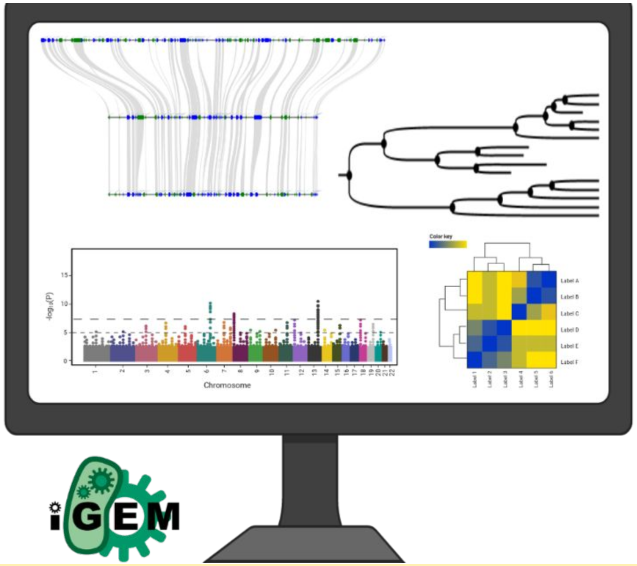
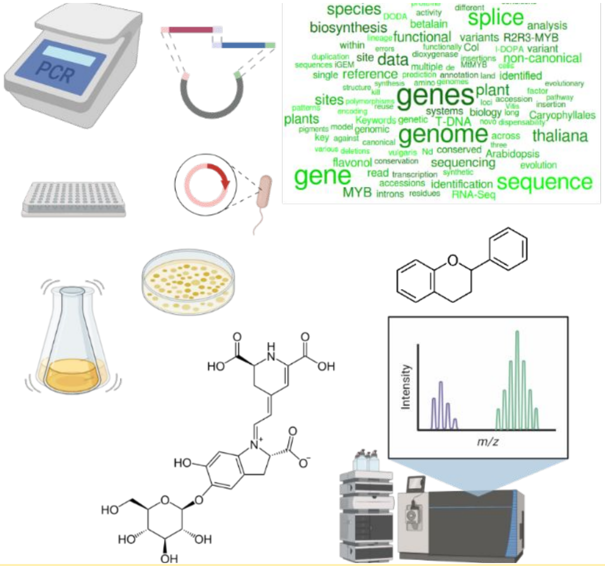
Technische Universität Braunschweig





Gene Expression & Coexpression Analyses

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

Availability of slides

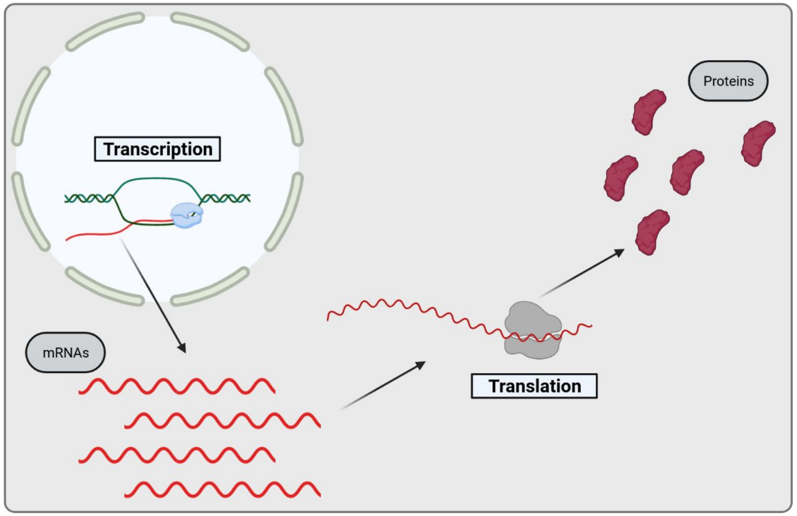
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: Lecture: Grundlagen der Biochemie und Bioinformatik der Pflanzen (Bio-MB 09)
  - Skype: (link shared via email)
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

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

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)

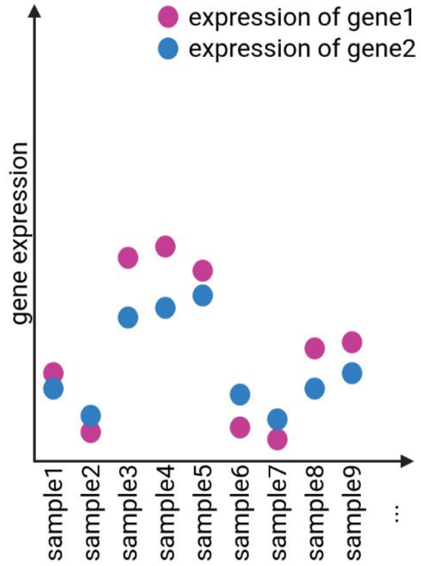


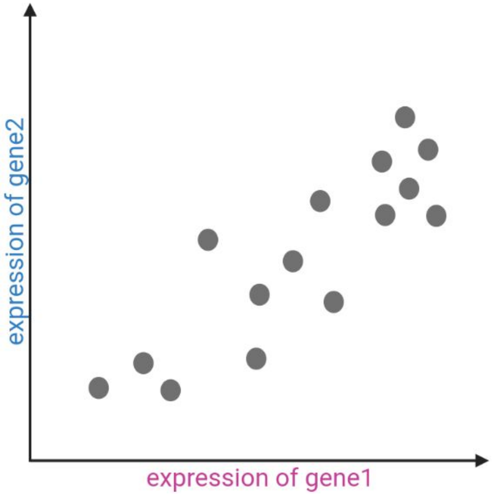
Concept of coexpression

- 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



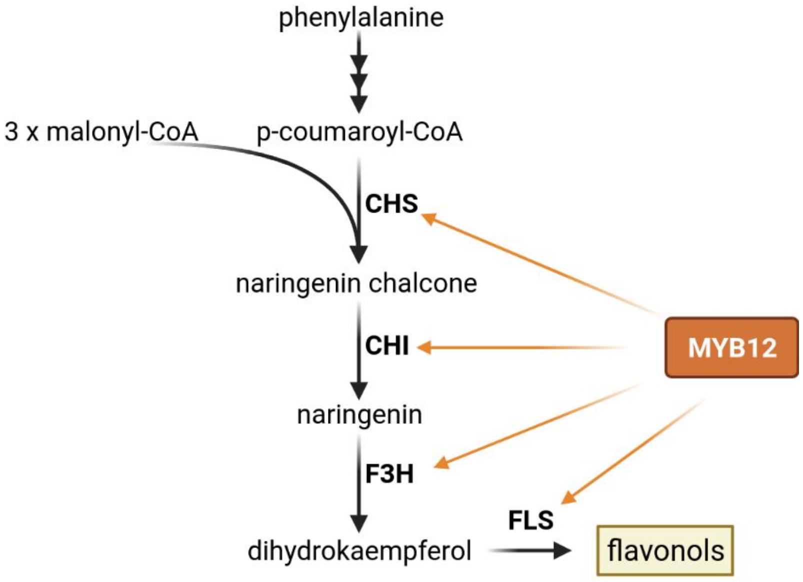


Molecular basis of coexpression

- Shared transcription factor can explain similar expression patterns

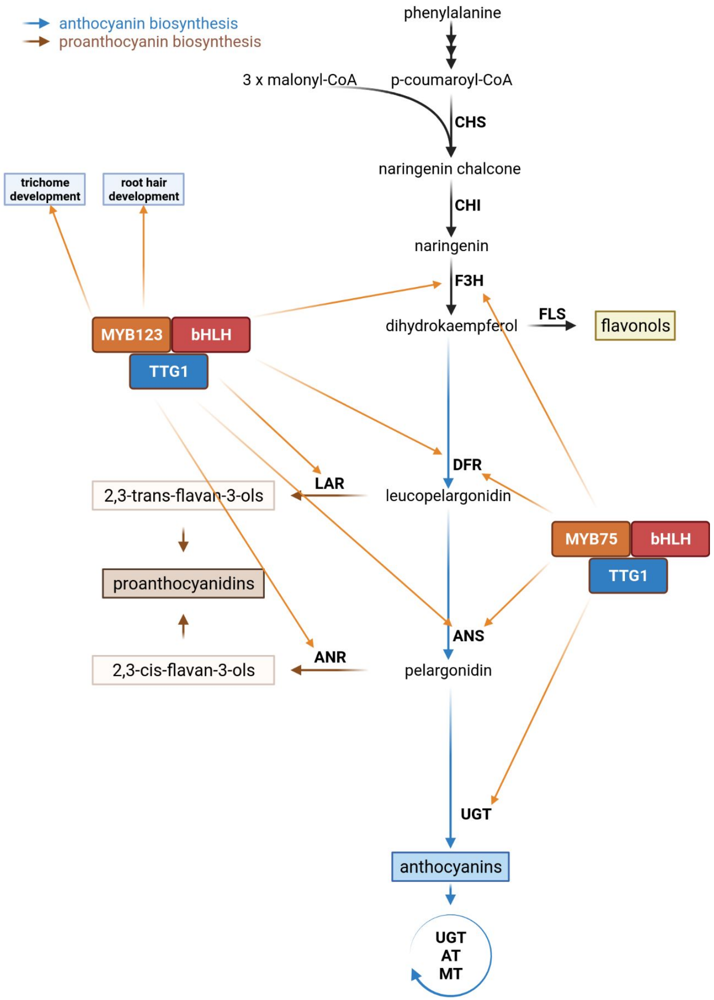
- Example: MYB12 controls the flavonol biosynthesis through activation of *CHS*, *CHI*, *F3H*, and *FLS*

- Expectation: *CHS*, *CHI*, *F3H*, and *FLS* should show a similar expression pattern



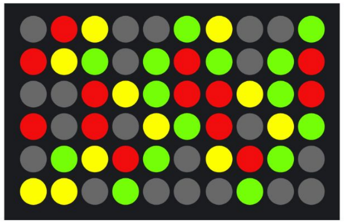
Nothing is perfect

- 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

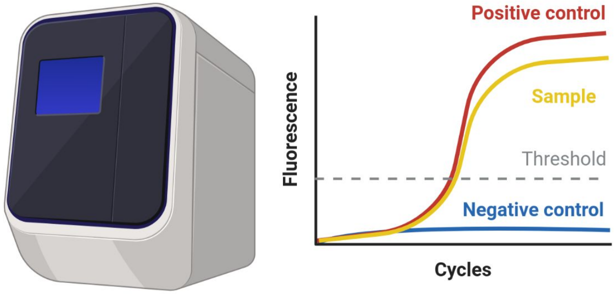


Types of expression data

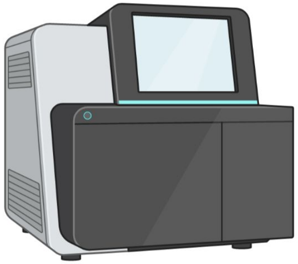
- Microarray



- RT-qPCR

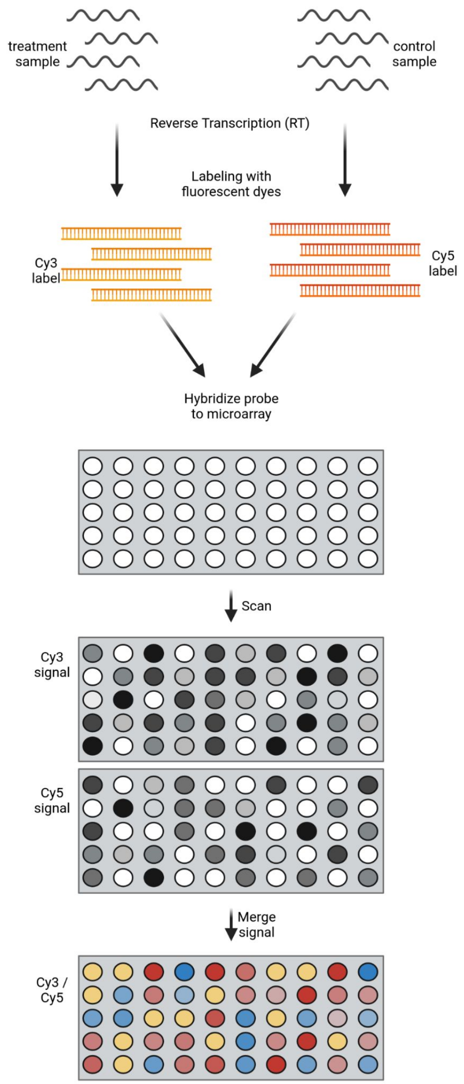


- RNA-Seq



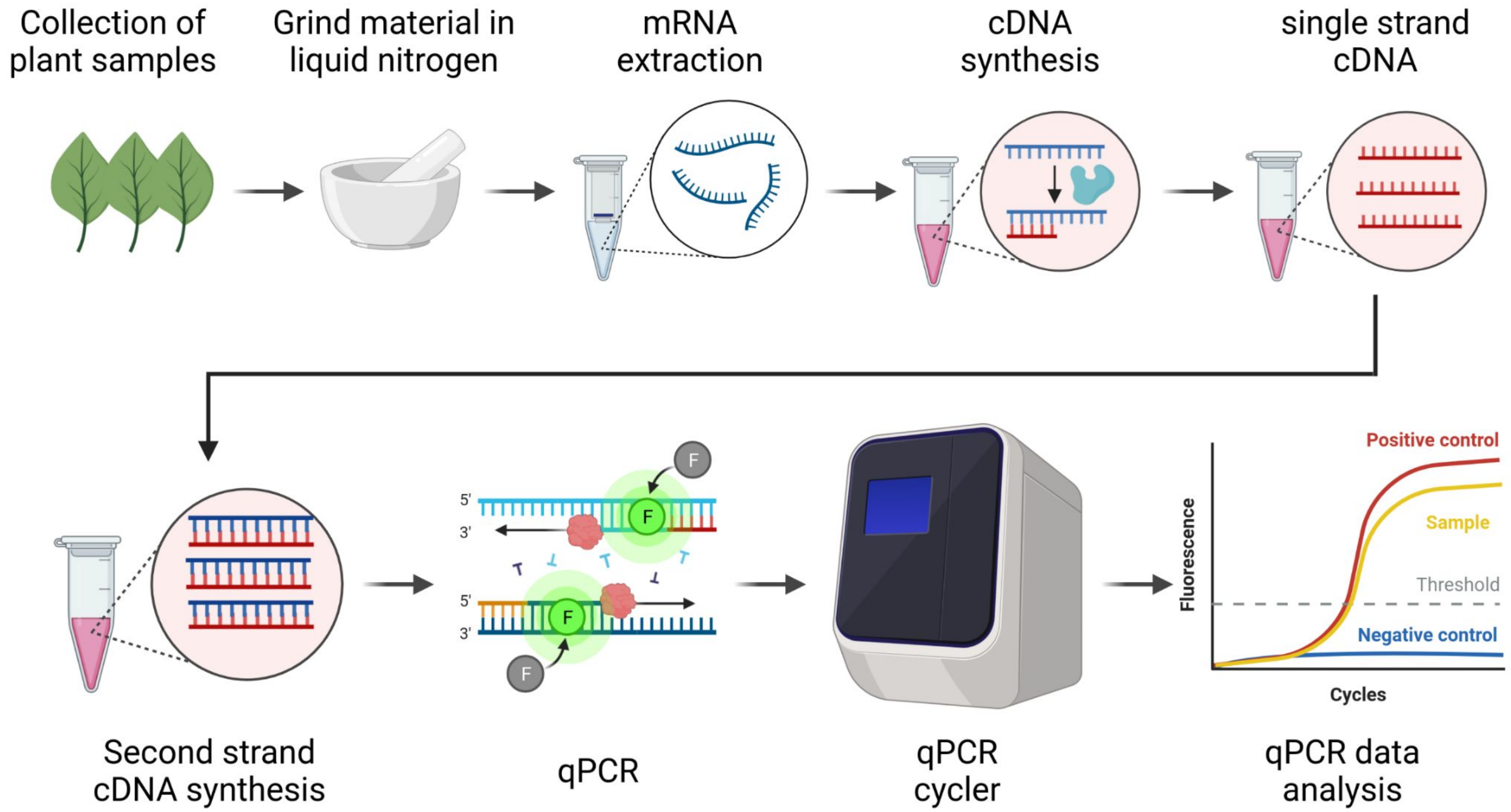
Microarray

- 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



Collection of plant samples

Grind material in liquid nitrogen

mRNA extraction

cDNA synthesis

single strand cDNA

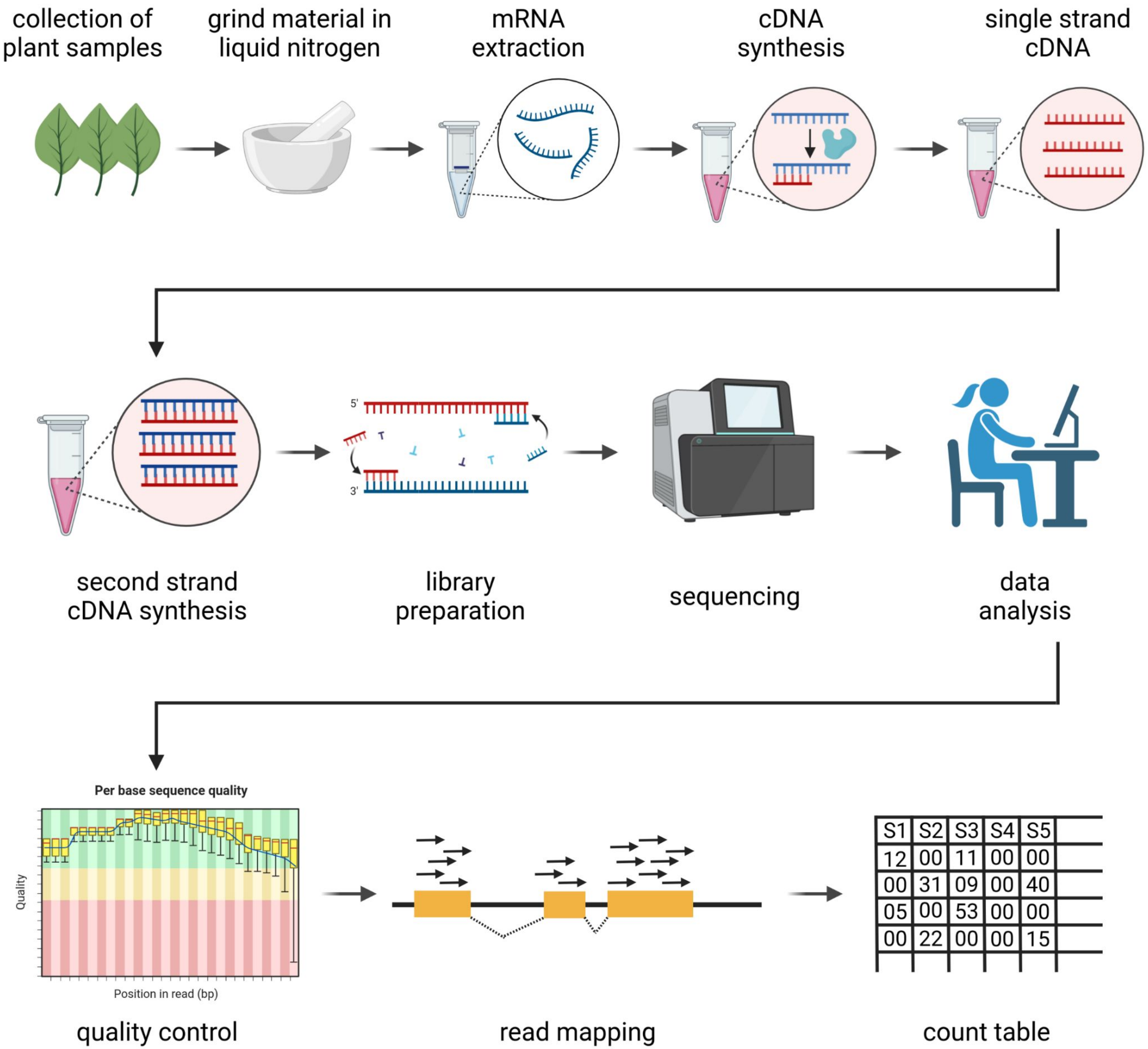
Second strand cDNA synthesis

qPCR

qPCR cycler

qPCR data analysis

RNA-Seq



collection of plant samples

grind material in liquid nitrogen

mRNA extraction

cDNA synthesis

single strand cDNA

second strand cDNA synthesis

library preparation

sequencing

data analysis

quality control

read mapping

count table

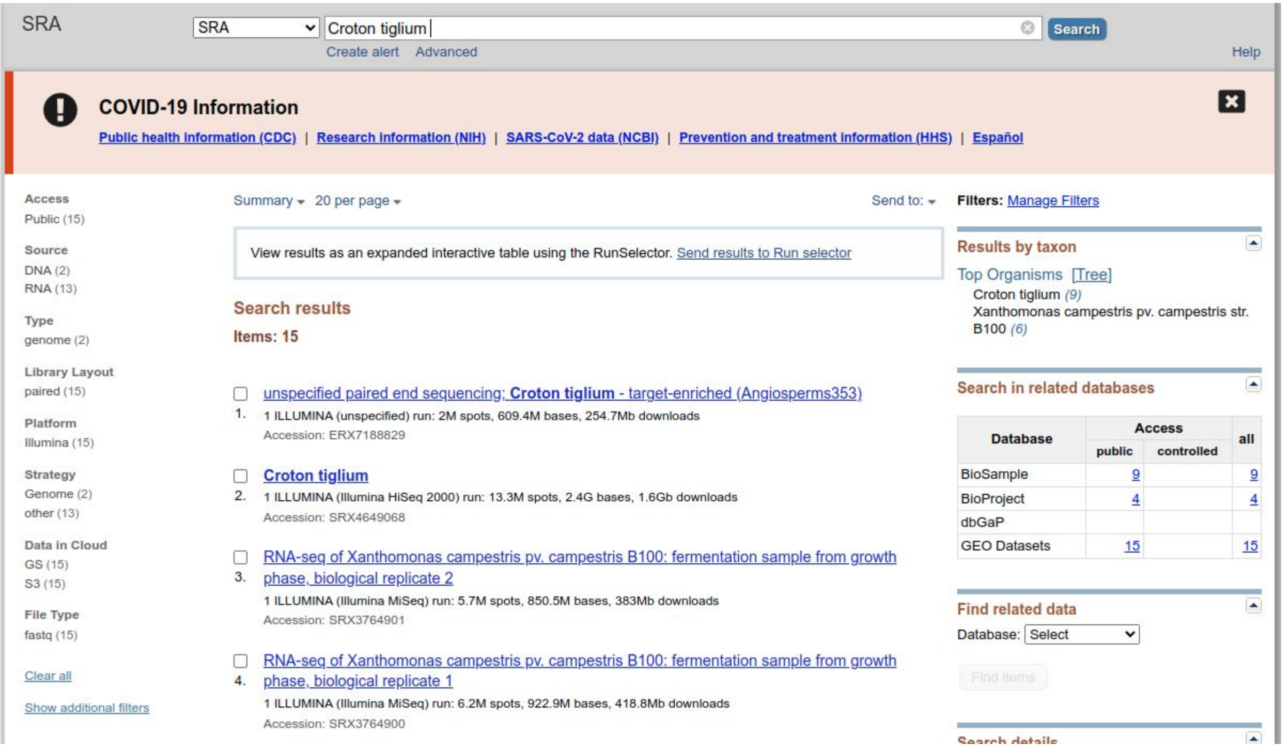
S1	S2	S3	S4	S5
12	00	11	00	00
00	31	09	00	40
05	00	53	00	00
00	22	00	00	15

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



SRA: Croton tiglium

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Access
Public (15)

Source
DNA (2)
RNA (13)

Type
genome (2)

Library Layout
paired (15)

Platform
Illumina (15)

Strategy
Genome (2)
other (13)

Data in Cloud
GS (15)
S3 (15)

File Type
fastq (15)

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 15

1. □ unspecified paired end sequencing; **Croton tiglium** - target-enriched (Angiosperms353)
   1 ILLUMINA (unspecified) run: 2M spots, 609.4M bases, 254.7Mb downloads
   Accession: ERX7188829
2. □ **Croton tiglium**
   1 ILLUMINA (Illumina HiSeq 2000) run: 13.3M spots, 2.4G bases, 1.6Gb downloads
   Accession: SRX4649068
3. □ RNA-seq of Xanthomonas campestris pv. campestris B100: fermentation sample from growth phase, biological replicate 2
   1 ILLUMINA (Illumina MiSeq) run: 5.7M spots, 850.5M bases, 383Mb downloads
   Accession: SRX3764901
4. □ RNA-seq of Xanthomonas campestris pv. campestris B100: fermentation sample from growth phase, biological replicate 1
   1 ILLUMINA (Illumina MiSeq) run: 6.2M spots, 922.9M bases, 418.8Mb downloads
   Accession: SRX3764900

Results by taxon

Top Organisms [Tree]
Croton tiglium (9)
Xanthomonas campestris pv. campestris str. B100 (6)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	9		9
BioProject	4		4
dbGaP			
GEO Datasets	15		15

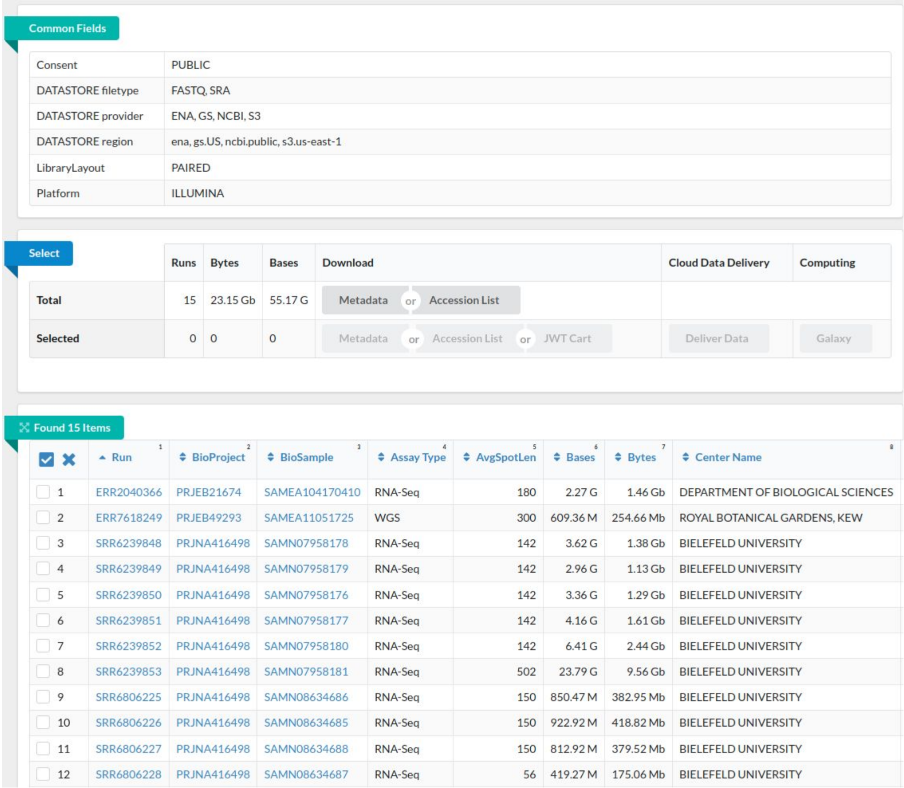
Find related data

Search details

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



Common Fields

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

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	15	23.15 Gb	55.17 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 15 Items

☑ ✖	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	SAMN07958178	RNA-Seq	142	3.62 G	1.38 Gb	BIELEFELD UNIVERSITY
□ 4	SRR6239849	PRJNA416498	SAMN07958179	RNA-Seq	142	2.96 G	1.13 Gb	BIELEFELD UNIVERSITY
□ 5	SRR6239850	PRJNA416498	SAMN07958176	RNA-Seq	142	3.36 G	1.29 Gb	BIELEFELD UNIVERSITY
□ 6	SRR6239851	PRJNA416498	SAMN07958177	RNA-Seq	142	4.16 G	1.61 Gb	BIELEFELD UNIVERSITY
□ 7	SRR6239852	PRJNA416498	SAMN07958180	RNA-Seq	142	6.41 G	2.44 Gb	BIELEFELD UNIVERSITY
□ 8	SRR6239853	PRJNA416498	SAMN07958181	RNA-Seq	502	23.79 G	9.56 Gb	BIELEFELD UNIVERSITY
□ 9	SRR6806225	PRJNA416498	SAMN08634686	RNA-Seq	150	850.47 M	382.95 Mb	BIELEFELD UNIVERSITY
□ 10	SRR6806226	PRJNA416498	SAMN08634685	RNA-Seq	150	922.92 M	418.82 Mb	BIELEFELD UNIVERSITY
□ 11	SRR6806227	PRJNA416498	SAMN08634688	RNA-Seq	150	812.92 M	379.52 Mb	BIELEFELD UNIVERSITY
□ 12	SRR6806228	PRJNA416498	SAMN08634687	RNA-Seq	56	419.27 M	175.06 Mb	BIELEFELD UNIVERSITY

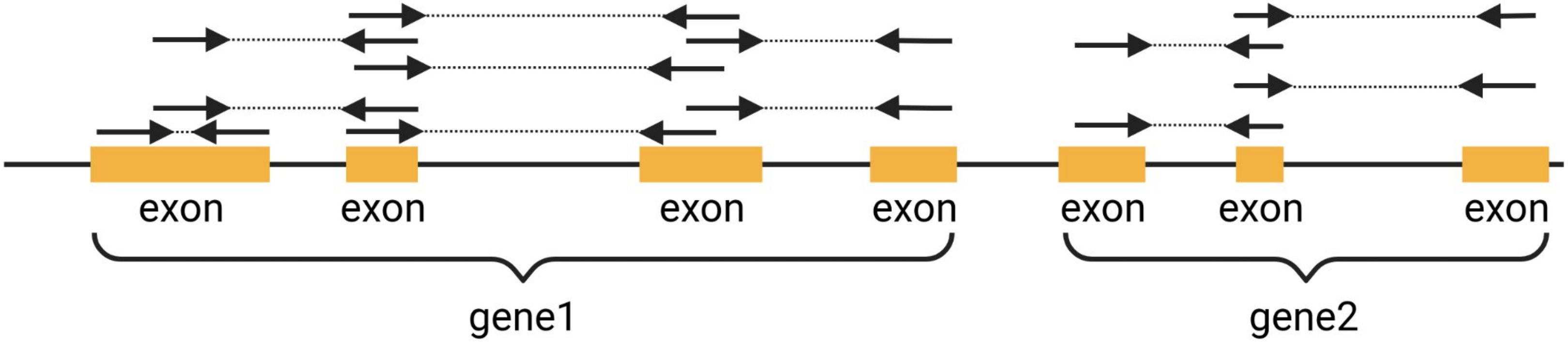
https://www.ncbi.nlm.nih.gov/sra

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

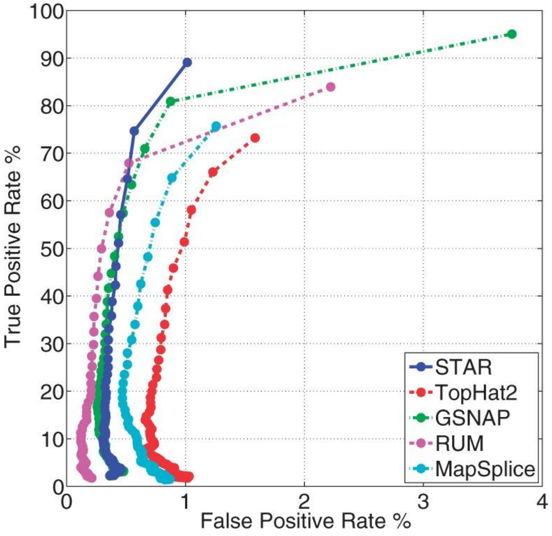
Concept of gene expression quantification

- 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

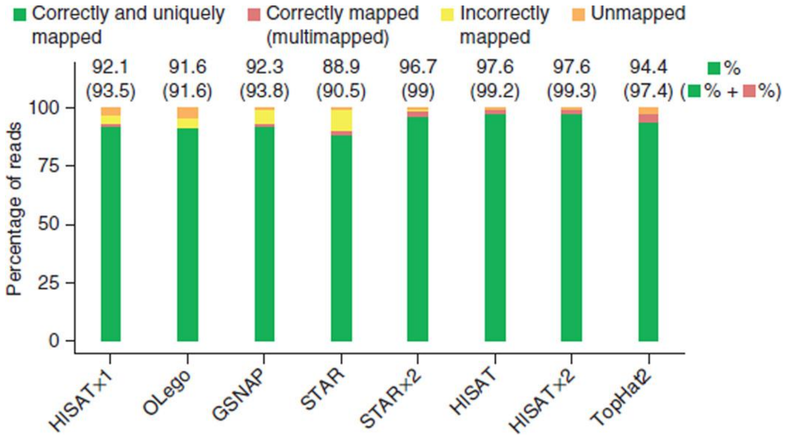


Processing expression data

- 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:

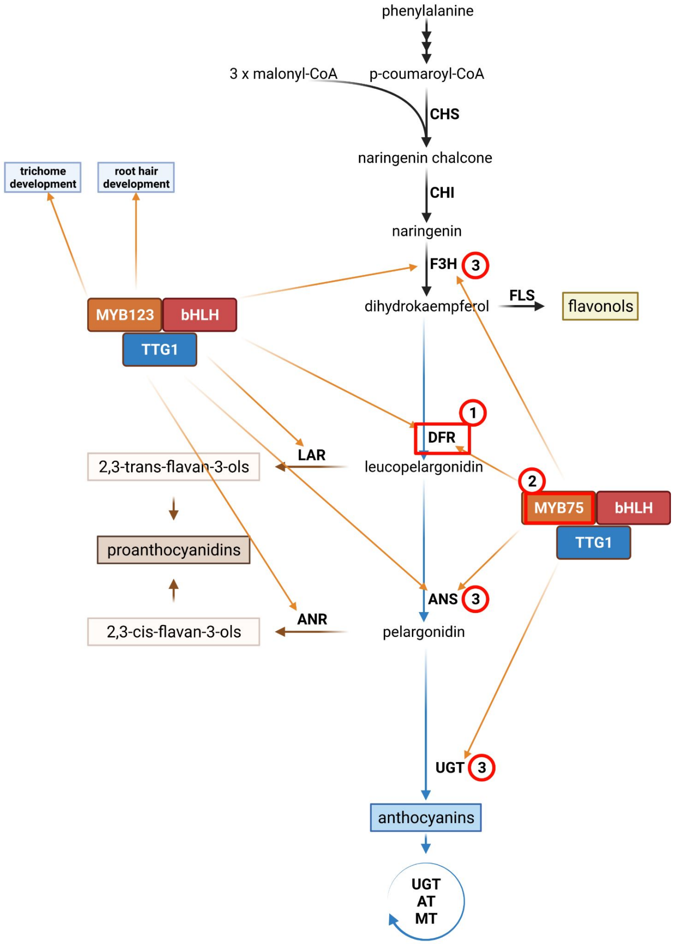
- Counts: gene1=12, gene2=3, gene3=5
- Transcript lengths: gene1=1.5kb, gene2=1kb, gene3=3kb
- TPMs (simplified approximation):
    - gene1 = 12 / ((12+3+5)/1000000)
    - gene2 = 3 / ((12+3+5)/1000000)
- RPKMs:
    - gene1 = 12 / (1.5 * ((12+3+5)/1000000))
- FPKMs:
    - same as RPKM, but for paired-end

Baits for coexpression analyses

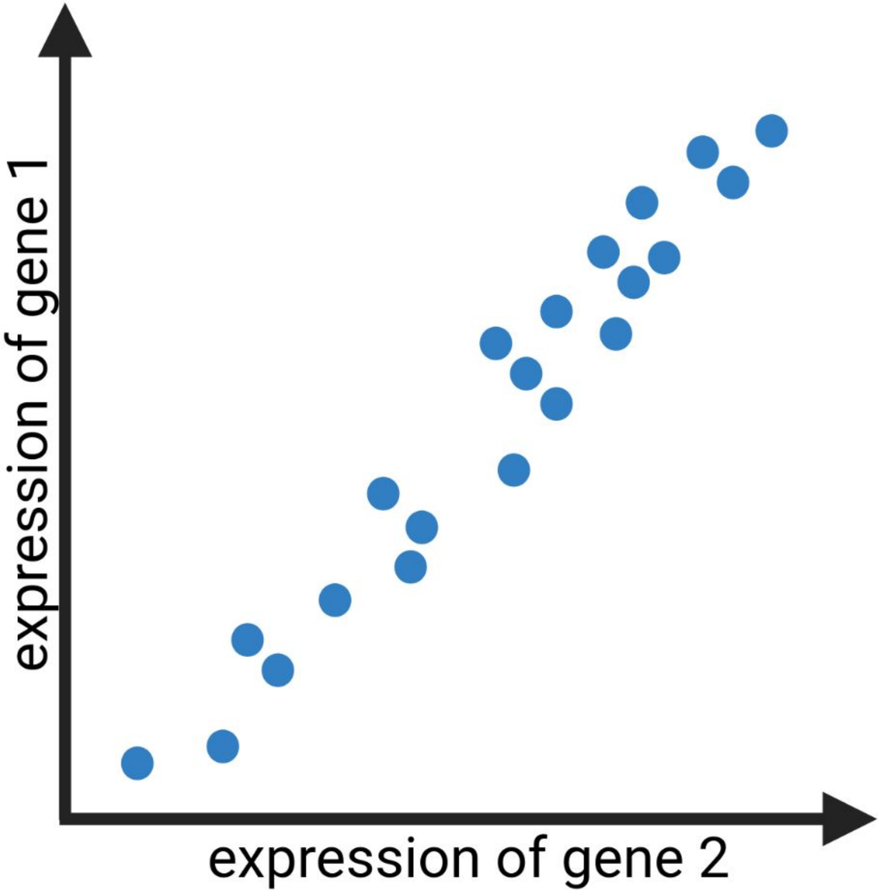
- Bait genes are previously characterized genes with a function of interest e.g. encode an enzyme in the same biosynthesis pathway

- Shared transcription factors of a pathway can be helpful to identify all structural genes of a pathway

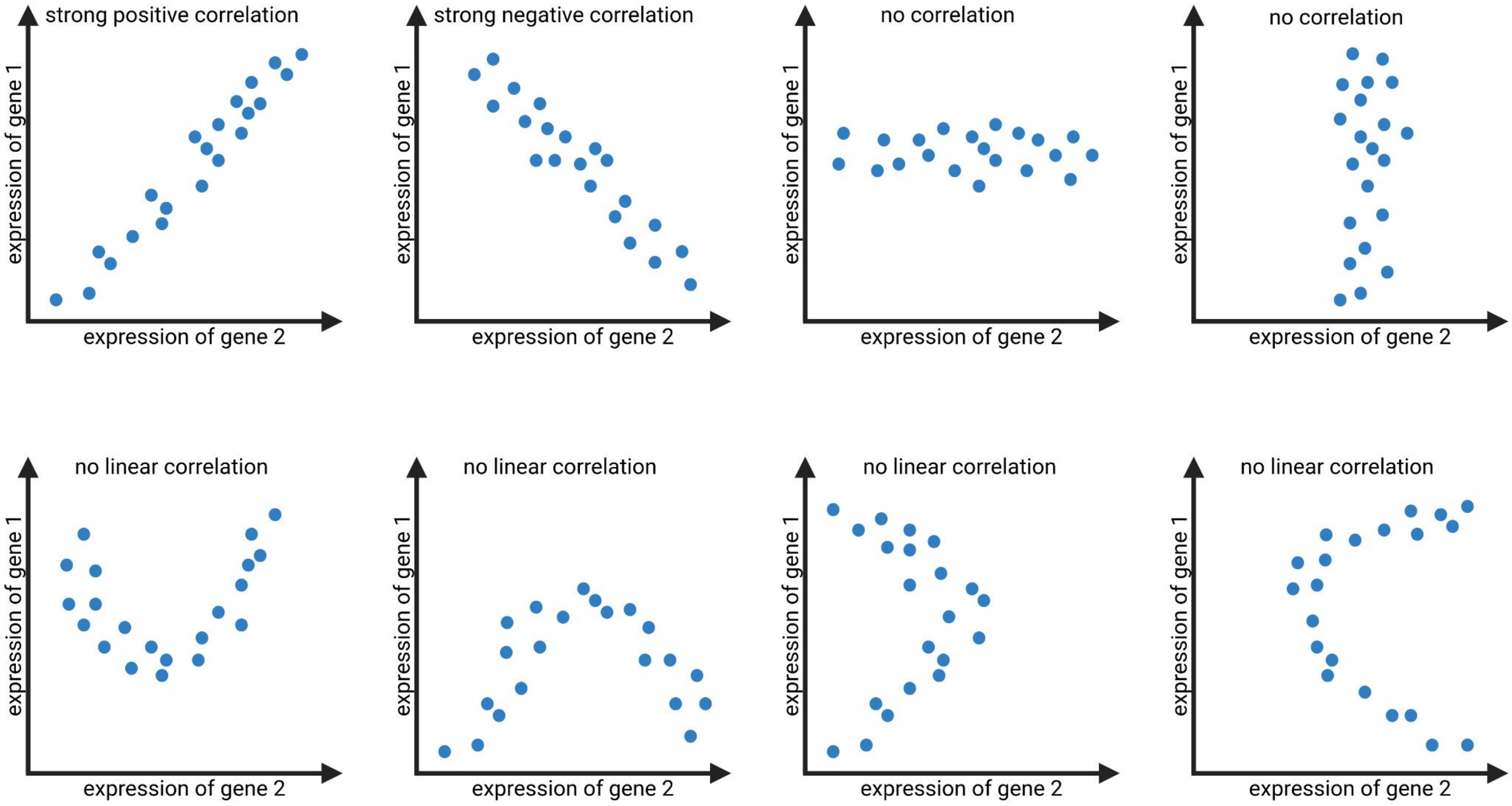
- Knowledge from other species can be applied in this step (details in later section)



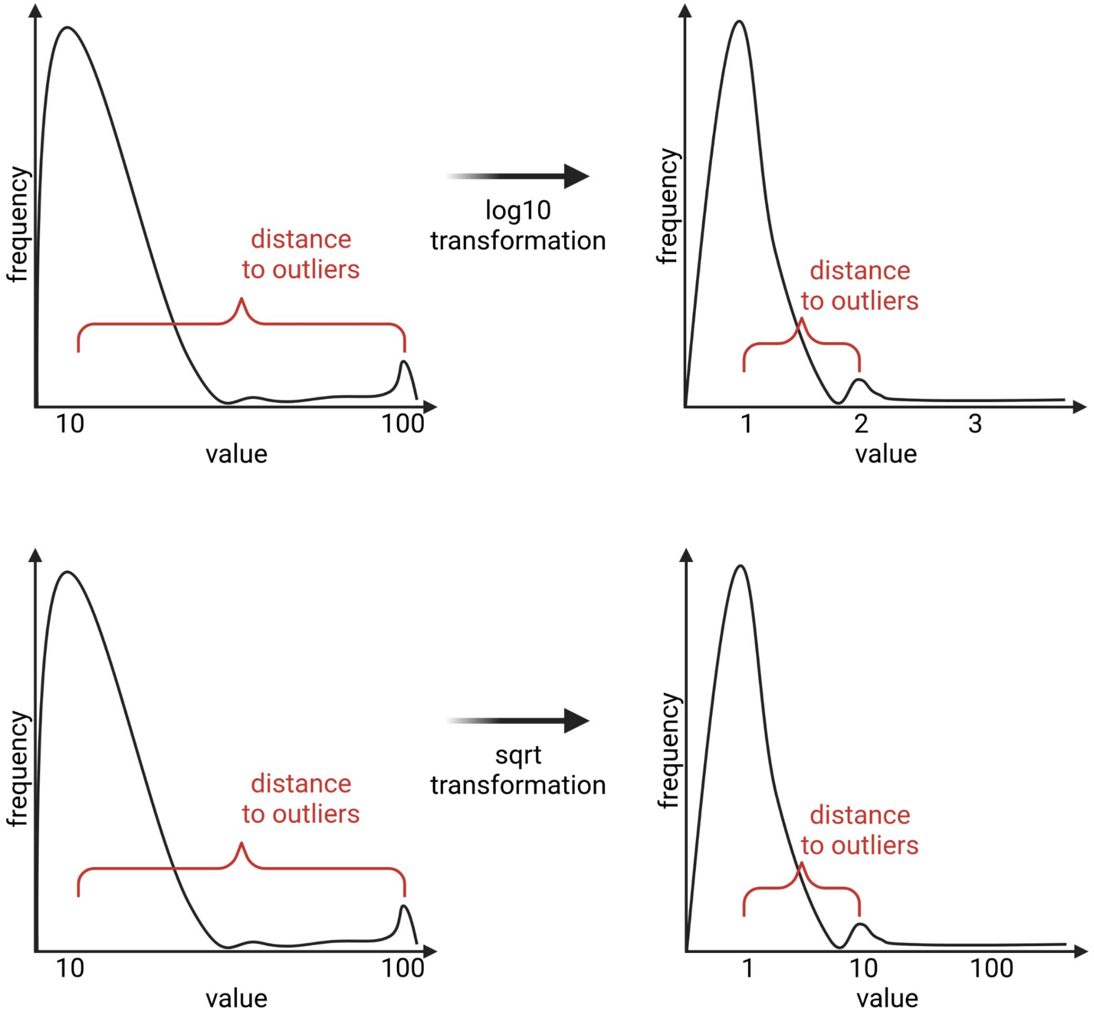
Coexpression



Correlation - examples



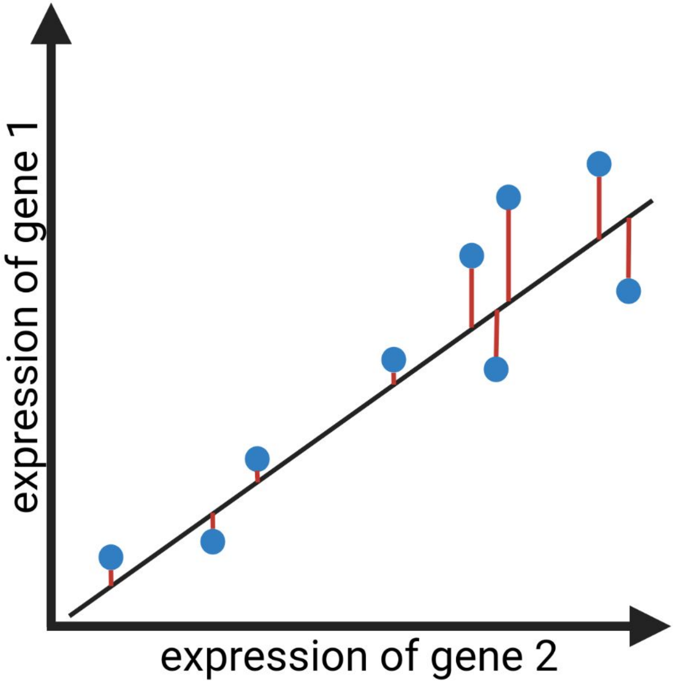
Normalization



Pearson correlation coefficient

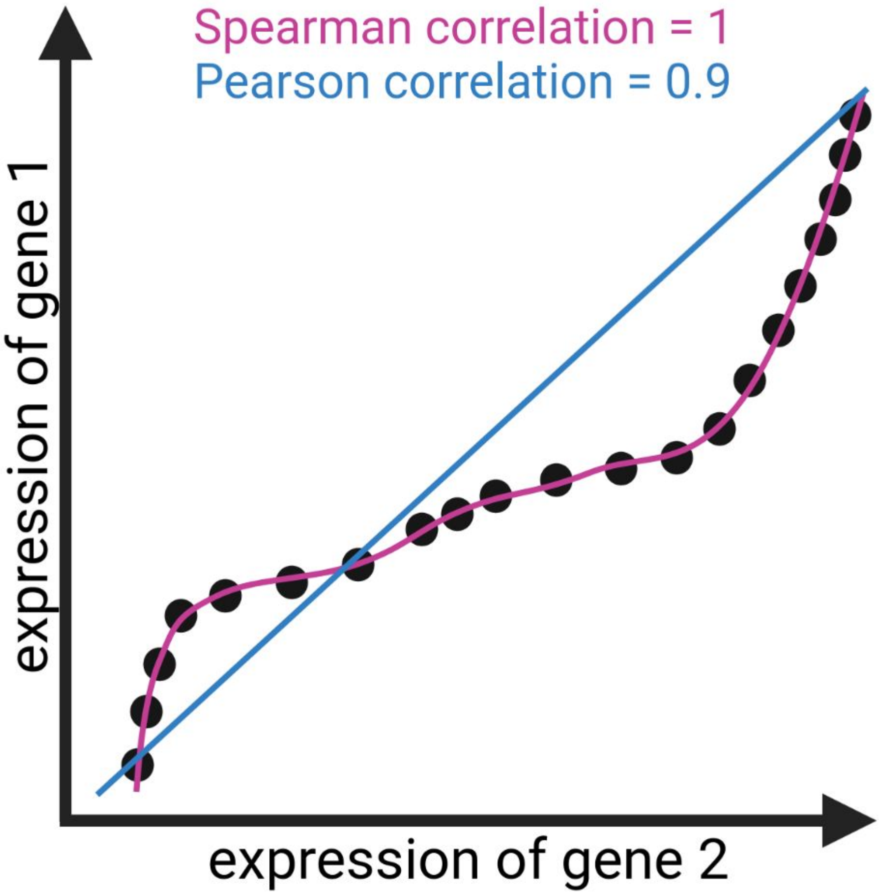
- Line is fitted to achieve minimal distance of all data points to the line

- 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



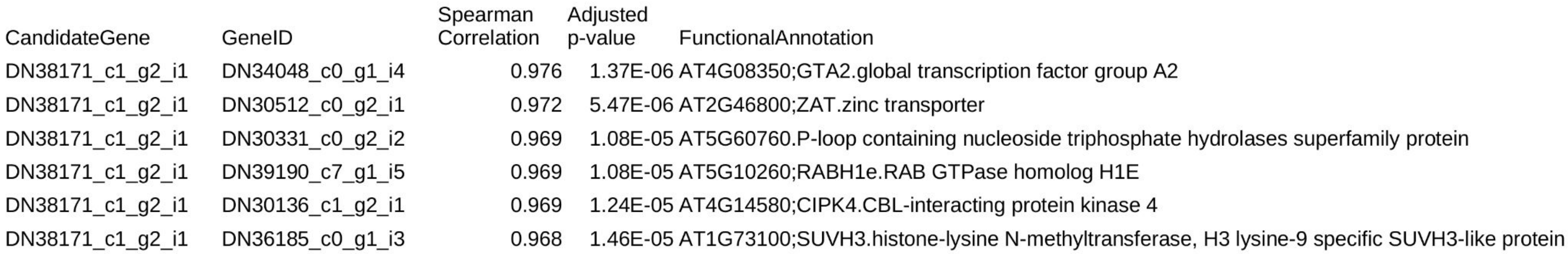
Simple coexpression analysis

- 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*

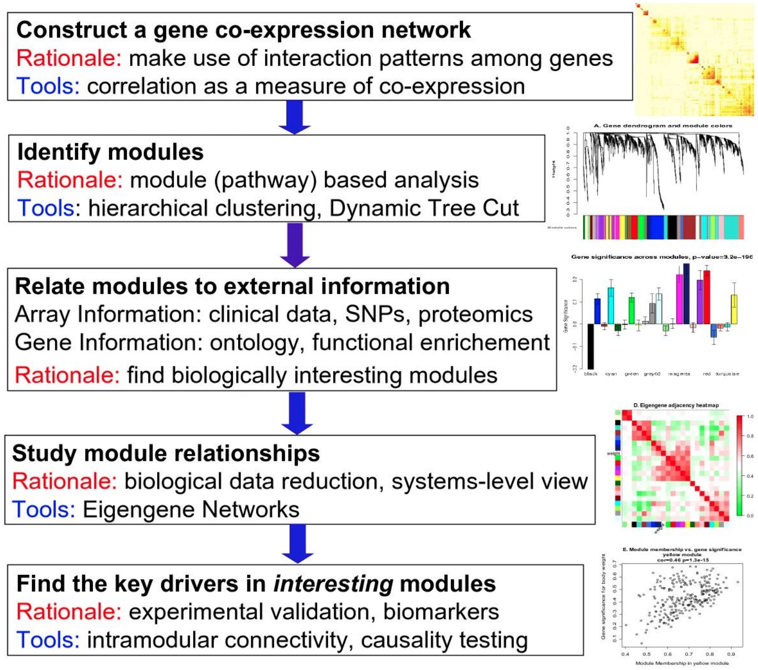


CandidateGene	GeneID	Spearman Correlation	Adjusted p-value	FunctionalAnnotation
DN38171_c1_g2_i1	DN34048_c0_g1_i4	0.976	1.37E-06	AT4G08350;GTA2.global transcription factor group A2
DN38171_c1_g2_i1	DN30512_c0_g2_i1	0.972	5.47E-06	AT2G46800;ZAT.zinc transporter
DN38171_c1_g2_i1	DN30331_c0_g2_i2	0.969	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



Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression

Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut

Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules

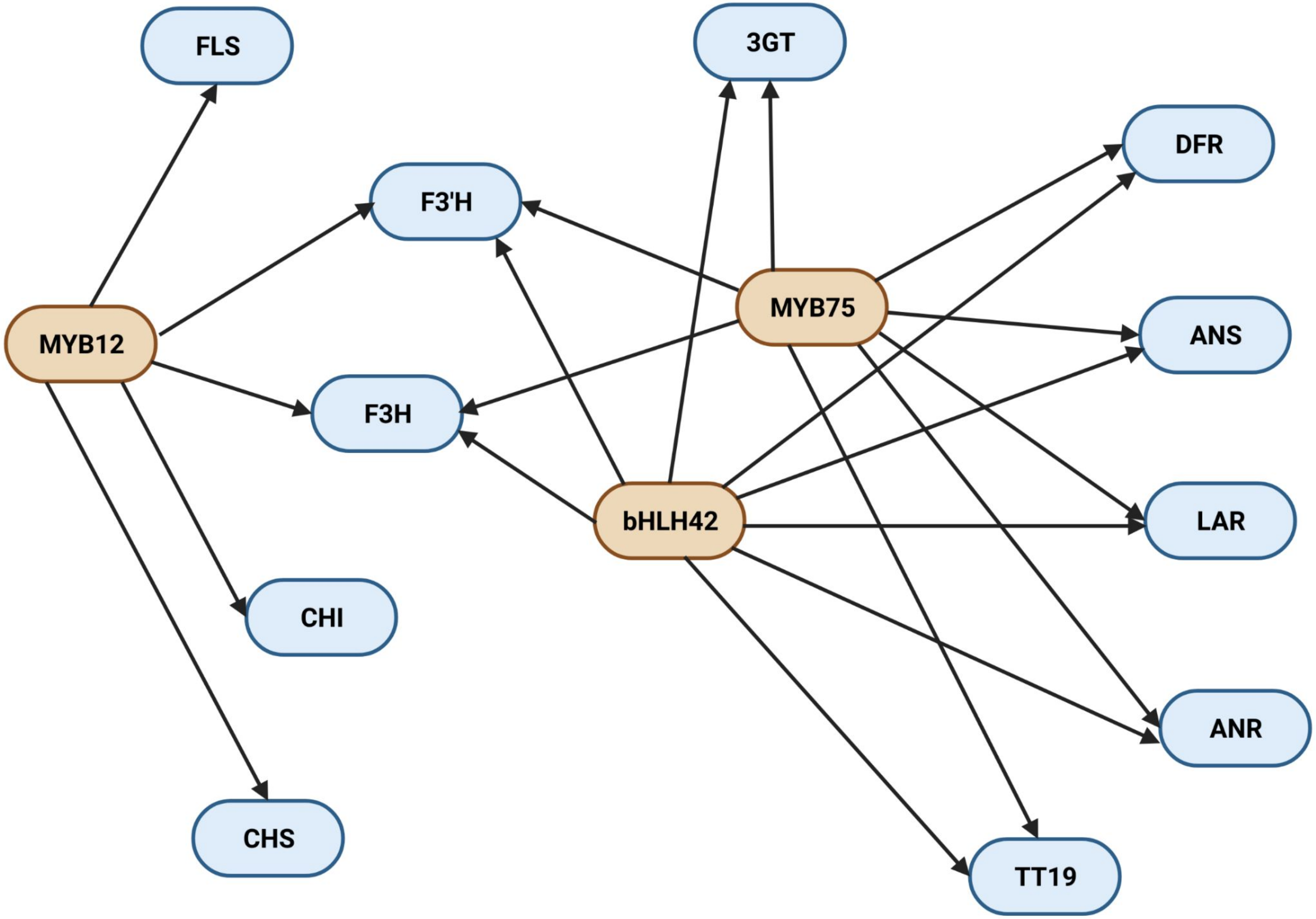
Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks

Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing

A. Gene dendrogram and module colors
Height
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Module colors
Gene significance across modules, p-value=3.2e-196
Gene Significance
-0.2 -0.1 0.0 0.1 0.2
black cyan green grey60 magenta red turquoise
D. Eigengene adjacency heatmap
weight
1.0 0.8 0.6 0.4 0.2 0.0
E. Module membership vs. gene significance
yellow module
cor=0.46 p=1.3e-15
Gene significance for body weight
0.1 0.2 0.3 0.4 0.5 0.6 0.7
0.4 0.5 0.6 0.7 0.8 0.9
Module Membership in yellow module

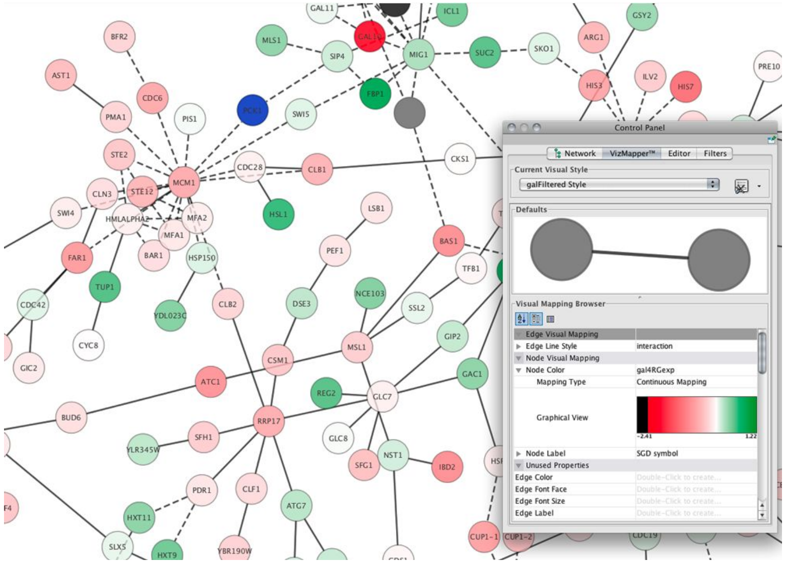
Langfelder & Horvath, 2008: 10.1186/1471-2105-9-559

Coexpression network example



Cytoscape

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

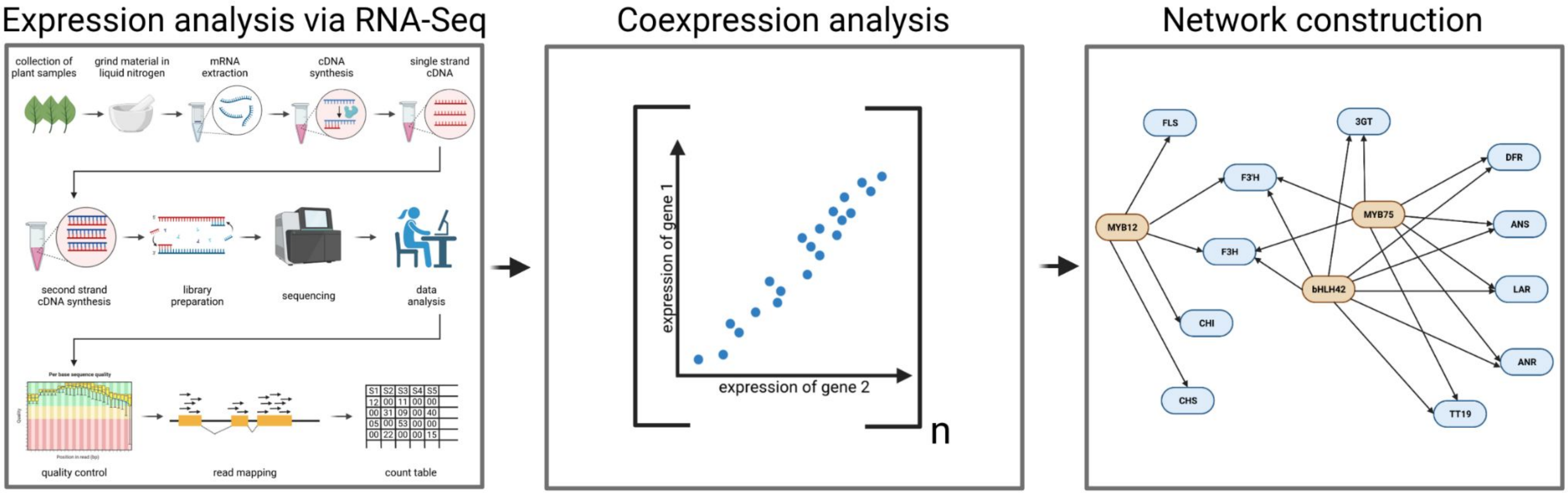


Current Visual Style: galFiltered Style

Edge Visual Mapping	
Edge Line Style	interaction
Node Visual Mapping	
Node Color	gal4RGexp
Mapping Type	Continuous Mapping
Graphical View	
Node Label	SGD symbol
Unused Properties	
Edge Color	Double-Click to create...
Edge Font Face	Double-Click to create...
Edge Font Size	Double-Click to create...
Edge Label	Double-Click to create...

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

Summary of the process



S1	S2	S3	S4	S5
12	00	11	00	00
00	31	09	00	40
05	00	53	00	00
00	22	00	00	15

Thank you!

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?