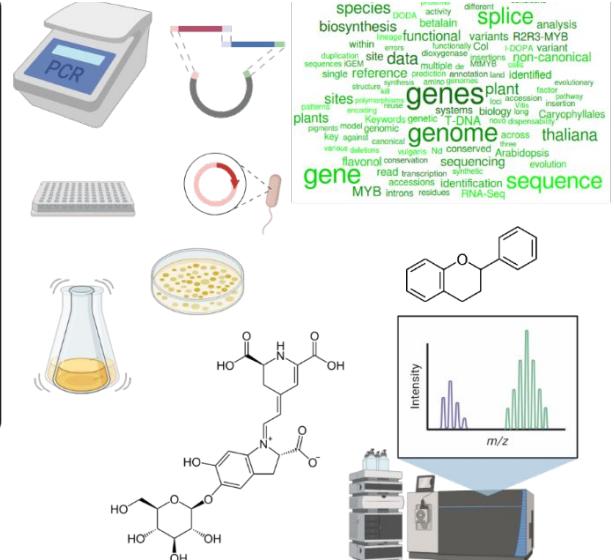
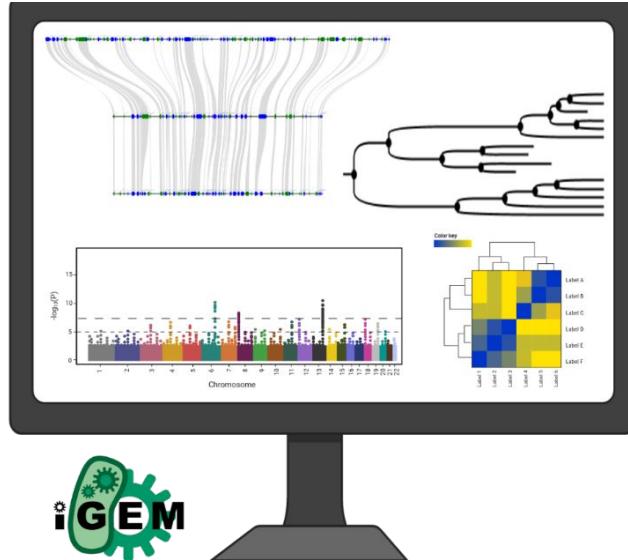
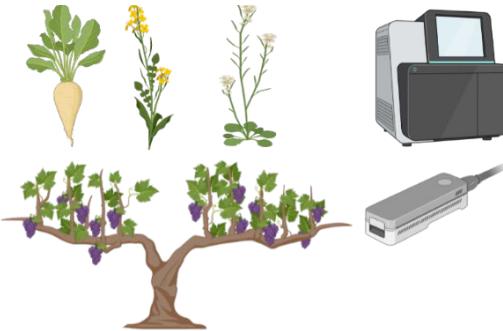




Technische
Universität
Braunschweig



Data analysis & basic bioinformatics

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

Availability of slides

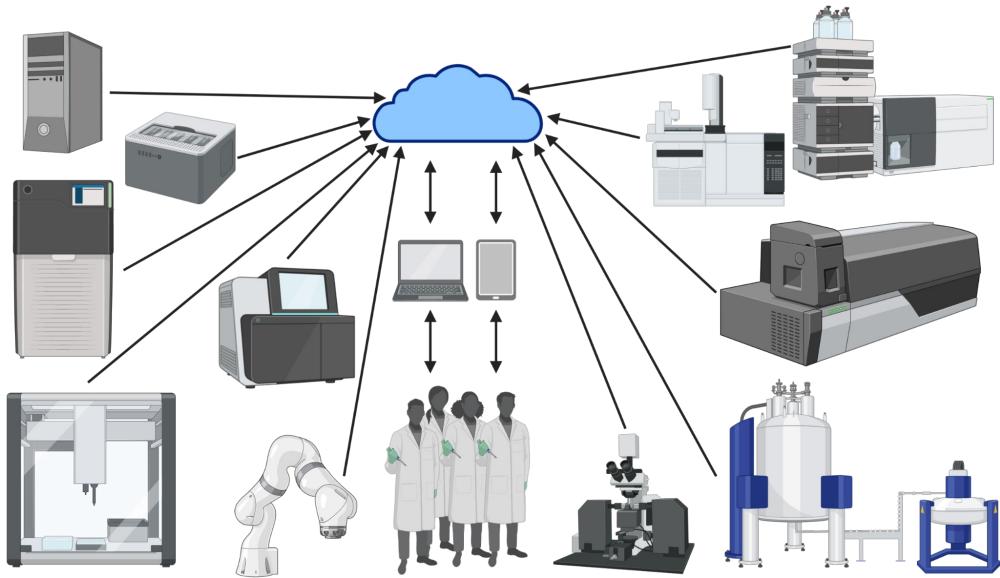
- All materials are freely available (CC BY) - after the lectures:
 - StudIP: LMChemBSc12
 - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: b.pucker[a]tu-bs.de

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.

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 ...

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

ElasticBLAST 1.0.0 is Now available!
ElasticBLAST version 1.0.0 has support for faster cheaper disks at AWS and better supports Kubernetes on GCP!

Mon, 09 Jan 2023 [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

<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

Multiple Sequence Alignment

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.

We have recently changed the default parameter settings for MAFFT. Alignments should run much more quickly and larger DNA alignments can be carried out by default. Please click the 'More options' button to review the defaults and change them if required.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of
AUTOMATIC

sequences in any supported format:

Or upload a file: No file chosen

Use a [example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your Parameters

OUTPUT FORMAT
Pearson/FASTA

The default settings will fulfill the needs of most users.

[More options...](#) (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

<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 tried (3x forward, 3x reverse)

EMBOSS Transeq

EMBOSS Transeq translates nucleic acid sequences to their corresponding peptide sequences. It can translate to the three forward and three reverse frames, and output multiple frame translations at once.

STEP 1 - Enter your input sequence

Enter or paste a DNA/RNA sequence in any supported format:

Or, upload a file: No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Select Parameters

FRAME	CODON TABLE
1	Standard Code

The default settings will fulfill the needs of most users.
 More options... (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

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



Select a plant



Arabidopsis



Tomato



Potato



Poplar

GenBank

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



The screenshot shows the official GenBank website. At the top, the NIH logo and "National Library of Medicine" are displayed, along with the subtitle "National Center for Biotechnology Information". Below this is a search bar with "GenBank" and "Nucleotide" selected. A navigation menu includes links for GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, Documentation, and Help. The main content area is titled "GenBank Overview". It features a section on "What is GenBank?", explaining it as the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. It mentions the International Nucleotide Sequence Database Collaboration involving DDBJ, ENA, and NCBI, and notes daily data exchange. It also describes the two-month release cycle, available via an ftp site, and provides release notes and growth statistics. An example GenBank record for *Saccharomyces cerevisiae* is mentioned.

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



RNA-seq databases

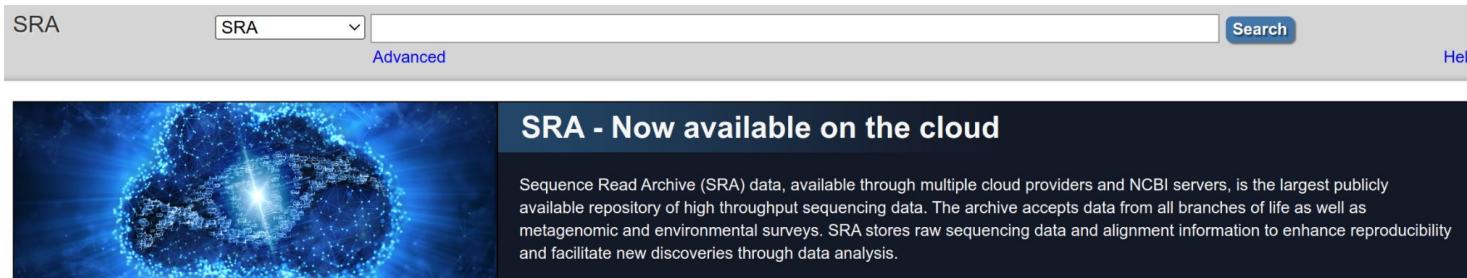
- Gene Expression Omnibus (GEO)

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 SRA Advanced Search Help

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.

PubMed

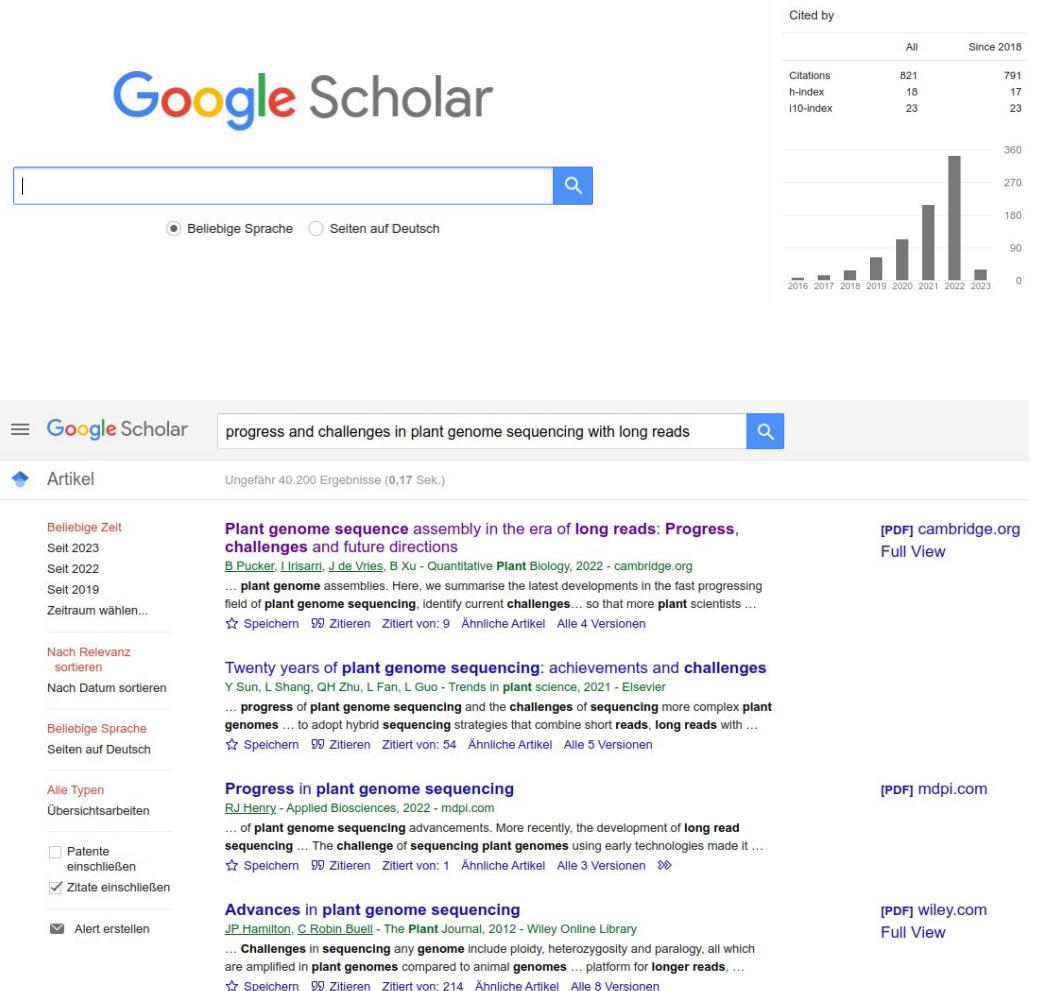
- Collection of scientific papers in journals with connection to medicine
- Many full text directly available through PubMed
- “Citations” and “Cited by” provide connections to related articles



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

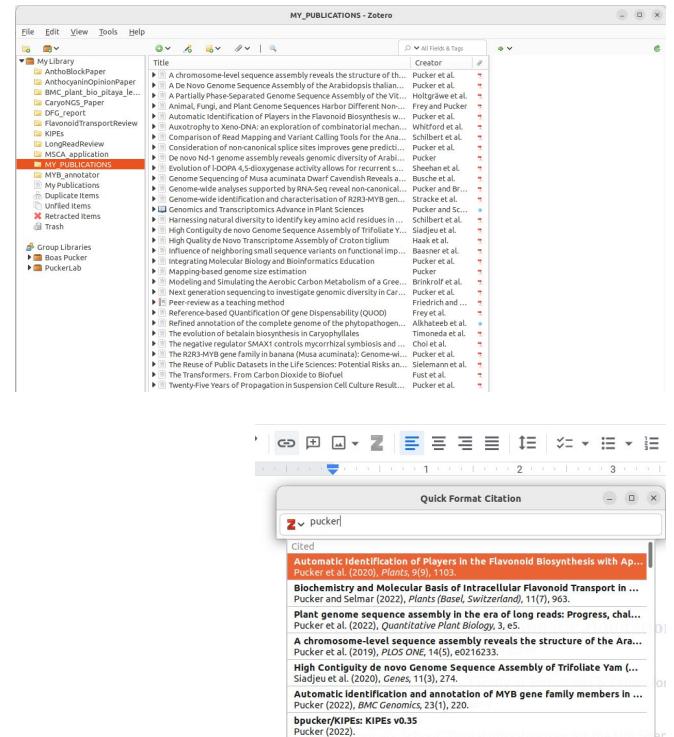
Google Scholar

- 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



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

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

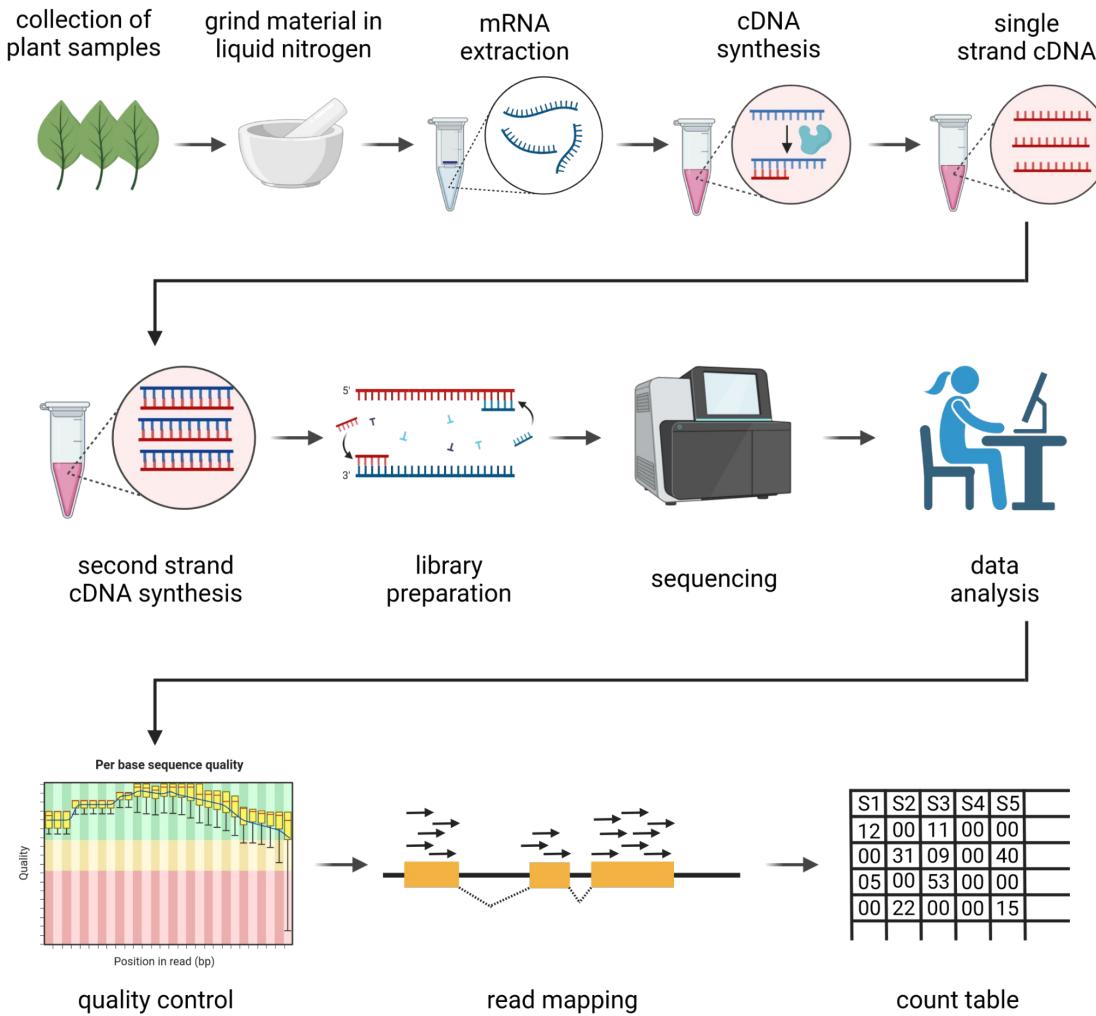
The screenshot shows the ORCID homepage with a dark green background featuring a network of connections. At the top right are links for "SIGN IN/REGISTER" and "English". Below is a search bar. A navigation bar includes "ABOUT", "FOR RESEARCHERS", "MEMBERSHIP", "DOCUMENTATION", "RESOURCES", and "NEWS & EVENTS". The main heading "Distinguish yourself in three easy steps" is centered, with a green "iD" icon above it. Below the heading is a brief description of what ORCID is and its benefits. A blue button says "FIND OUT MORE ABOUT OUR MISSION AND VALUES". The three steps are numbered 1, 2, and 3 in green circles:

- 1 REGISTER**
Get your unique ORCID identifier. It's free and only takes a minute, so [register now!](#)
- 2 USE YOUR ORCID ID**
Use your ID, when prompted, in systems and platforms from grant application to manuscript submission and beyond, to ensure you get credit for your contributions.
- 3 SHARE YOUR ORCID iD**
The more information connected to your ORCID record, the more you'll benefit from sharing your ID - so give the organizations you trust permission to update your record as well as adding your affiliations, emails, other names you're known by, and more.

Gene expression analysis

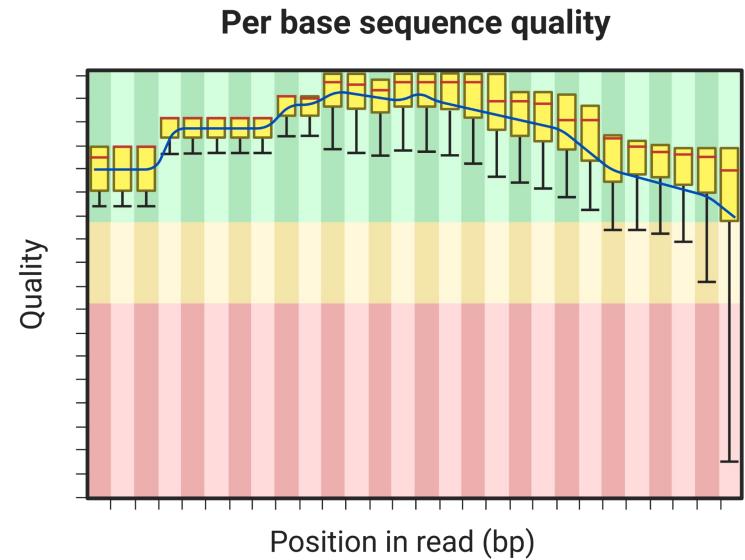


RNA-seq workflow (reminder)

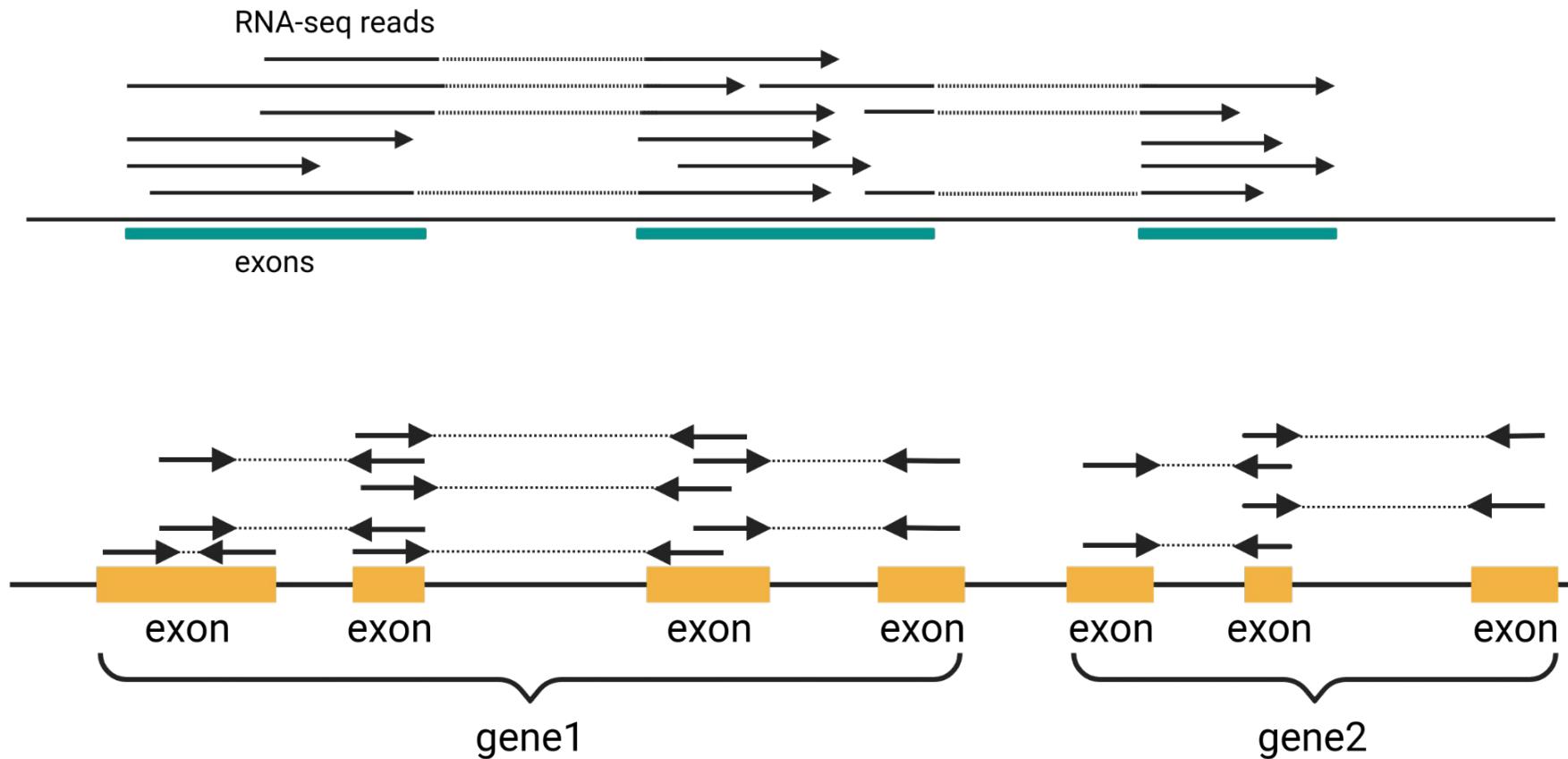


Quality checks

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

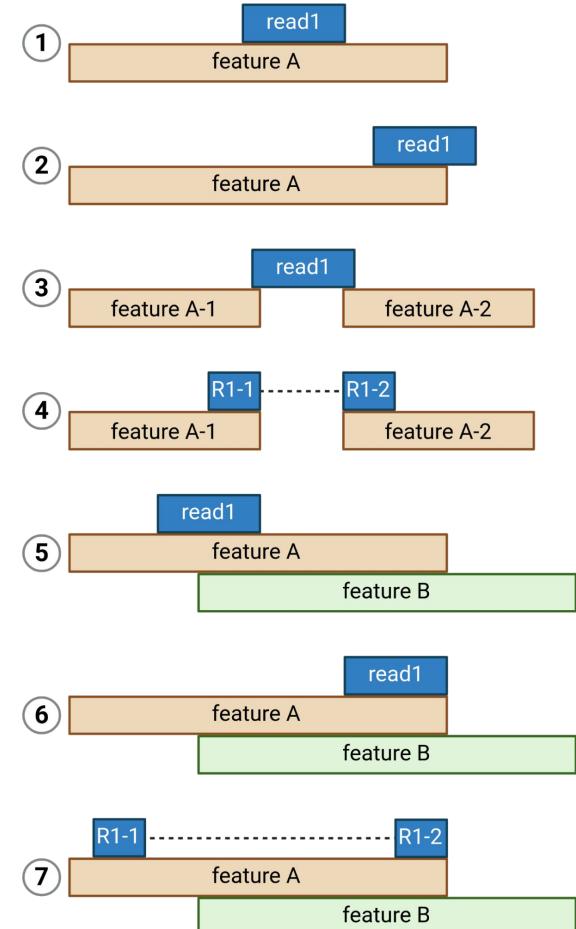


Read mapping



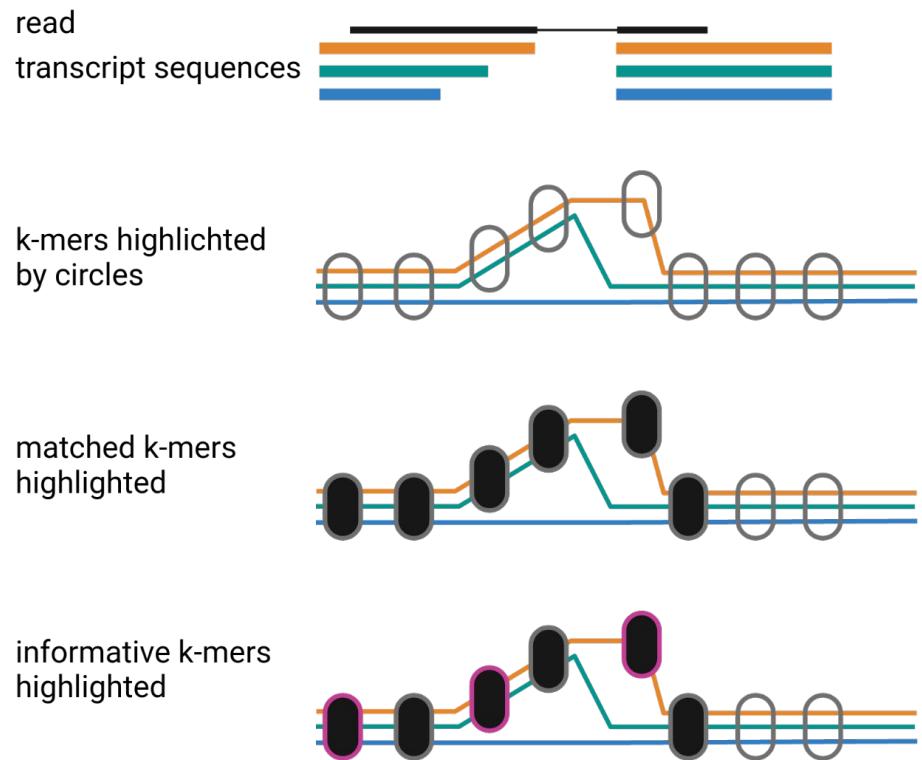
Counting mapped reads

- Reads can be counted in different ways
- Unique assignment to one feature
- Proportional assignment to different features



Mapping-free quantification: kallisto

- Much faster than proper 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

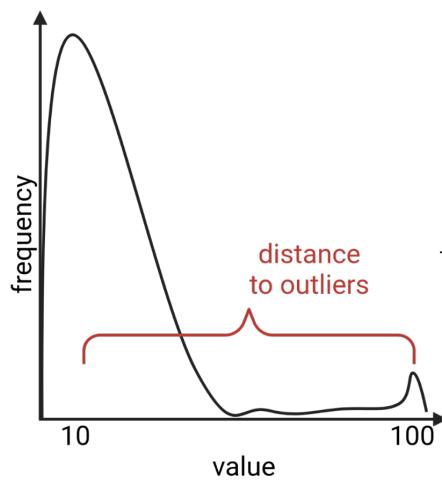
gene	SRR11603178	SRR11603179	SRR11603180	SRR11603181	SRR11603182	SRR11603183	SRR11603184	SRR11603185
TRINITY_DN10003_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10005_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
TRINITY_DN10005_c1_g1_i1	0.0	0.0	0.0	0.0	1.0	2.0	0.0	1.0
TRINITY_DN10008_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
TRINITY_DN10014_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10016_c0_g1_i1	0.0	2.0	0.0	0.0	1.0	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	1.0
TRINITY_DN10018_c0_g1_i1	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	2.0
TRINITY_DN10029_c0_g1_i1	12.0	4.0	3.0	1.0	9.0	2.0	63.0	1.0
TRINITY_DN1002_c0_g1_i1_0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0
TRINITY_DN10032_c0_g1_i1	3.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TRINITY_DN10032_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
TRINITY_DN10034_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10037_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
TRINITY_DN10038_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10040_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10041_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10043_c0_g1_i1	0.0	0.0	0.0	0.0	2.0	3.0	1.0	0.0
TRINITY_DN10049_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0
TRINITY_DN10050_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10052_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10053_c0_g1_i1	2.0	0.0	0.0	1.0	0.0	0.0	4.0	0.0
TRINITY_DN10054_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10054_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10056_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10057_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
TRINITY_DN10058_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
TRINITY_DN10058_c0_g1_i2	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

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

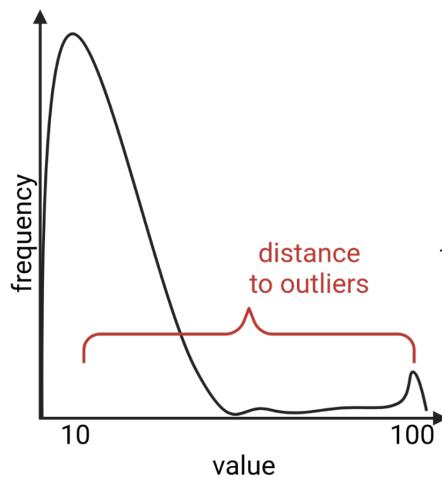
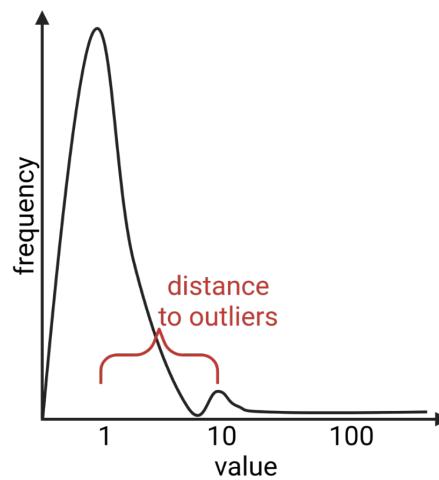
Normalized gene expression (units of gene expression)

- 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

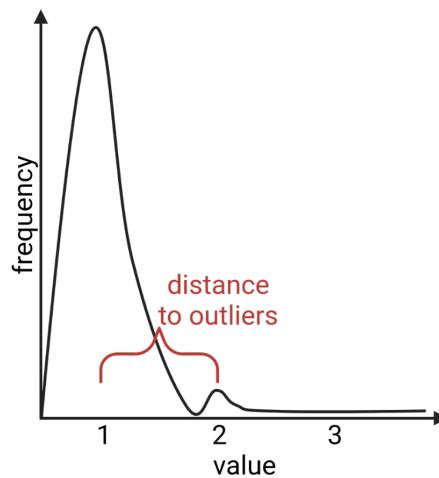
Other normalization methods



sqrt
transformation



log10
transformation



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

DESeq2

platforms all rank 29 / 2183 support 244 / 254 in Bioc 9.5 years
build ok updated before release dependencies 92

DOI: [10.18129/B9.bioc.DESeq2](https://doi.org/10.18129/B9.bioc.DESeq2) [f](#) [t](#)

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.16)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Author: Michael Love [aut, cre], Constantin Ahlmann-Eltze [ctb], Kwame Forbes [ctb], Simon Anders [aut, ctb], Wolfgang Huber [aut, ctb], RADIANT EU FP7 [fnd], NIH NHGRI [fnd], CZI [fnd]

Maintainer: Michael Love < michaelisaiahlove at gmail.com >

Citation (from within R, enter `citation("DESeq2")`):

Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550. doi: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).

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](https://doi.org/10.18129/B9.bioc.edgeR) [f](#) [t](#)

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@wehi.edu.au>, Gordon Smyth <smyth@wehi.edu.au>, Aaron Lun <infinite.monkeys.with.keys@gmail.com>, Mark Robinson <mark.robinson@imls.uzh.ch>

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

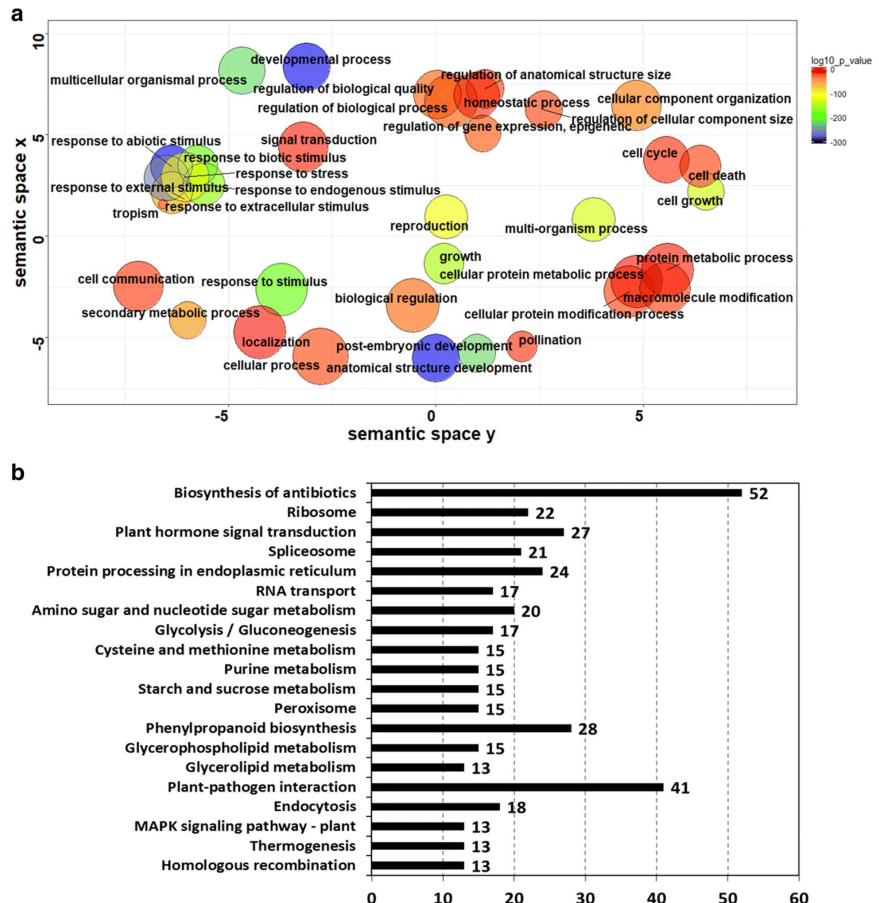
Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140. doi: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616).

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](https://doi.org/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](https://doi.org/10.12688/f1000research.8987.2).

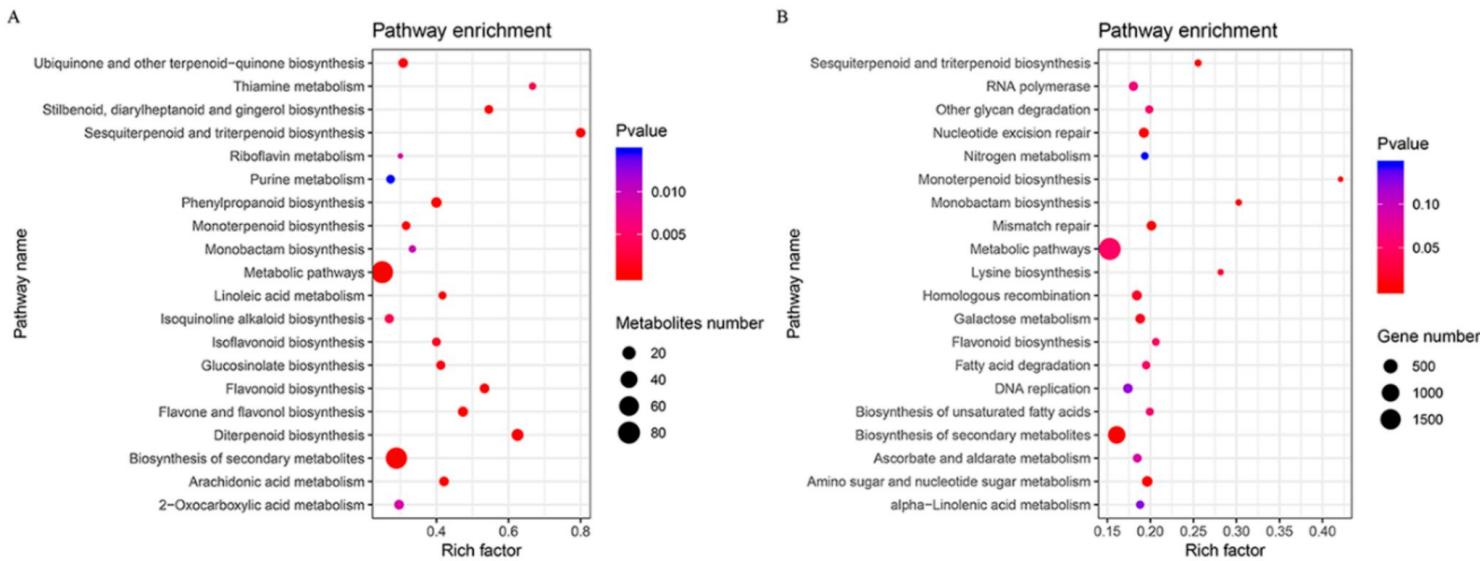
GO enrichment

- 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, ...



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

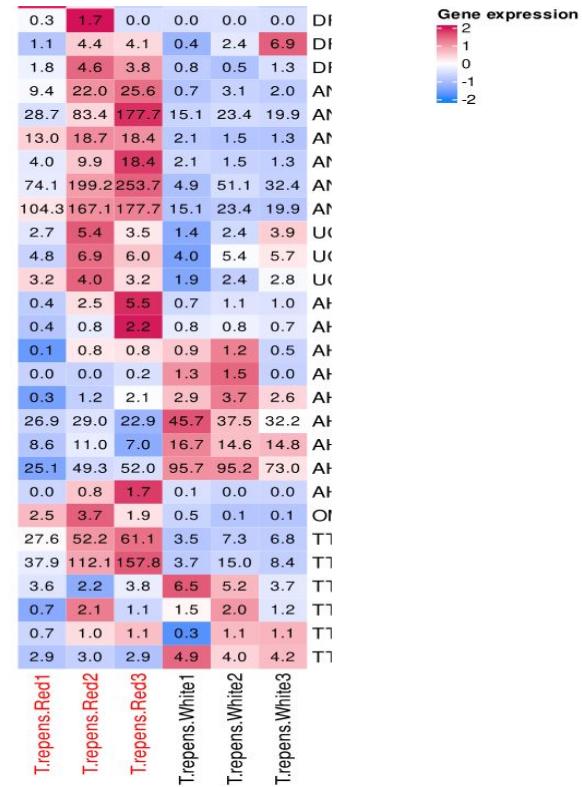


<https://doi.org/10.3390/genes11020187>



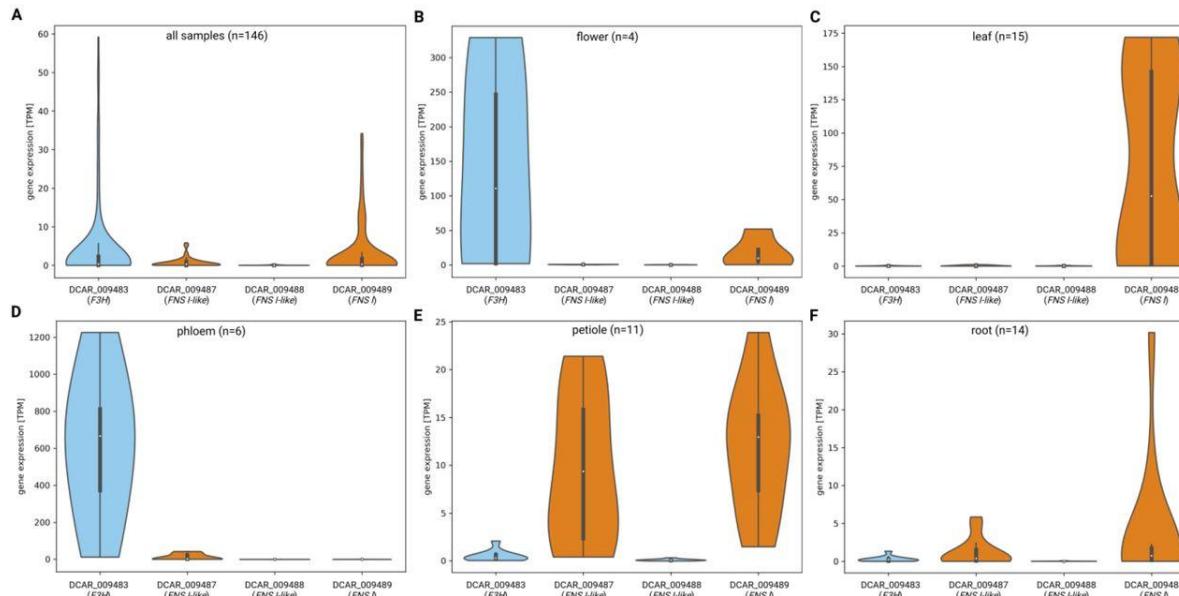
Heatmaps

- 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



Violin plots

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

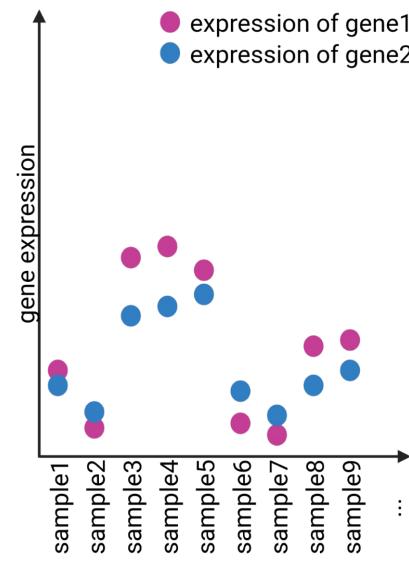
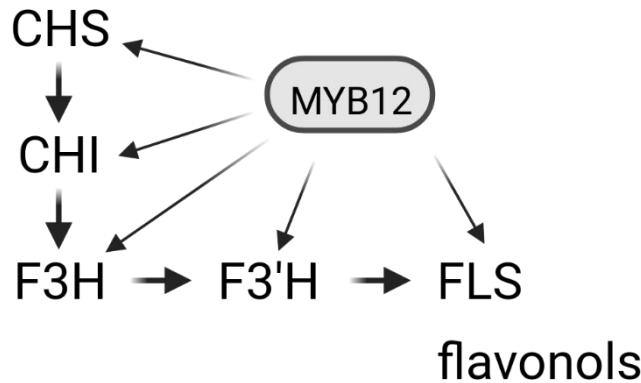


<https://doi.org/10.1371/journal.pone.0280155>



Co-expression analysis

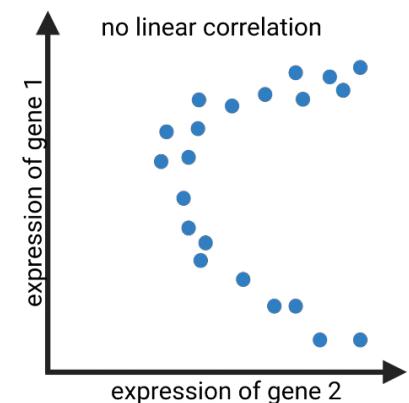
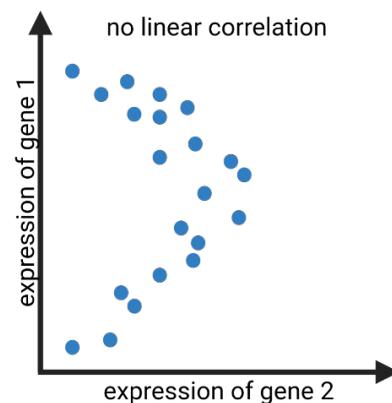
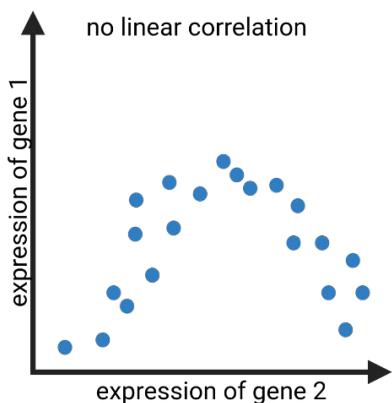
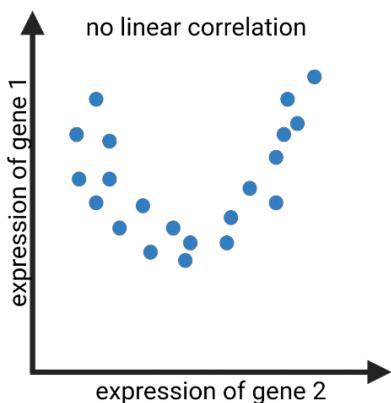
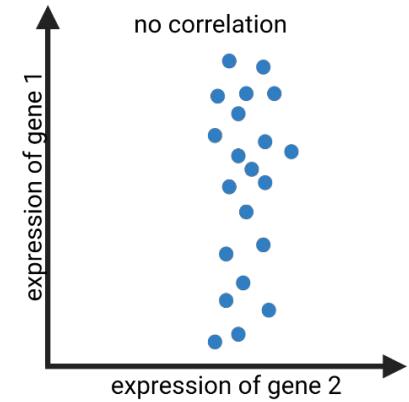
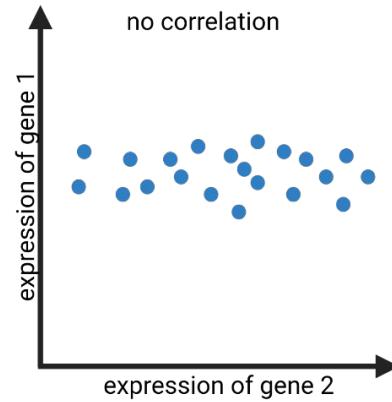
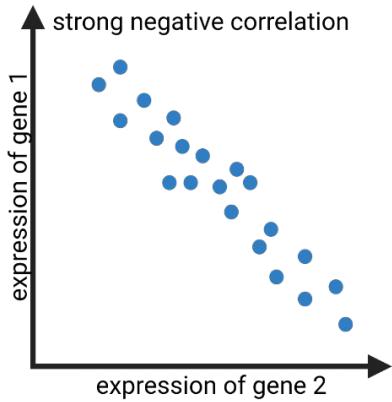
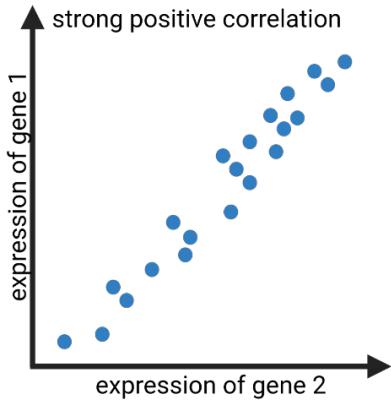
- Genes of the same biosynthesis pathway show similar expression patterns
- Co-expression is a very powerful indication for connection between genes



Correlation vs. causation

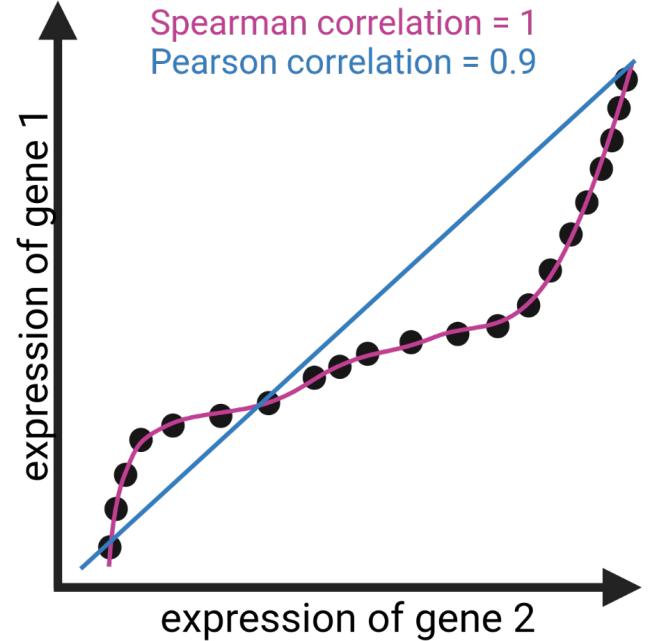
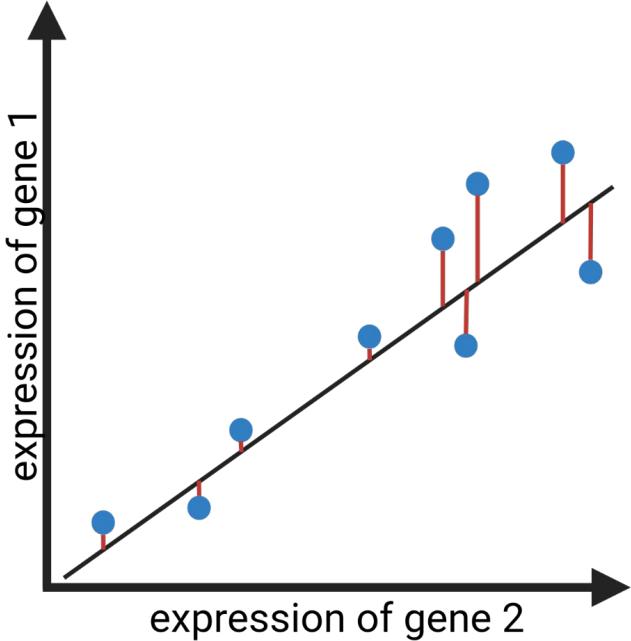
- Correlation does not allow conclusions about the direction i.e. cause and consequence
- Guilt-by-association is frequently used to find genes in a pathway/network
- Correlation can be caused by a direct or indirect connection

Examples



Types of correlation coefficients

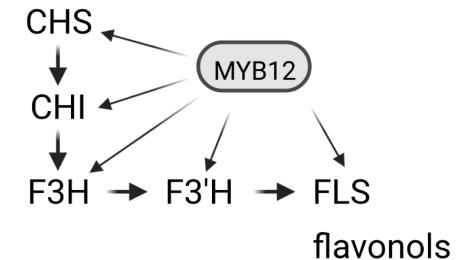
- Pearson is only suitable for linear correlation
- Spearman is more tolerant towards outliers and non-linear correlations



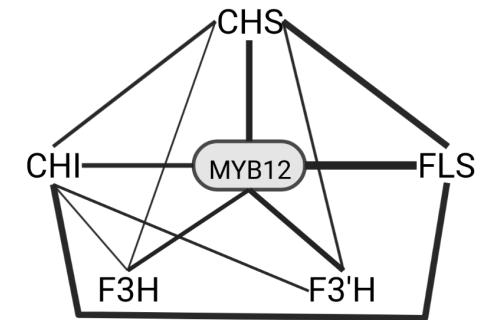
Coexpression network connection

- 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



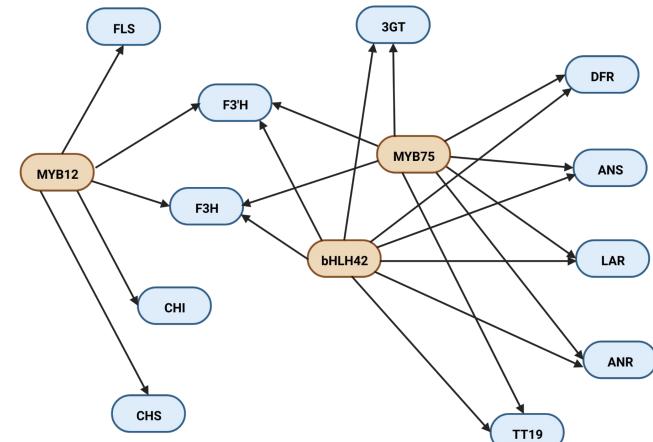
CHS	
genes	correlation
CHI	0.4
F3H	0.2
F3'H	0.3
FLS	0.7



CHI	
genes	correlation
F3H	0.2
F3'H	0.3
FLS	0.7

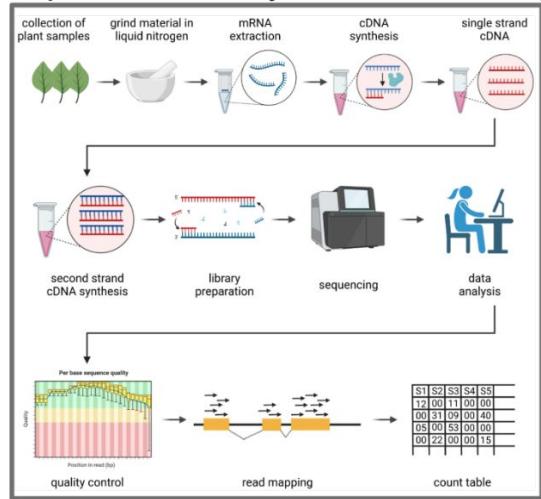
Gene expression networks

- 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

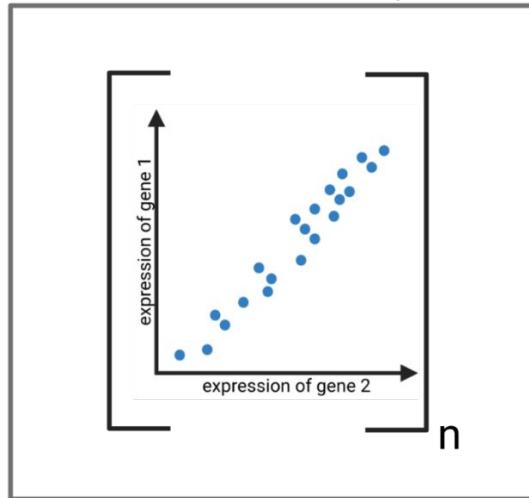


Summary: co-expression workflow

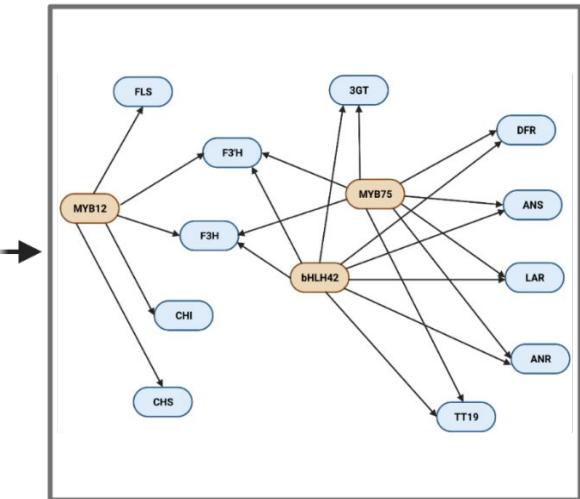
Expression analysis via RNA-Seq



Coexpression analysis



Network construction



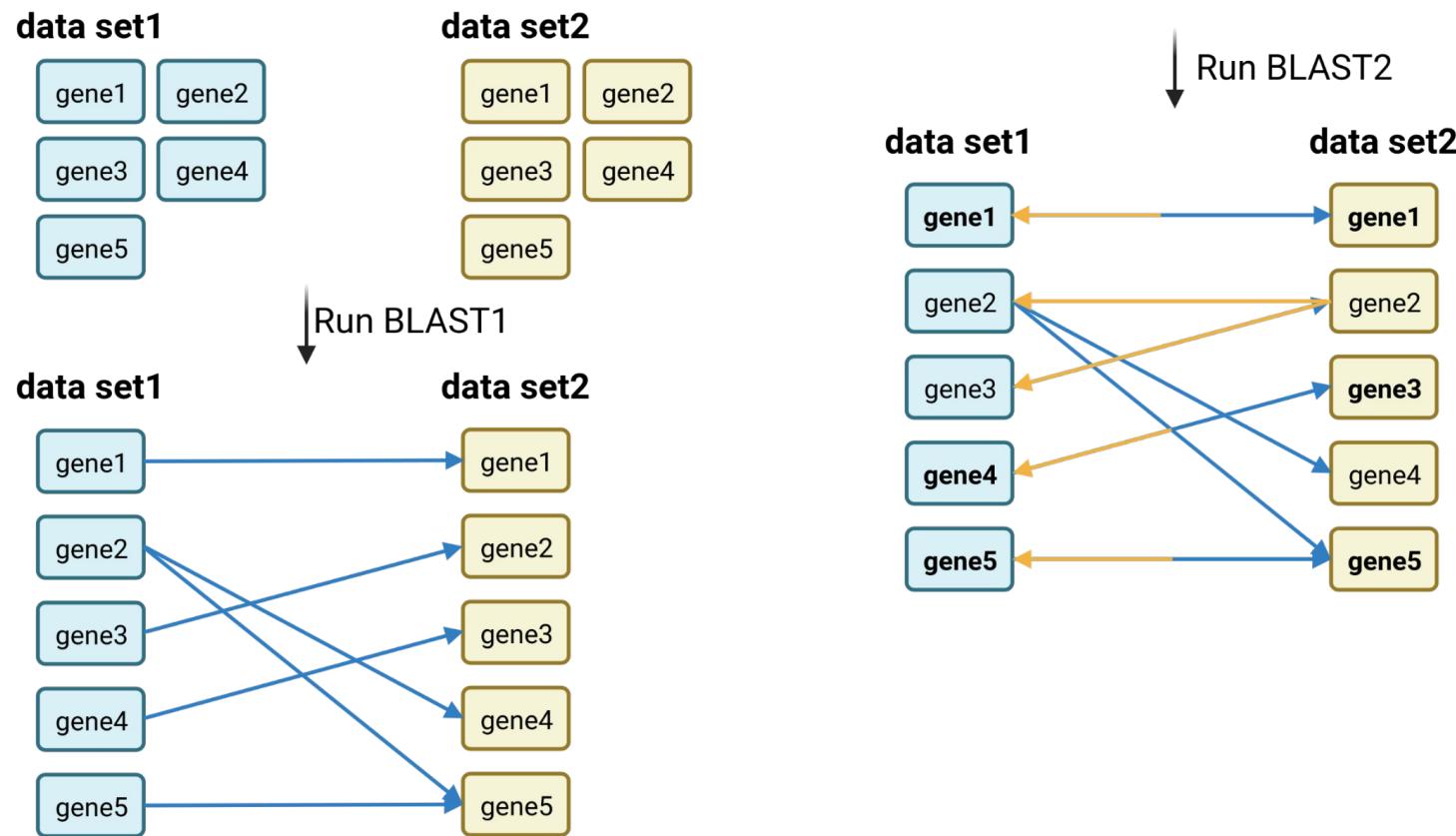
Assigning functional annotation



Functional annotation

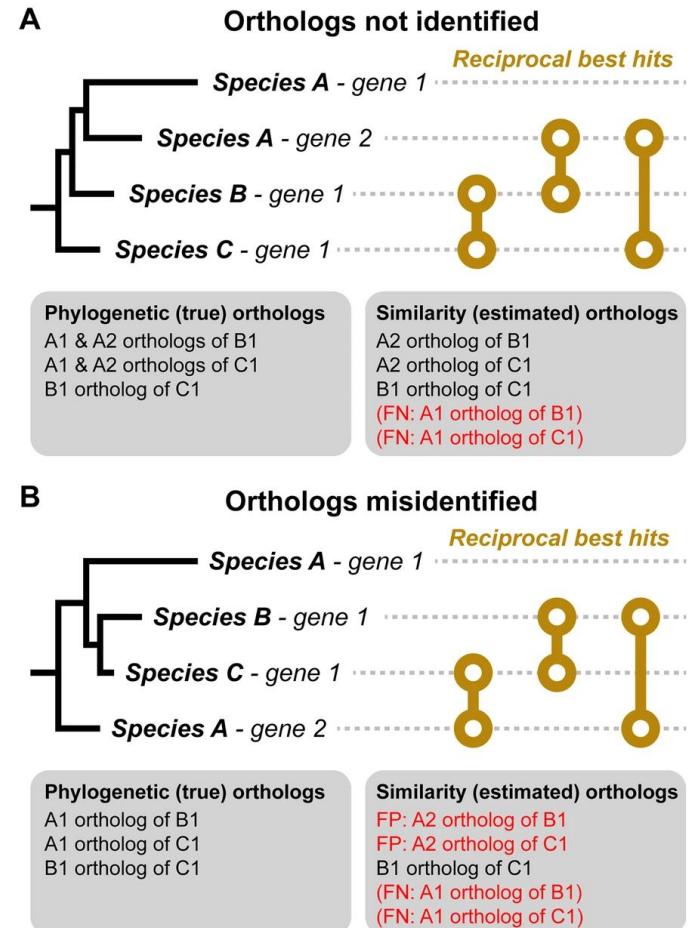
- Reciprocal Best BLAST Hits (RBHs)
- OrthoFinder2
- Knowledge-based Identification of Pathway Enzymes (KIPES)
- Mercator
- InterProScan5

Reciprocal Best BLAST Hits (RBHs)



OrthoFinder2

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

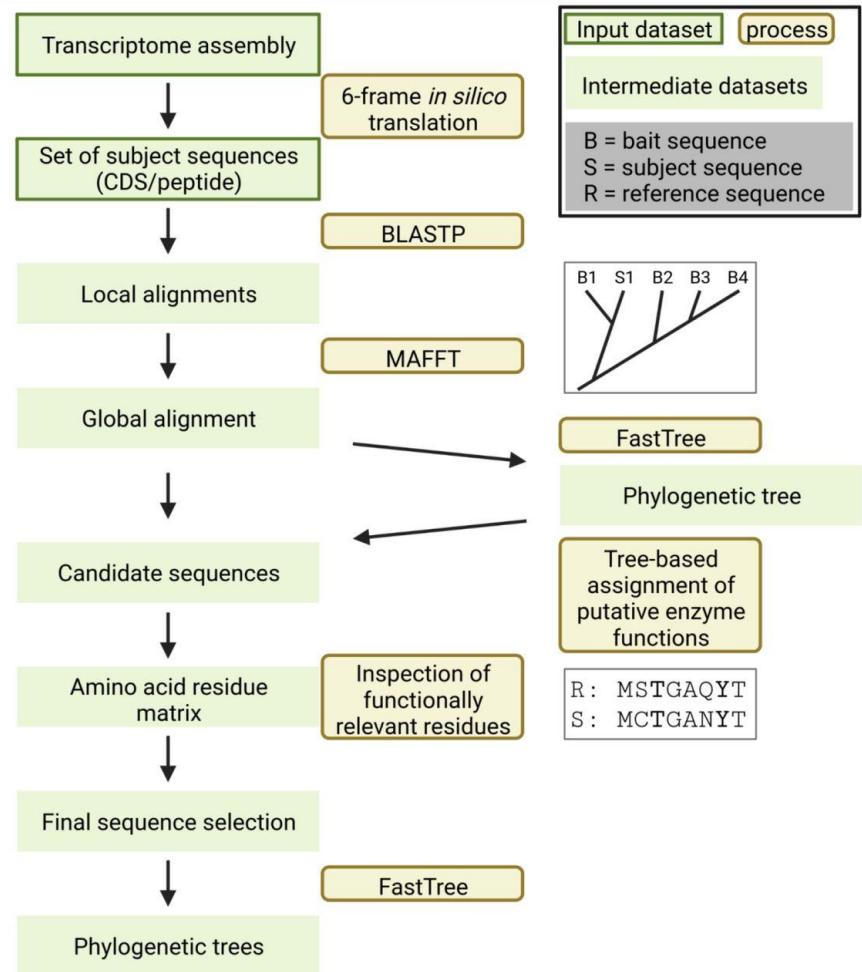


<https://doi.org/10.1186/s13059-019-1832-y>



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



Mercator

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



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

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?