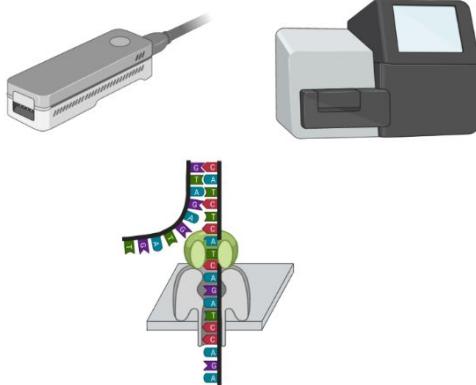
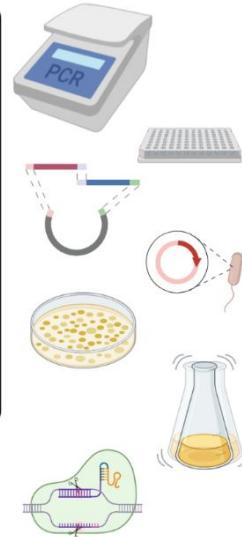
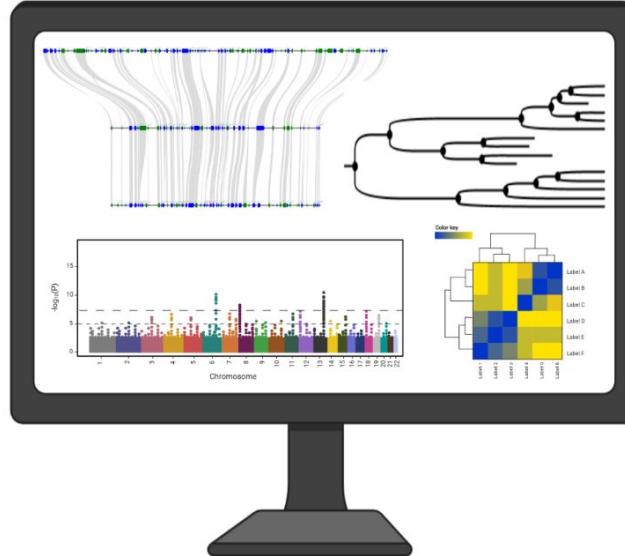




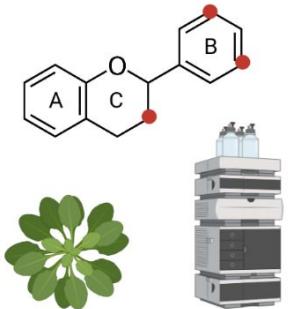
Technische
Universität
Braunschweig



Plant Biotechnology
and Bioinformatics



species biosynthesis proteins analysis different conditions
biosynthesis DODA bellman variants H2R3-MYB
within genes site data Col loco variant
dispersed site data dioxigenases bHLH non-canonical
sequences KEGG multiple protein annotation level identified
single reference structure synthesis amino acid evolutionary
sites interactions pathways loci accession pathway
plants accessions systems biology long evolution
pigments model genome Koyvanta genes Arabidopsis
key against canonical for conserved Canophylales
flavonoid conservation sequencing Arabidopsis
accessions identification sequence thaliana
gene read transcription synthesis evolution
MYB introns residues RNA-Seq



Gene Expression & Coexpression Analyses

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

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
 - StudIP: [Lecture: Grundlagen der Biochemie und Bioinformatik der Pflanzen \(Bio-MB 09\)](#)
 - 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.

What causes these pigmentation patterns?



https://commons.wikimedia.org/wiki/File:Blood_orange_sliced.jpg. CC-BY-SA 4.0

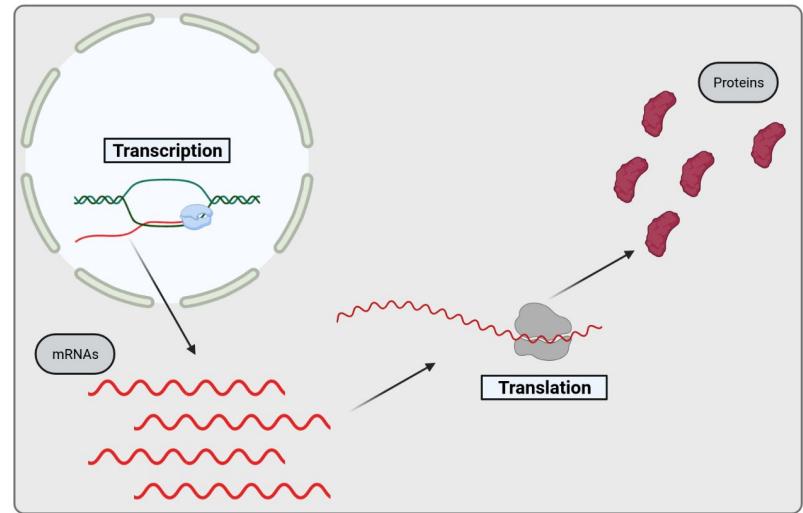


https://commons.wikimedia.org/wiki/File:Pitaya_Colors.jpg. CC-BY-SA 4.0



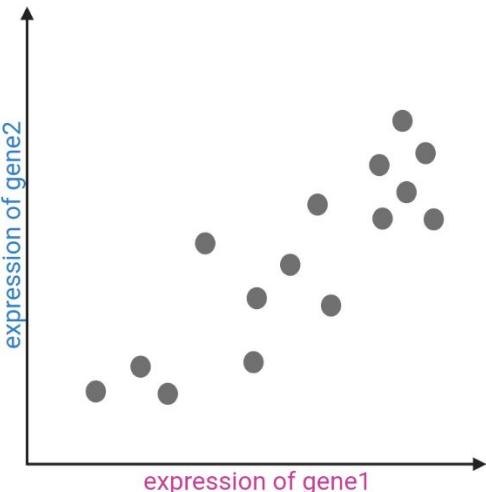
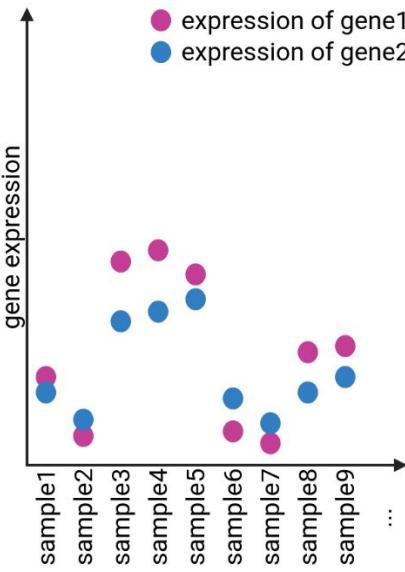
What is gene expression?

- Gene expression = formation of gene product (i.e. a protein)
- Transcription of DNA by RNA polymerase and translation of mRNAs by polymerase
- Transcript abundance is often used as proxy (=gene expression)



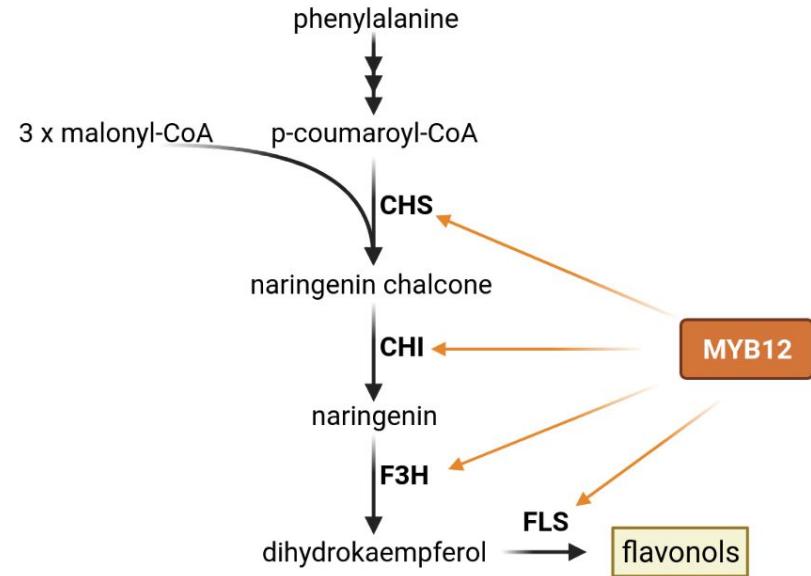
Concept of coexpression

- Genes can show similar expression values across numerous samples
- Reality usually results in similar, but not identical patterns
- Different samples could be different plant parts of plants cultivated under different conditions



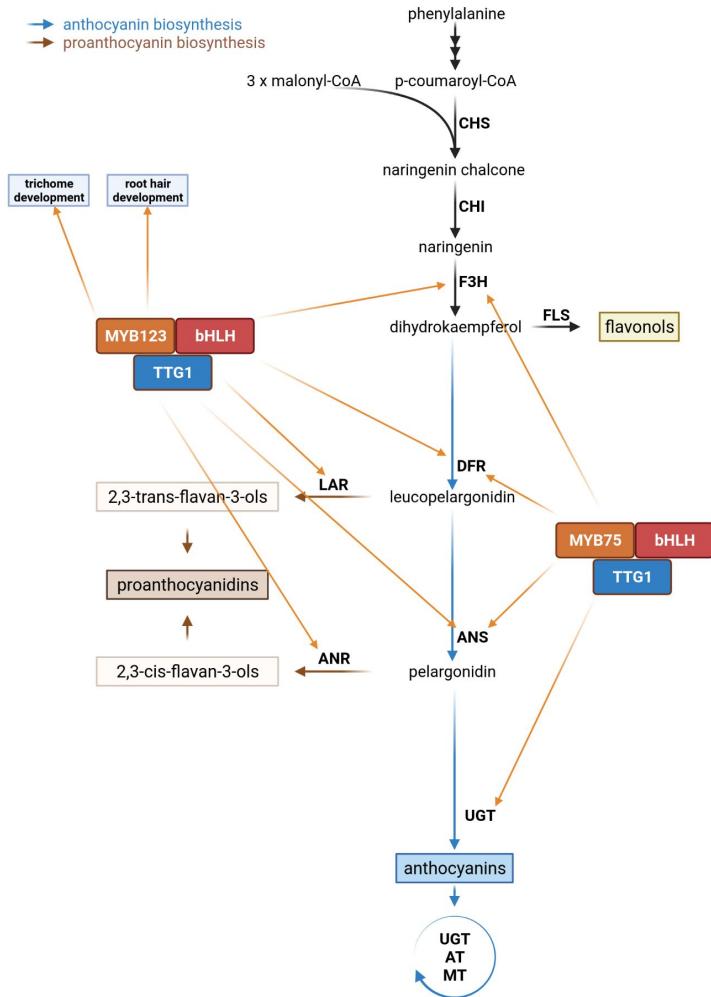
Molecular basis of coexpression

- Shared transcription factor can explain similar expression patterns
- Example: MYB12 controls the flavonol biosynthesis through activation of *CHS*, *CHI*, *F3H*, and *FLS*
- Expectation: *CHS*, *CHI*, *F3H*, and *FLS* should show a similar expression pattern



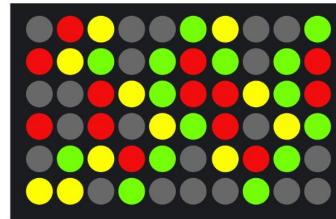
Nothing is perfect

- Genes can be regulated by multiple TFs (e.g. *DFR* by MYB123 and MYB75)
- TFs can control different processes (e.g. proanthocyanidins, trichome development, root hair development)
- Co-expression of TFs and structural genes in pathways is not perfect

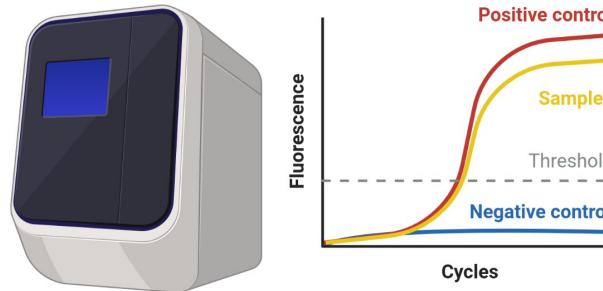


Types of expression data

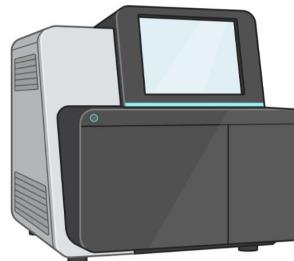
- Microarray



- RT-qPCR

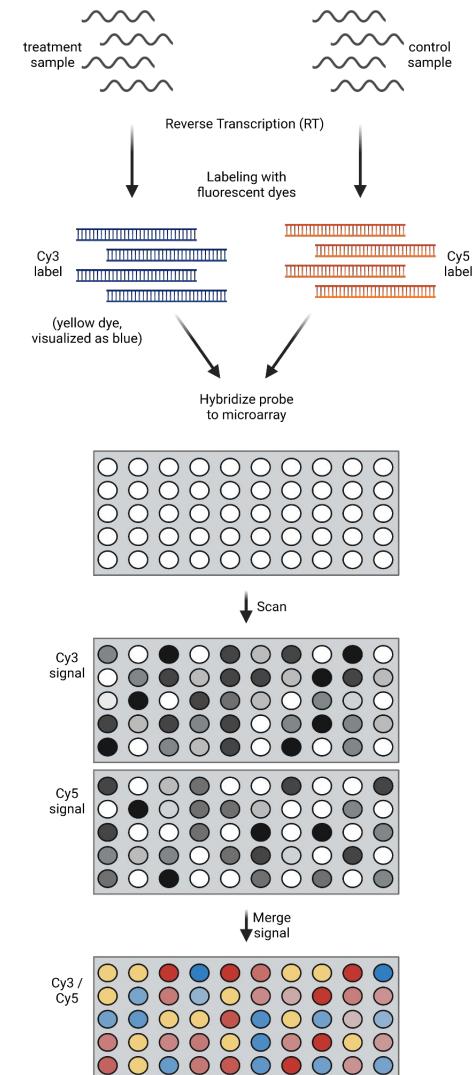


- RNA-Seq



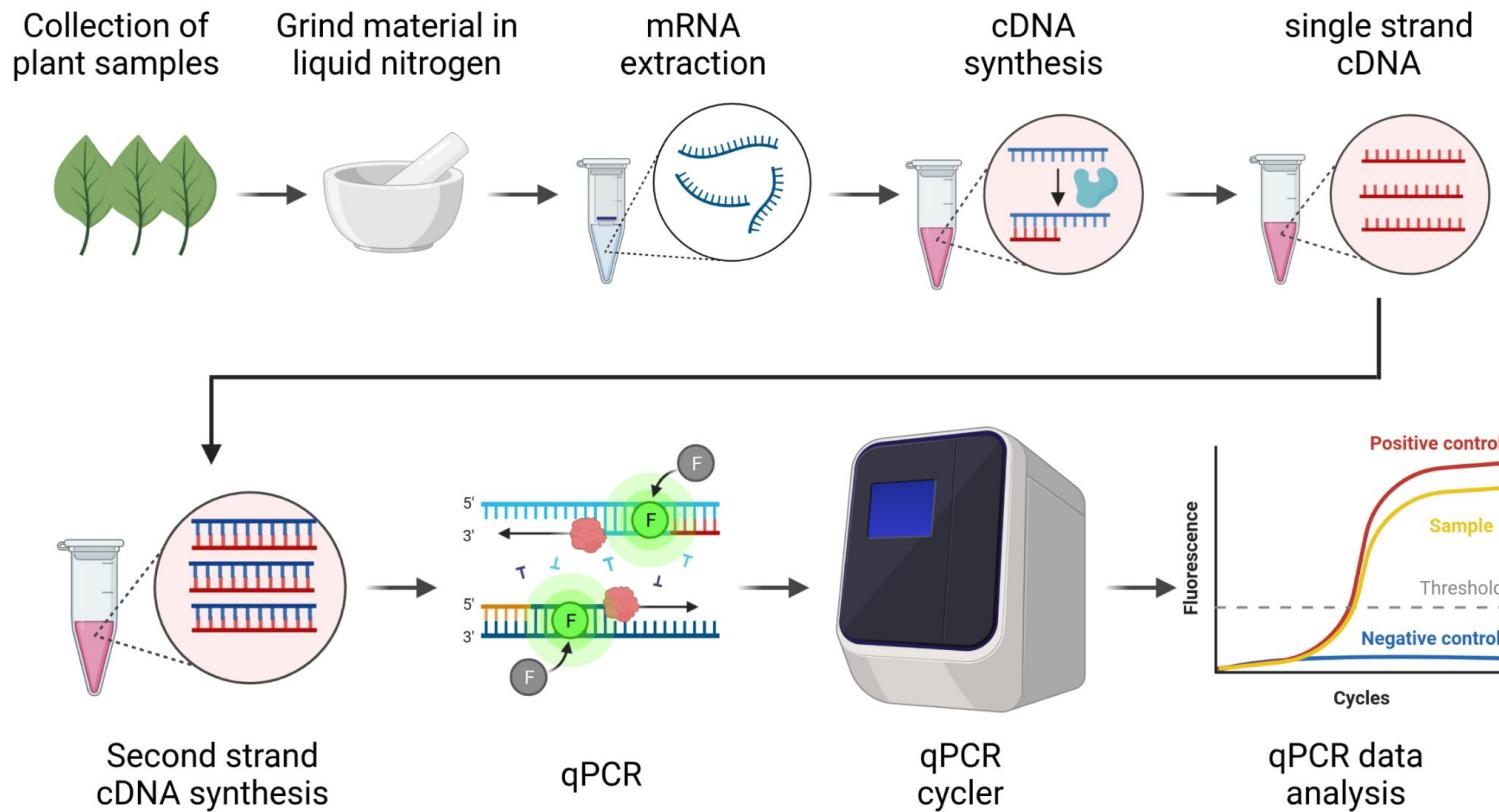
Microarray

- Transcript abundances are compared
- Cy3 and Cy5 are fluorescent labels
- Fluorescence intensity indicates transcript abundances
- Dynamic range is small due to saturation of signal
- Only genes represented on the microarray can be studied
- High investment costs for microarray generation

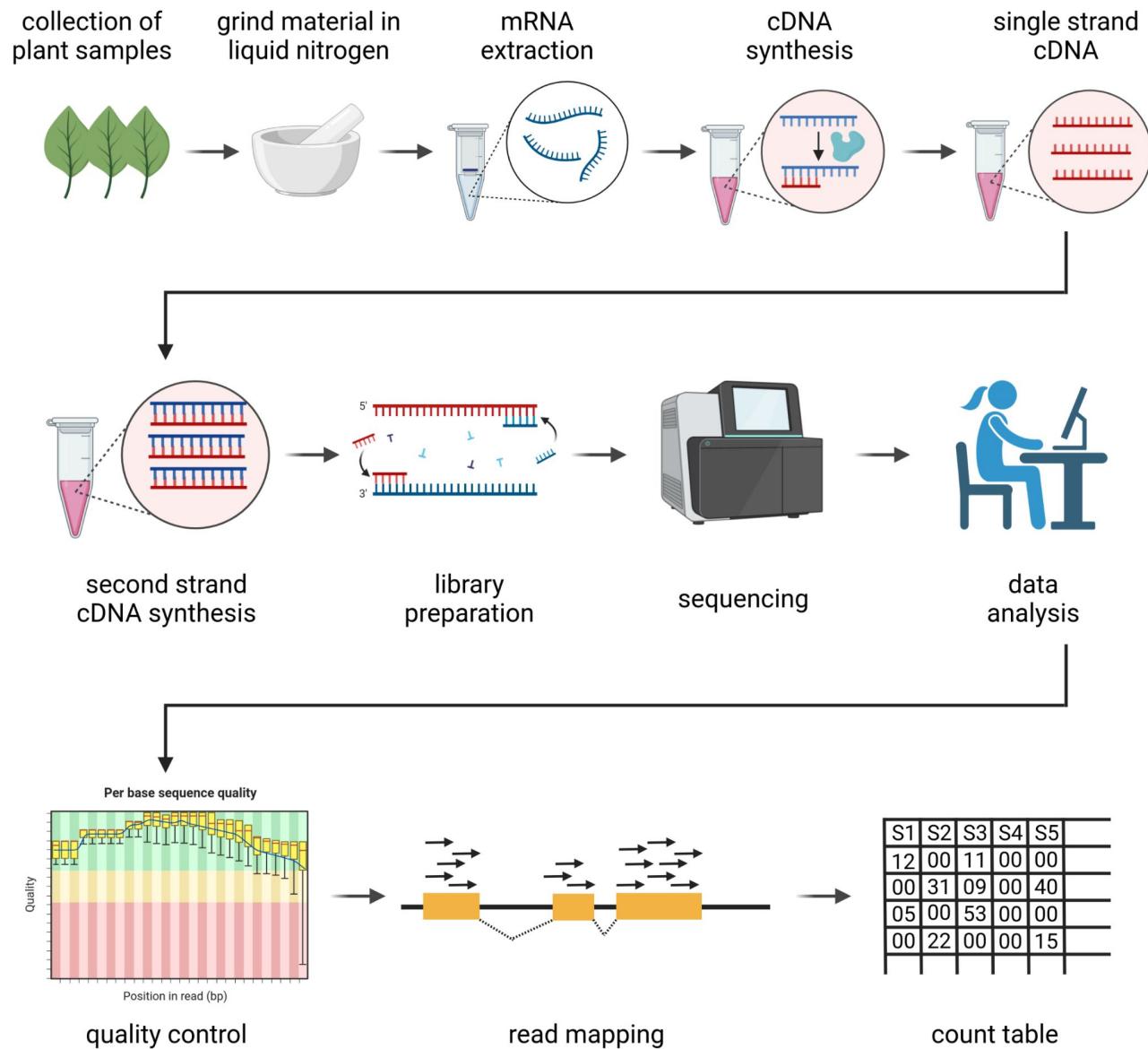


RT-qPCR

- Quantification of cDNA based on incorporation of fluorescent dyes



RNA-Seq



Gene expression databases

- GEO: Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>)
- SRA/ENA: Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>)
- ArrayExpress: microarray database (<https://www.ebi.ac.uk/arrayexpress/>)

How to find the right dataset? (1)

- Search for the species of interest
- Additional keywords e.g. specific tissues are possible
- Filter by species (panel on the right)
- Filter by 'RNA', 'paired' (?), and sequencing technology

The screenshot shows the NCBI SRA search results page for the query "Croton tiglium". The search bar at the top contains "SRA" and "Croton tiglium". Below the search bar, there are links for "Create alert" and "Advanced". On the left, there is a sidebar with filters for "Access", "Source", "Type", "Library Layout", "Platform", "Strategy", "Data in Cloud", "File Type", and "Clear all". The main search results area displays 15 items. Each result includes a checkbox, a title, and a brief description. For example, the first result is "unspecified paired end sequencing: Croton tiglium - target-enriched (Angiosperms353)". The right side of the page features a panel titled "Results by taxon" showing "Top Organisms [Tree]" with "Croton tiglium (9)", "Xanthomonas campestris pv. campestris str. B100 (6)", and a "Search in related databases" section with tables for BioSample, BioProject, dbGaP, and GEO Datasets. A "Find related data" section with a dropdown menu for "Database" and a "Find items" button is also present.

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

How to find the right dataset? (2)

- Send pre-filtered results to ‘RunSelector’
- Download ‘Metadata’ and ‘AccessionList’
 - Metadata = table with details about samples
 - AccessionList = text file with one run ID per line

The screenshot shows the NCBI SRA Run Selector interface. At the top, there's a table for 'Common Fields' with the following data:

Consent	PUBLIC
DATASTORE filetype	FASTQ, SRA
DATASTORE provider	ENA, GS, NCBI, S3
DATASTORE region	ena, gs-US, ncbi-public, s3-us-east-1
LibraryLayout	PAIRED
Platform	ILLUMINA

Below this is a 'Select' section with tabs for 'Runs', 'Bytes', 'Bases', and 'Download'. The 'Download' tab is selected, showing options for 'Metadata' or 'Accession List'. There are also buttons for 'Cloud Data Delivery' and 'Computing'.

Underneath is a table titled 'Found 15 Items' with columns: Run, BioProject, BioSample, Assay Type, AvgSpotLen, Bases, Bytes, and Center Name. The data is as follows:

	Run	BioProject	BioSample	Assay Type	AvgSpotLen	Bases	Bytes	Center Name
1	ERR2040366	PRJEB21674	SAMEA104170410	RNA-Seq	180	2.27 G	1.46 Gb	DEPARTMENT OF BIOLOGICAL SCIENCES
2	ERR7618249	PRJEB49293	SAMEA11051725	WGS	300	609.36 M	254.66 Mb	ROYAL BOTANICAL GARDENS, KEW
3	SRR6239848	PRJNA416498	SAMNO07958178	RNA-Seq	142	3.62 G	1.38 Gb	BIELEFELD UNIVERSITY
4	SRR6239849	PRJNA416498	SAMNO07958179	RNA-Seq	142	2.96 G	1.13 Gb	BIELEFELD UNIVERSITY
5	SRR6239850	PRJNA416498	SAMNO07958176	RNA-Seq	142	3.36 G	1.29 Gb	BIELEFELD UNIVERSITY
6	SRR6239851	PRJNA416498	SAMNO07958177	RNA-Seq	142	4.16 G	1.61 Gb	BIELEFELD UNIVERSITY
7	SRR6239852	PRJNA416498	SAMNO07958180	RNA-Seq	142	6.41 G	2.44 Gb	BIELEFELD UNIVERSITY
8	SRR6239853	PRJNA416498	SAMNO07958181	RNA-Seq	502	23.79 G	9.56 Gb	BIELEFELD UNIVERSITY
9	SRR6806225	PRJNA416498	SAMNO08634686	RNA-Seq	150	850.47 M	382.95 Mb	BIELEFELD UNIVERSITY
10	SRR6806226	PRJNA416498	SAMNO08634685	RNA-Seq	150	922.92 M	418.82 Mb	BIELEFELD UNIVERSITY
11	SRR6806227	PRJNA416498	SAMNO08634688	RNA-Seq	150	812.92 M	379.52 Mb	BIELEFELD UNIVERSITY
12	SRR6806228	PRJNA416498	SAMNO08634687	RNA-Seq	56	419.27 M	175.06 Mb	BIELEFELD UNIVERSITY

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

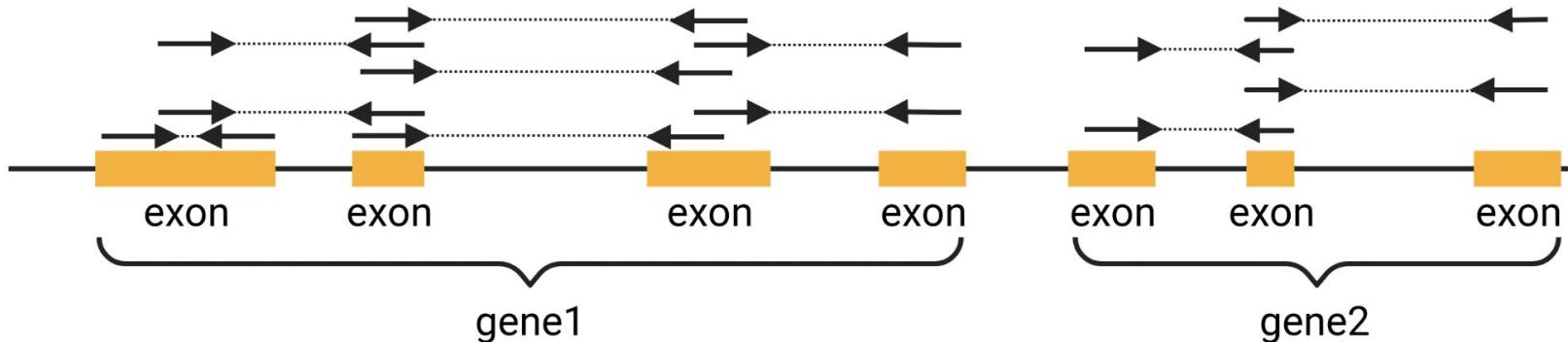


Retrieving data

- Various tools available for large data set download
- Fastq-dump: <https://rnnh.github.io/bioinfo-notebook/docs/fastq-dump.html>
- Wget: <https://www.gnu.org/software/wget/>
- Web browser-based download is no longer supported by most repositories

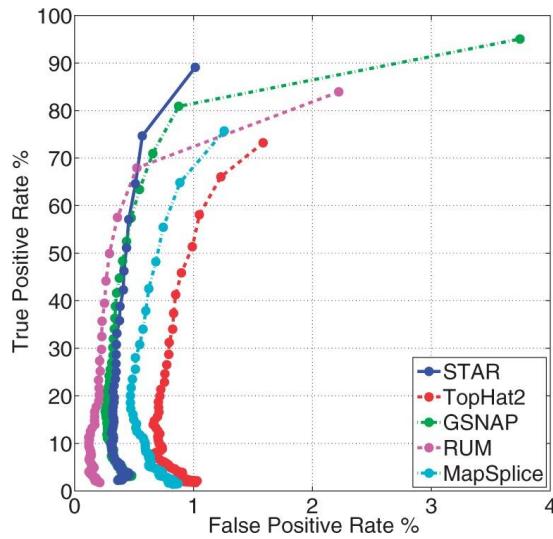
Concept of gene expression quantification

- Reads can be aligned to a reference genome sequence or transcriptome assembly
- Pseudo-alignments are an alternative
- Reads per gene serve as basis for relative gene expression calculation
- Normalization for sequencing depth of all samples

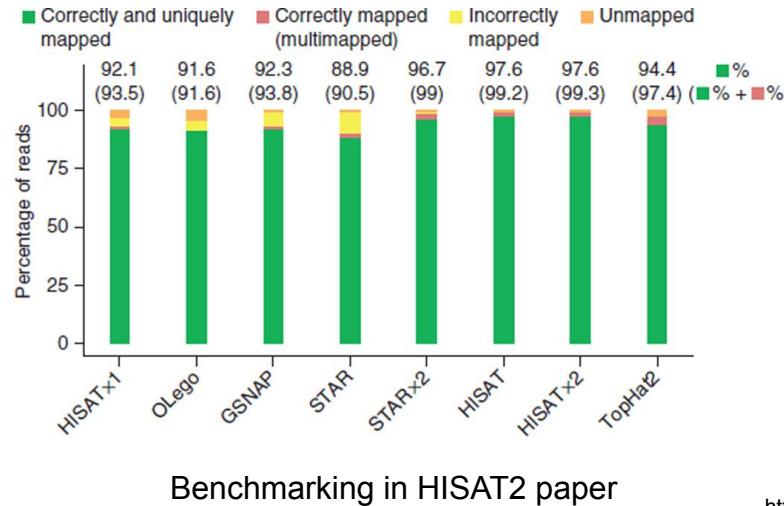


Processing expression data

- Kallisto: alignment-free analysis approach; very fast, but slightly less precise
- STAR: split read alignment; very memory intensive
- HISATII: split read alignment



Benchmarking in STAR paper



Benchmarking in HISAT2 paper

<https://github.com/pachterlab/kallisto>
Bray et al., 2016: 10.1038/nbt.3519
<https://github.com/alexdobin/STAR>
Dobin et al., 2013: 10.1093/bioinformatics/bts635
<http://daehwankimlab.github.io/hisat2/>
Kim et al., 2019: 10.1038/s41587-019-0201-4

Counts, TPMs, and FPKMs

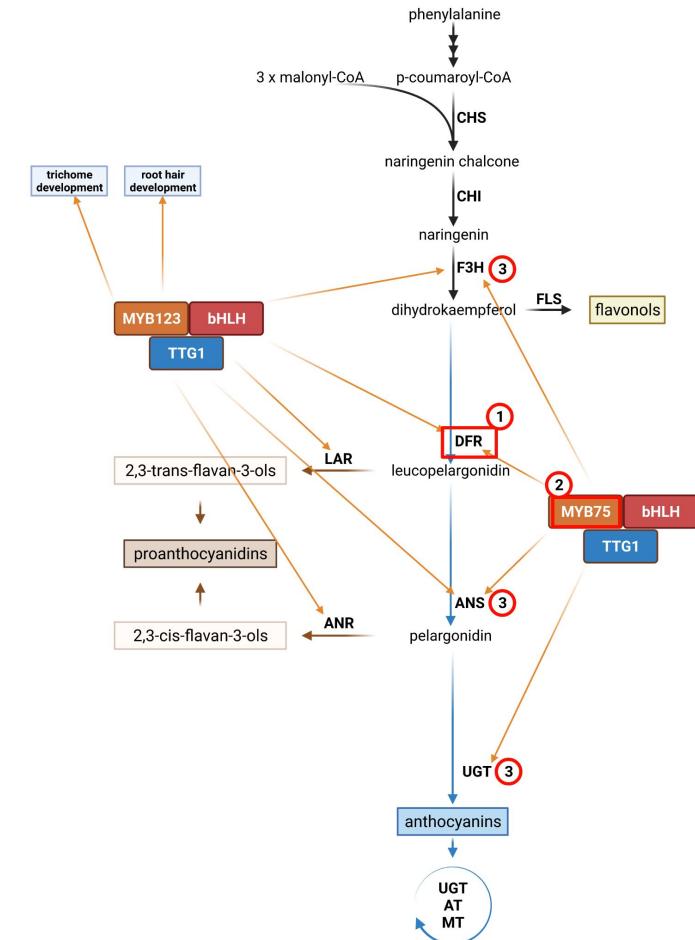
- Counts = Number of reads that are assigned to a feature (gene, exon, transcript isoform, ...)
- TPMs = Transcripts Per Million Transcripts
- RPKMs = Reads Per Kb exon per Million reads (single-end reads)
- FPKMs = Fragments Per Kb exon per Million fragments (paired-end reads)

Example:

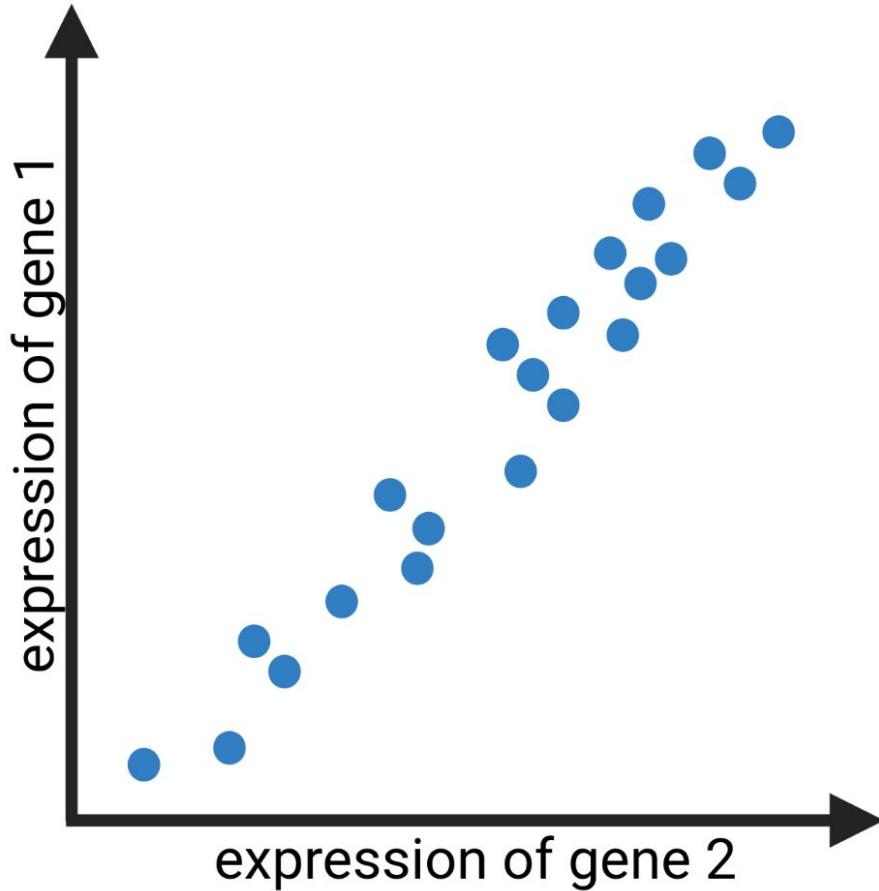
- Counts: gene1=12, gene2=3, gene3=5
- Transcript lengths: gene1=1.5kb, gene2=1kb, gene3=3kb
- TPMs (simplified approximation):
 - $\text{gene1} = 12 / ((12+3+5)/1000000)$
 - $\text{gene2} = 3 / ((12+3+5)/1000000)$
- RPKMs:
 - $\text{gene1} = 12 / (1.5 * ((12+3+5)/1000000))$
- FPKMs:
 - same as RPKM, but for paired-end

Baits for coexpression analyses

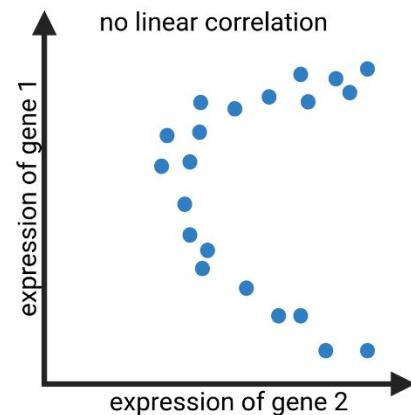
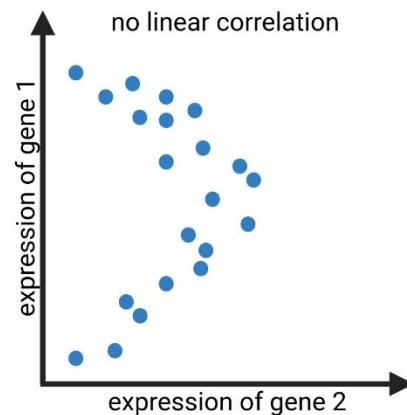
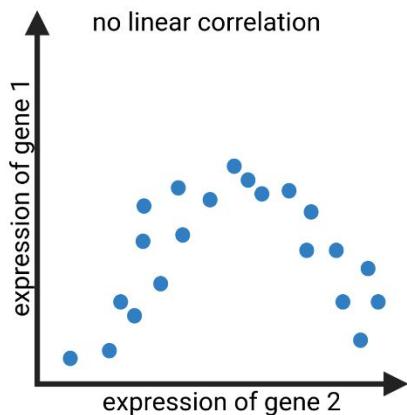
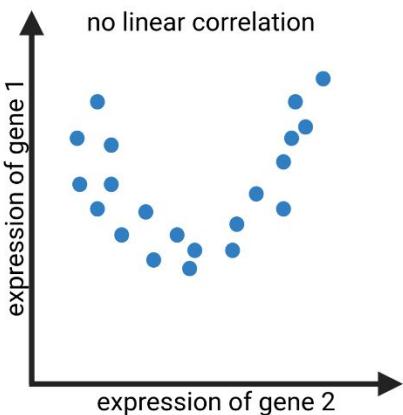
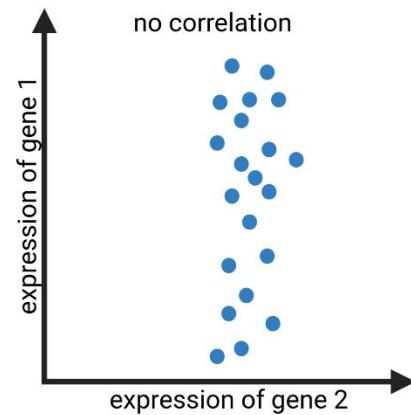
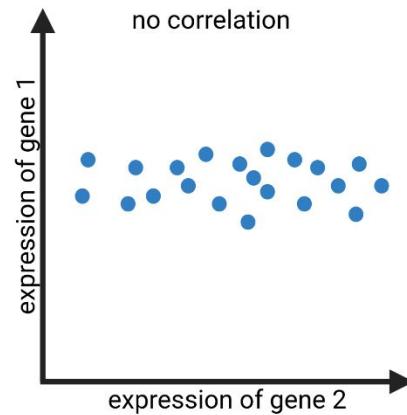
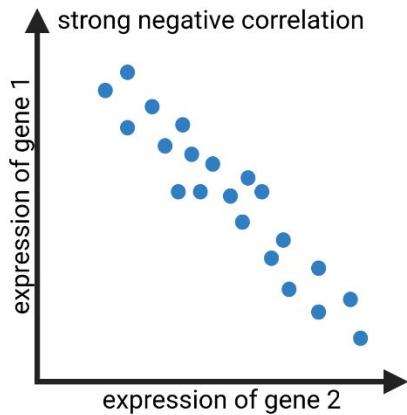
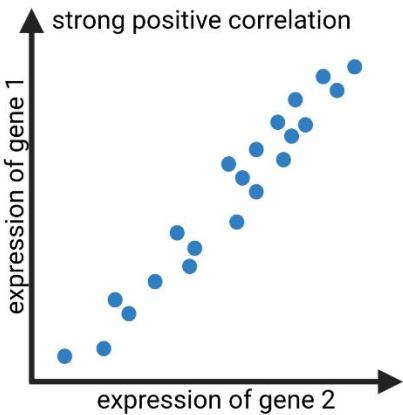
- Bait genes are previously characterized genes with a function of interest e.g. encode an enzyme in the same biosynthesis pathway
- Shared transcription factors of a pathway can be helpful to identify all structural genes of a pathway
- Knowledge from other species can be applied in this step (details in later section)



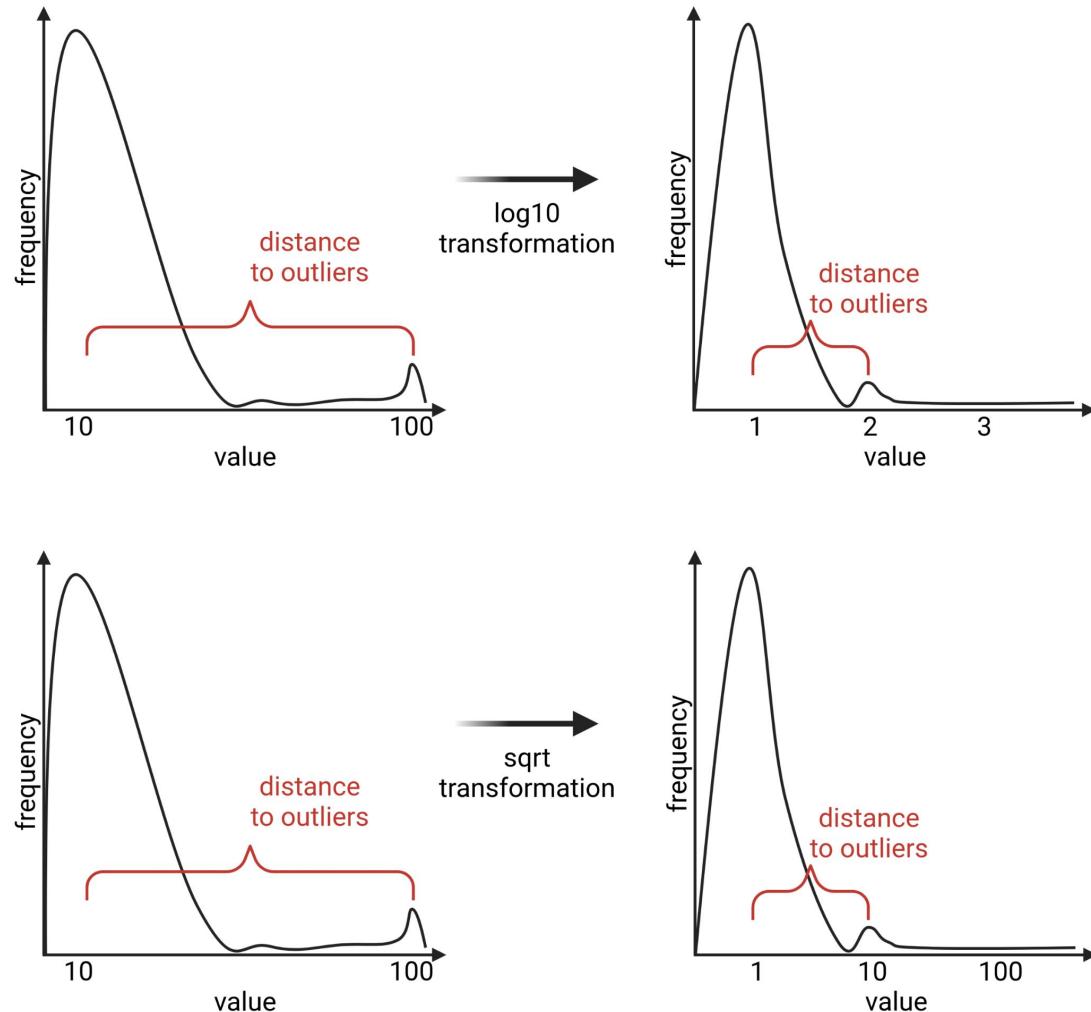
Coexpression



Correlation - examples

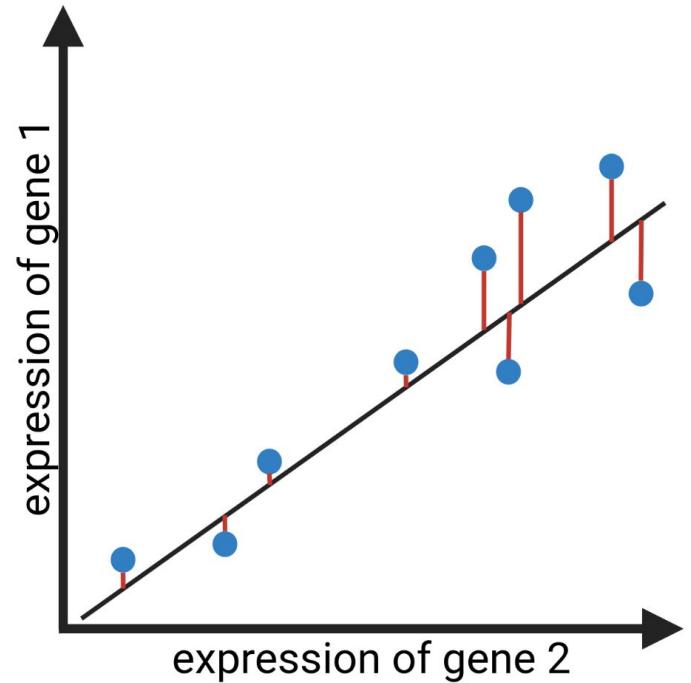


Normalization



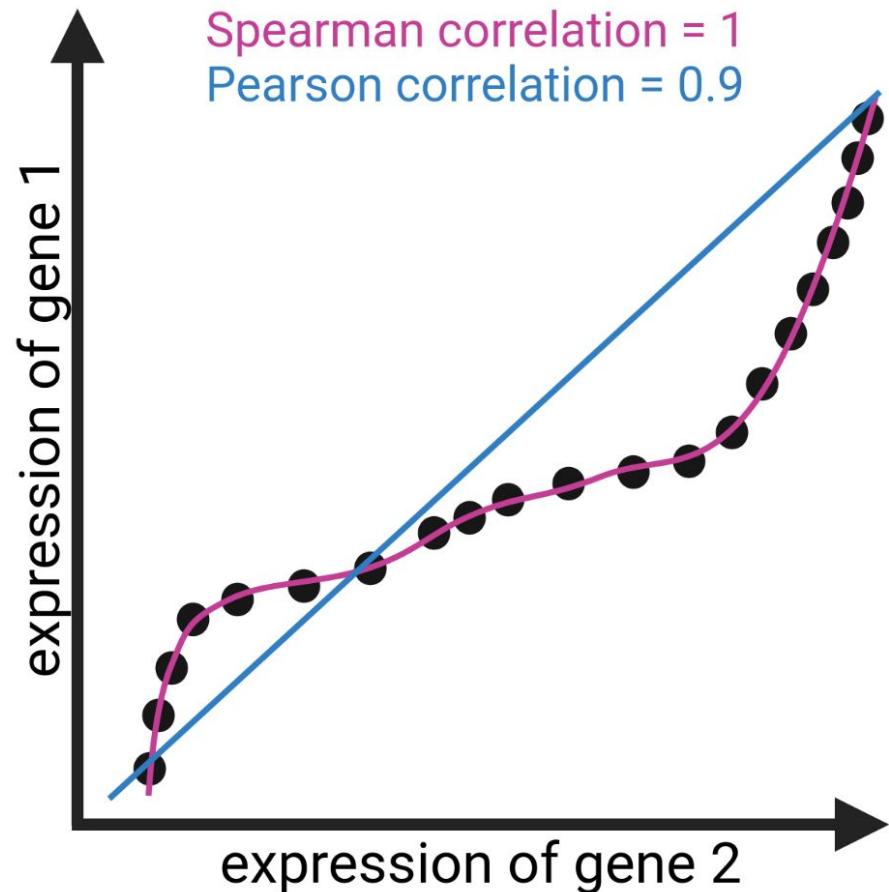
Pearson correlation coefficient

- Line is fitted to achieve minimal distance of all data points to the line
- Only good for linear correlation



Spearman correlation coefficient

- Rank-based correlation coefficient
- Not restricted to linear correlation
- More appropriate for gene expression which might not show linear correlation



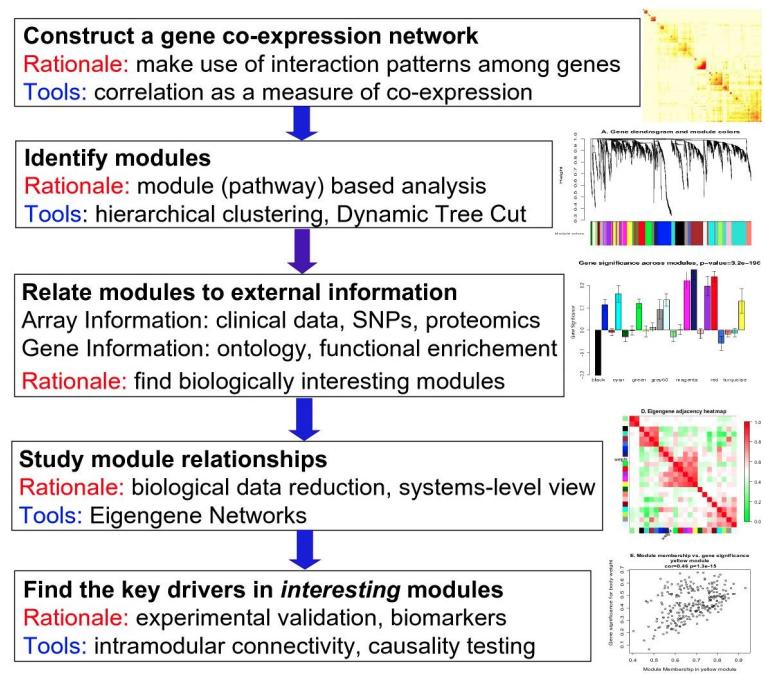
Simple coexpression analysis

- Coexpression analysis of DN38171_c1_g2_i1 (ID of sequence in Trinity transcriptome assembly)
- Correlation coefficient between 0 and 1
- Adjusted p-value describes how well correlation fits the data points
- Annotation is based on *Arabidopsis thaliana*

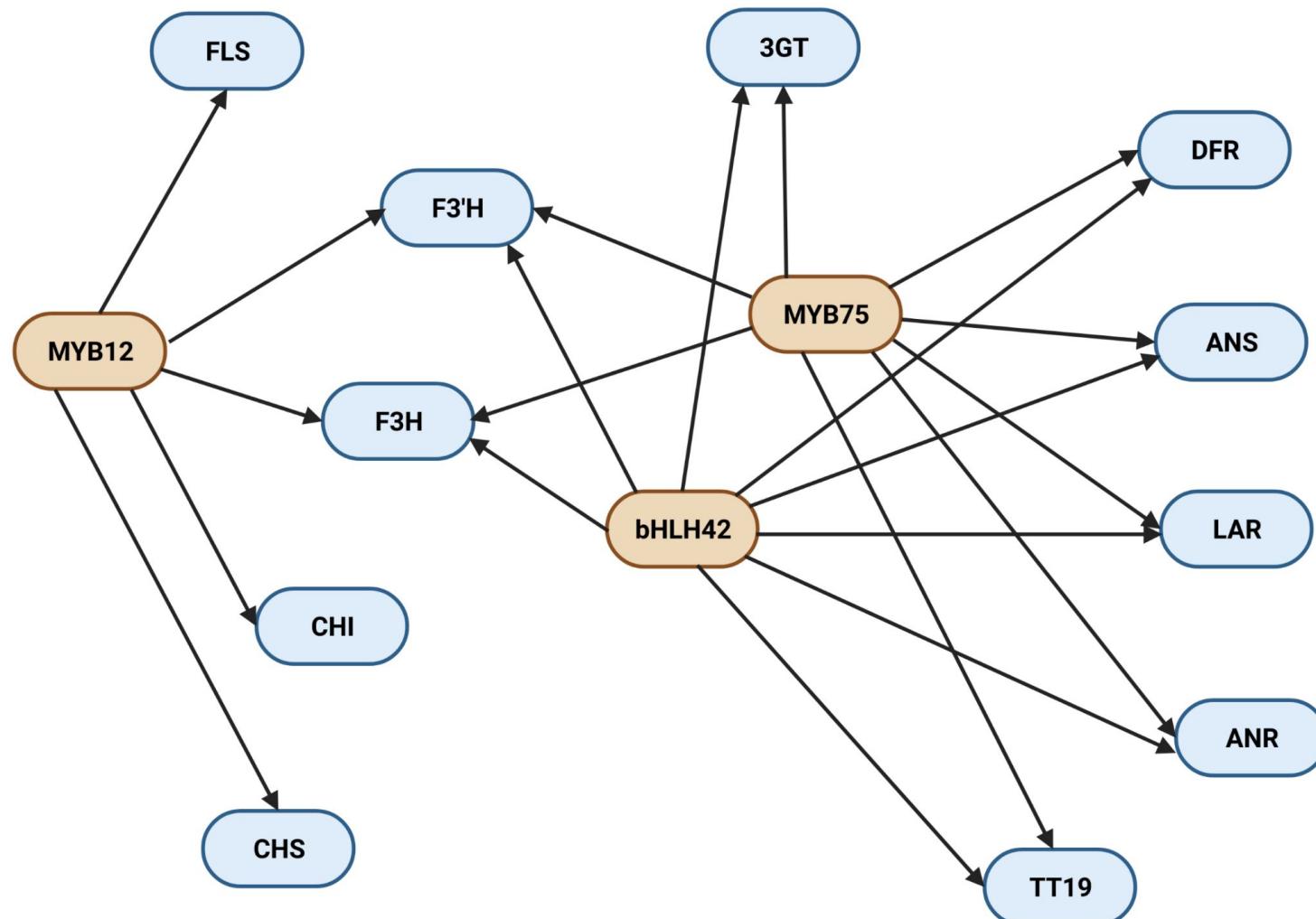
CandidateGene	GeneID	Spearman Correlation	Adjusted p-value	FunctionalAnnotation
DN38171_c1_g2_i1	DN34048_c0_g1_i4	0.976	1.37E-06	AT4G08350;GTA2.global transcription factor group A2
DN38171_c1_g2_i1	DN30512_c0_g2_i1	0.972	5.47E-06	AT2G46800;ZAT.zinc transporter
DN38171_c1_g2_i1	DN30331_c0_g2_i2	0.969	1.08E-05	AT5G60760.P-loop containing nucleoside triphosphate hydrolases superfamily protein
DN38171_c1_g2_i1	DN39190_c7_g1_i5	0.969	1.08E-05	AT5G10260;RABH1e.RAB GTPase homolog H1E
DN38171_c1_g2_i1	DN30136_c1_g2_i1	0.969	1.24E-05	AT4G14580;CIPK4.CBL-interacting protein kinase 4
DN38171_c1_g2_i1	DN36185_c0_g1_i3	0.968	1.46E-05	AT1G73100;SUVH3.histone-lysine N-methyltransferase, H3 lysine-9 specific SUVH3-like protein

WGCNA

- WGCNA = Weighted Gene Correlation Network Analysis
- Expression of genes is controlled by multiple TFs > not only linear correlation

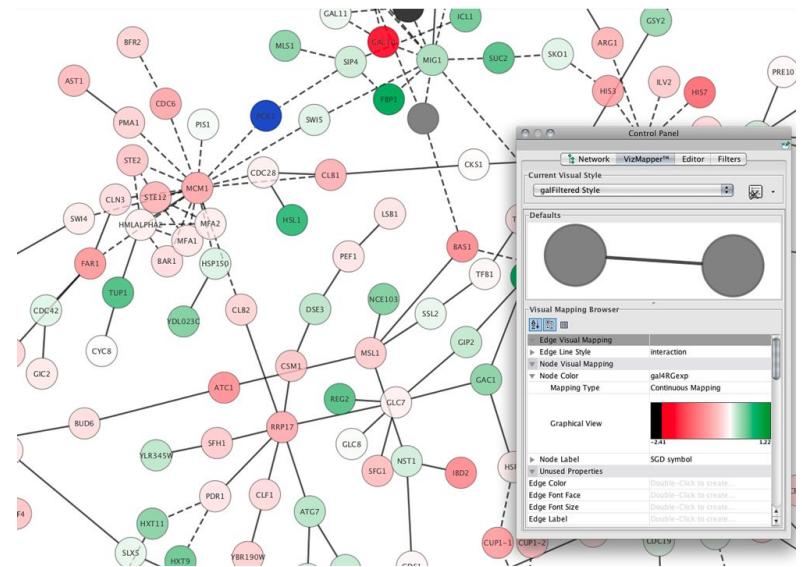


Coexpression network example



Cytoscape

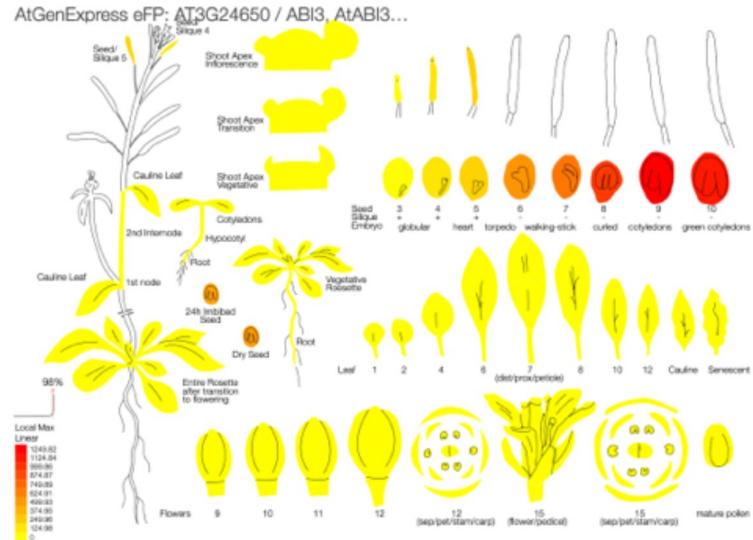
- Cytoscape can be used for illustration of regulatory networks
- Mapping of expression values (heatmap)
- Freely available open source software



<https://cytoscape.org/>
Shannon et al., 2003: 10.1101/gr.1239303

ePlant

- Database integrates various (expression) data of plants
- Visualization of gene expression in different tissues/cell types and conditions
- Broad collection of different plant species covered

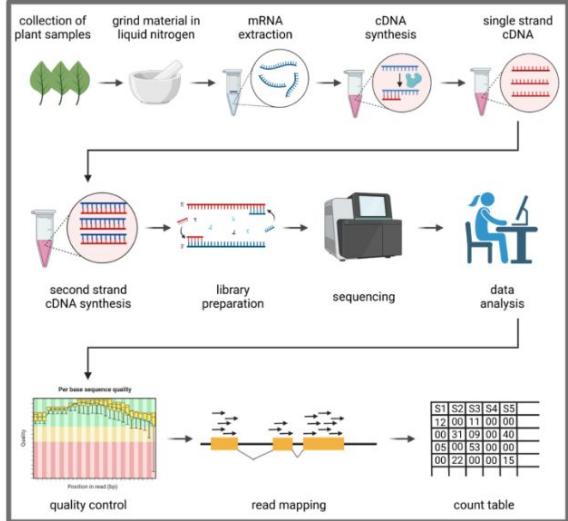


Plant eFP viewers

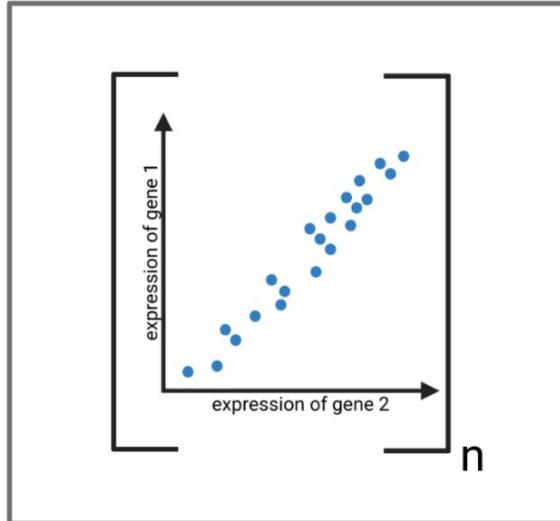
<http://bar.utoronto.ca/>

Summary of the process

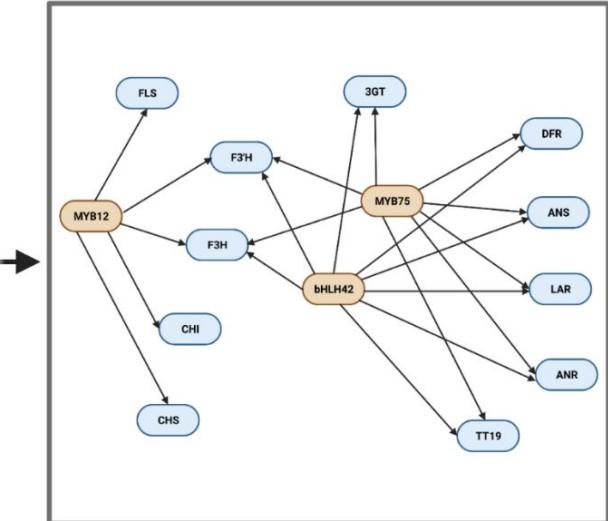
Expression analysis via RNA-Seq



Coexpression analysis



Network construction



Time for questions!



Questions

1. What is gene expression?
2. Why are genes co-expressed?
3. Which methods can be used to measure/approximate gene expression?
4. What are the important steps of an RNA-Seq experiment?
5. Where can you find transcriptomic data sets?
6. What are TPM and RPKM/FPKM?
7. What are the differences between Pearson and Spearman correlation coefficients?
8. How can you normalize expression data prior to co-expression analyses?

EXAMPLE: competition of FLS and DFR

Nancy project



EXAMPLE: discovery of novel anthocyanin transporters

Michael project