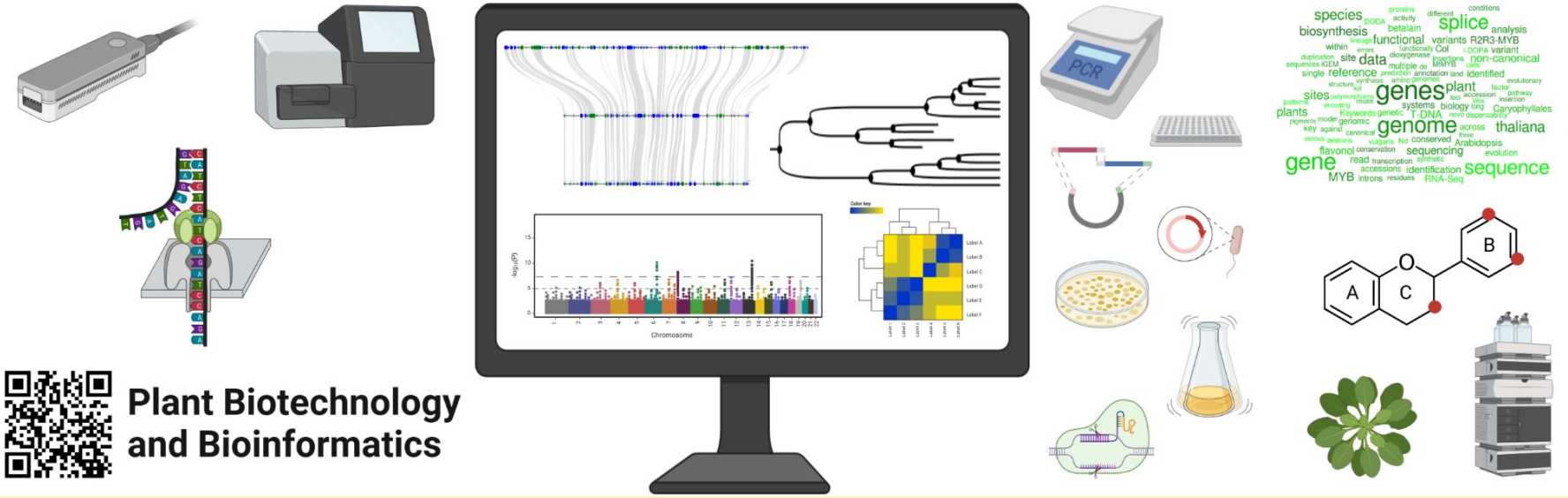
Technische Universität Braunschweig



Plant Biotechnology and Bioinformatics

Transcriptomics - Big Data Analytics in Life Sciences

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

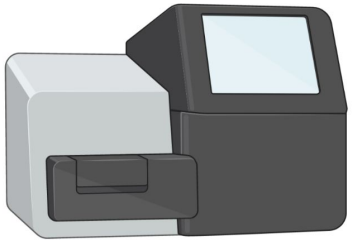
Outline - Transcriptomics

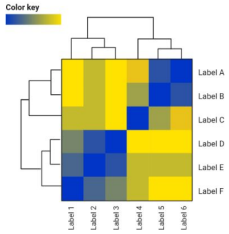
- Concept of RNA-seq & workflow (wet lab)
- Experimental design considerations
- Read mapping & quantification
- De novo transcriptome assembly
- PCA, Heatmaps, DEG identification
- Re-using public RNA-seq datasets
- Direct RNA sequencing & scRNA-seq

What causes these pigmentation patterns?

What is transcriptomics?

- Study of the transcriptome with high-throughput methods
- Characterized by technological progress and big data analyses





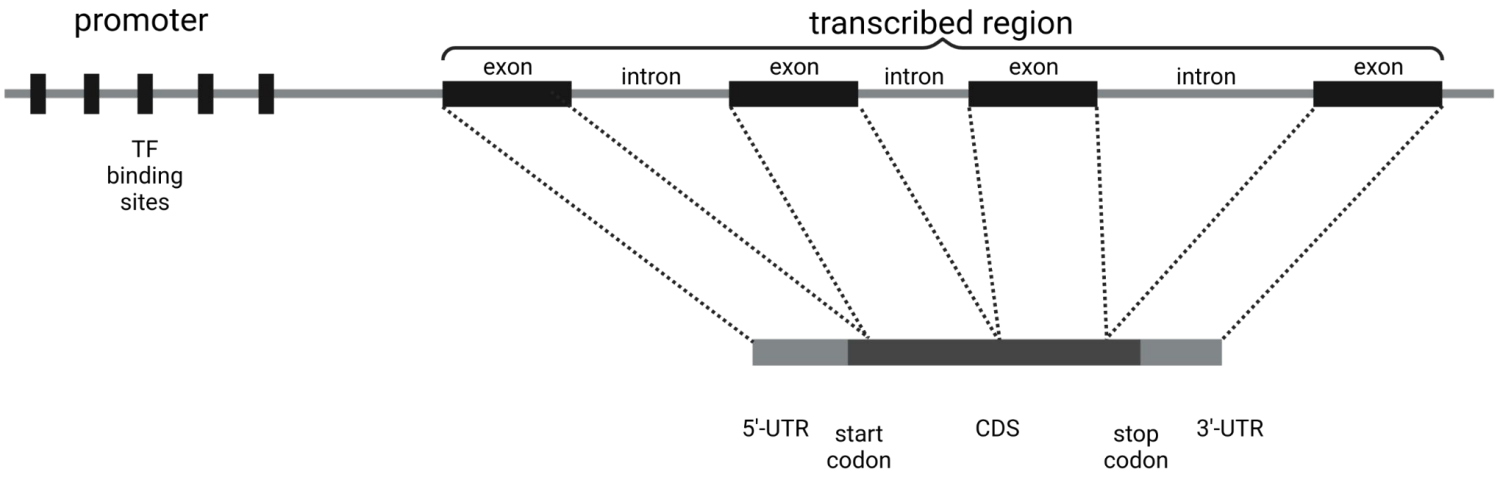
What is the transcriptome?

- “The set of all transcripts and their abundances in a defined cell/tissue/organism under defined conditions at a certain time point.”
- Highly variable over time
- Responding to various stimuli

What is a gene?

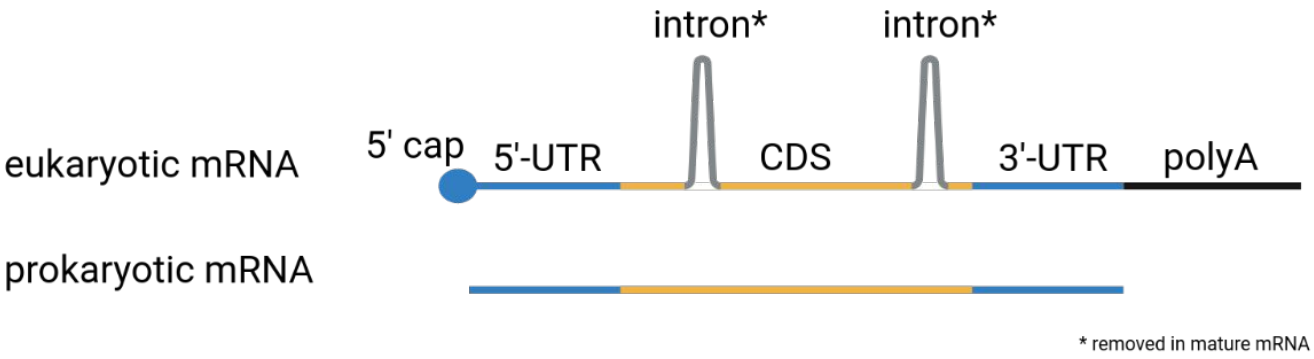
- No perfect and universal definition

- Restrict to protein coding gene in plants:
  - promoter region
  - UTRs
  - coding sequence
  - introns



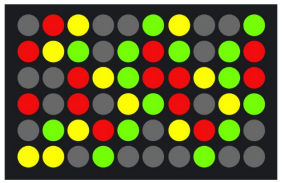
Prokaryotes vs. eukaryotes - transcripts

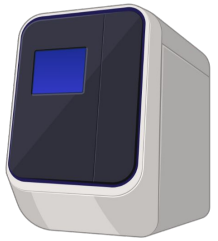
- Eukaryotic genes can harbour introns
- Eukaryotic genes have a 5’-cap
- Eukaryotic genes have a poly-A tail

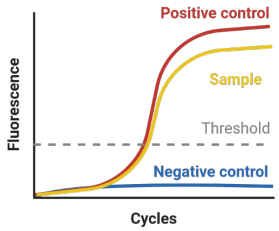


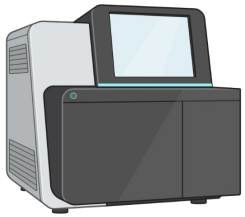
How to measure gene expression?

- Northern blot
- Reverse transcription quantitative PCR (RT-qPCR)
- Microarray
- Expressed Sequence Tags (ESTs)
- Serial Analysis of Gene Expression (SAGE)
- RNA-seq (not RNA sequencing!!)
- Droplet digital PCR (ddPCR)
- direct RNA sequencing

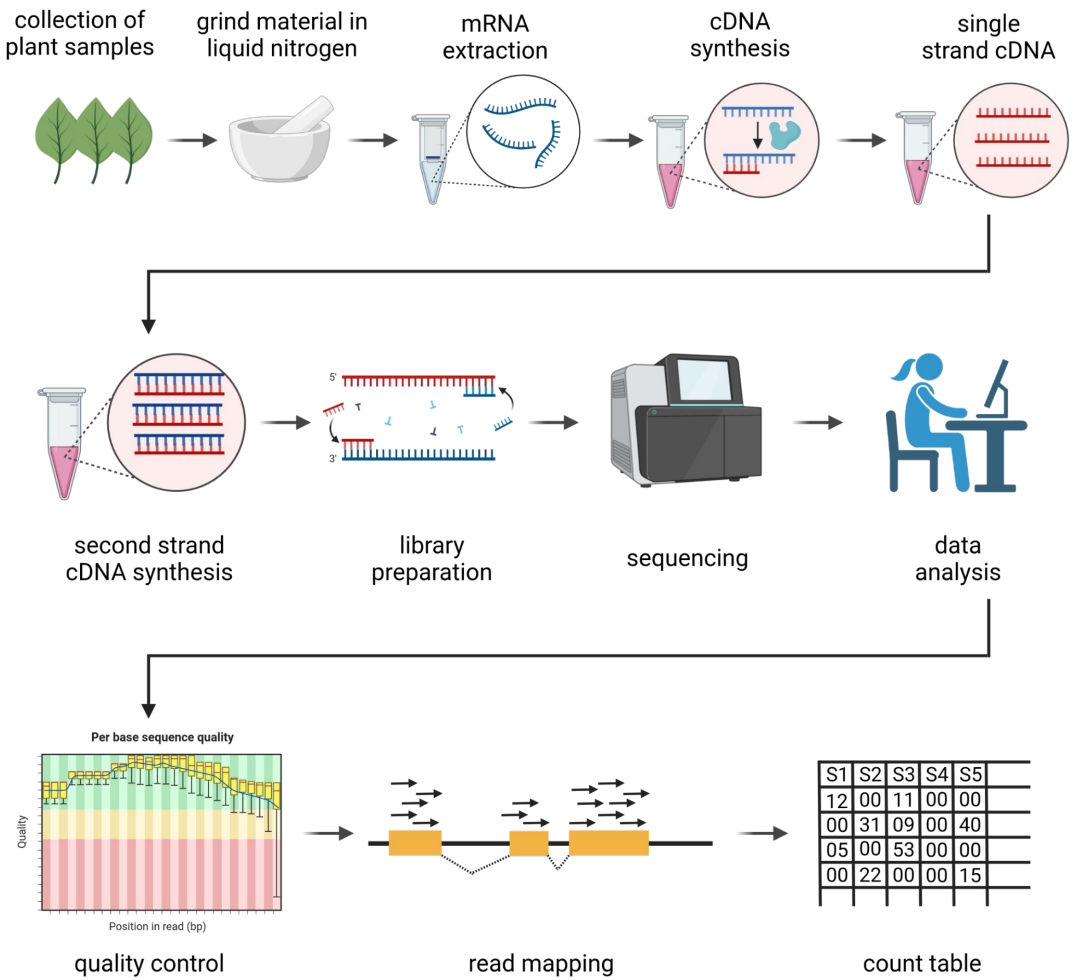






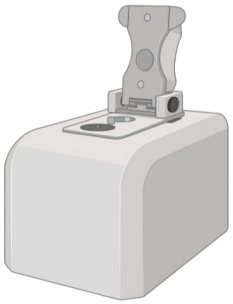


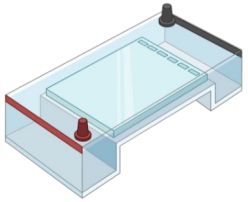
RNA-seq (not RNA sequencing!)

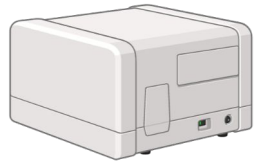


RNA quality assessment

- Evaluate quality through photometric measurement
- Check quality on gel
- Measure with fluorescence
- Check RNA integrity via Agilent chip





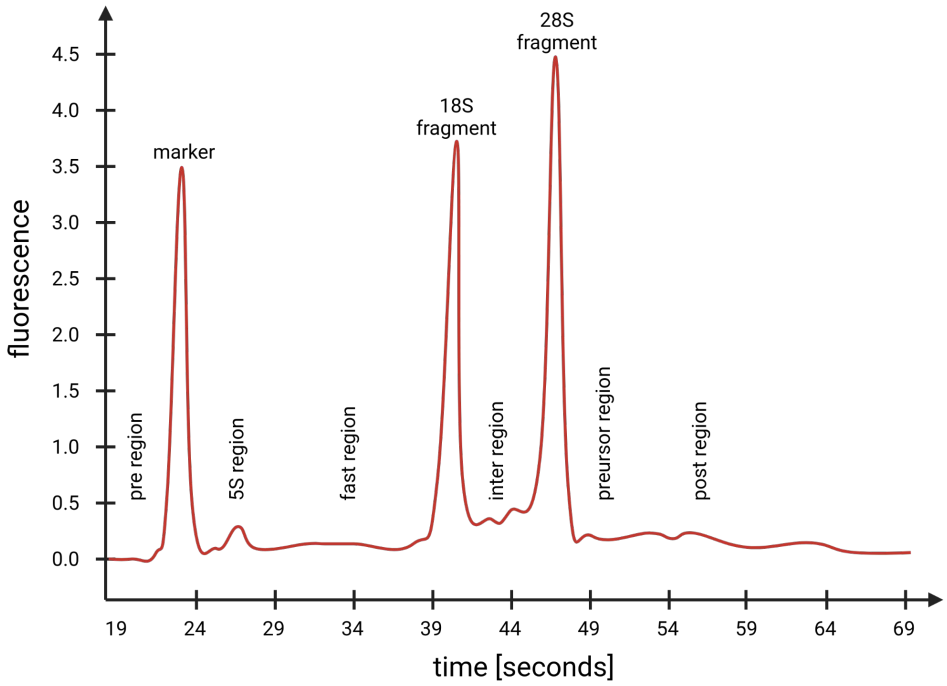


NanoDrop: photometric measurement

- Photometric measurement in tiny volumen
- No dilution required
- Analysis of RNA, DNA, and protein possible
- OD260/OD230 = contamination with small fragments or phenolic compounds
- OD260/OD280 = contamination with protein

RNA Integrity Number (RIN)

- Previously: ratio of 28S to 18S rRNA as indicator
- mRNA degrades faster than rRNA and in non-linear way
- RIN: inference of mRNA integrity based on overall RNA integrity
- Chip for analysis is like running a gel with higher resolution



Shipping

- Shipping to (international) service provider
- Dry ice (solid CO_2) is used to keep RNA samples frozen
- RNA-seq costs per sample: < $200

Depletion of rRNA

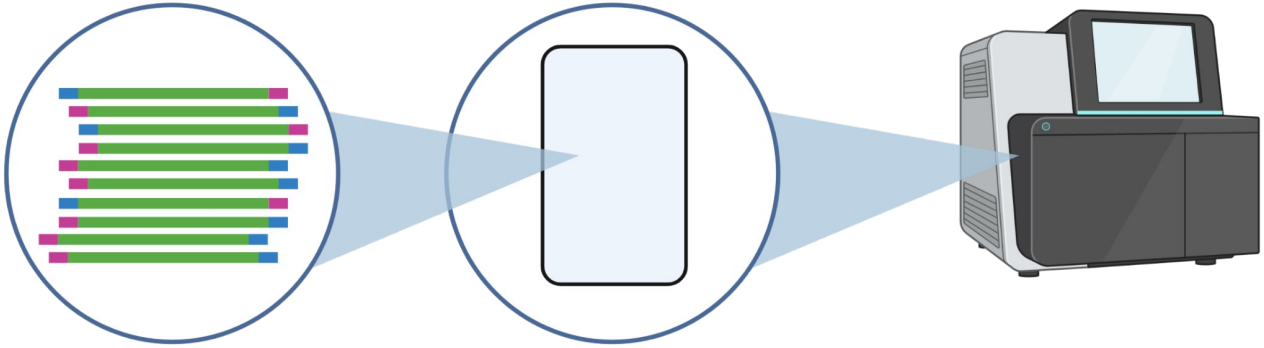
- rRNAs account for >80% of all RNAs
- rRNA probes are bound to magnetic beads
- Binding of probe to rRNA enables pull down and removal of rRNA
- Risk: pull down and removal might be incomplete

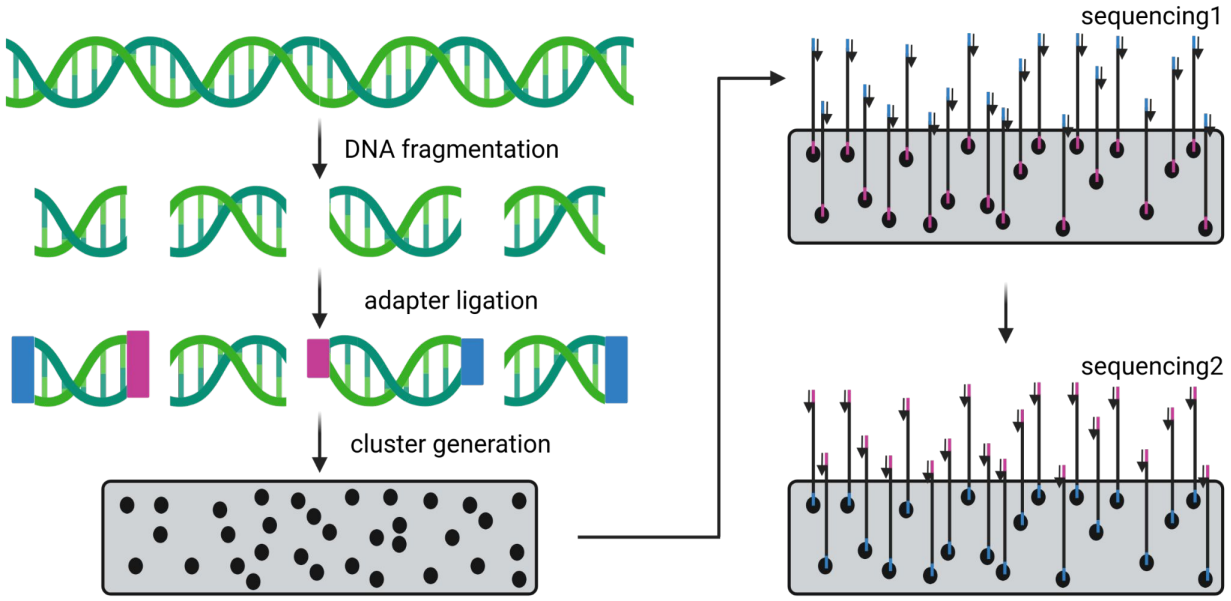
Enrichment of mRNA

- mRNAs are characterized by long poly-A tails
- Binding of mRNAs to oligo-T beads/columns
- Risk1: Degraded mRNAs (without poly-A) are lost; strong bias for 3'-end
- Risk2: Other RNAs with long A stretches might bind

cDNA synthesis

Illumina sequencing of cDNAs





Experimental design

- Number of replicates determines power of analysis
- Experiments can have one or multiple factors (e.g. time, temperature, infection)
- Genotypes of individuals must not introduce a bias

It is all about comparison

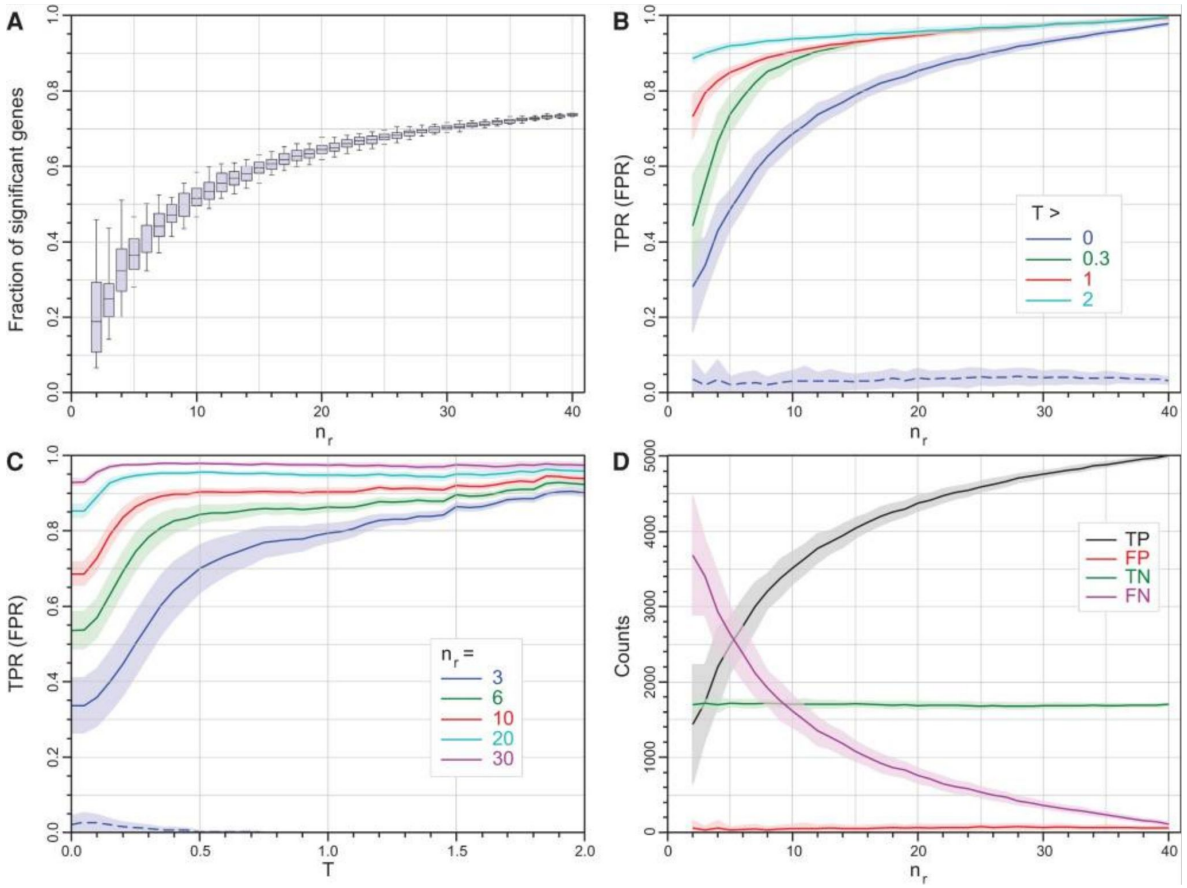
- Gene expression is usually analyzed in a comparative way
  - Between samples: fold change differences
  - Between transcript of same samples: transcript per million transcripts

- Absolute measurement of transcript numbers is extremely tricky

- Differentially expressed genes (DEGs)

Replicates vs. sequencing depth

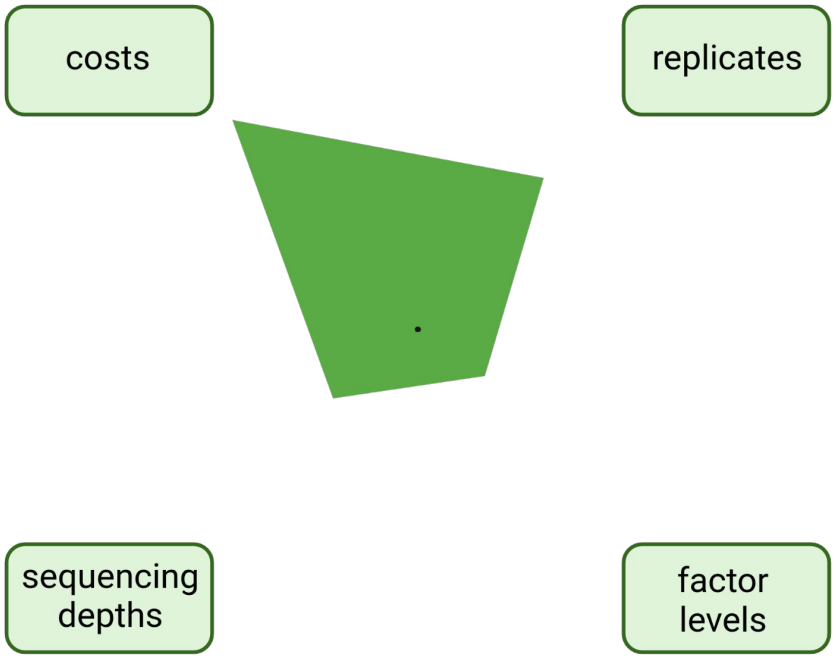
- Number of replicates determines power of analysis
- 3 biological replicates is considered minimum
- 12 biological replicates are recommended
- Technical replicates are not required for RNA-seq



Schurch et al., 2016: 10.1261/rna.053959.115

Trade offs

- Limited resources determine experimental setup
- Trade-offs between costs, factor levels, number of replicates, and sequencing depth



Importance of growth conditions

- Transcriptome responds to environmental conditions
- Development might be ongoing
- Conditions need to be the same for all replicates
- Regular rotation of study objectives avoids position bias; randomized positioning
- Precise documentation of temperature, humidity, provided substrate, infections, …

Nagoya protocol & ABS laws

- Original intention of Nagoya protocol was to protect genetic resources
- Currently one of the biggest obstacles to research projects
- German plants can be studied everywhere in the world
- Plants from abroad might require additional permissions and registration
- Getting permissions is often complicated
- Solution: only work with ‘free plants’

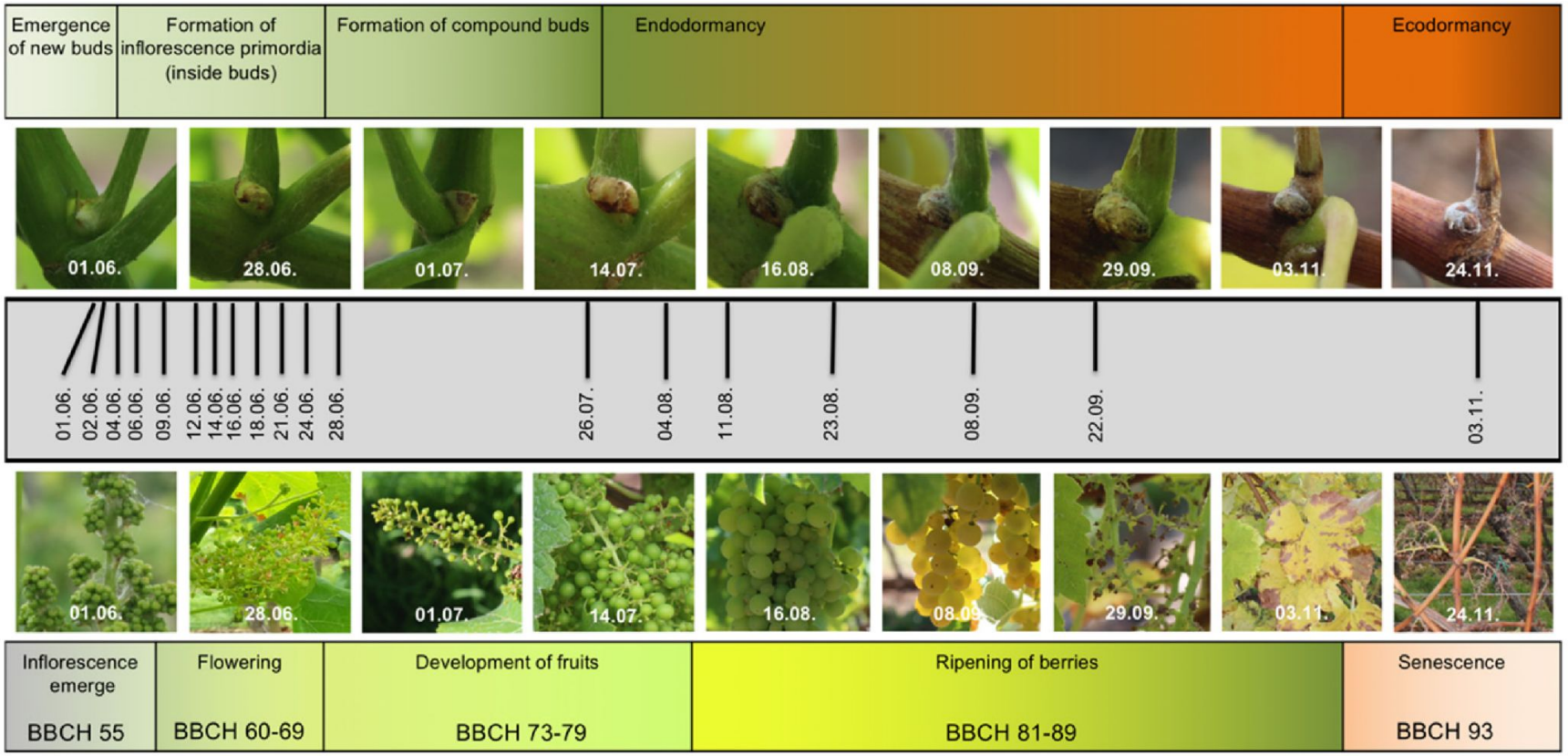
Phenotyping

- RNA-seq is often performed to investigate trait
- Phenotypic information about samples is important
- Phenotype examples:
  - presence of certain metabolite
  - visible color
  - morphological structure
  - resistance against pathogens

Metadata

- Growth conditions:
  - Light
  - Temperature
  - Humidity
  - Soil
- Precise time point of harvest (date, day time)
- Harvested plant part
- RNA extraction protocol
- Library construction protocol
- Details about sequencing

Example: time course experiments



https://doi.org/10.3390/plants9111548

RNA-seq data: FASTQ files

- Standard format for sequences with associated quality information
- Results of RNA-seq experiments are 1-2 FASTQ files (1 = SE, 2=PE)

- Four lines per entry:
  - Header starts with @ (title + description)
  - Sequence
  - + (optional repetition of header)
  - Quality (phred encoded in ASCII character)

- Different versions exist that use different quality values offsets

- Example:

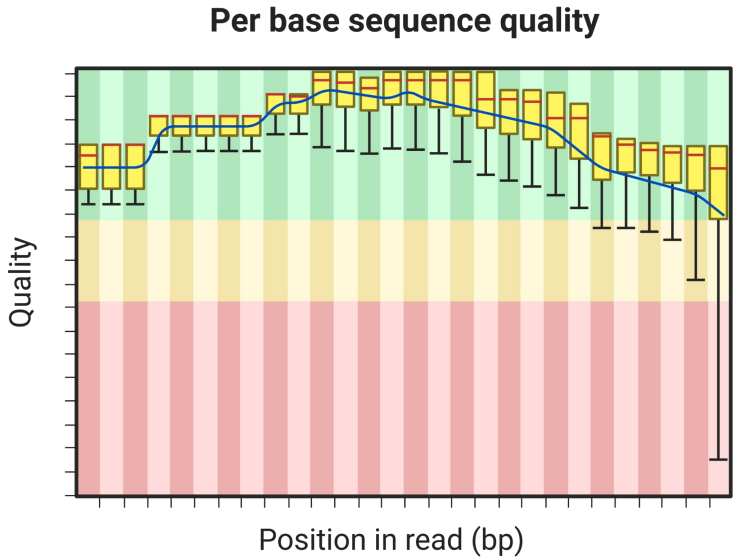
```
@seq1
ACGTACGTACGT
+
“”?FCB””:DC”
```

FASTQ example

```
@A00125:59:H3W7MDSXX:4:1101:4155:1016 1:N:0:ACCATCAC+TGATCTCC
GNCCCGTTCTCGGTATGCCGGTAGCTATGCTTGCATGTATGTGCAGATAGGTTAGGGGGGAAGCAACACTGTCTACGATCAAGTTGAACATATCTTGCTAC
+
F#FFFFFFFFF:FFF,FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF
@A00125:59:H3W7MDSXX:4:1101:8260:1016 1:N:0:ACCATCAC+TGATCTCC
CNGAATCAAGGTCGCCGAGGGACGACGATGGACACGGTAGTCATACATTAACAACAGCAGGAGGTTCTGAAGTTGCAGGAGCTAGCCTCTTTGGTTTAGCT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F::FFF:FFFFFF
@A00125:59:H3W7MDSXX:4:1101:10773:1016 1:N:0:ACCATCAC+TGATCTCC
CNCCTCTCTCTGCAACTAGCTACTGCCAATGACTAATACTCTAATCTTTTCTTCTTCATCAAAACTTTCTTTCTTACTTTTTCTATTTTATTTTTACCAGT
```

Quality checks

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

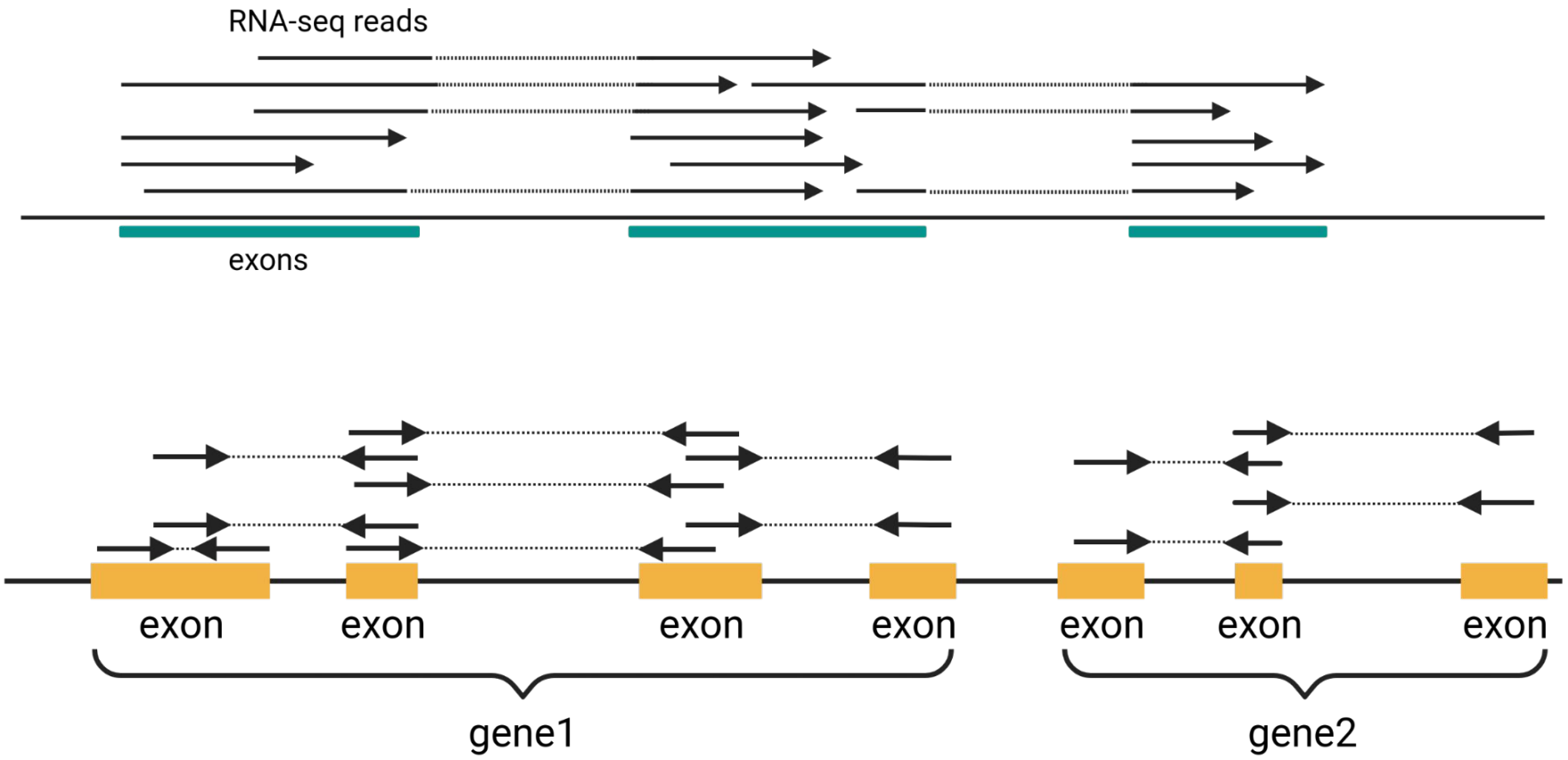


Phred score

- Phred score indicates the per base quality in an efficient way
- Single ASCII character is used to show the quality
- Formula: $Q = -10 \log_{10} P$ $P = 10^{-Q/10}$

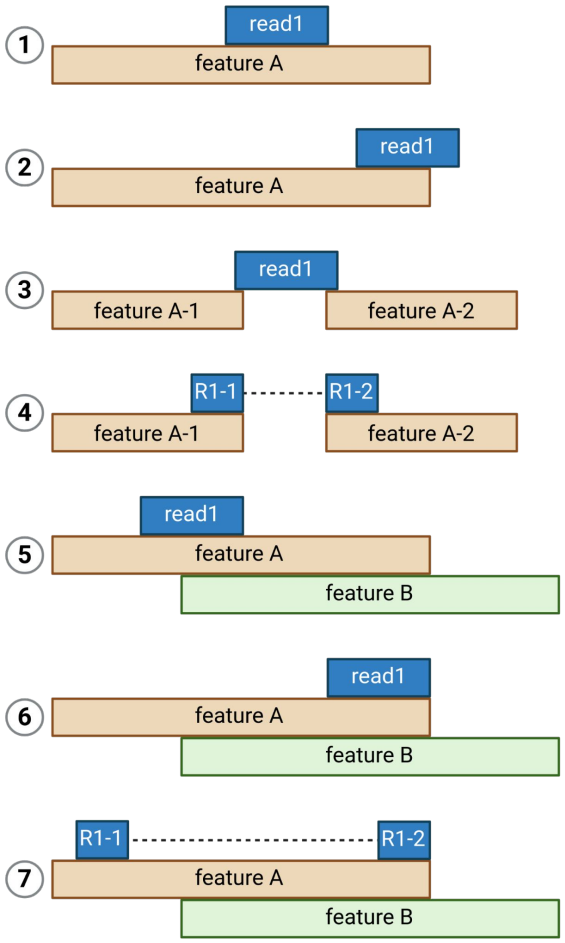
Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Read mapping



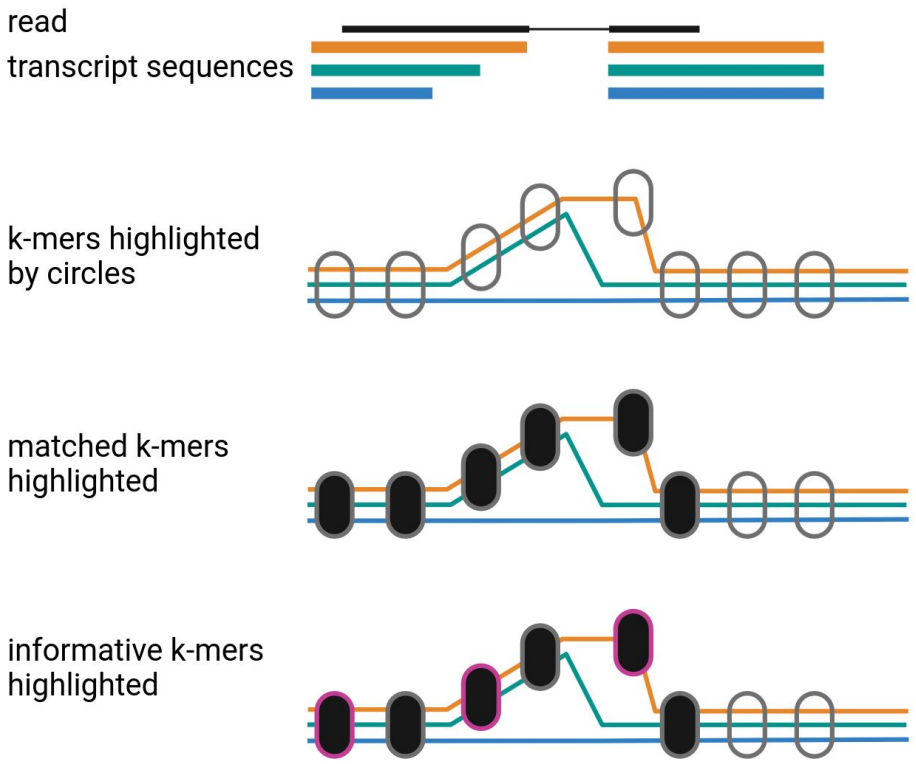
Counting mapped reads

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



Mapping-free quantification: kallisto

- Much faster than 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

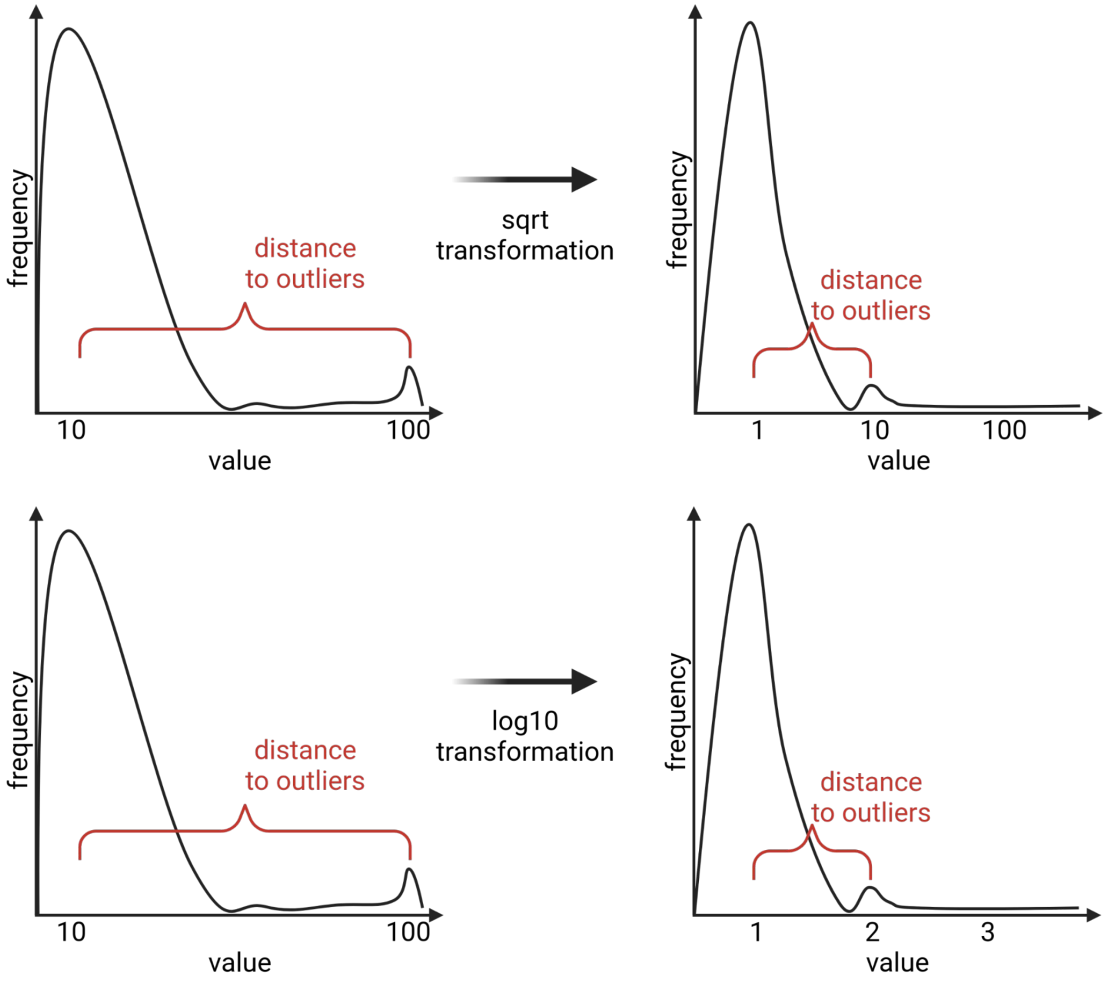
gene	SRR11603178	SRR11603179	SRR11603180	SRR11603181	SRR11603182	SRR11603183	SRR11603184	SRR11603185
TRINITY_DN10003_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10005_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	2.0
TRINITY_DN10005_c1_g1_i1	0.0	0.0	0.0	0.0	1.0	1.0	2.0	0.0	0.0	1.0	1.0	0.0	0.0
TRINITY_DN10008_c0_g1_i1	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



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

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.

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.

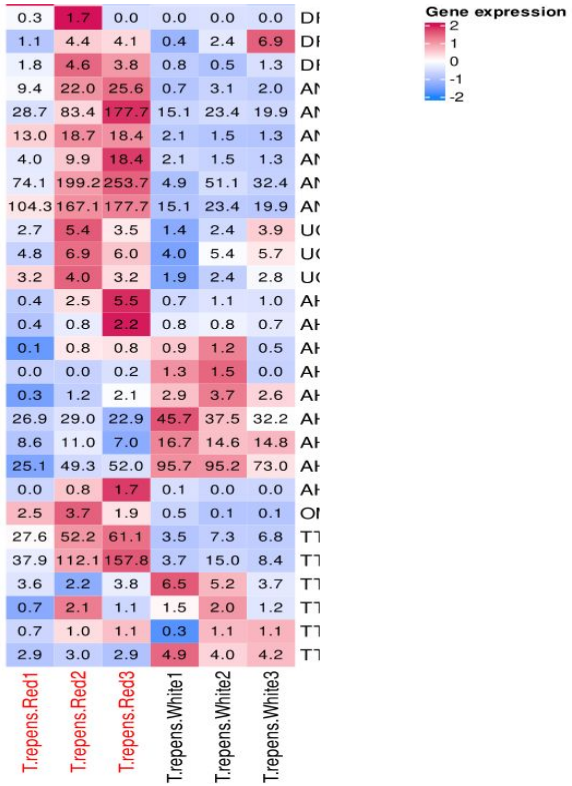
Chen Y, Lun AAT, Smyth GK (2016). "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline." *F1000Research*, **5**, 1438. doi: 10.12688/f1000research.8987.2.

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

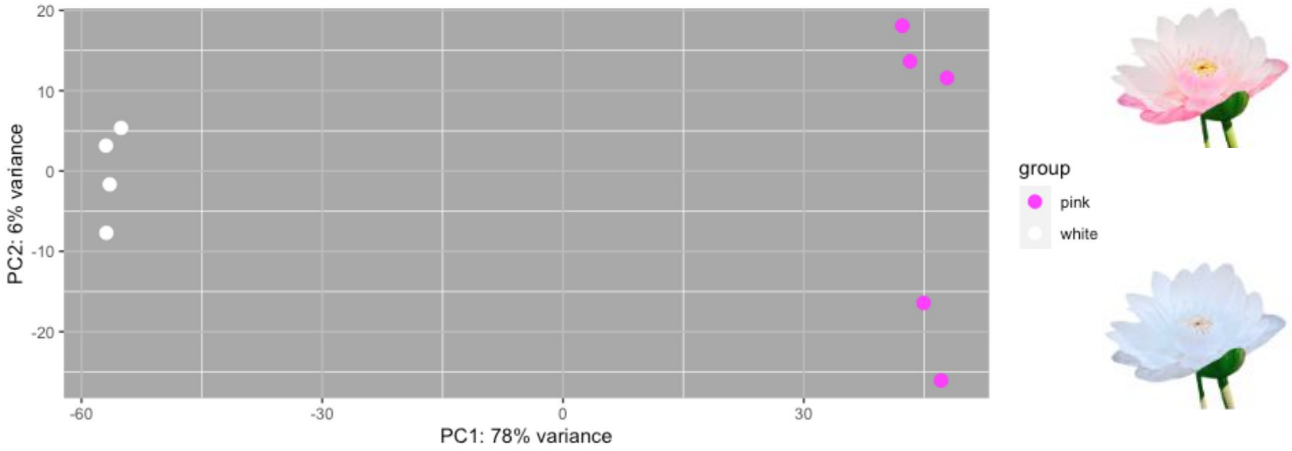
Heatmaps and visualization of gene expression

- Transcriptomic data sets are complex
- Simplified illustration required for interpretation
- Colors help to make data intuitive



Principal Component Analysis (PCA)

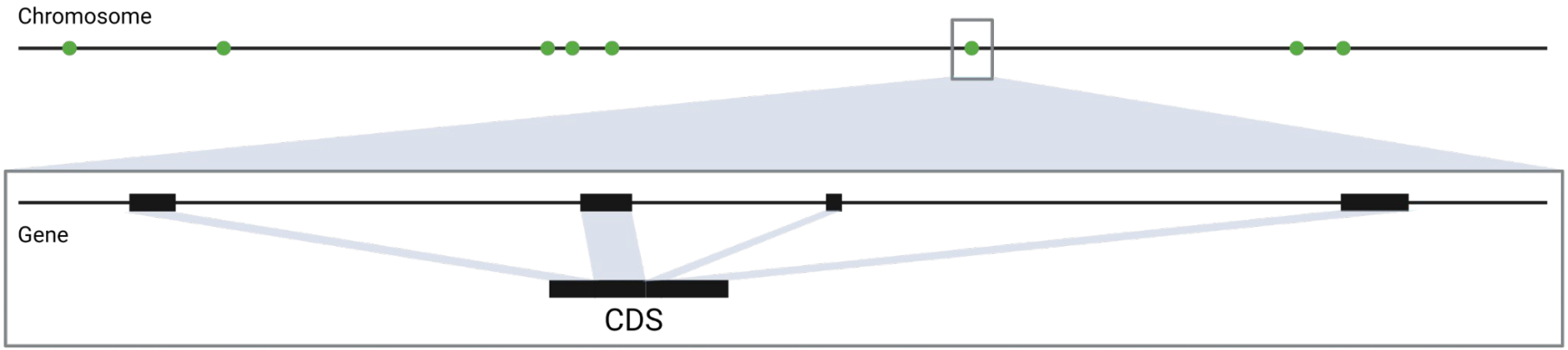
- Separation of large data sets with numerous features based on variation
- PC1 explains high proportion of variance
- Axes (PCs) are NOT biologically meaningful



Nowak et al., 2024: 10.1101/2024.06.15.599162

De novo transcriptome assembly

- Coding sequences account for tiny fraction of plant genome
- Transcriptome analysis can be cost-effective compared to genome sequencing
- Reveals genes relevant under certain conditions/at certain time point
- Analysis can be faster and straight forward



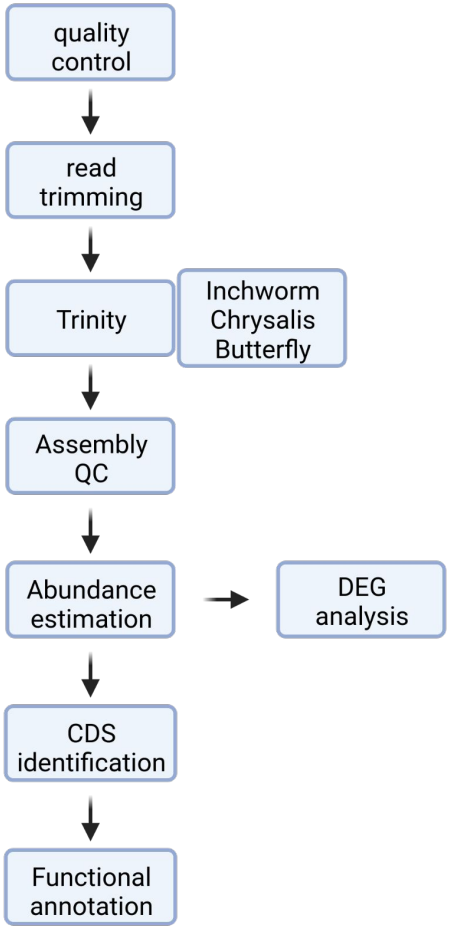
De novo transcriptome assemblers

- Trinity: assembly process comprises three steps
  - Most frequently deployed transcriptome assembler
  - https://github.com/trinityrnaseq/trinityrnaseq/wiki

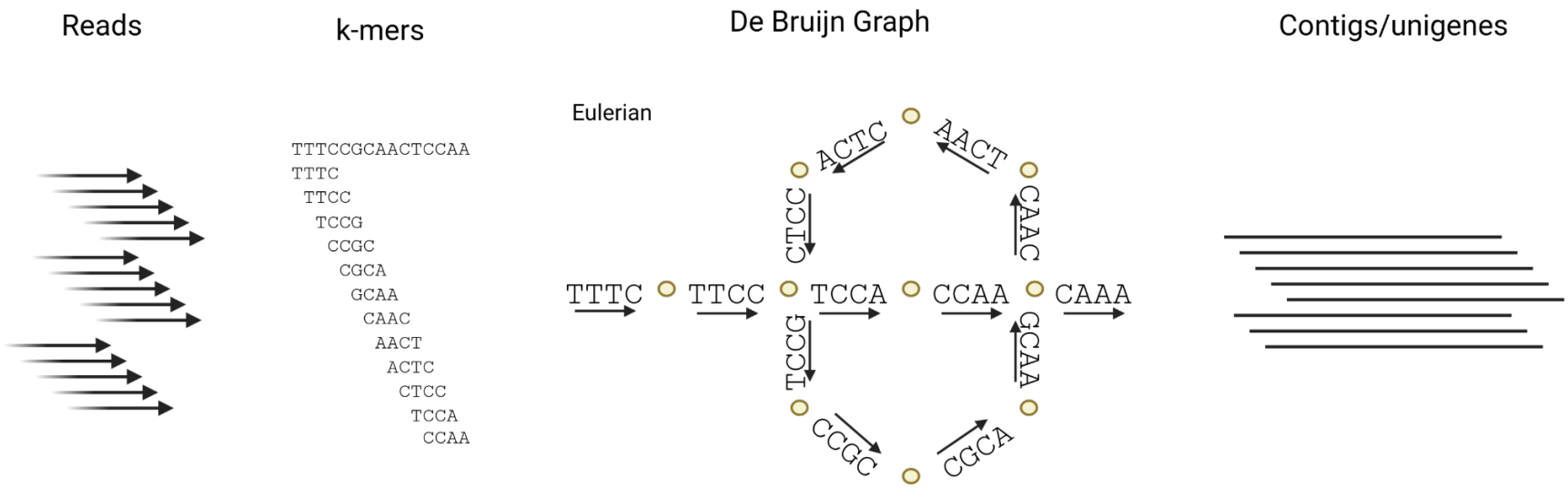
- rnaSPAdes:
  - Recent assembler with support for long reads
  - https://cab.spbu.ru/software/rnaspades/

Trinity *de novo* transcriptome assembly - overview

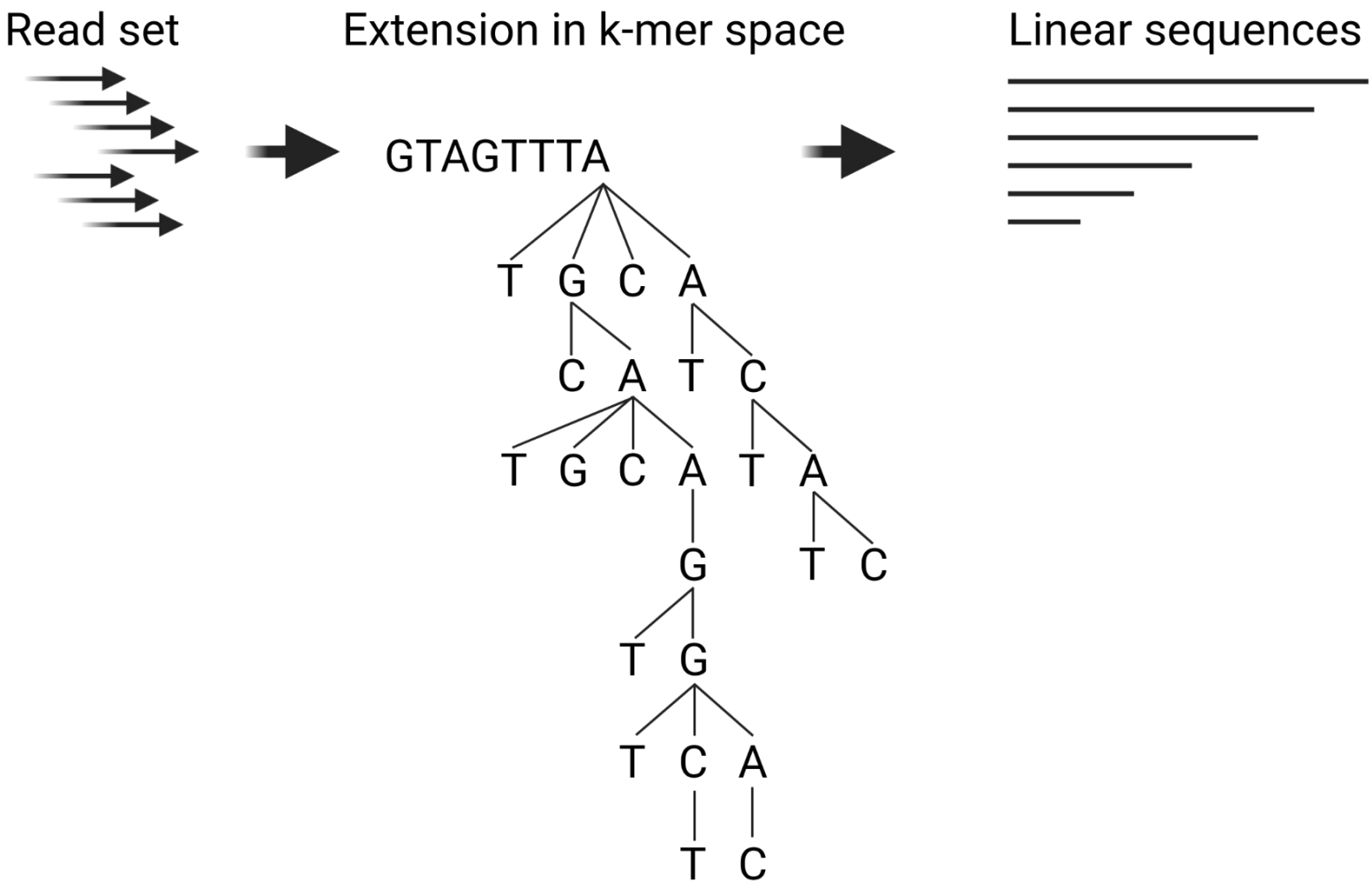
- Quality of reads should be checked prior to the assembly process
- Adapters and low quality parts need to be removed from the reads
- (Read name adjustment)
- Normalization of reads prior to assembly
- Assembly comprises multiple internal steps
- Many downstream analyses are possible depending on the research questions



De Bruijn Graph

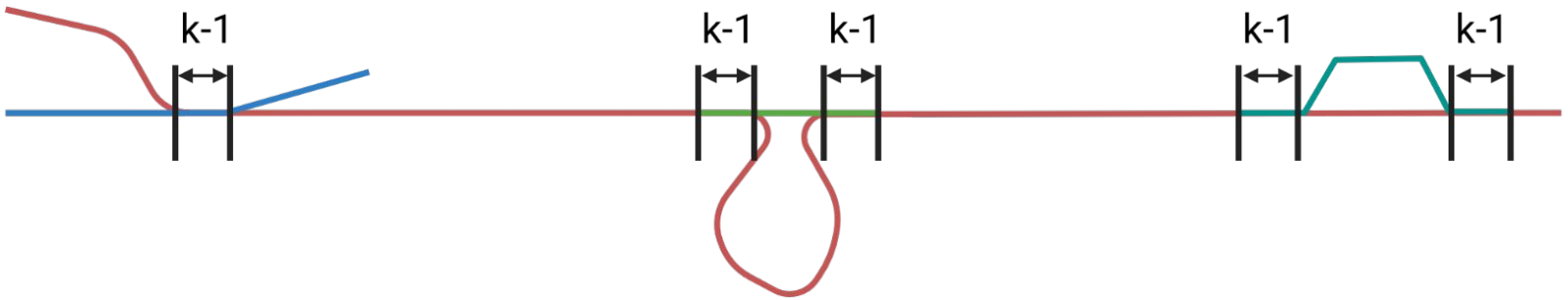


Stages of Trinity: Inchworm



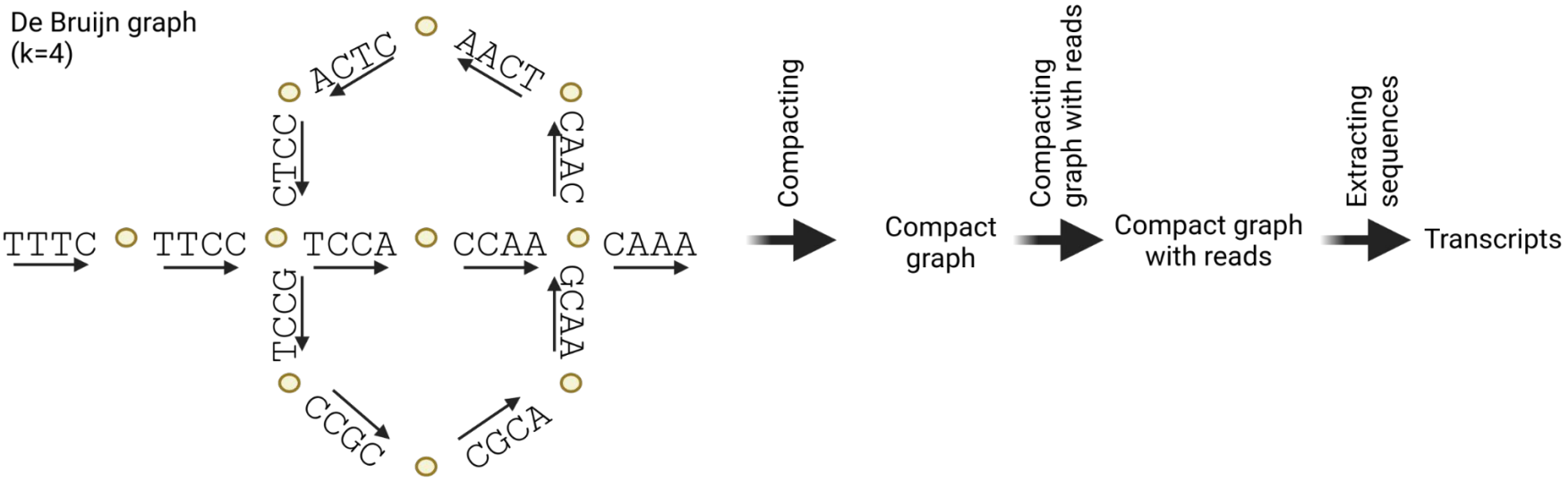
Stages of Trinity: Chrysalis

- Identify differences between sequences and reads
- Transcript isoforms differ only slightly
- Paralogs: differ by single nucleotide variants
- Isoforms: differ by exons



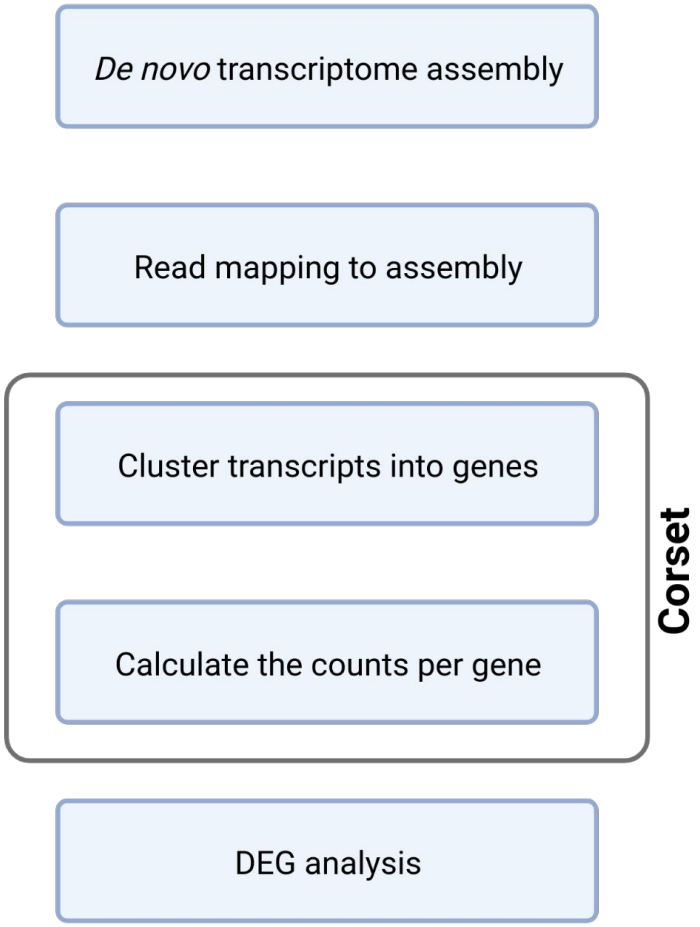
Stages of Trinity: Butterfly

- Extraction of transcript sequences from graph
- Resulting transcript sequences resolve many different isoforms per gene



Corset - merging contigs

- Merges isoforms based on shared multi-mapped reads
- Calculated abundances for individual isoforms and for clusters of isoforms



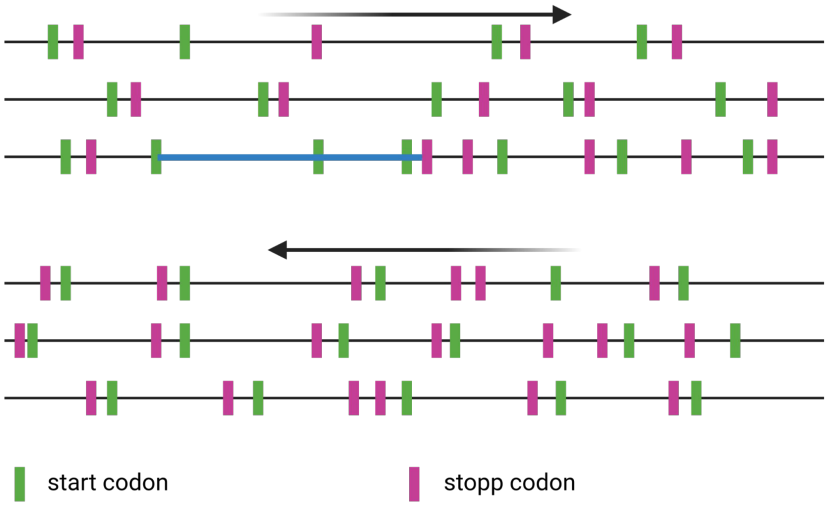
https://github.com/Oshlack/Corset
https://doi.org/10.1186/s13059-014-0410-6

Contamination removal

- Length filter (exclude contigs <200bp)
- GC content
  - plant genomes have a low GC content (usually <40%)
  - GC content in coding sequences is determined by codon usage
- Comparison against sequence database
- Identification of fusion transcripts

CDS identification

- Identify all potential Open Reading Frames (ORFs) in contig
- Six possible reading frames to consider (+1,+2,+3,-1,-2,-3)
- Start codon and stop codon define ends of coding sequences



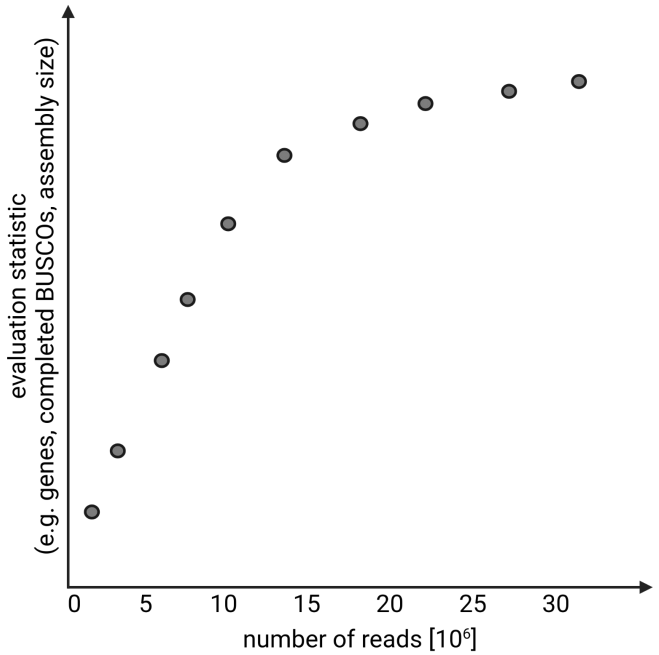
https://github.com/bpucker/PBBtools

Trinity - advanced options/considerations

- Normalization of reads: exclude reads of highly abundant transcripts
  - Minimal and maximal k-mer coverage used as filter
- Stranded reads allow better quality assembly (avoids antisense transcript issues)
- Gene dense genomes pose problem of fusion transcripts (not relevant for plants)

Required data set size

- Number of required reads depends on application
- *De novo* transcriptome assemblies with subsets
- Sufficient number of reads indicated by saturation of evaluation statistics



Required computational resources (Trinity)

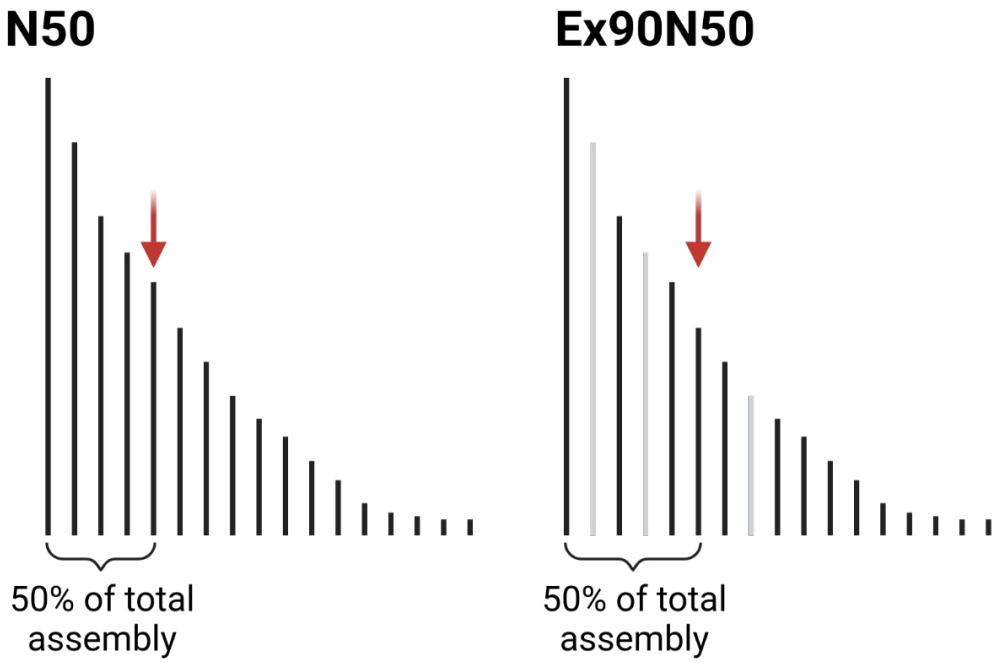
- Example: 20 million fragments; 1+2+4 hours (20 CPUs)
- Example: 200 million fragments: 1+120+7 hours (20 CPUs)
- Trinity is restartable if the process is interrupted
- Restart failed Butterfly jobs (see log file) with reduced CPU number
  - Optional settings: --bflyCPU 20 --bflyHeapSpaceMax 4G

Assembly evaluation

- Number of contigs
- Assembly size (more constant across plant species)
- Mapping of reads (equal coverage, properly paired)
- N50 & Ex90N50

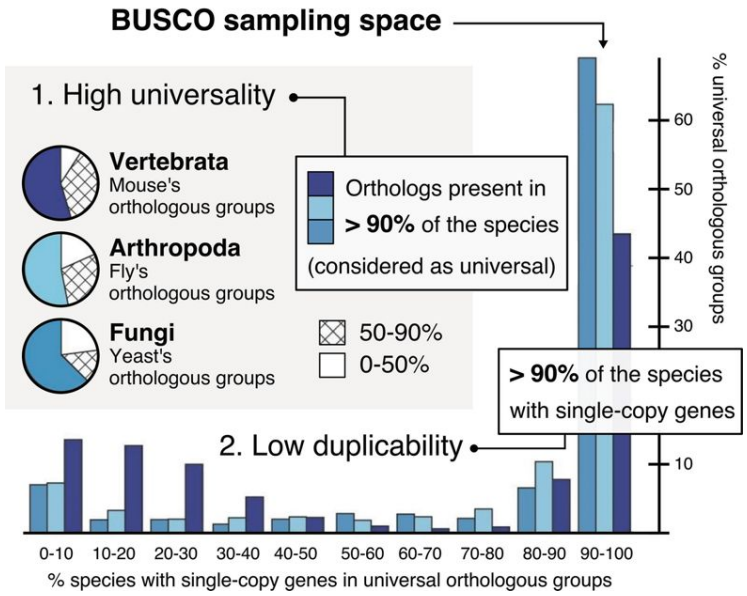
N50 & Ex90N50

- N50 = contig at the border of the contig set that account for 50% of the total assembly size
- Ex90N50 = N50 value of all contigs among the top expressed 90% contigs
- Ex90N50 excludes contigs with low abundances i.e. potential artifacts



BUSCO

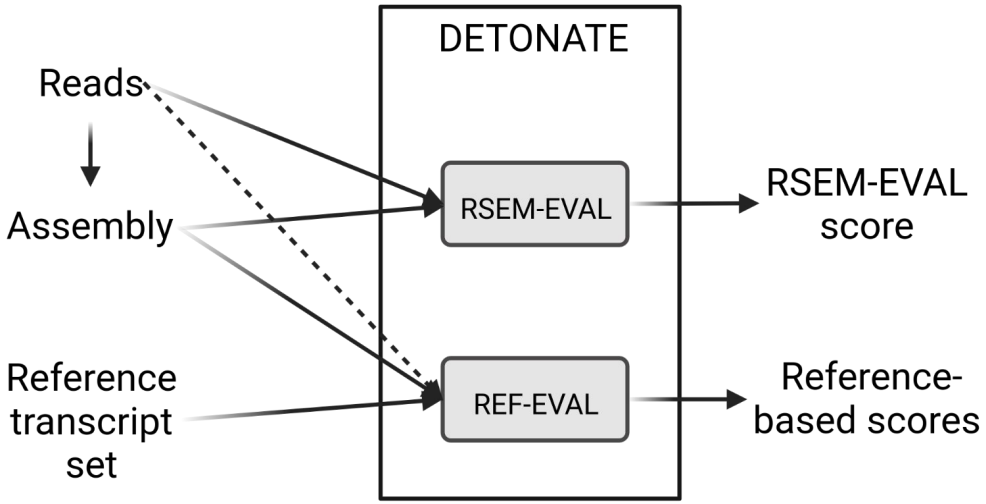
- BUSCO = Benchmarking Universal Single-Copy Orthologs
- Measuring assembly completeness based on presence of conserved genes
- Assessment of annotation quality



https://busco.ezlab.org/

DETONATE

- **DE** novo **T**ranscript**O**me r**N**a-seq **A**ssembly with or without the **T**ruth **E**valuation
- RSEM-EVAL: evaluation score based on reads and assembly
- REF-EVAL: calculates an evaluation score based on a reference transcript set



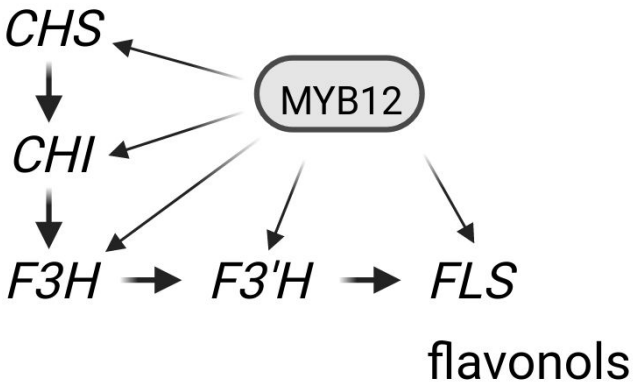
https://github.com/deweylab/detonate
https://doi.org/10.1186/s13059-014-0553-5

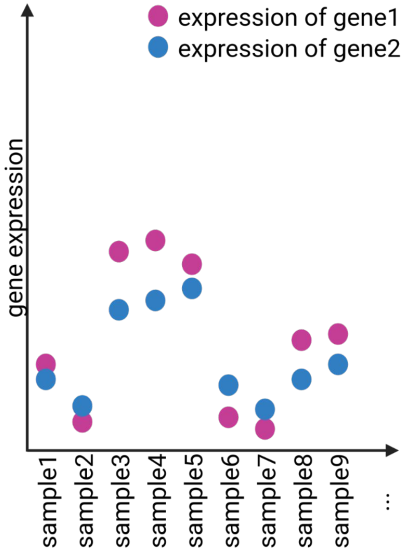
Functional annotation

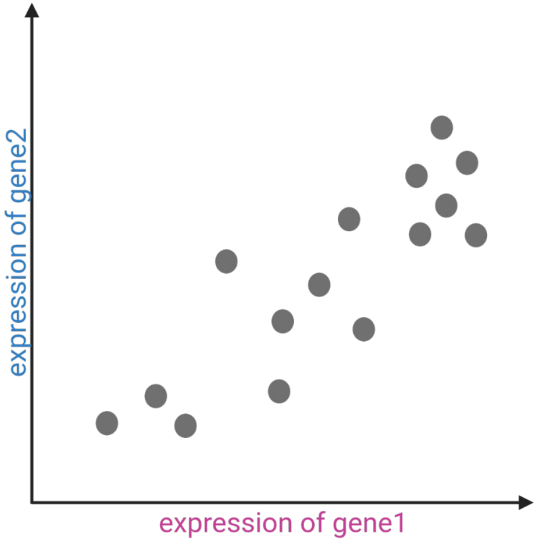
- Reciprocal Best BLAST Hits (RBHs)
- OrthoFinder2
- Knowledge-based Identification of Pathway Enzymes (KIPEs)
- Mercator
- InterProScan5

Co-expression analysis

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



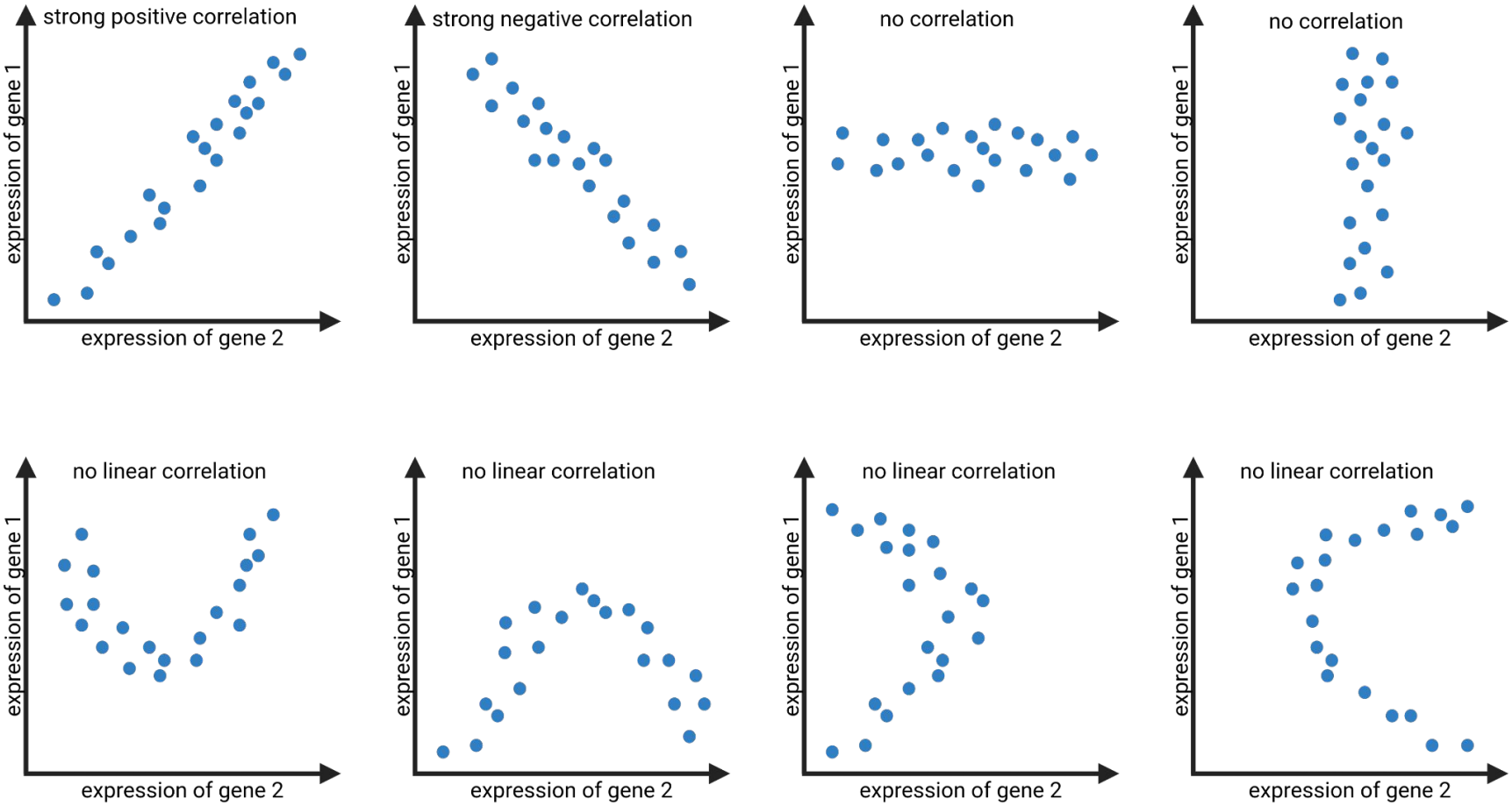




Correlation vs. causation

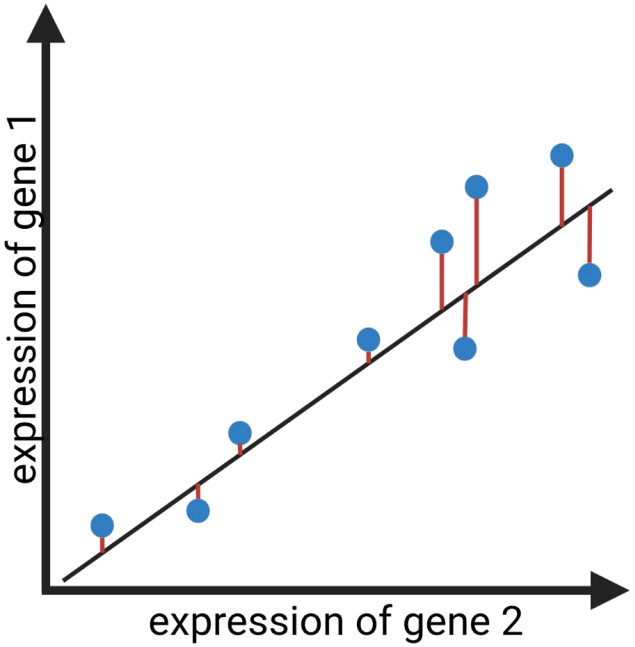
- Correlation does not allow conclusions about the direction i.e. cause and consequence
- Guilt-by-association is frequently used to find genes in a pathway/network
- Correlation can be caused by a direct or indirect connection

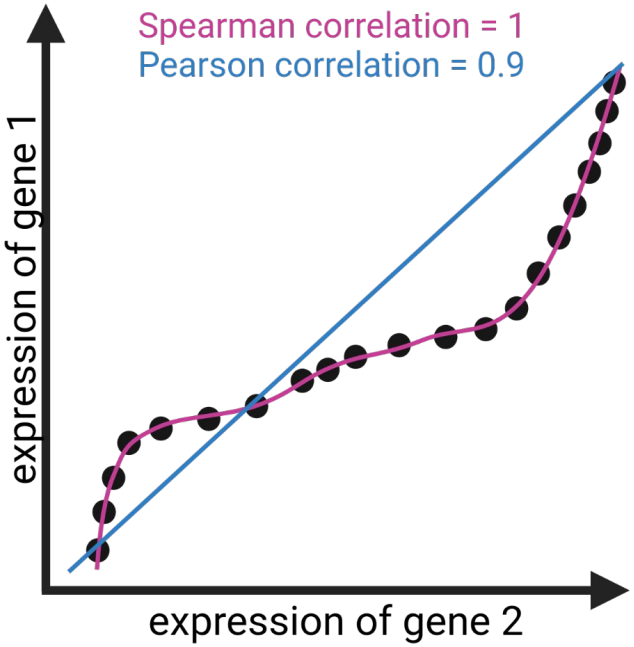
Examples


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

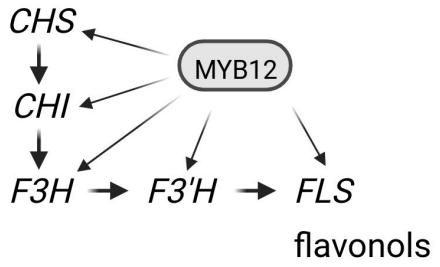
genes	correlation
CHS	0.7
CHI	0.5
F3H	0.45
F3'H	0.6
FLS	0.9

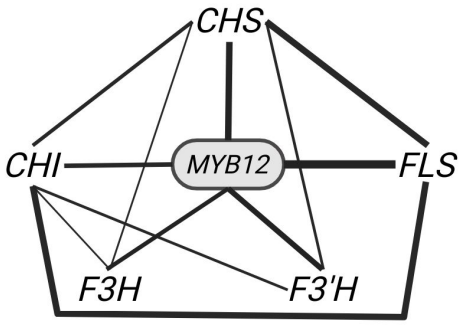
CHS

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

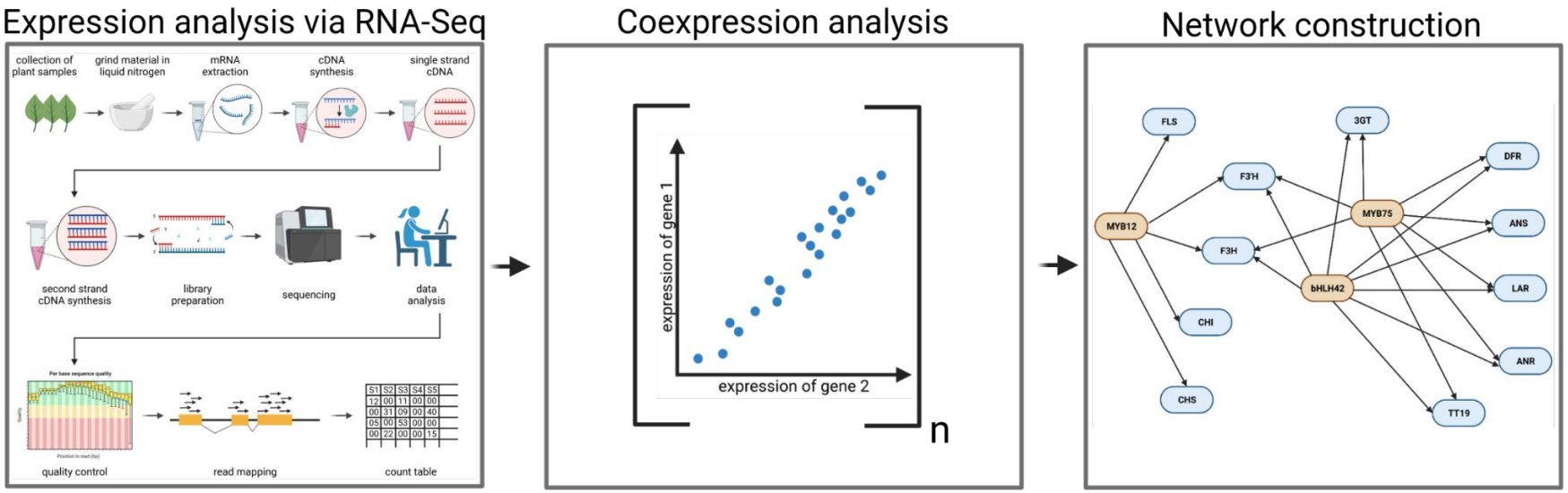
CHI

genes	correlation
F3H	0.2
F3'H	0.3
FLS	0.7

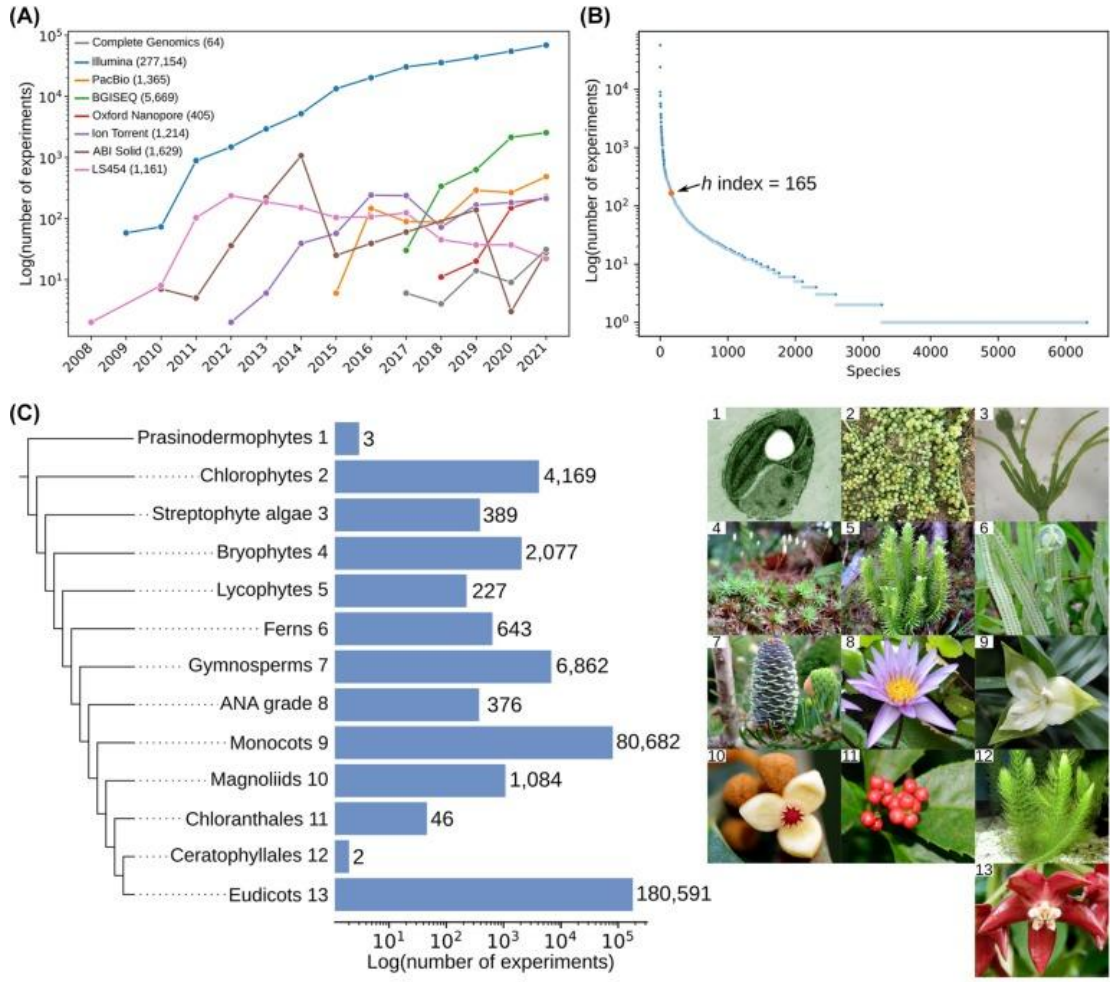




Summary: co-expression workflow



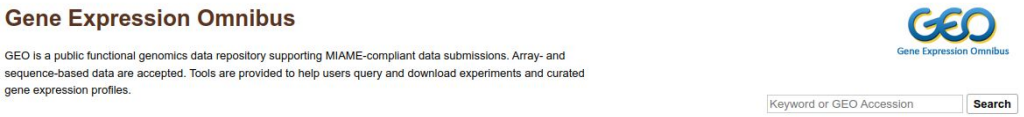
Available plant RNA-seq data sets



Julca et al., 2022: 10.1016/j.tplants.2022.09.007

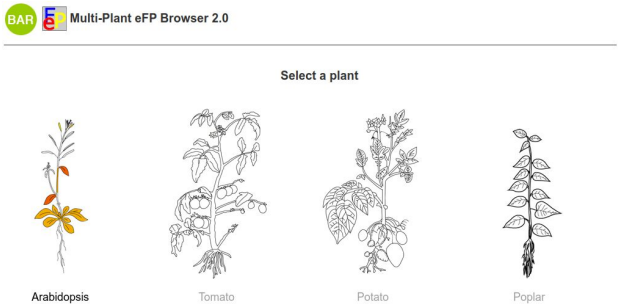
Data sources

- Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/



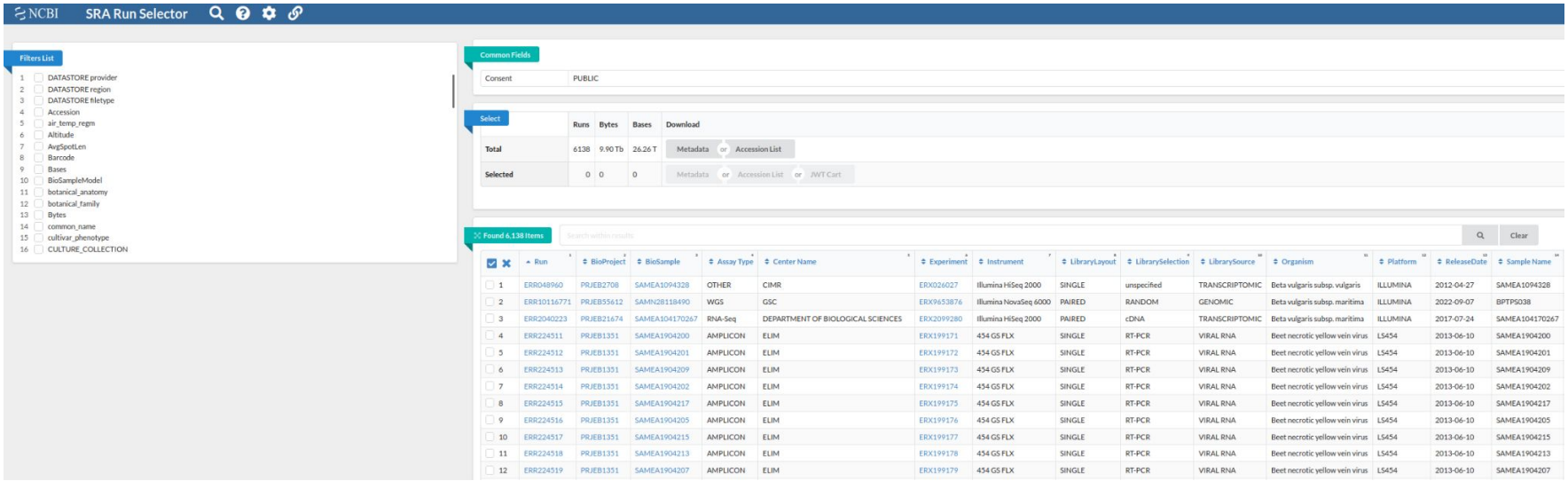
- Sequence Read Archive (SRA): https://www.ncbi.nlm.nih.gov/sra

- eFP browser: http://bar.utoronto.ca/efp2/



SRA Run Selector

- Search based on various parameters and filtering
- Download of data set IDs and metadata



Retrieving expression data

- Fastq-dump: automatic download of data set (FASTQ) based on ID
- Part of SRA tools; Faster alternatives are available
- GEO: download of count tables
- Latest development: bring analysis to data i.e. run analyses in cloud

Metadata

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

Checking/filtering RNA-seq data sets

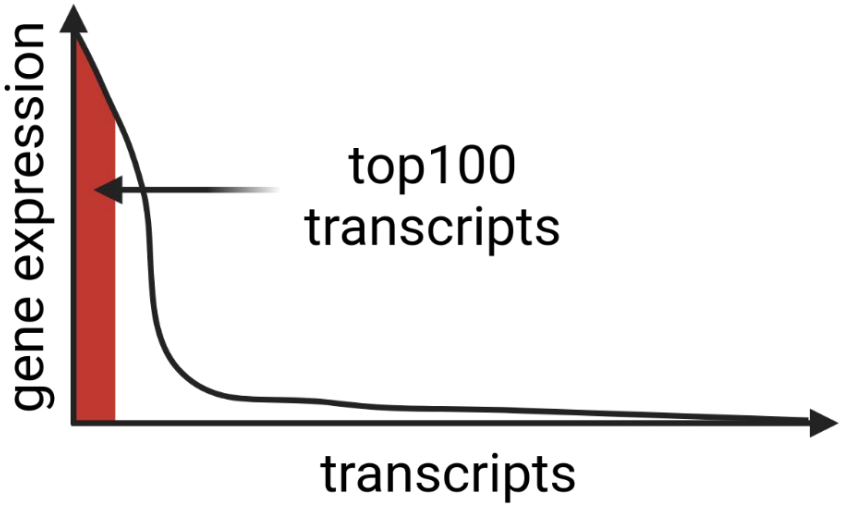
- Size of data set: count number of reads or read pairs
  - >5 million recommended

- Check species identity
  - Exclude mislabeled samples e.g. symbiosis

- Check tissue identity
  - Photosynthesis genes should be expressed in green tissue
  - No photosynthesis gene expression in roots
  - …

Plausibility checks

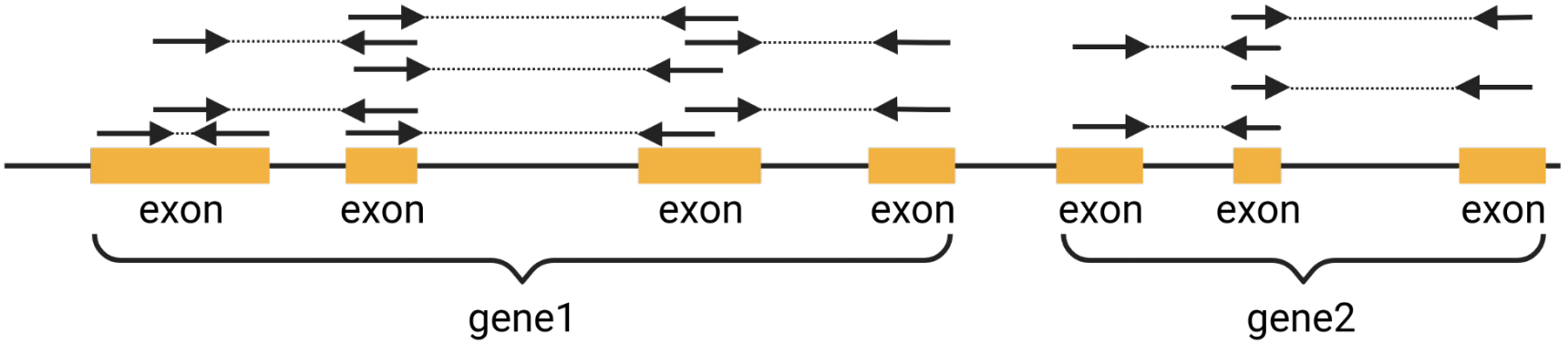
- Check for RNA character of samples (validate metadata)
- Exclude mislabeled DNA data sets
- Top100 transcripts should account for half of all transcripts (reads)



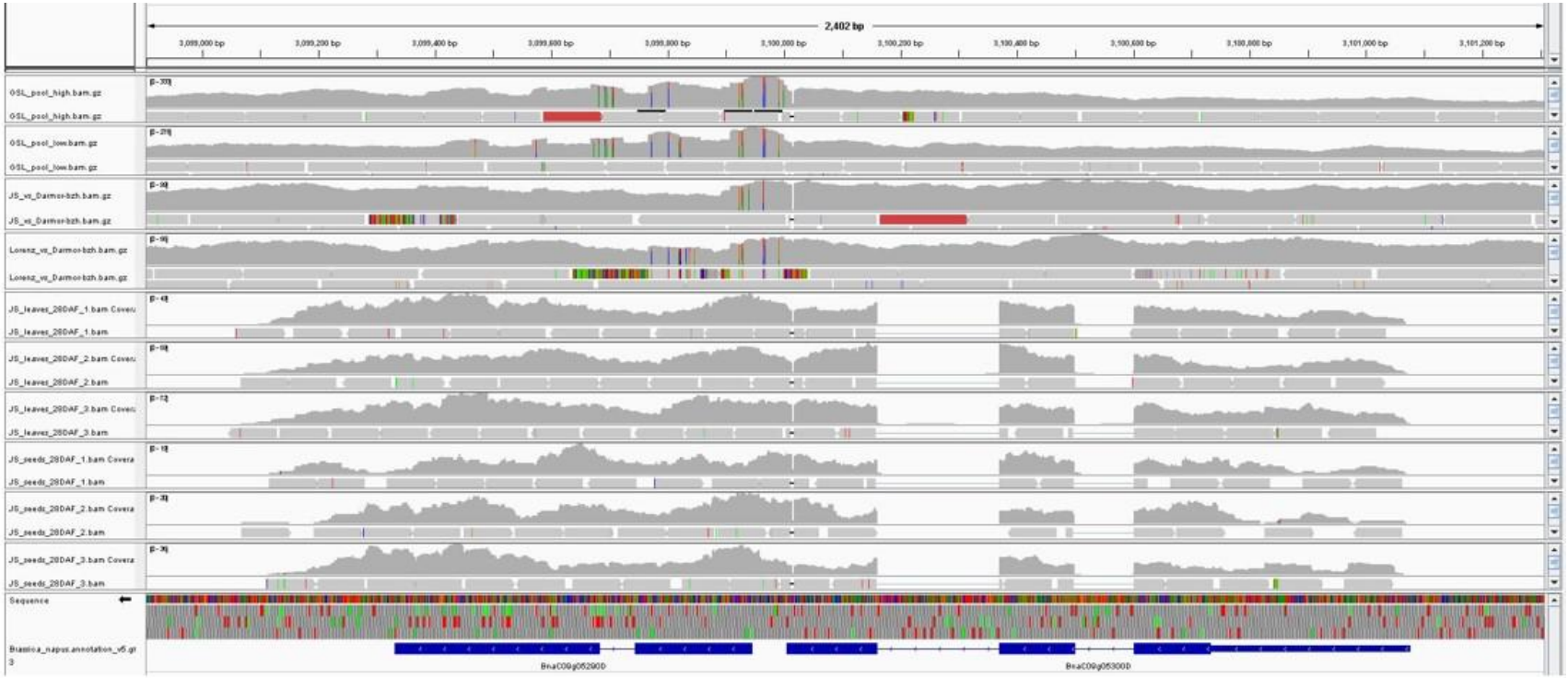
https://github.com/bpucker/CoExp

Hints for gene prediction

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

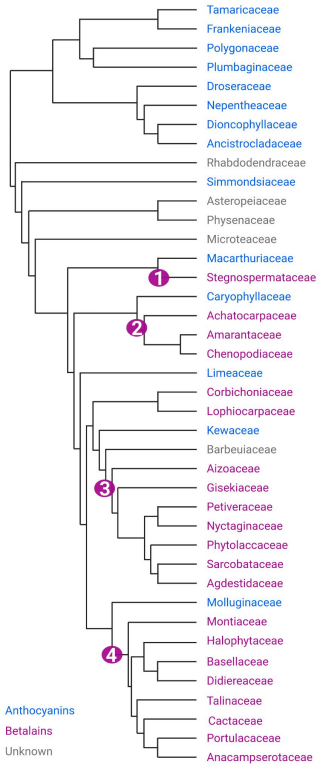


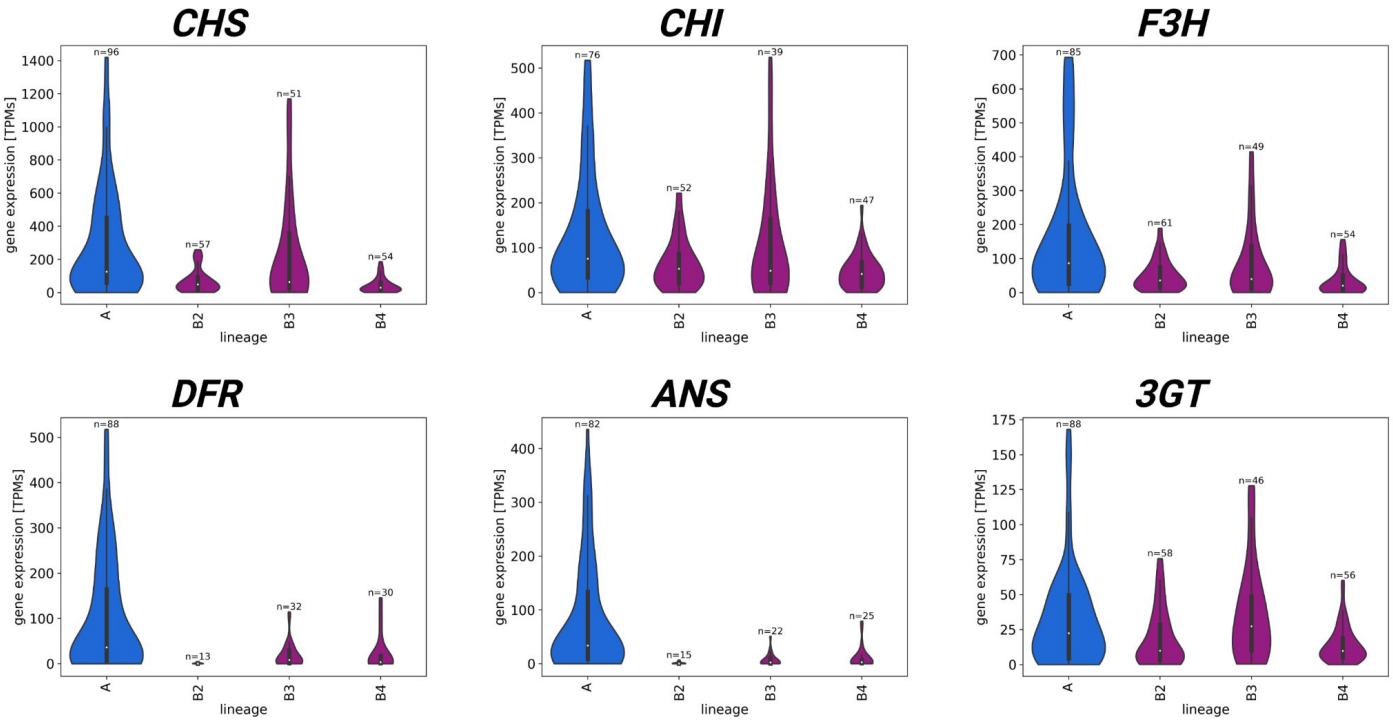
Annotation of *Bna*MYB28



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

Cross-species transcriptomics

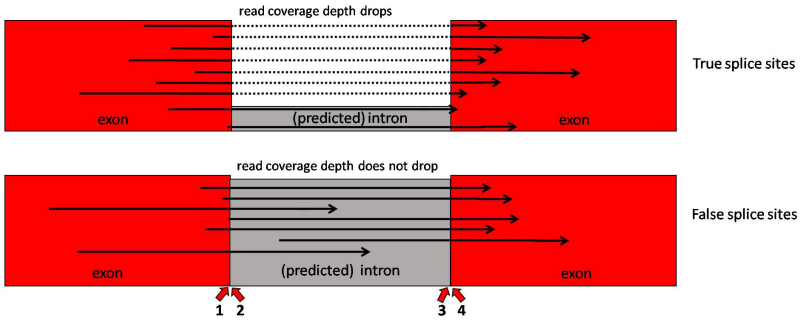


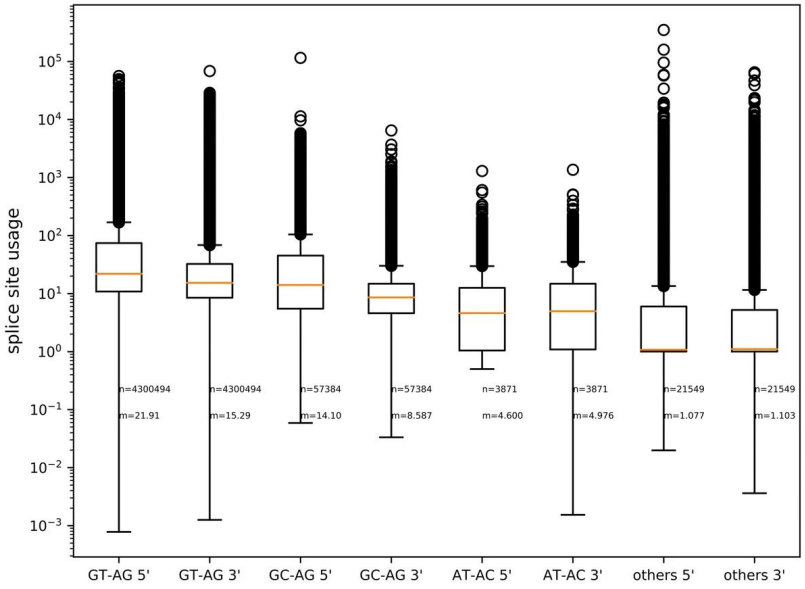


Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

Non-canonical splice sites

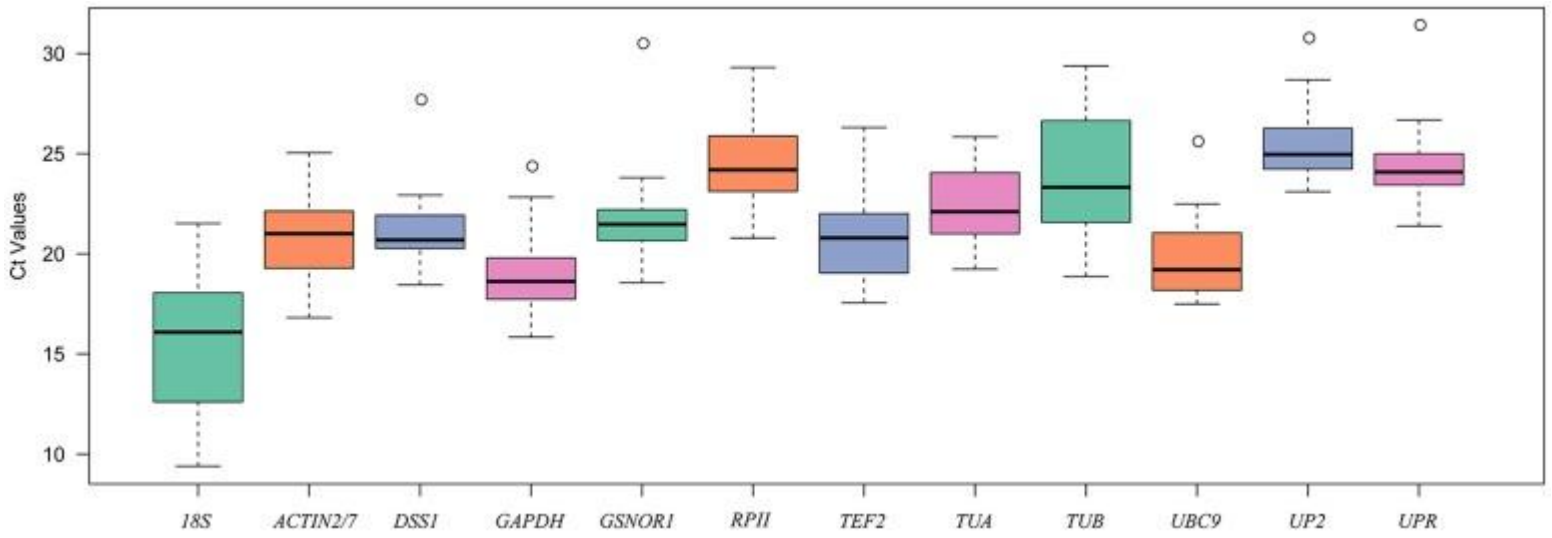
- RNA-seq reads can be used to study the usage of splice sites
- Splitting of reads in the alignments is crucial to investigate introns
- STAR/HISAT2 are suitable read mappers to generate RNA-seq read alignments
- Splice site usage inferred from difference between terminal exon and terminal intron coverage





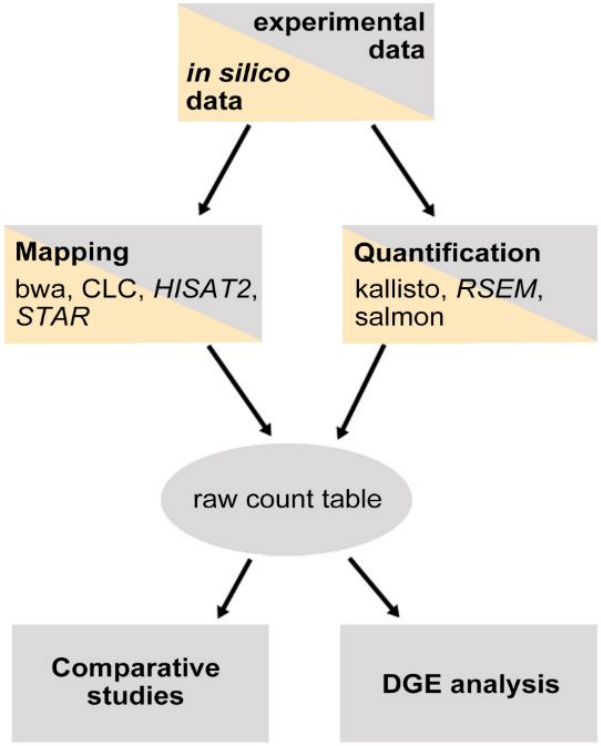
Pucker & Brockington, 2019: 10.1186/s12864-018-5360-z

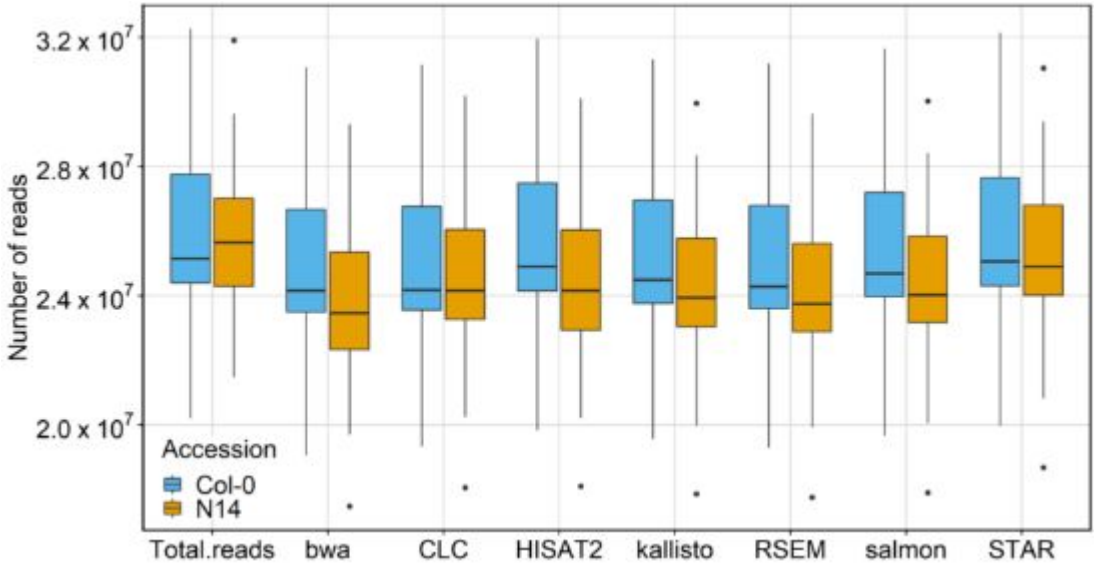
Identification of qPCR reference genes



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

RNA-seq read mapper benchmarking

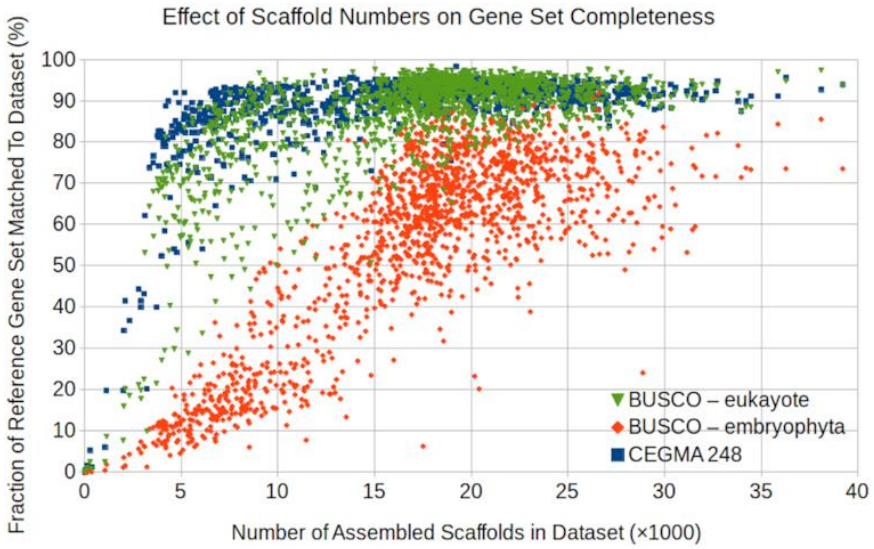




Schaarschmidt et al., 2020: 10.3390/ijms21051720

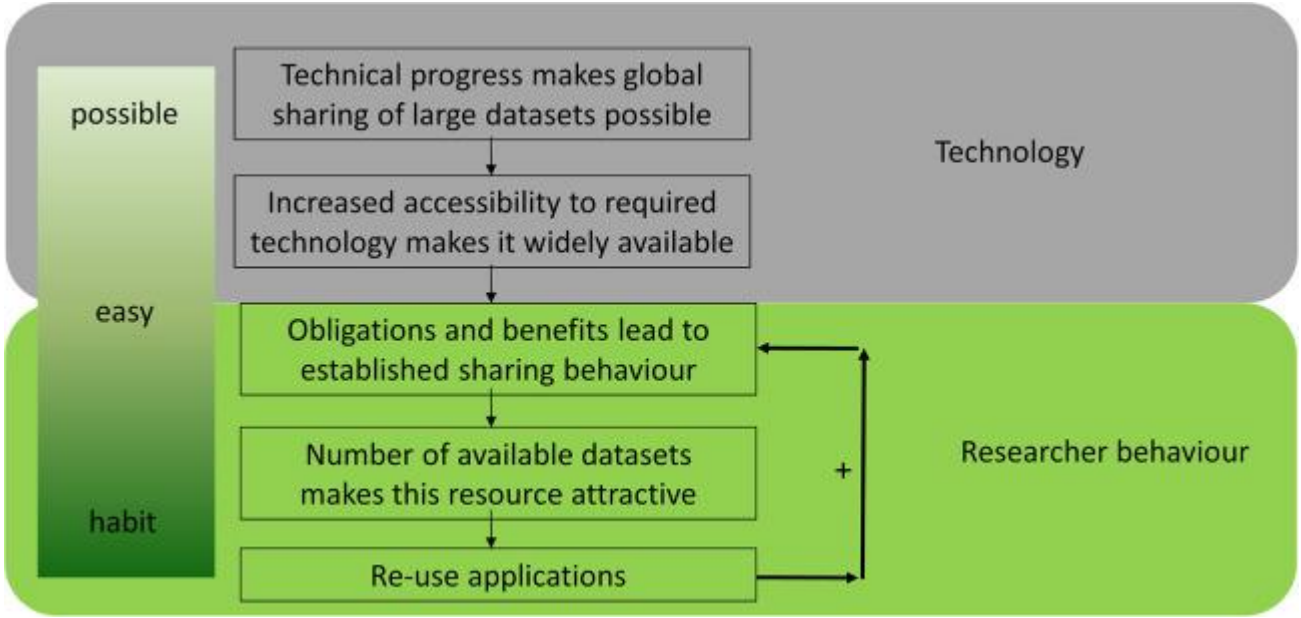
1KP

- Phylogenetics project to cover taxonomic width of plants
- Transcriptome assemblies for 1000 (1k) plants
- Large, international consortium for sequencing projects



http://www.onekp.com/samples/list.php
https://doi.org/10.1093/gigascience/giz126

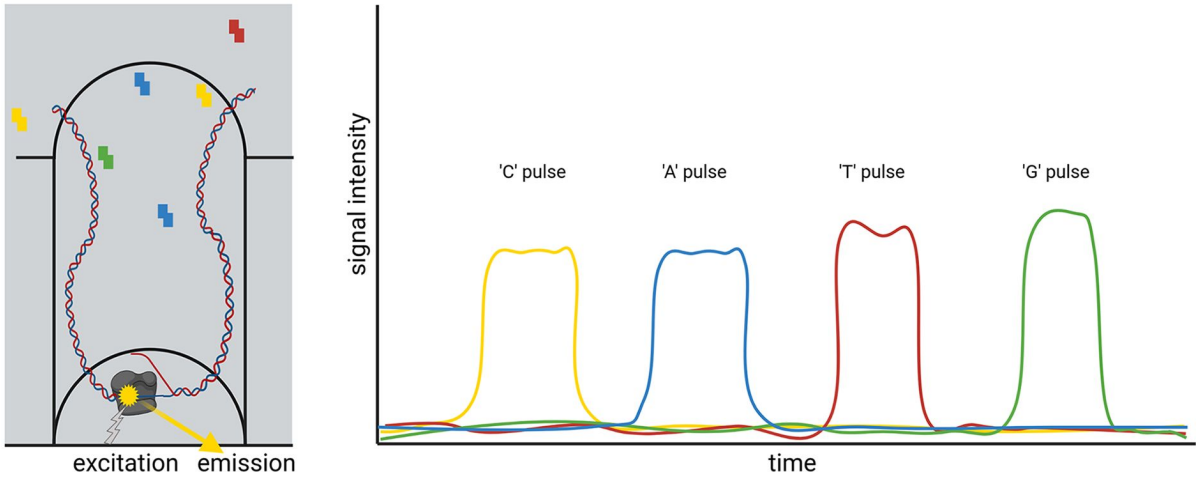
How to facilitate data reuse?



Sielemann *et al.*, 2020: 10.7717/peerj.9954

PacBio (SMRT) sequencing

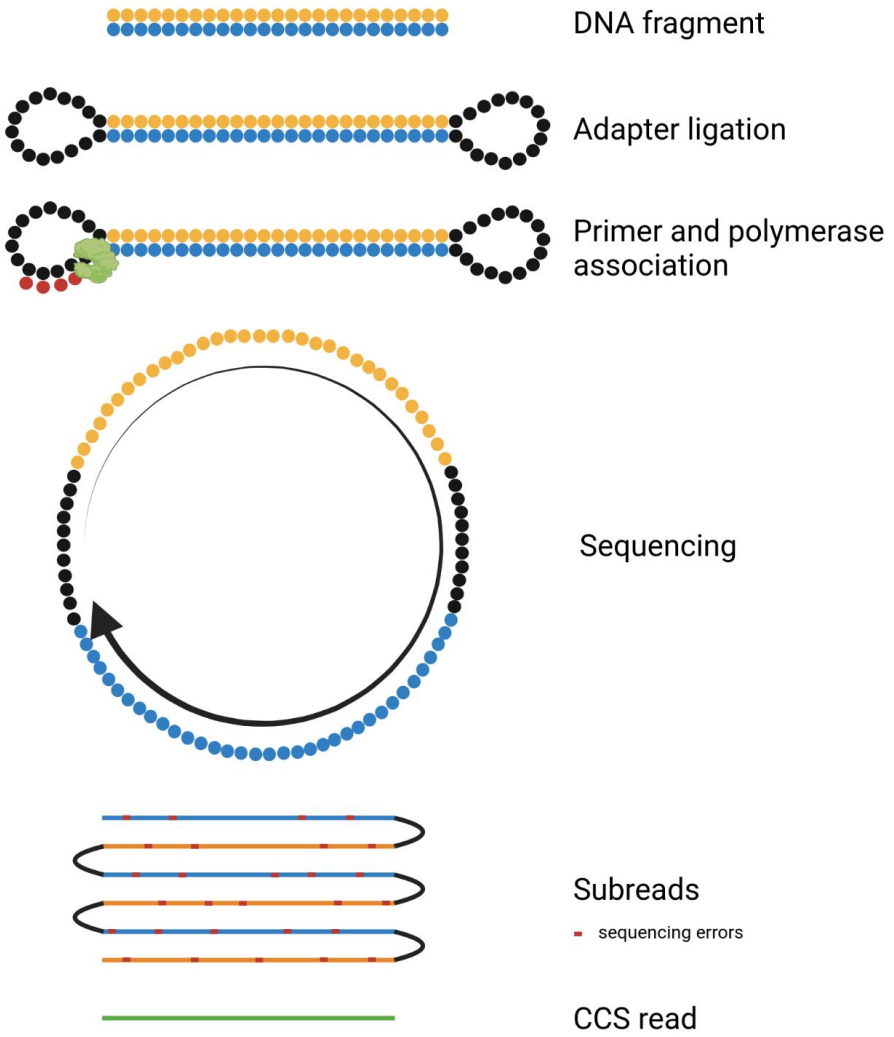
- Single DNA molecule is processed
- SMRT = single molecule real time sequencing
- DNA polymerase synthesizes new strand
- Nucleotides are labeled with fluorescence



https://doi.org/10.1017/qpb.2021.18

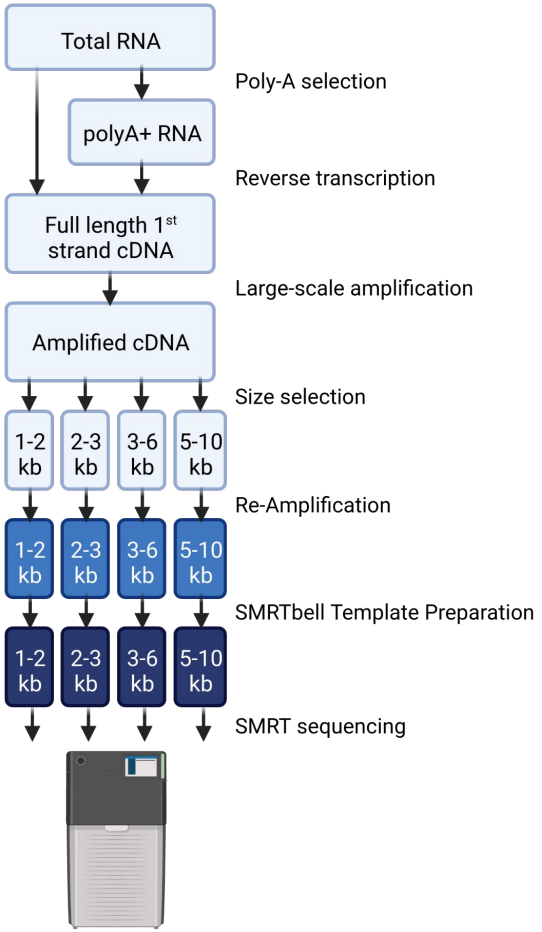
PacBio - HiFi

- HiFi = high fidelity
- Multiple rounds of sequencing increases circular consensus sequence (CCS) accuracy



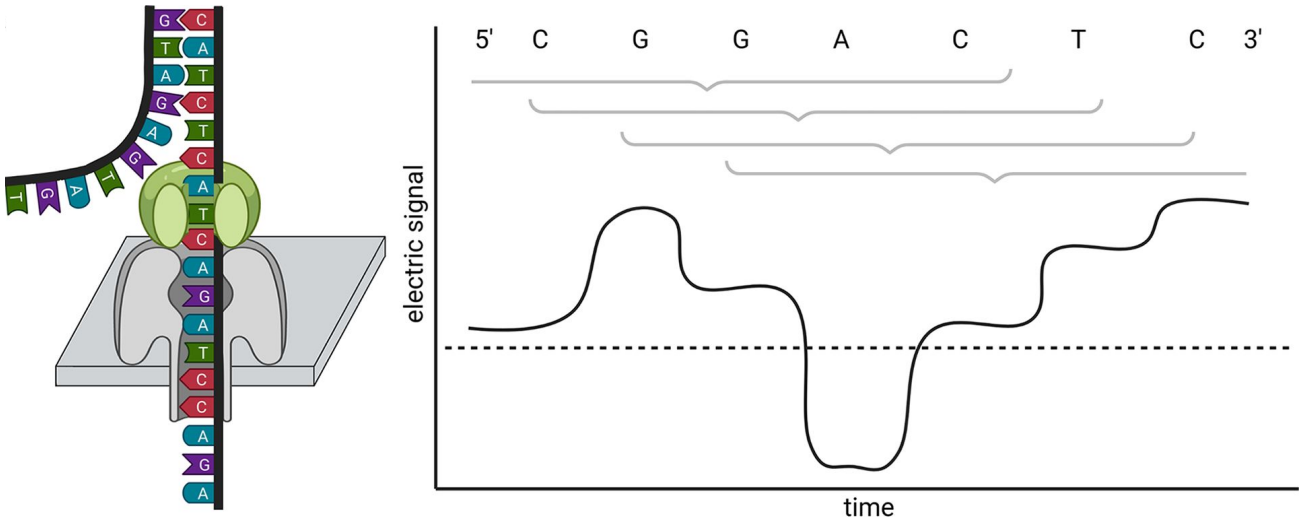
PacBio Iso-Seq (cDNA sequencing)

- SMRT = Single Molecule Real Time sequencing
- Full length cDNA sequencing is beneficial for gene prediction
- Iso-Seq generates several kb long reads and not only 2x300bp



Nanopore sequencing

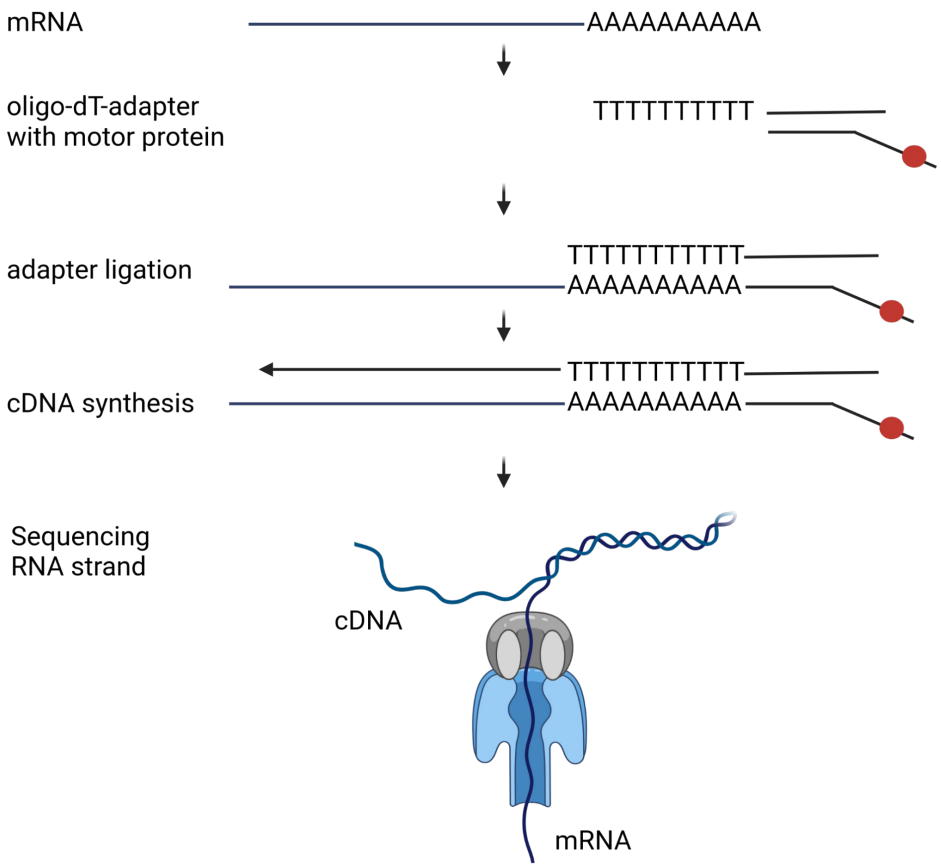
- Single DNA/RNA strand passes through a nanopore in a synthetic membrane
- Blockage of the pore is measured (electric signal)
- Sequence is inferred from changes in electric signal over time



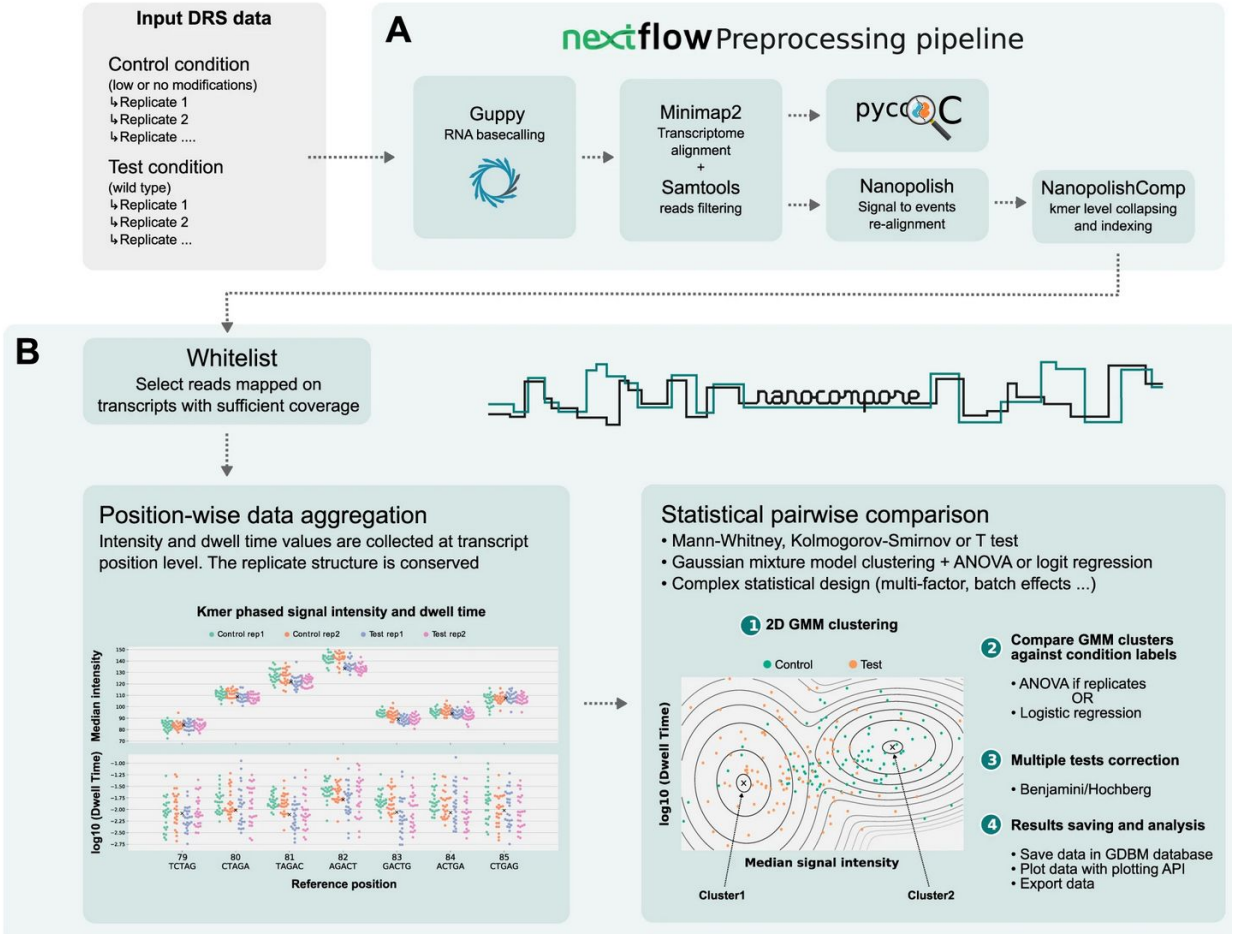
Direct RNA sequencing

- Only sequencing technology to analyze RNA directly at high throughput
- RNA sequencing requires adjusted data processing
- Full length sequences of RNAs are generated

(photo credit: Melina Nowak)



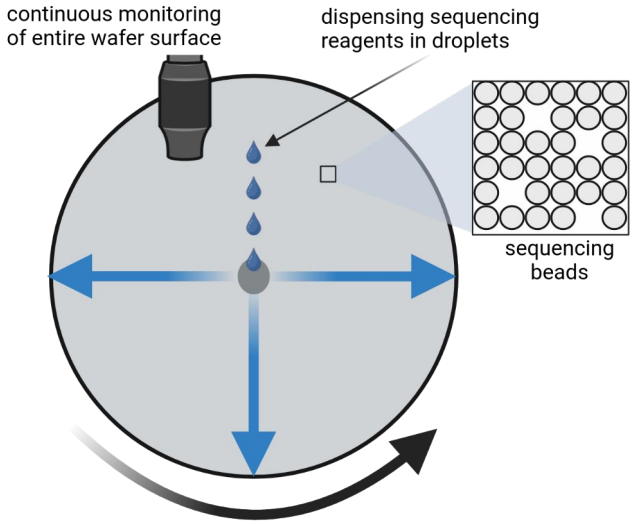
RNA modification analysis

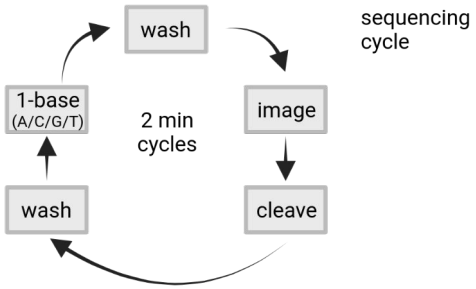


https://doi.org/10.1038/s41467-021-27393-3

New seq technology

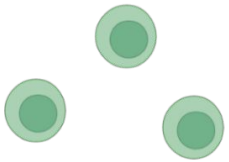
- About 300 nt read length possible
- Base quality: Q15-Q40
- Homopolymers pose an issue (up to 4 possible)
- More natural than labeled nucleotides supplied per sequencing cycle



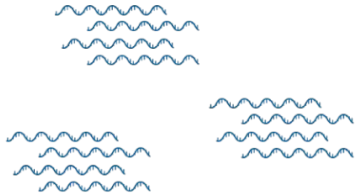


https://doi.org/10.1101/2022.05.29.493900

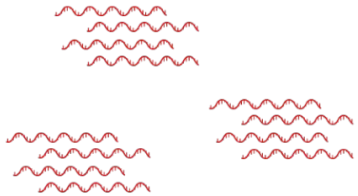
Single cell transcriptomics (scRNA-seq)



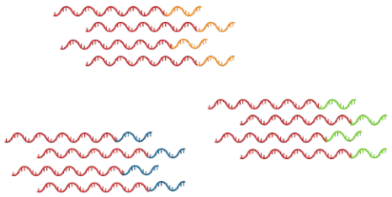
separating cells



RNA extraction

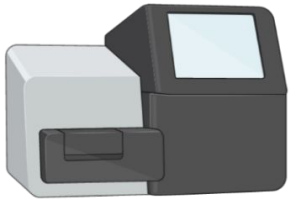


cDNA conversion

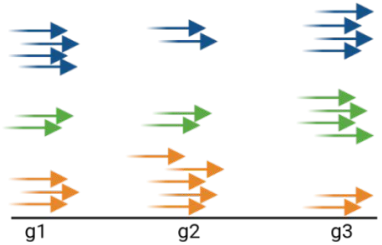


barcoding (UMI)

UMI = Unique Molecular Identifiers



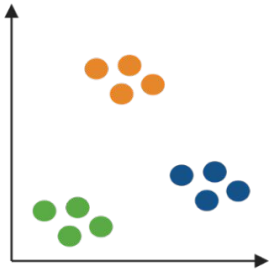
sequencing



read mapping

	s1	s2	s3
g1	0.6	0.3	0.6
g2	0.8	0.8	1.6
g3	0.8	1.2	0.6

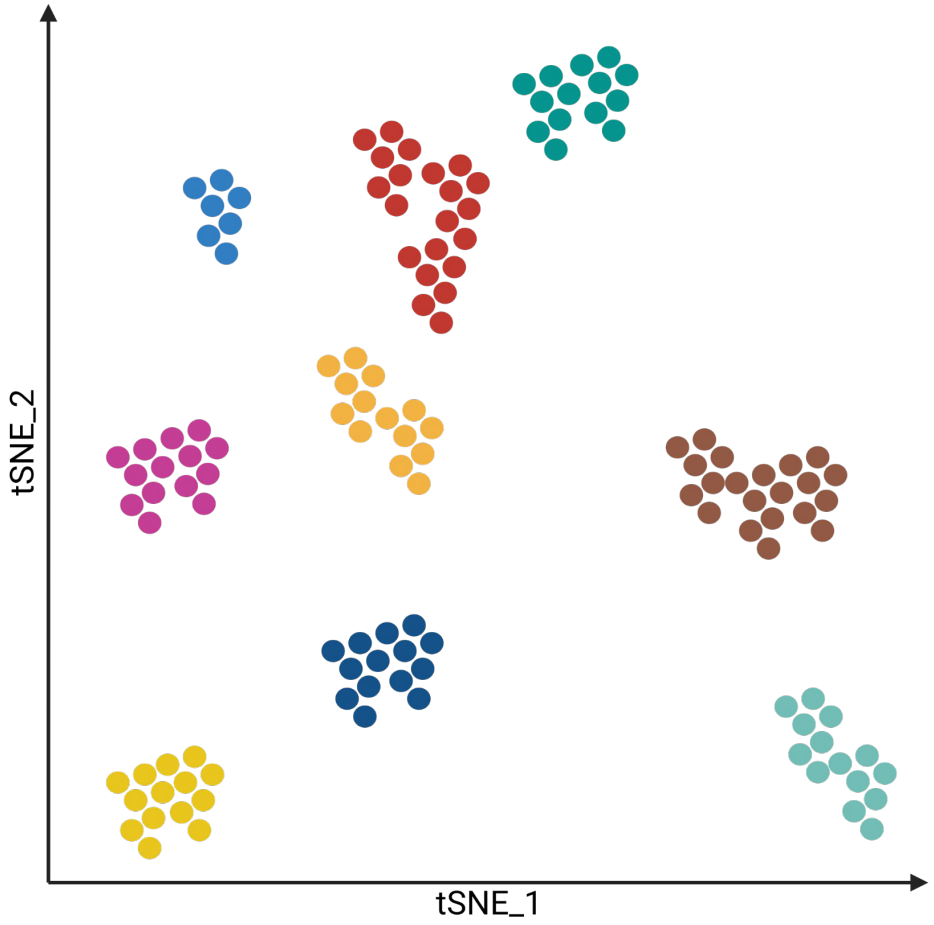
normalization



clustering

Molecular relationships among individual plant cells

t-distributed stochastic neighbor embedding (tSNE)



Inspired by Ryu et al., 2019: 10.1104/pp.18.01482

Summary

- Concept of RNA-seq & workflow (wet lab)
- Experimental design considerations
- Read mapping & quantification
- De novo transcriptome assembly
- PCA, Heatmaps, DEG identification
- Re-using public RNA-seq datasets
- Direct RNA sequencing & scRNA-seq

Time for questions!

Practical part: outline day4

- QC and trimming of reads (fastQC, Trimmomatic)
- De novo transcriptome assembly (Trinity)
- Split read mapping (STAR, HISAT2)
- Quantification (kallisto)
- Identification of DEGs (DESeq2)
- Co-expression analysis (ppb-tools.de)

Example data set: dragon fruit color differences

- Dragon fruits (pitaya) come with different colors: green, yellow, red
- Pitaya belongs to the Caryophyllales which are known for betalain pigmentation in several families
- Pitaya pigmentation is based on betalains (not anthocyanins)

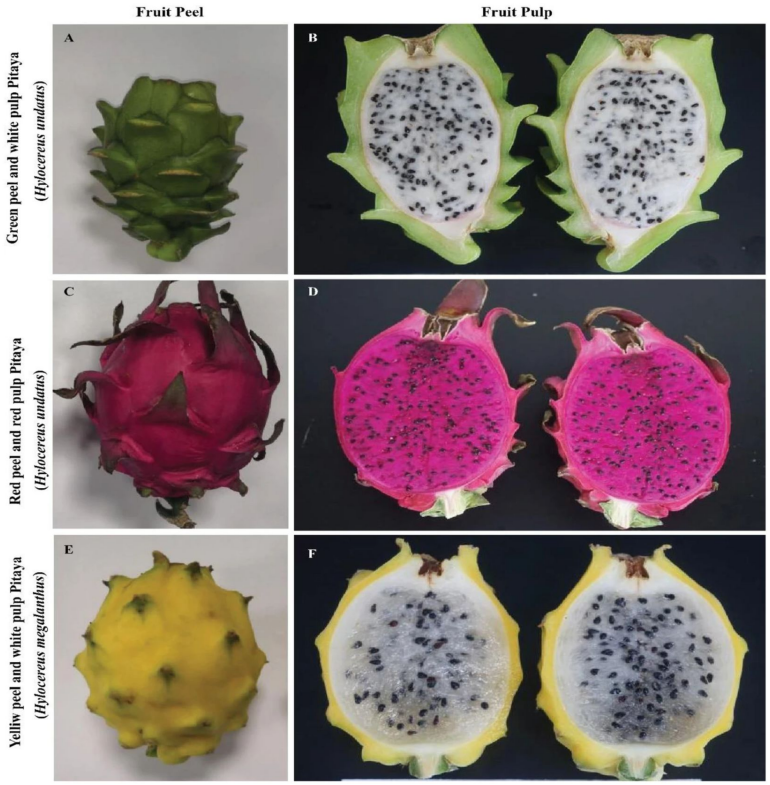


Figure taken from 10.1186/s12864-020-07133-5

Exercises: 4-1

- Retrieve RNA-seq data sets from SRA/PUB
- Check the quality of RNA-seq data with fastQC
- Trimm low quality parts of the reads and adapters with Trimmomatic
- Assemble a transcriptome assembly with Trinity (requires renaming the reads)

Running fastQC

- Quality assessment of Illumina reads
- Availability: https://github.com/s-andrews/FastQC
- Usage: `fastqc /path/to/your_file.fastq`
- Output:

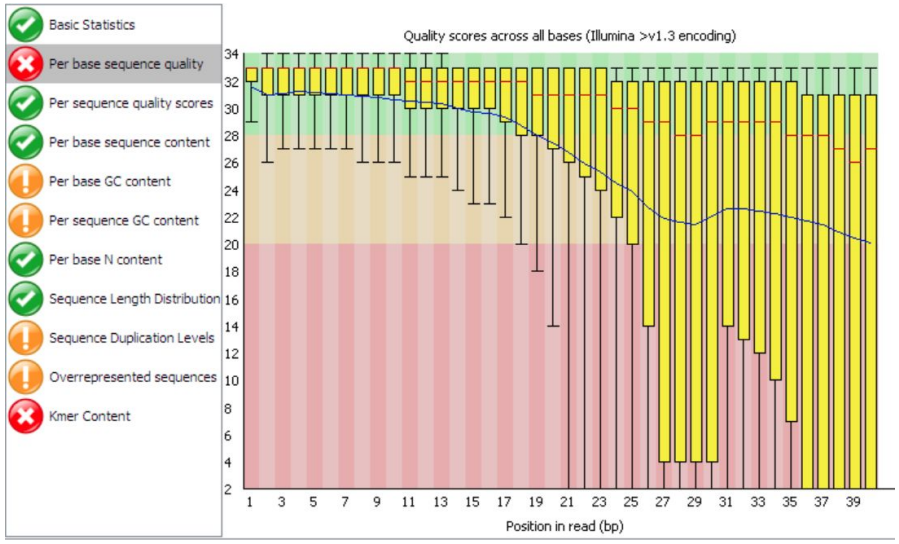


Figure taken from https://github.com/s-andrews/FastQC

Trimming reads with Trimmomatic

- Removal of low quality parts of the read ends
- Removal of adapter sequences
- Availability: http://www.usadellab.org/cms/?page=trimmomatic
- Usage:

```
java -jar trimmomatic-0.39.jar PE \
input_forward.fq.gz \
input_reverse.fq.gz \
output_forward_paired.fq.gz \
output_forward_unpaired.fq.gz \
output_reverse_paired.fq.gz \
output_reverse_unpaired.fq.gz \
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True \
LEADING:3 TRAILING:3 MINLEN:50
```

De novo transcriptome assembly with Trinity

- Short reads can be connected to form contigs representing transcripts
- Availability: https://github.com/trinityrnaseq/trinityrnaseq/wiki/
- Script for renaming reads:
- Usage:

```
Trinity \
--seqType fq \
--left reads_1.fq \
--right reads_2.fq \
--CPU 10 --max_memory 100G
```

Exercises: 4-2

- Align RNA-seq reads with a split read mapper STAR/HISAT2 and count the assigned reads with featureCounts

- Perform a gene expression quantification with kallisto

STAR/HISAT2 for RNA-seq read mapping

- HISAT2 explained on day #2
- STAR can perform a split read alignment required for RNA-seq data analysis
- Availability: https://github.com/alexdobin/STAR
- Usage:

```
STAR \
--runMode genomeGenerate \
--genomeDir /path/to/genome/index \
--genomeFastaFiles /path/to/reference/genome.fasta

STAR \
--genomeDir /path/to/genome/index \
--readFilesIn /path/to/reads/reads_R1.fastq /path/to/reads/reads_R2.fastq \
--outFileNamePrefix /path/to/output/directory/output_prefix
```

Read counting with featureCounts

- Counting number of reads mapped to each gene for expression quantification
- Availability: https://subread.sourceforge.net/featureCounts.html
- Usage:

```
featureCounts \
-a annotation.gtf \
-o counts.txt \
-t mRNA \
-g gene_id \
RNA_seq_mapping.sorted.bam
```

Exercises: 4-3

- Perform a differential gene expression analysis with DESeq2
- Perform a co-expression analysis with pbb-tools.de

Differential gene expression analysis with DESeq2

- DESeq2 is one of the best tools for differential gene expression analysis
- R package that needs to be utilized through R (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)
- Availability: https://github.com/bpucker/PBBtools/blob/main/collection/DGE_analysis.R
- Usage:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
      install.packages("BiocManager")
BiocManager::install("DESeq2")

library(DESeq2)
countData <- read.csv("count_matrix.csv", row.names = 1)
sampleMetadata <- read.csv("sample_metadata.csv", row.names = 1)

dds <- DESeqDataSetFromMatrix(countData, colData = sampleMetadata, design = ~ condition)
dds <- DESeq(dds)
res <- results(dds)
```

Coexpression analysis via pbb-tools

- Coexpression can reveal genes associated with the same pathway/process
- Availability: https://pbb-tools.de/CoExp/
- Usage:
  - Text file containing gene IDs as baits
  - Count table containing all TPMs (genes in rows, samples in columns)
  - File with functional annotation terms of all genes

Time for questions!