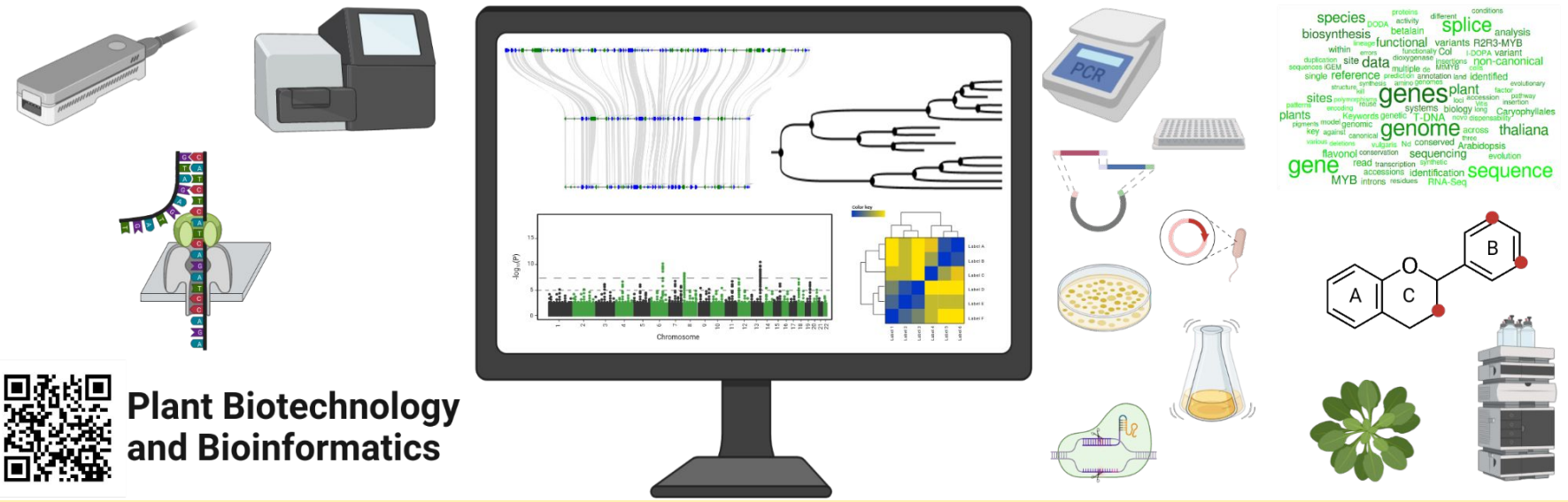
Technische Universität Braunschweig



Plant Biotechnology and Bioinformatics

Data analysis & basic bioinformatics

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

Availability of slides

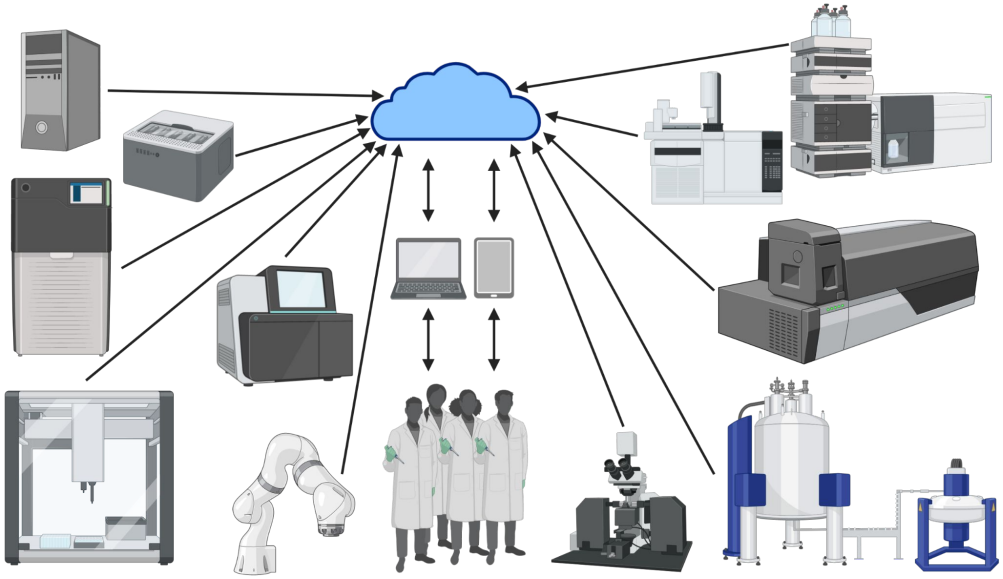
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: LMChemBSc12
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-bs.de

Big Data & Lab 4.0

https://commons.wikimedia.org/wiki/File:BigData_2267x1146_white.png, CC-BY-SA 3.0



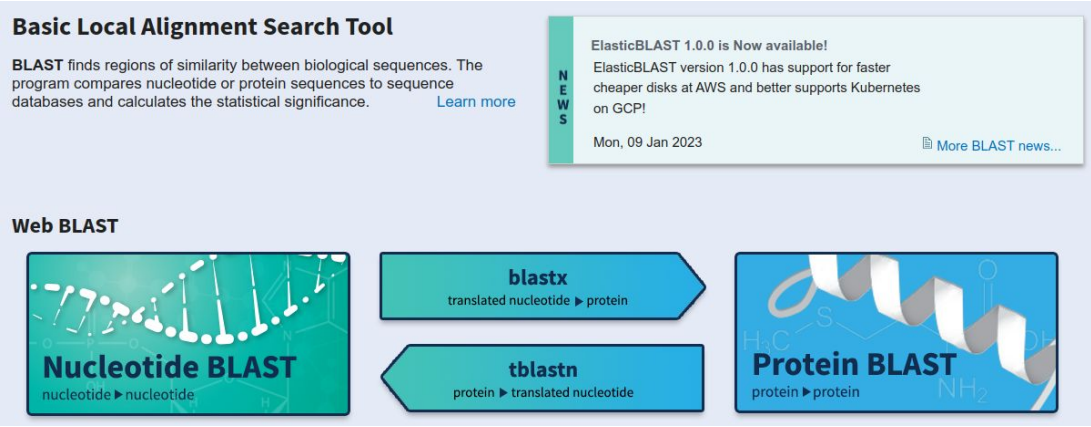
Databases and online services

BLAST

- BLAST = Basic Local Alignment Search Tool
- Most frequently used and cited tool in bioinformatics

Basic local alignment search tool - PubMed

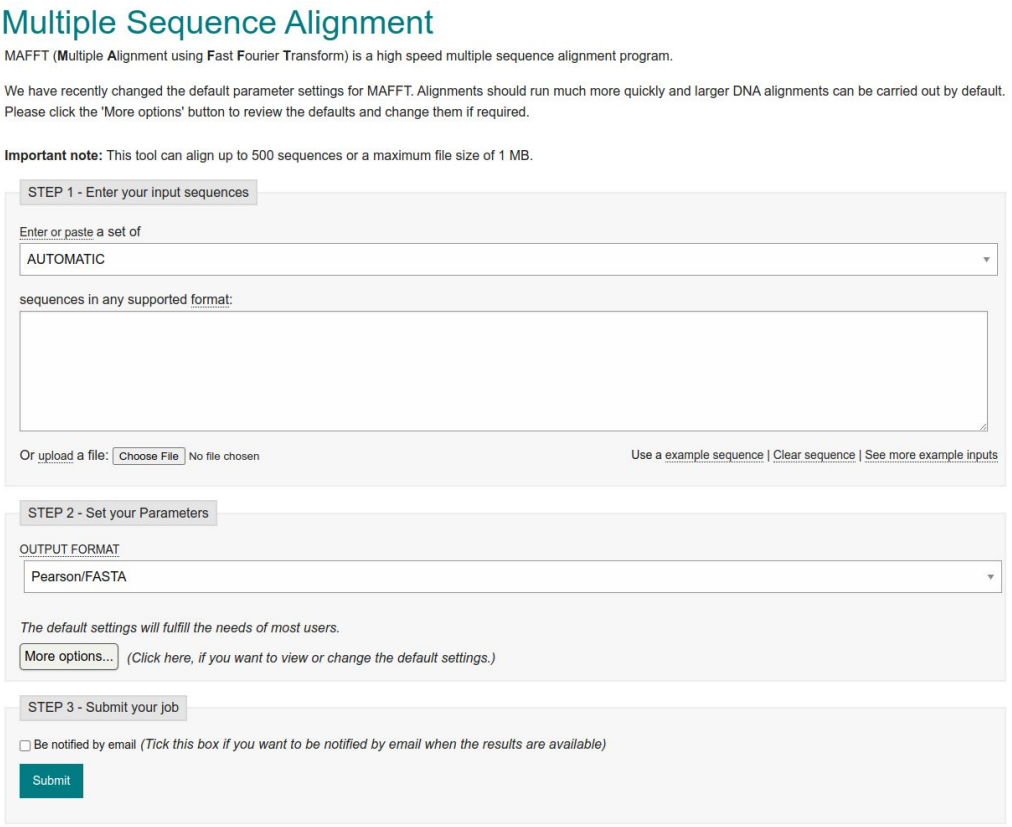
von SF Altschul · 1990 · Zitiert von: 102645 — A new approach to rapid sequence comparison, basic local alignment search tool (**BLAST**), directly approximates alignments that optimize a ...



https://blast.ncbi.nlm.nih.gov/Blast.cgi

MAFFT

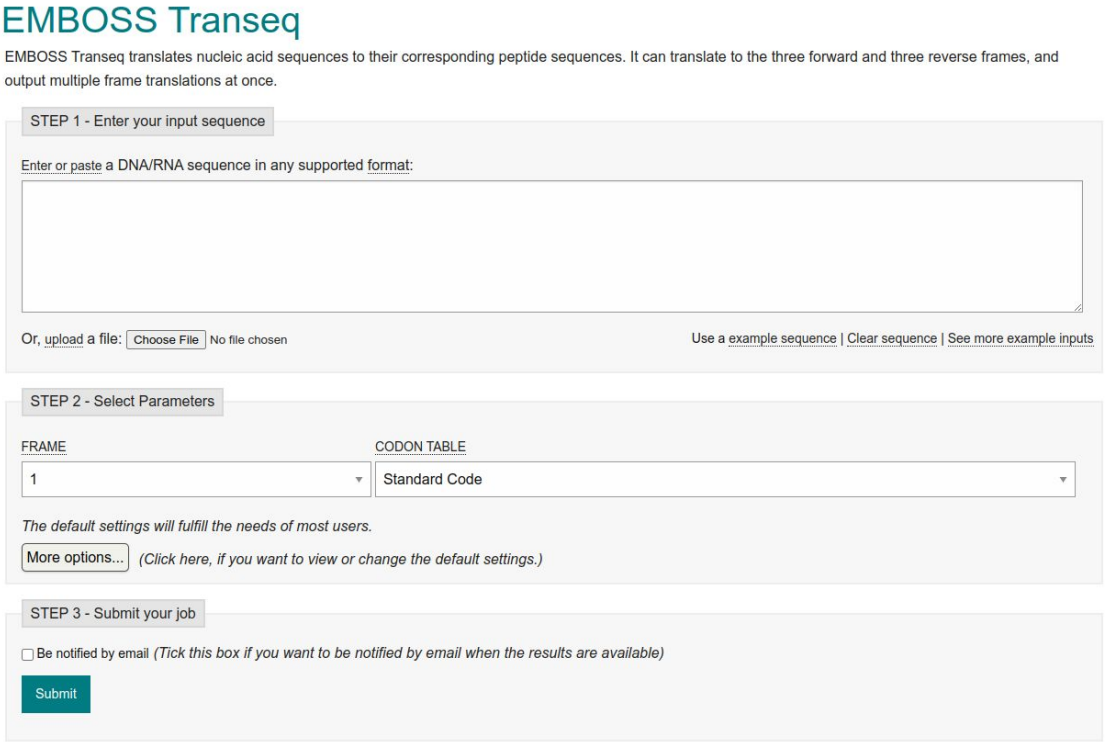
- MAFFT = Multiple Alignment using Fast Fourier Transformation
- Highly efficient for extremely large sequence collections
- Available as command line tool and online service



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

Transeq

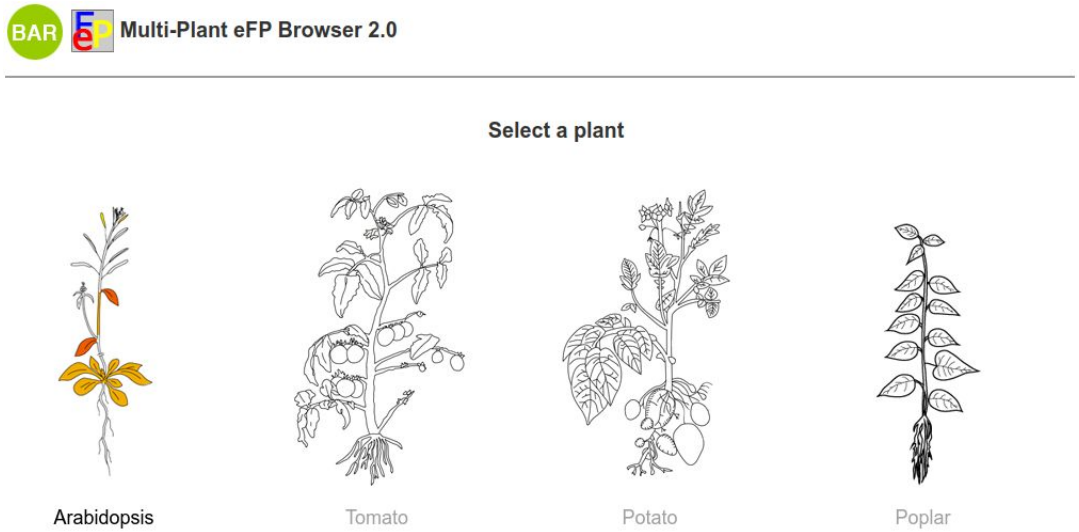
- Transeq translates a given nucleotide sequence into a polypeptide sequence
- Different genetic codes can be used
- Six different frames can be explored (3x forward, 3x reverse)



https://www.ebi.ac.uk/Tools/st/emboss_transeq/

eFP browser

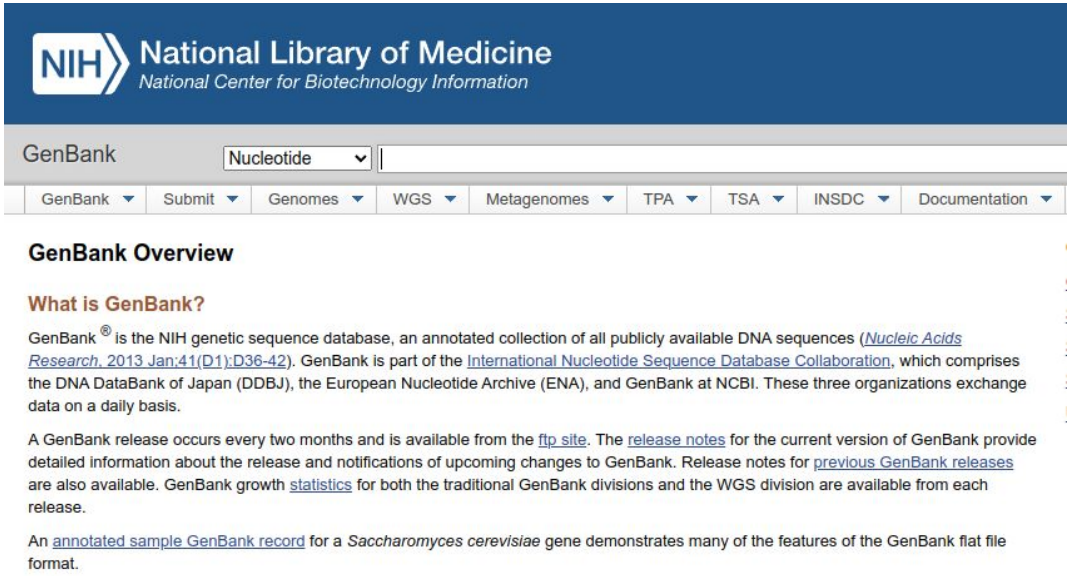
- eFP = electronic Fluorescent Pictographic
- Visualization of gene expression with organ/tissue resolution
- Multiple plant species are covered



http://bar.utoronto.ca/efp2/

GenBank

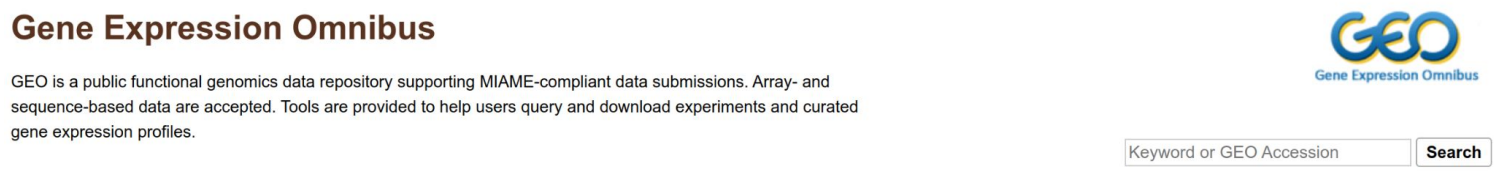
- Collection of sequences described in scientific publications
- Freely accessible database
- Partners: European Nucleotide Archive (ENA) and DNA DataBank of Japan (DDBJ)



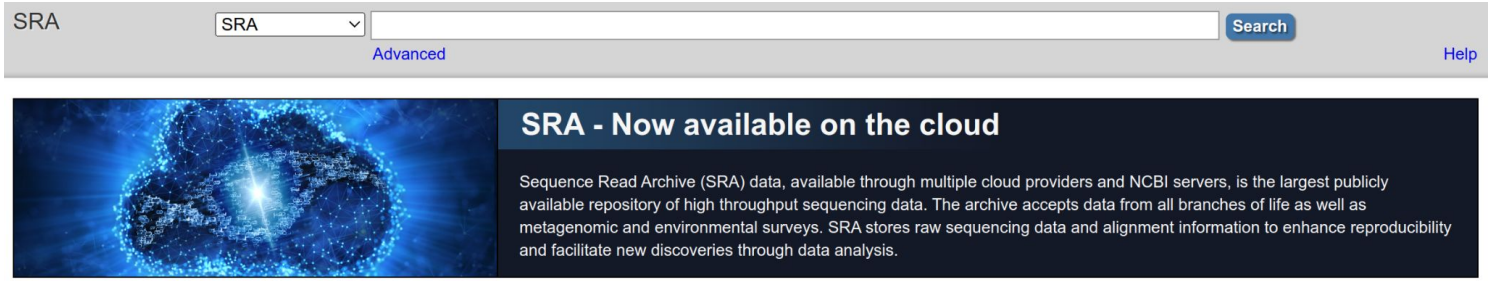
https://www.ncbi.nlm.nih.gov/genbank/

RNA-seq databases

- Gene Expression Omnibus (GEO)

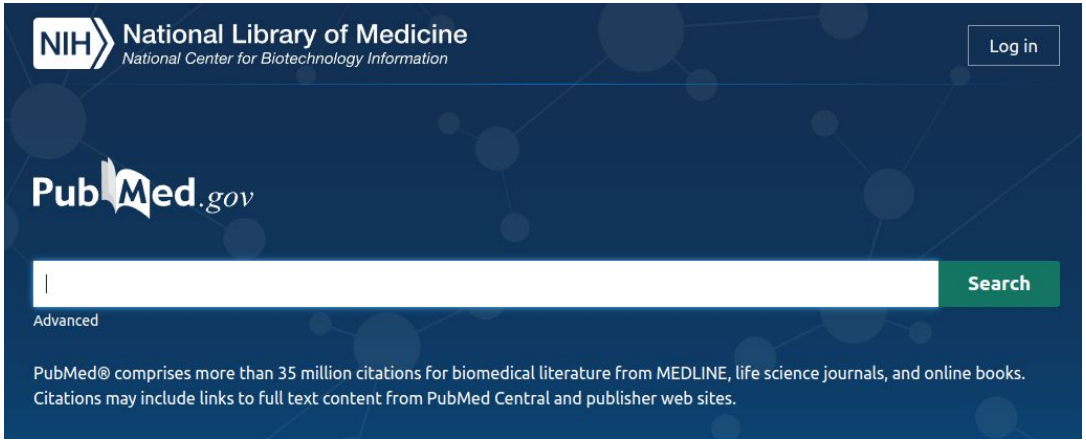


- Sequence Read Archive (SRA)



PubMed

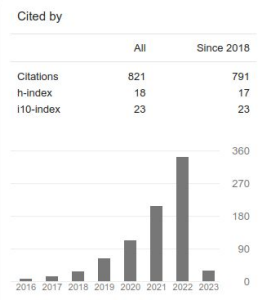
- Collection of scientific papers in journals with connection to medicine
- Many full text directly available through PubMed
- “Citations” and “Cited by” provide connections to related articles

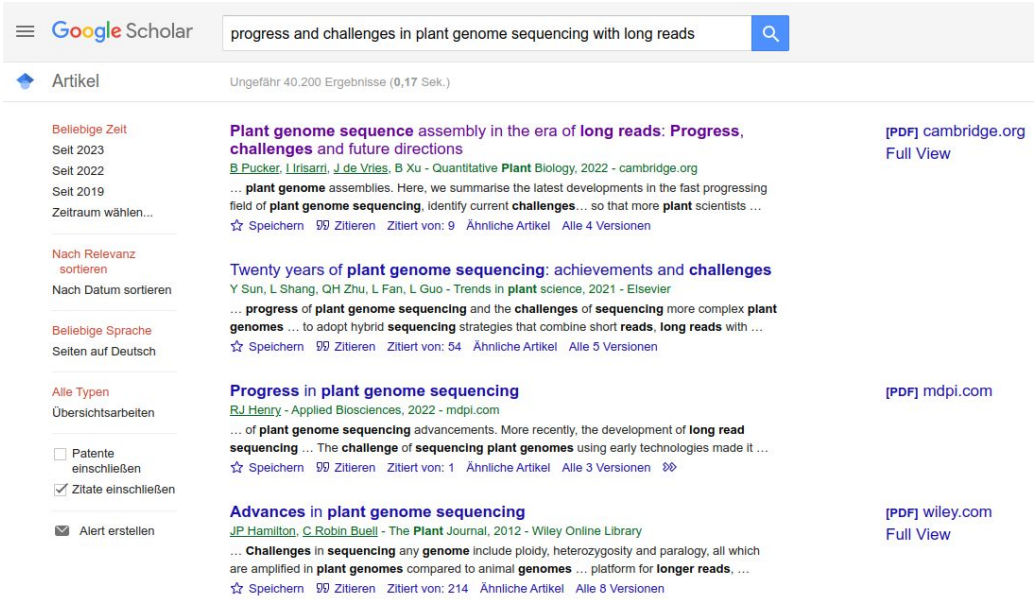


https://pubmed.ncbi.nlm.nih.gov/

Google Scholar

- Excellent search algorithm to find relevant scientific publications
- Provides information about the number of citations
- Export options for citations
- Profiles for researchers allow quick access to publication lists

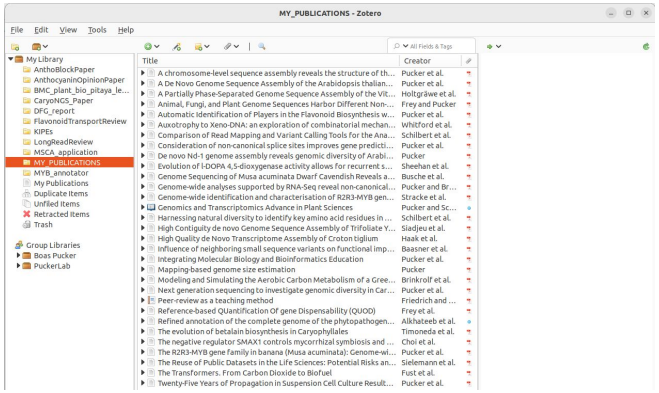




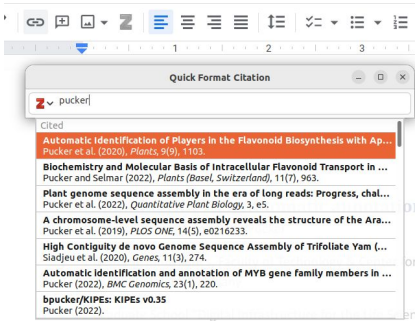
https://scholar.google.de/

Literature management tools

- Zotero (free, desktop client + web profile):
  - Web browser plugin for automatic saving of references
  - Processor plugin for LibreOffice, Google docs, Microsoft Office

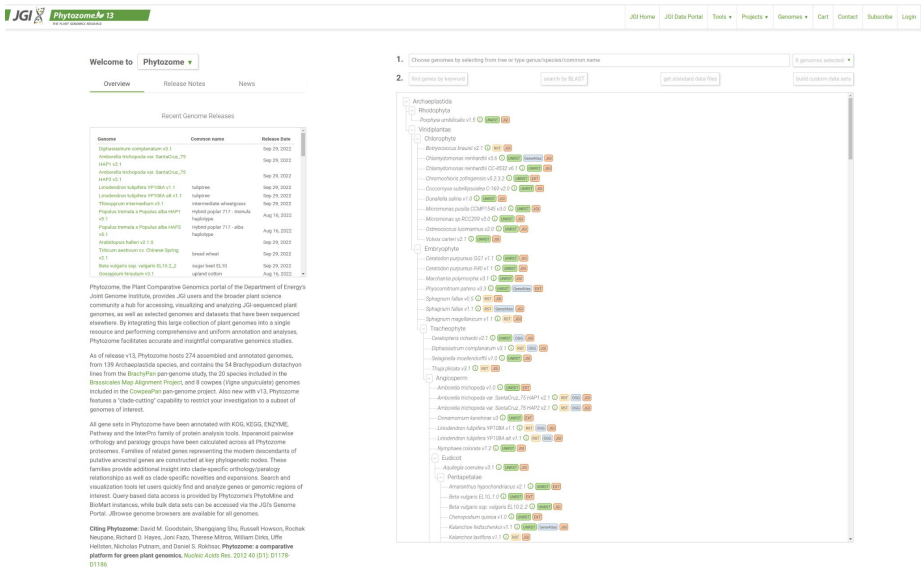


- Citavi (campus license)
- RefWorks (commercial, web-based)
- Mendeley (Elsevier, desktop client)
- EndNote (commercial, desktop client)
- BibTeX (free, desktop client)
- Read Cube Papers (commercial, desktop client)



Phytozome

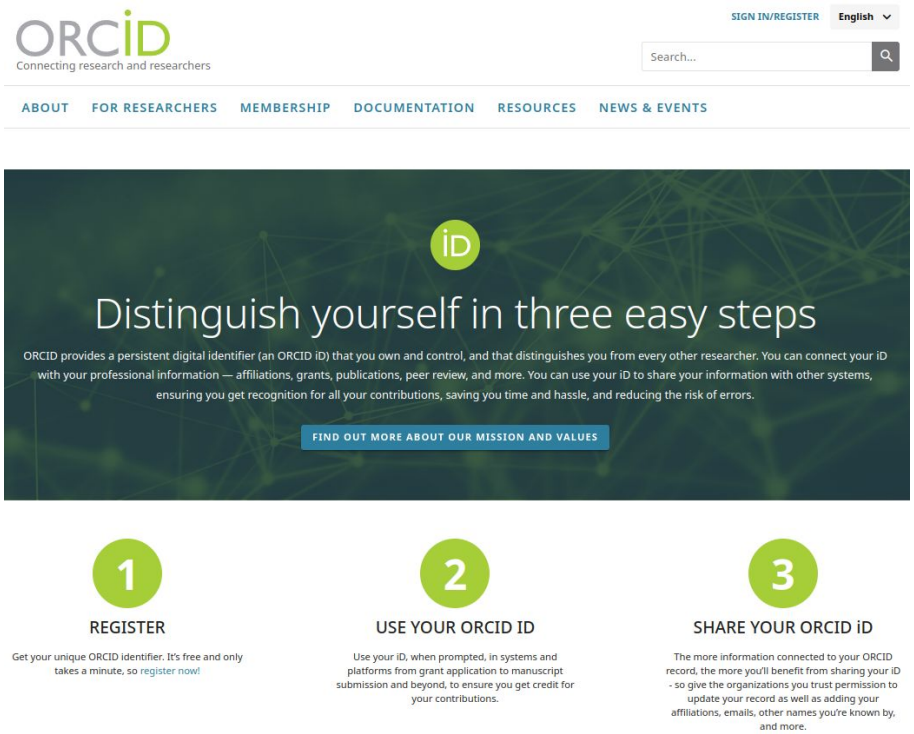
- Largest database of plant genome sequences and corresponding annotation
- Web portal provides convenient access to these data sets
- Comparative genomics support through website



https://phytozome-next.jgi.doe.gov/

ORCiD for authentication

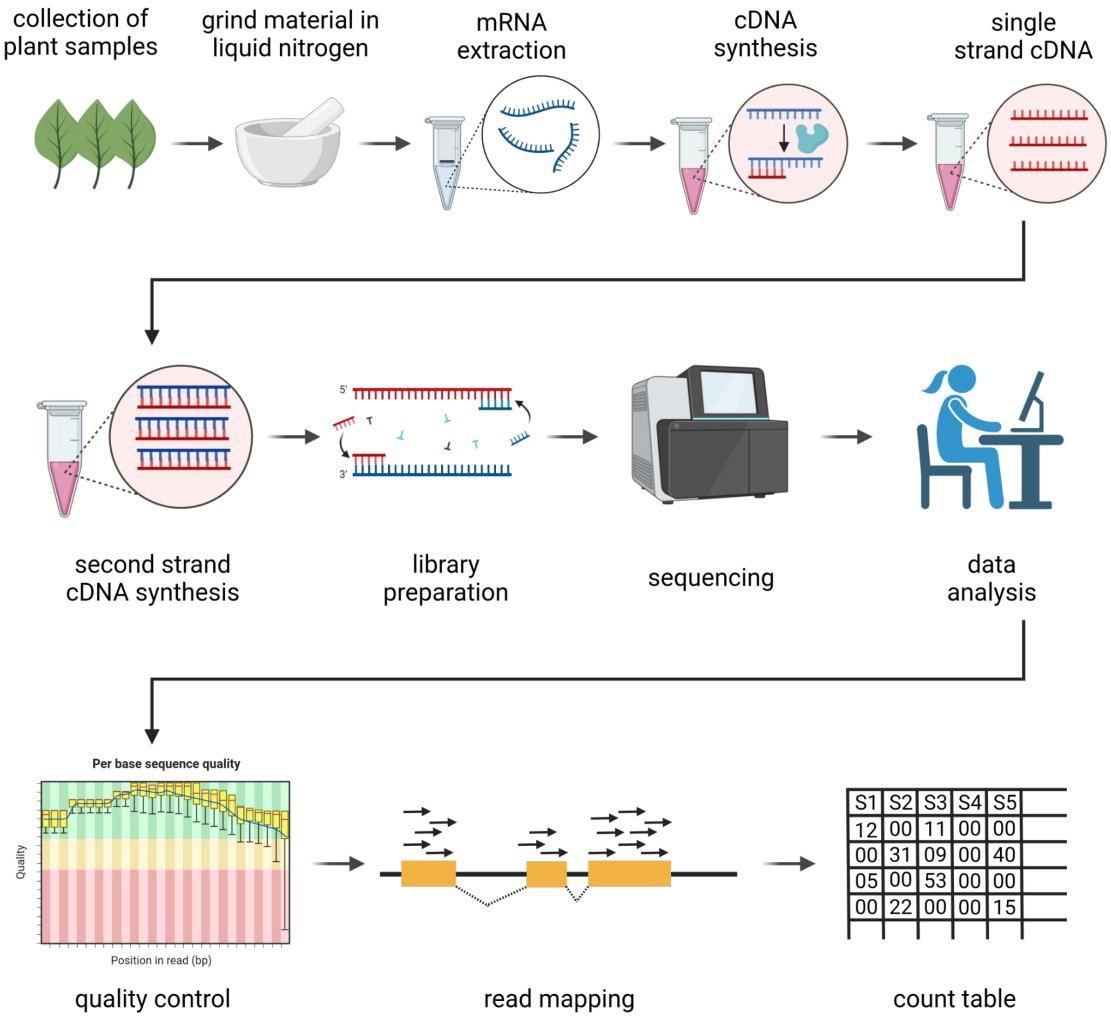
- ORCiD = Open Researcher and Contributor ID
- Invented as unique identifier for scientists to ensure proper assignment of publications
- Established as an authentication system for many journals and other science-related websites



https://orcid.org/

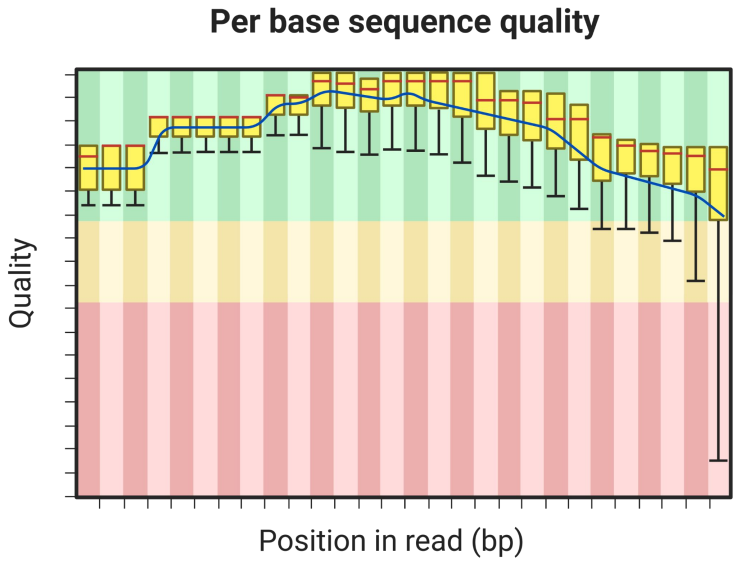
Gene expression analysis

RNA-seq workflow (reminder)

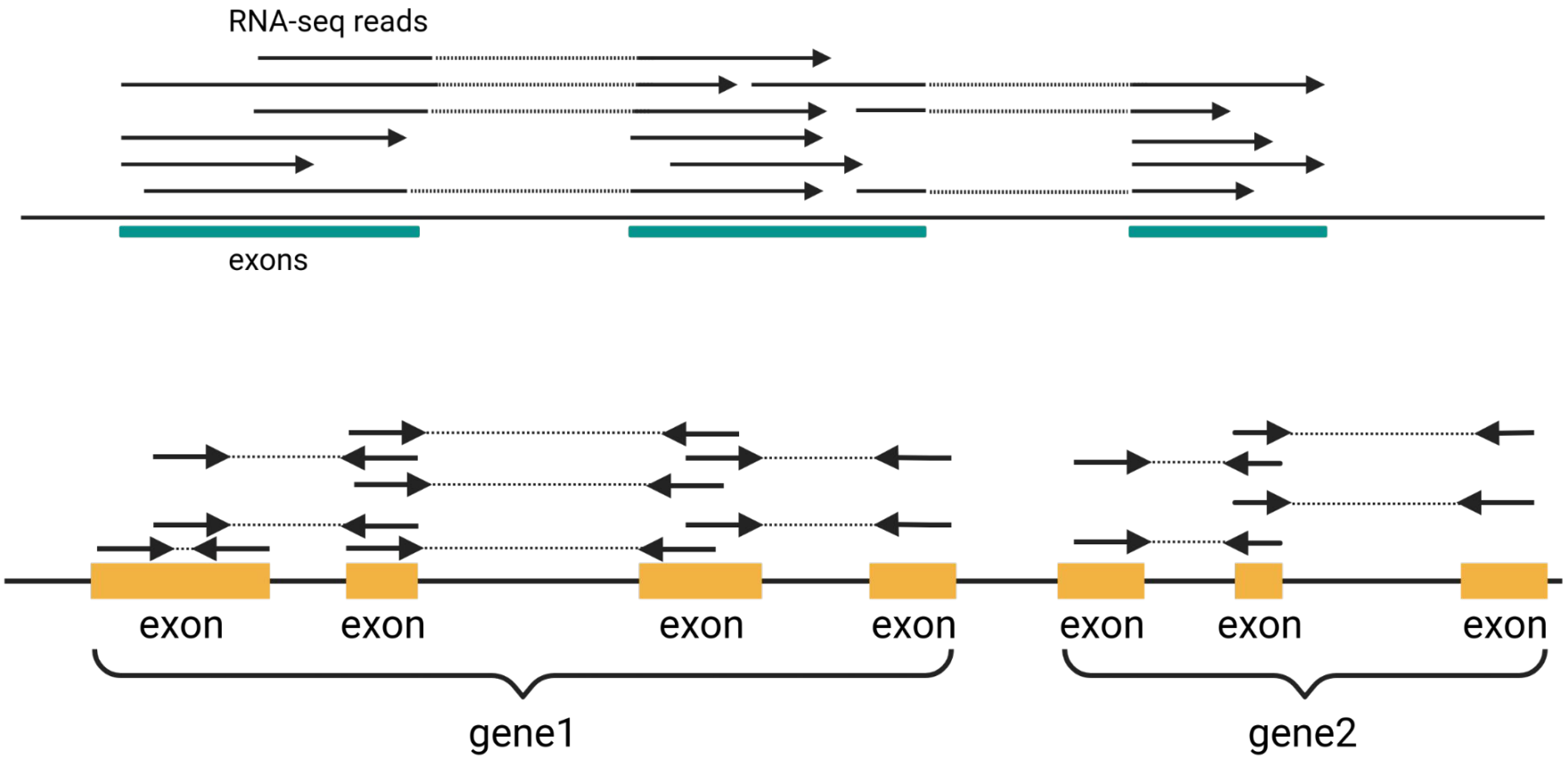


Quality checks

- Number of reads
- Adapter contamination
- Proportion of rRNA reads
- Contamination with genomic reads
- Quality of individual reads: Phred score

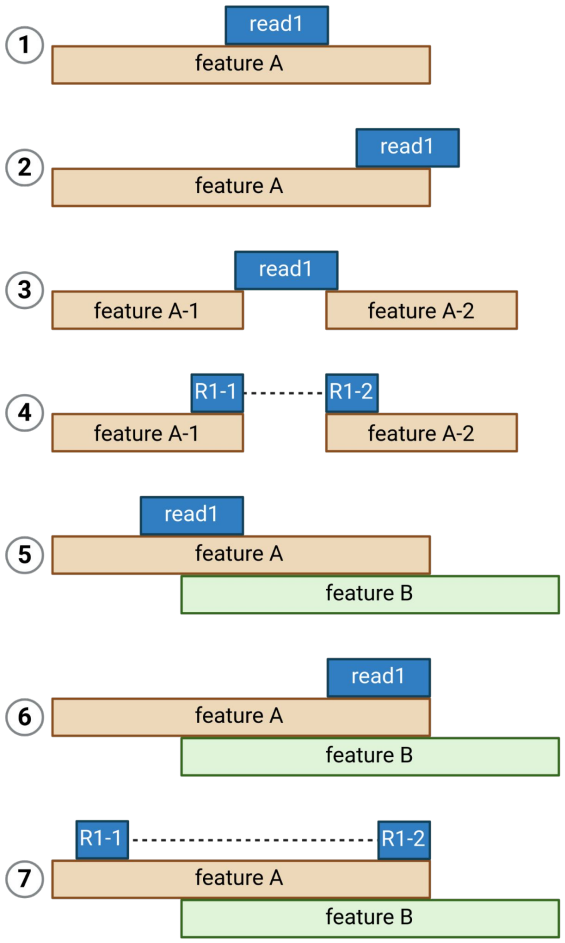


Read mapping



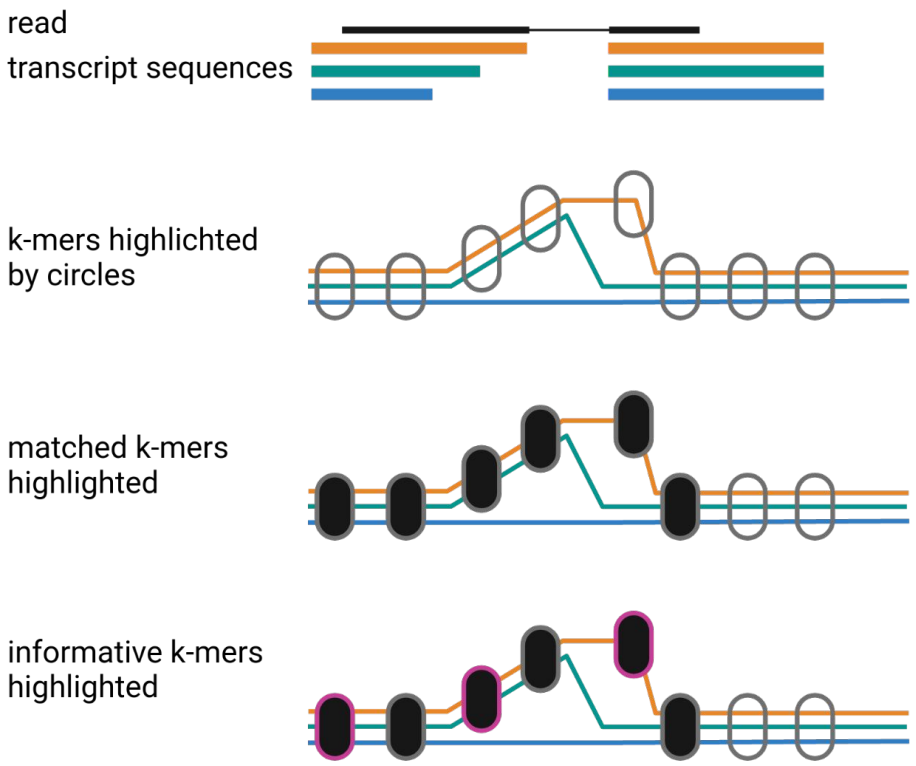
Counting mapped reads

- Reads can be counted in different ways
- Unique assignment to one feature
- Proportional assignment to different features
- Tools:
  - STAR/HISAT2 for RNA-seq read mapping
  - featureCounts for quantification



Mapping-free quantification: kallisto

- Much faster than propper read mappings with STAR/HISAT2
- Context of k-mers is considered and not only individual k-mers



Count tables

- Gene/transcript IDs in first column
- Sample IDs in first row
- Expression values in all fields of the table

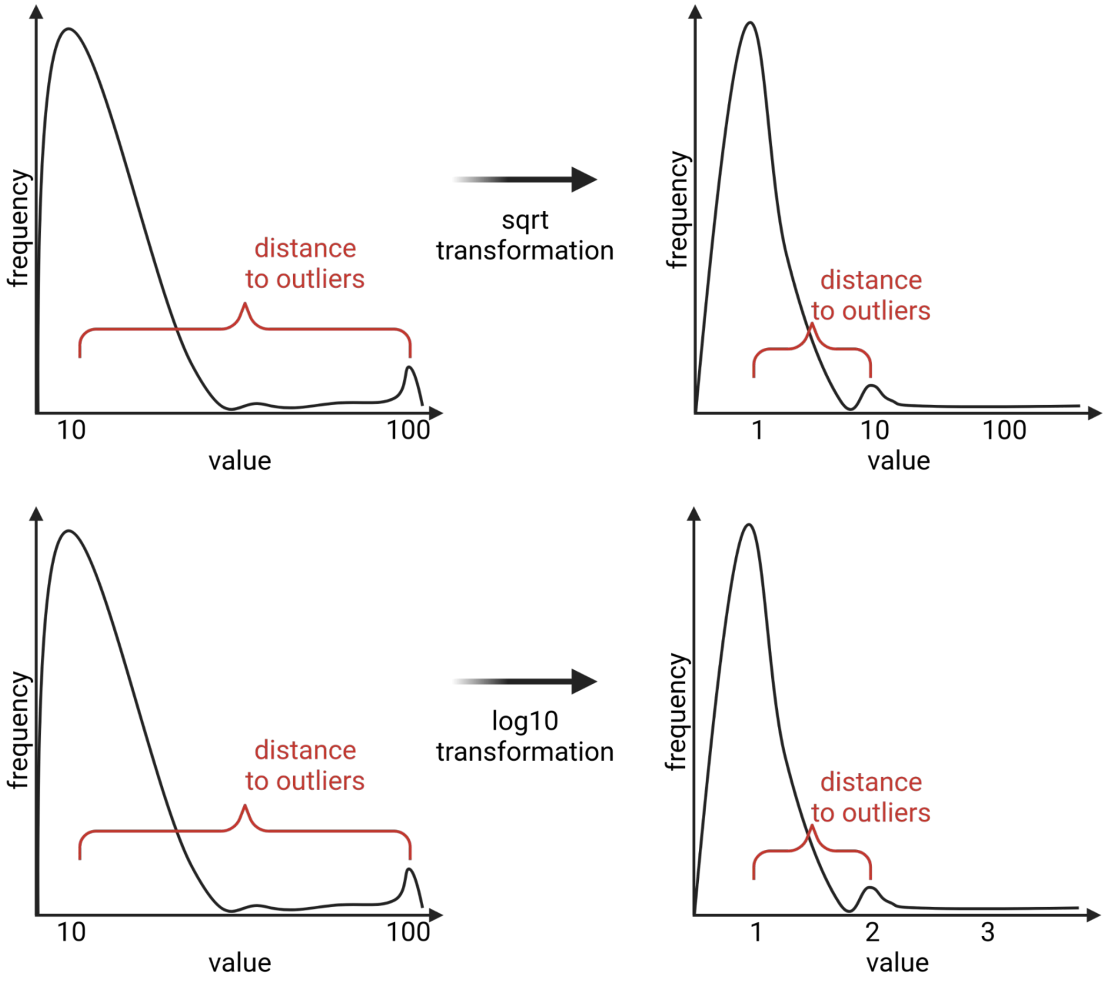
gene	SRR11603178	SRR11603179	SRR11603180	SRR11603181	SRR11603182	SRR11603183	SRR11603184	SRR11603185
TRINITY_DN10003_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10005_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	2.0
TRINITY_DN10005_c1_g1_i1	0.0	0.0	0.0	0.0	1.0	1.0	2.0	0.0	0.0	1.0	1.0	0.0	0.0
TRINITY_DN10008_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0
TRINITY_DN10014_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
TRINITY_DN10016_c0_g1_i1	0.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
TRINITY_DN10016_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
TRINITY_DN10018_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10019_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0
TRINITY_DN10029_c0_g1_i1	12.0	4.0	3.0	1.0	9.0	2.0	63.0	1.0	8.0	36.0	7.0	48.0	13.0
TRINITY_DN1002_c0_g1_i1 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	1.0	1.0	2.0	0.0
TRINITY_DN10032_c0_g1_i1	3.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	7.05039	11.2251
TRINITY_DN10032_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.94961	1.77489
TRINITY_DN10034_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0
TRINITY_DN10037_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	5.0
TRINITY_DN10038_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	2.0
TRINITY_DN10040_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0
TRINITY_DN10041_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	2.0
TRINITY_DN10043_c0_g1_i1	0.0	0.0	0.0	0.0	2.0	3.0	1.0	0.0	0.0	1.0	1.0	5.0	4.0
TRINITY_DN10049_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0
TRINITY_DN10050_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0
TRINITY_DN10052_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
TRINITY_DN10053_c0_g1_i1	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.0	0.0
TRINITY_DN10054_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
TRINITY_DN10054_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
TRINITY_DN10056_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0
TRINITY_DN10057_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	3.0	0.0	1.0
TRINITY_DN10058_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10058_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10059_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

https://pub.uni-bielefeld.de/record/2956788

Normalized gene expression (units of gene expression)

- Huge variation over orders of magnitude
- Raw counts = Number of reads assigned to a gene/transcript
- CPMs = counts per million counts (reads per million reads)
- RPKMs = Reads per kb exon per million reads
- FPKMs = Fragments per kb exon per million fragments
- TPMs = Transcripts per million transcripts

Other normalization methods

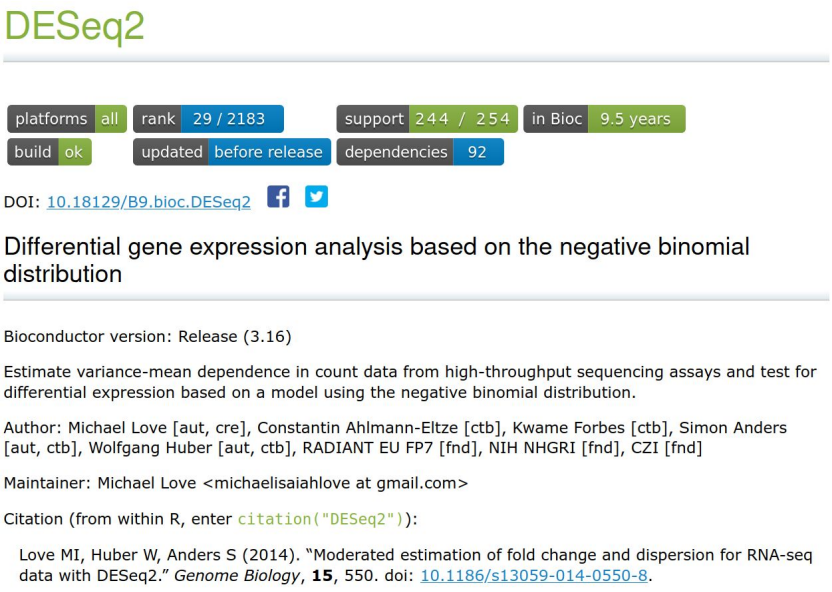


DEG identification

- DEG = Differentially Expressed Gene
- p-value describes chances of expression difference by chance
- logFC = log(Fold Change) describes difference between conditions
- Correction for multiple tests necessary in DEG analyses

DESeq2

- Frequently deployed R package for the identification of DEGs
- Do not confuse this with DEGseq (suspicious tool working without replicates)
- Internal normalization and calculations of statistics with necessary corrections
- Returns list of potential DEGs that can be filtered



edgeR

- R package for identification of differentially expressed genes
- More sensitive than DESeq2

edgeR

platforms all | rank 22 / 2183 | support 3 2 / 3 6 | in Bioc 14 years
build ok | updated before release | dependencies 10

DOI: 10.18129/B9.bioc.edgeR

Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.16)

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce read counts, including ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE and CAGE.

Author: Yunshun Chen, Aaron TL Lun, Davis J McCarthy, Matthew E Ritchie, Belinda Phipson, Yifang Hu, Xiaobei Zhou, Mark D Robinson, Gordon K Smyth

Maintainer: Yunshun Chen <yuchen at wehi.edu.au>, Gordon Smyth <smyth at wehi.edu.au>, Aaron Lun <infinite.monkeys.with.keyboards at gmail.com>, Mark Robinson <mark.robinson at imls.uzh.ch>

Citation (from within R, enter `citation("edgeR")`):

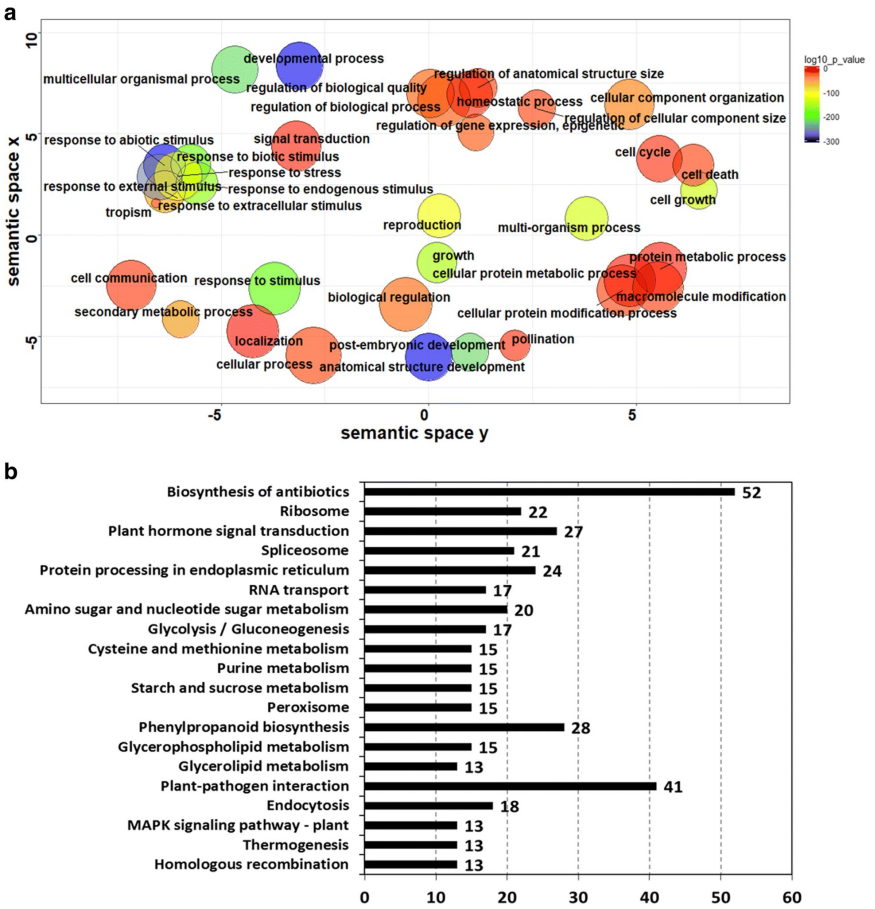
Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140. doi: 10.1093/bioinformatics/btp616.

McCarthy DJ, Chen Y, Smyth GK (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, **40**(10), 4288-4297. doi: 10.1093/nar/gks042.

Chen Y, Lun AAT, Smyth GK (2016). "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline." *F1000Research*, **5**, 1438. doi: 10.12688/f1000research.8987.2.

GO enrichment

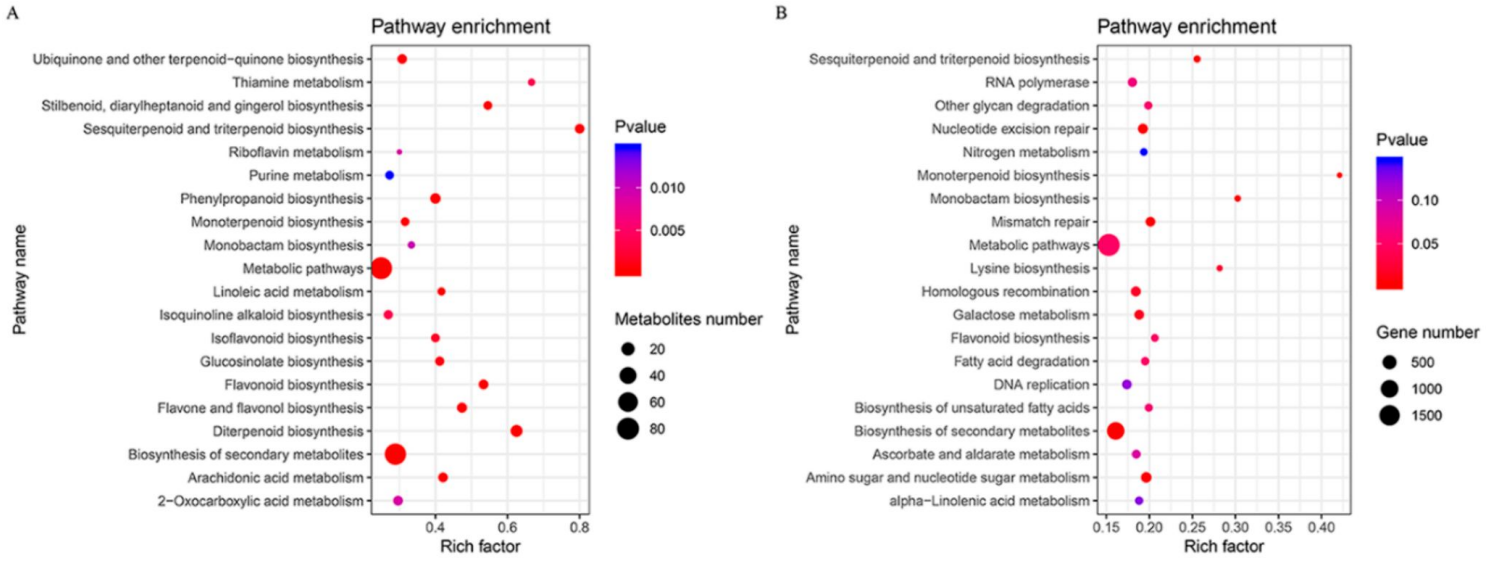
- GO = Gene Ontology
- GO term can be assigned to multiple genes (e.g. ‘flavonoid biosynthesis’)
- Up-regulated GO = larger proportion of genes with this GO term are up-regulated than expected by chance
- Tools: AmiGO, GOrilla, ShinyGO, GOnet, g:Profiler, …



https://doi.org/10.1186/s12864-018-5357-7

KEGG enrichment

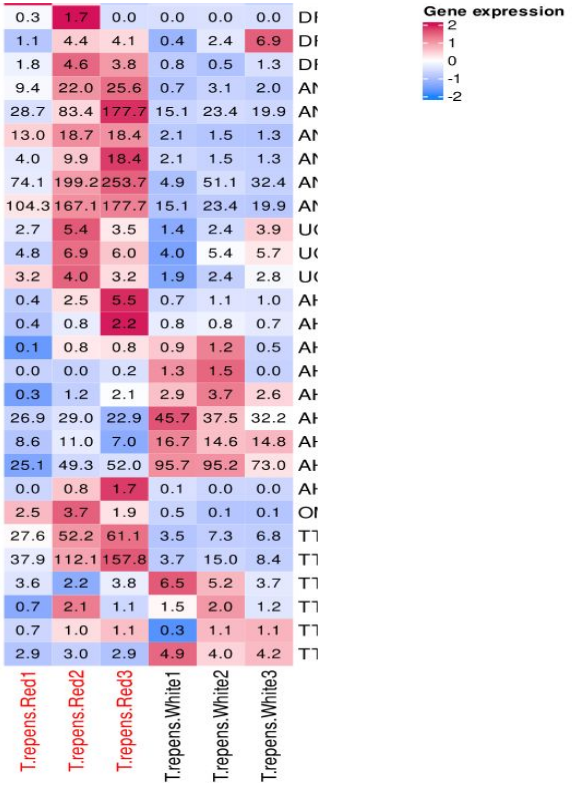
- Enrichment is based on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways
- Mercator can be used to assign genes to KEGG pathways
- Up-regulated pathway = large proportion of up-regulated genes in pathway



https://doi.org/10.3390/genes11020187

Heatmaps

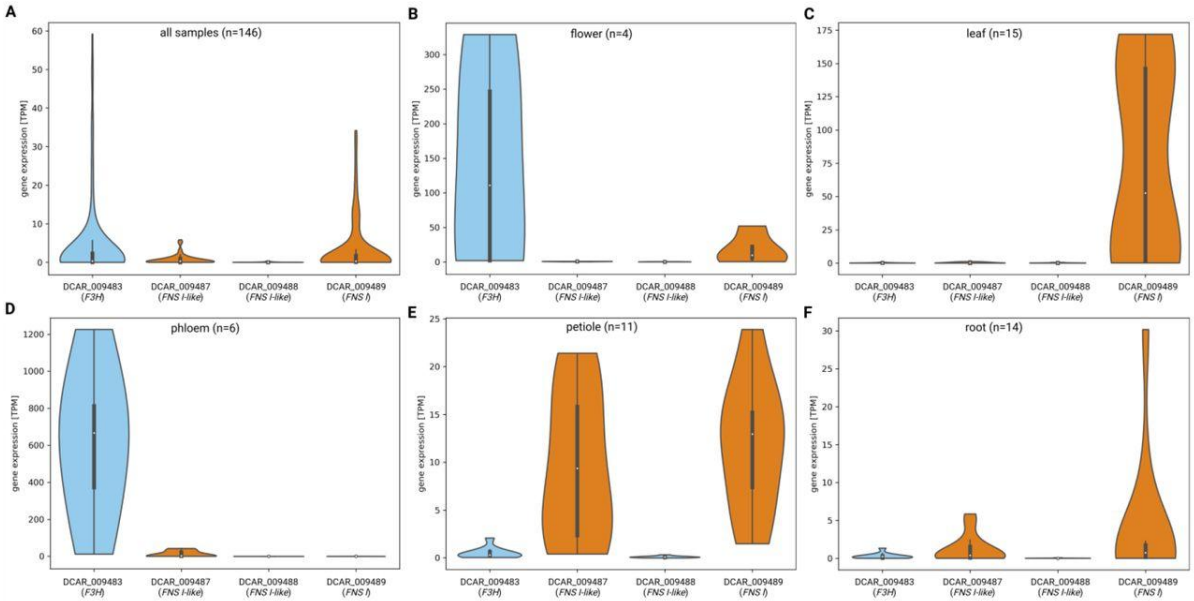
- Visualization of gene expression values based on color/color intensity
- Red (heat) often indication of high expression
- Different color ranges are available in Python/R
- Additional normalization per gene across all samples



Marin Recinos & Pucker, in preparation

Violin plots

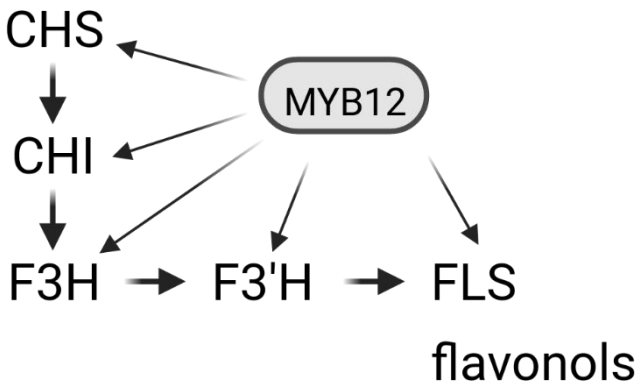
- Summarization of multiple data points
- Indication of data point distribution
- Mean and median can be displayed

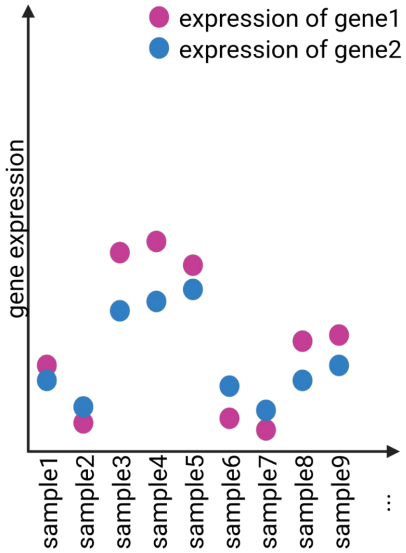


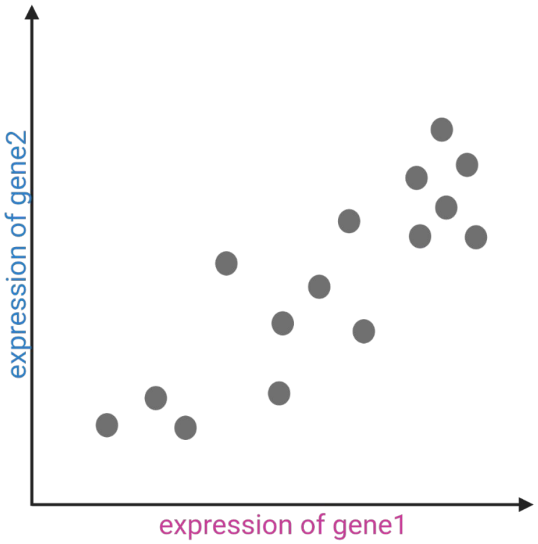
https://doi.org/10.1371/journal.pone.0280155

Co-expression analysis

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



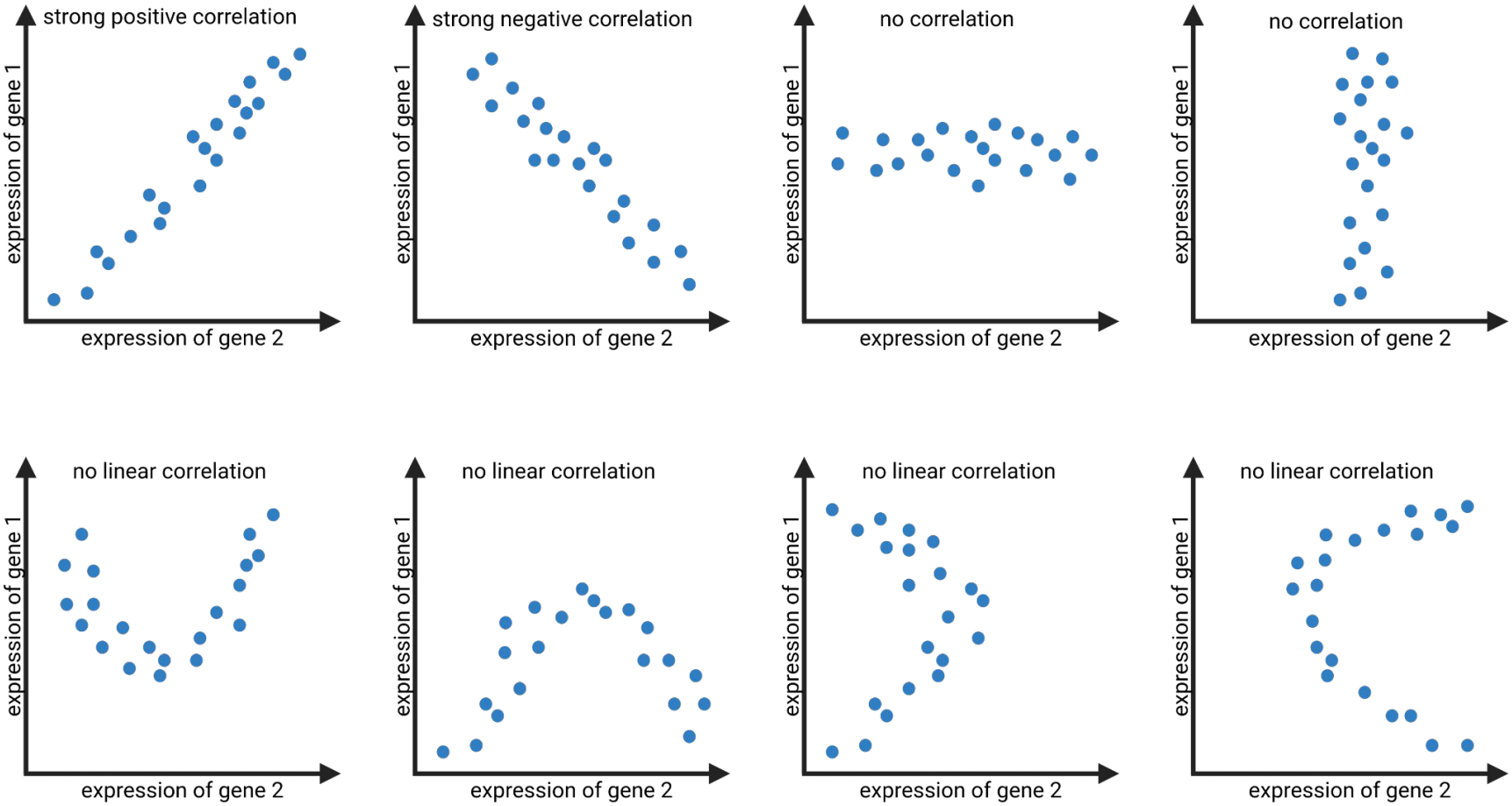




Correlation vs. causation

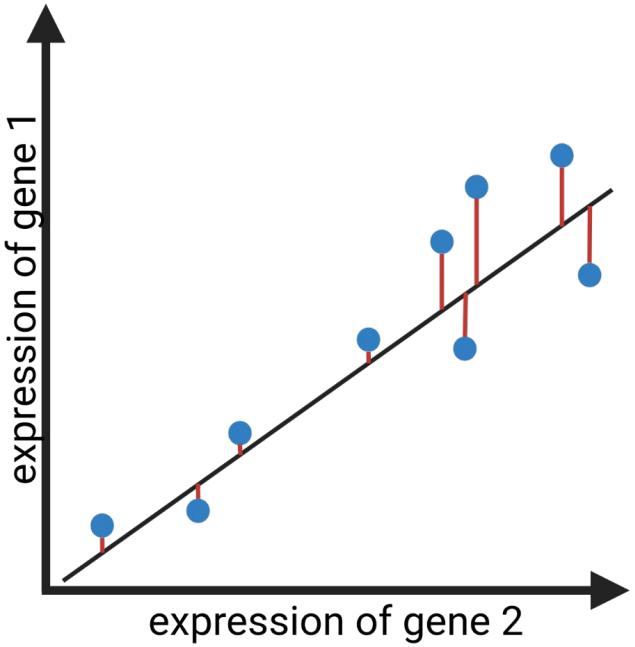
- Correlation does not allow conclusions about the direction i.e. cause and consequence
- Guilt-by-association is frequently used to find genes in a pathway/network
- Correlation can be caused by a direct or indirect connection

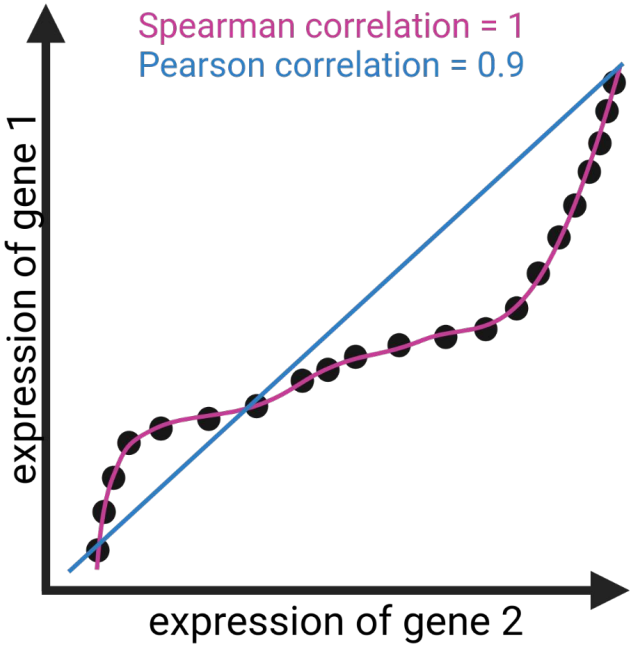
Examples


strong positive correlation
expression of gene 1
expression of gene 2
strong negative correlation
expression of gene 1
expression of gene 2
no correlation
expression of gene 1
expression of gene 2
no correlation
expression of gene 1
expression of gene 2
no linear correlation
expression of gene 1
expression of gene 2
no linear correlation
expression of gene 1
expression of gene 2
no linear correlation
expression of gene 1
expression of gene 2
no linear correlation
expression of gene 1
expression of gene 2

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

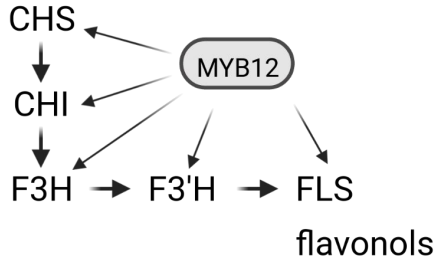
genes	correlation
CHS	0.7
CHI	0.5
F3H	0.45
F3'H	0.6
FLS	0.9

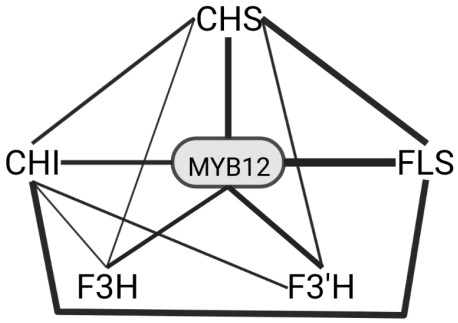
CHS

genes	correlation
CHI	0.4
F3H	0.2
F3'H	0.3
FLS	0.7

CHI

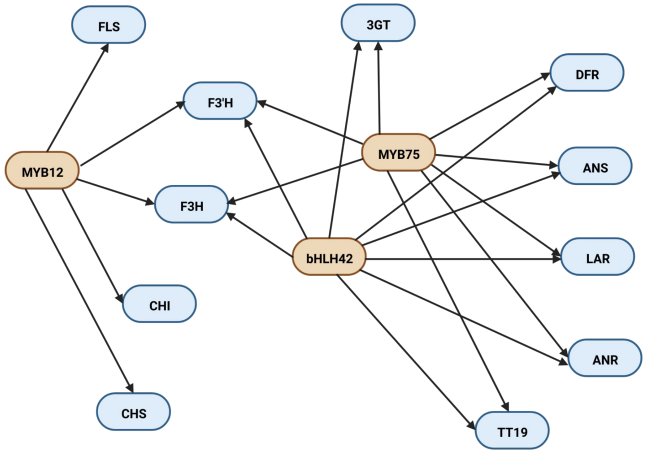
genes	correlation
F3H	0.2
F3'H	0.3
FLS	0.7





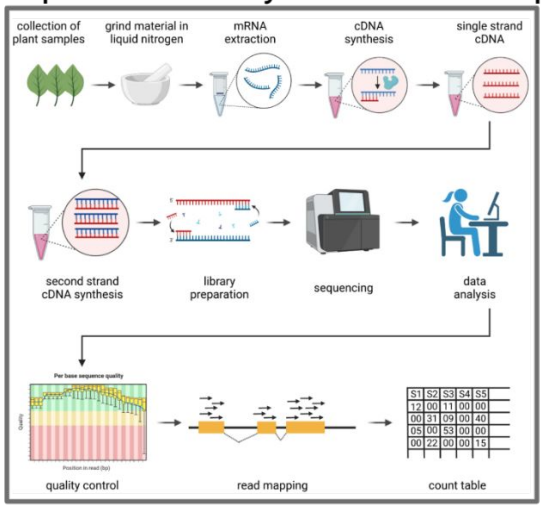
Gene expression networks

- Gene expression in multiple samples can be used to infer regulatory networks
- Edges in the network are based on co-expression
- Transcription factors represent central nodes
- Cytoscape can be used to visualize network files

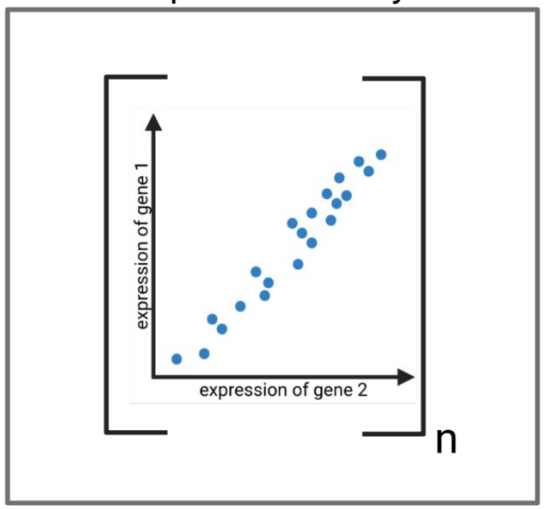


Summary: co-expression workflow

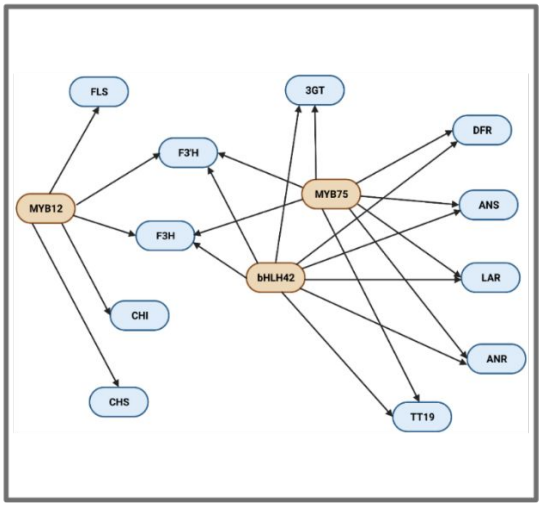
Expression analysis via RNA-Seq



Coexpression analysis



Network construction

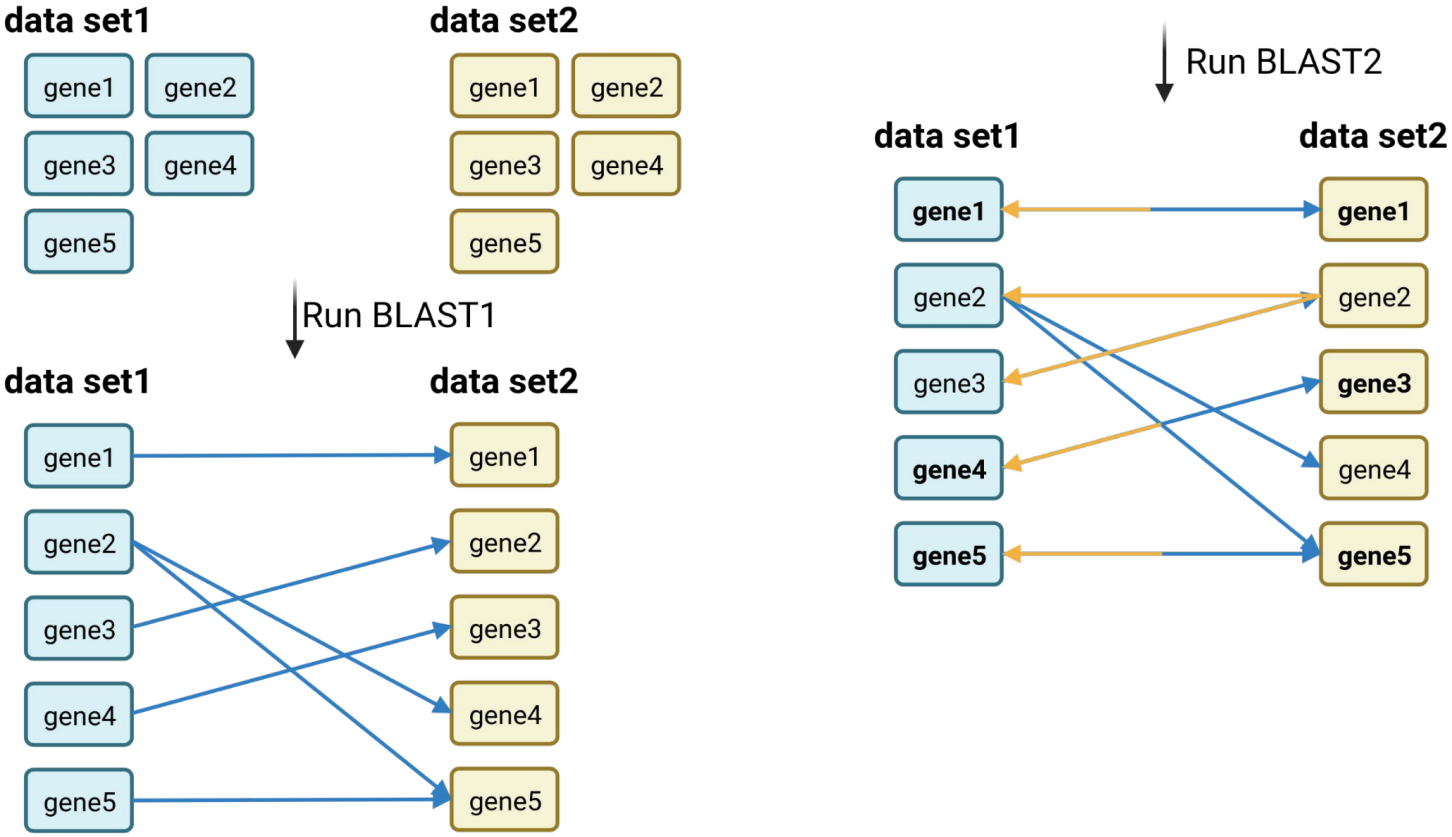


Assigning functional annotation

Functional annotation

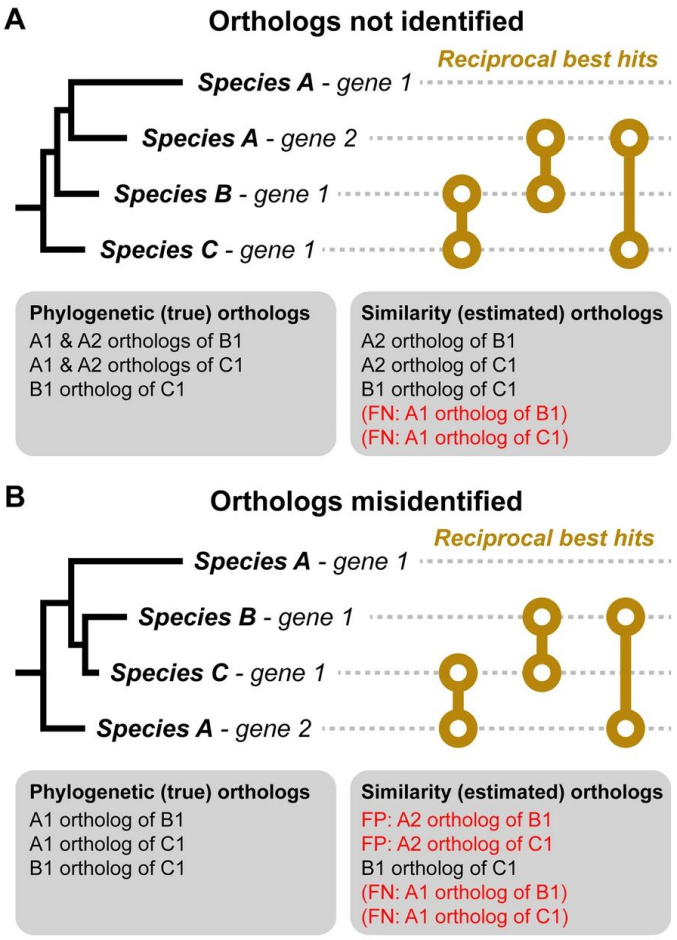
- Reciprocal Best BLAST Hits (RBHs)
- OrthoFinder2
- Knowledge-based Identification of Pathway Enzymes (KIPEs)
- Mercator
- InterProScan5

Reciprocal Best BLAST Hits (RBHs)



OrthoFinder2

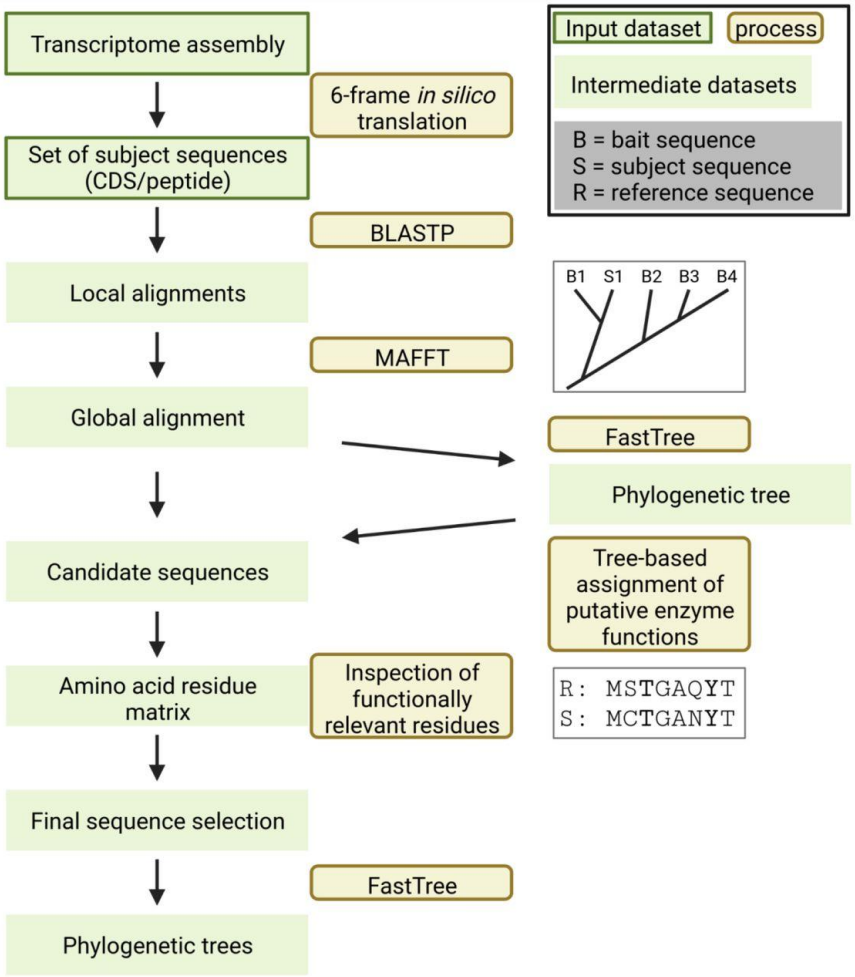
- Identification of orthologs often not possible
- Orthogroups are better reflection of reality
- Assumption: orthologs have the same function



https://doi.org/10.1186/s13059-019-1832-y

KIPEs

- KIPEs = Knowledge-based Identification of Pathway Enzymes
- Identification of genes involved in well-studied biosynthesis pathways
- Prediction about functionality of enzymes
- Requires existing knowledge from other species



https://doi.org/10.1101/2022.06.30.498365

Mercator

- Automatic assignment of functional annotation terms to peptide sequences
- Web server available (https://plabipd.de/portal/mercator-sequence-annotation)
- Basis for MapMan analysis

https://plabipd.de/portal/mercator-sequence-annotation
https://doi.org/10.1111/pce.12231

InterProScan

- Automatic annotation of sequences
- InterPro domain IDs are assigned to sequences
- KEGG pathway IDs and other information is included
- Search based on sequence similarity via DIAMOND

https://www.ebi.ac.uk/interpro/search/sequence/

Time for questions!

Questions

1. Which online tools are helpful for life scientists?
2. Which databases are important in molecular plant biology?
3. What are the important steps of an RNA-seq data analysis workflow?
4. Which correlation coefficients can be helpful in co-expression analyses?
5. How can gene expression data be visualized?
6. Which tools can be used to functionally annotate sequences?