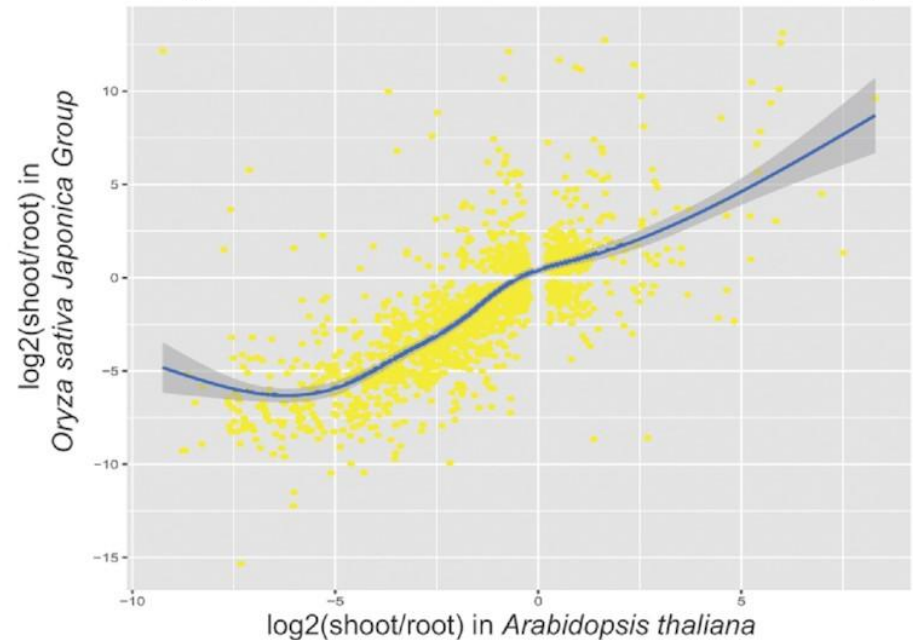


PlantExp: example plots (3)

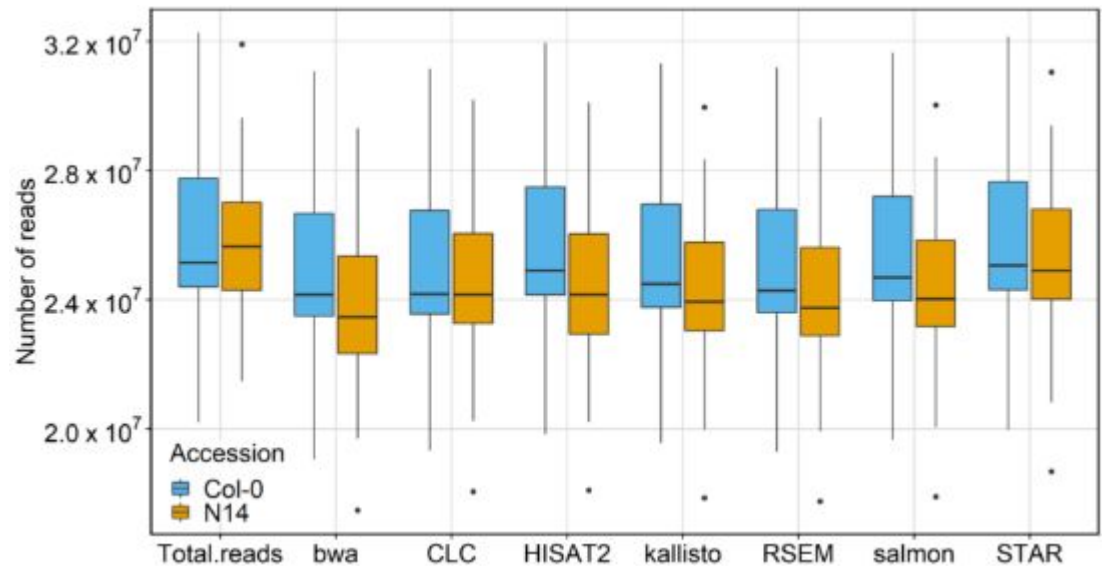
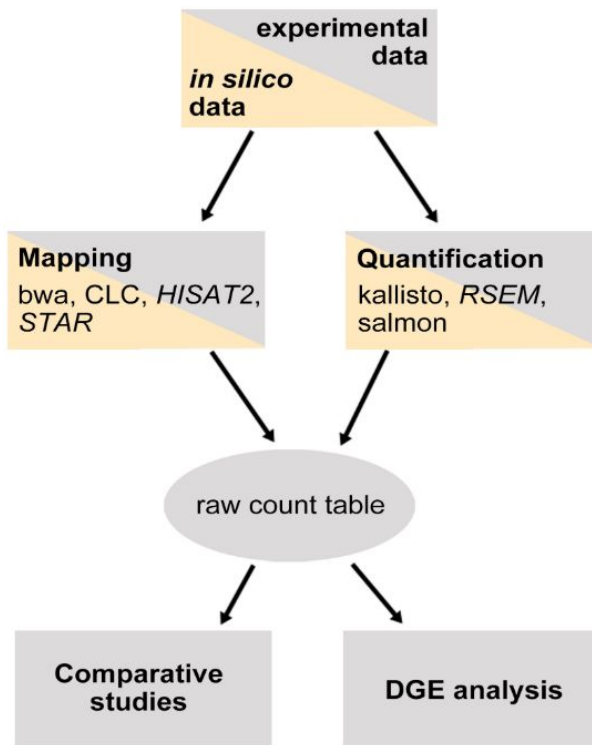
E Ortholog phylogenetic tree and expression profile comparison



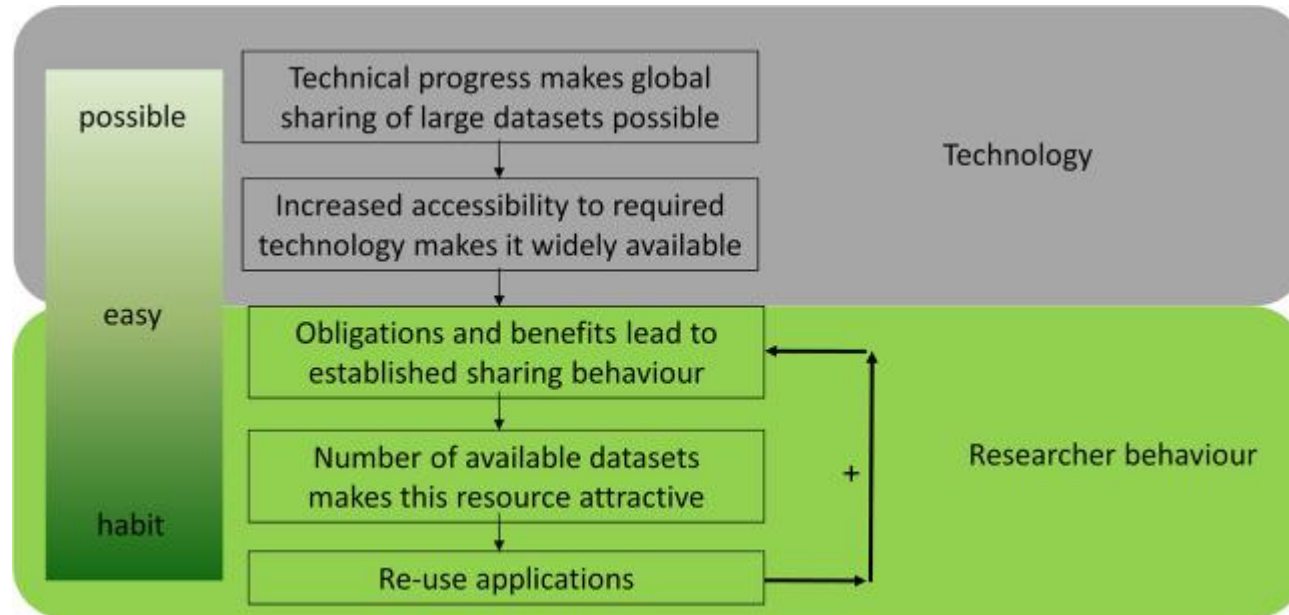
F Scatter plot of 1:1 ortholog expression changes for shoot Vs. root in rice and arabidopsis



RNA-seq read mapper benchmarking



How to facilitate data reuse?



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?