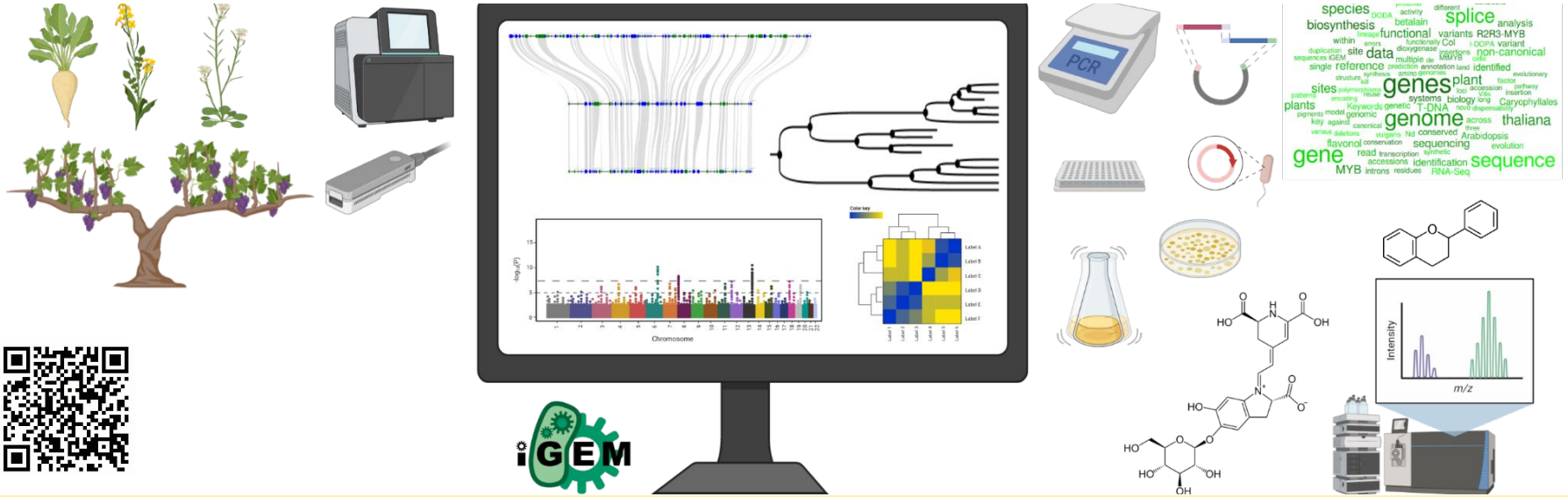
Technische Universität Braunschweig



RNA-seq (dry lab)

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

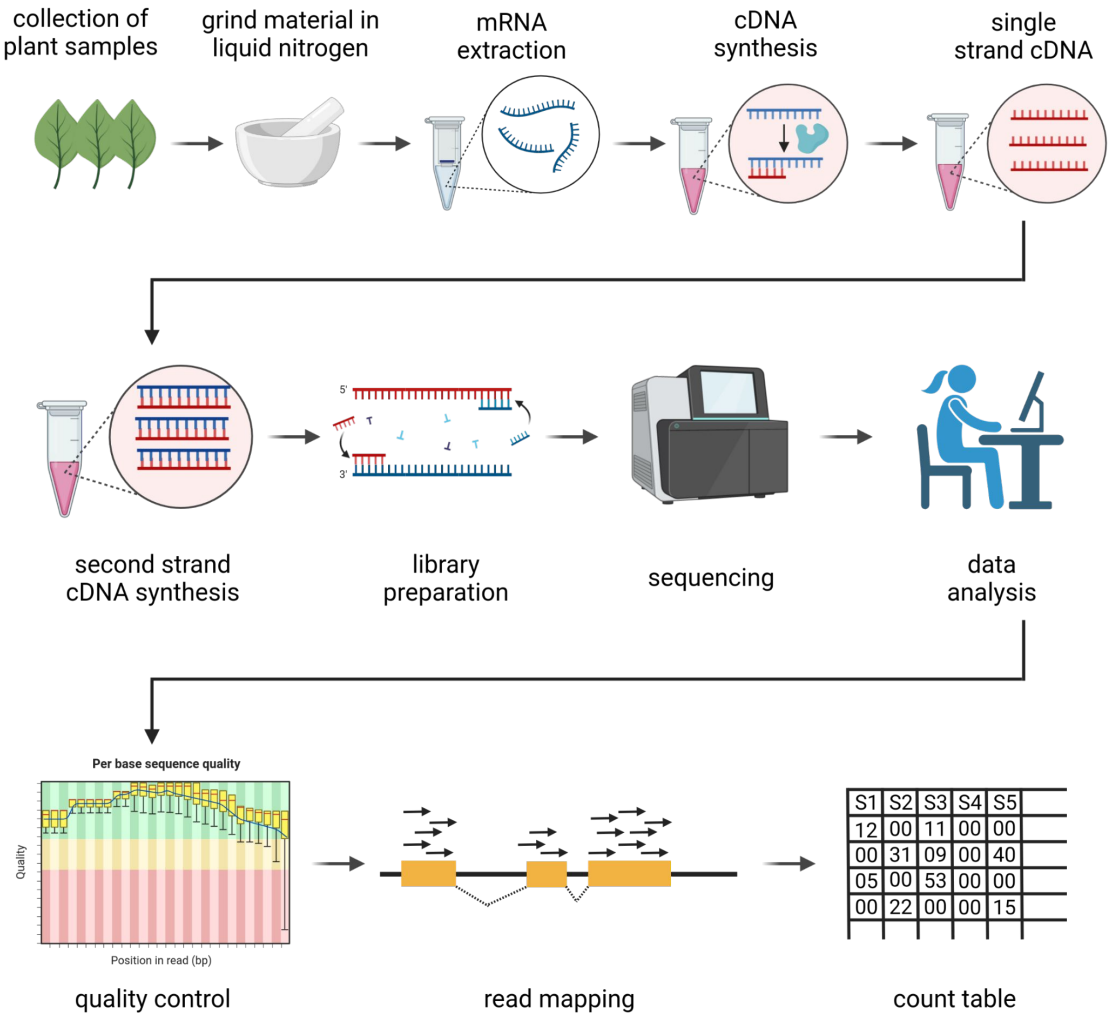
Availability of slides

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

- Questions: Feel free to ask at any time

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

RNA-seq workflow (reminder)



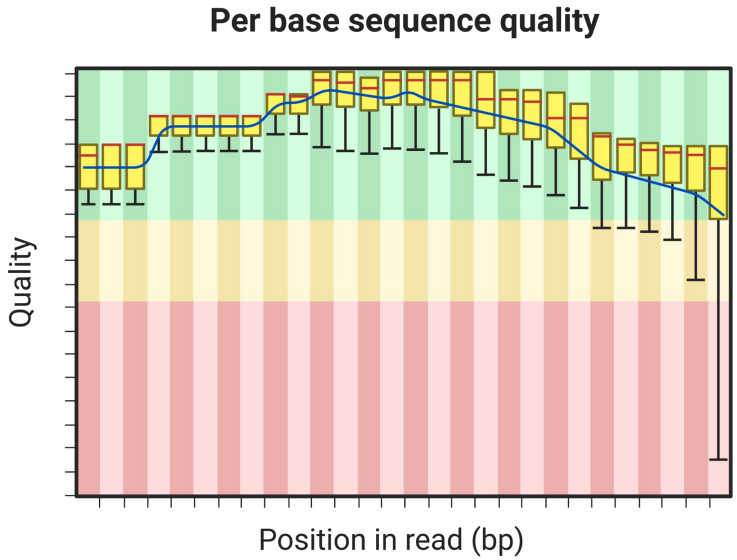
RNA-seq data: FASTQ files

- Results of RNA-seq experiments are 1-2 FASTQ files
- 1 FASTQ file: single end sequencing
- 2 FASTQ files: paired-end sequencing

```
@A00125:59:H3W7MDSXX:4:1101:4155:1016 1:N:0:ACCATCAC+TGATCTCC
GNCCCGTTCTCGGTATGCCGGTAGCTATGCTTGCATGTATGTGCAGATAGGTTAGGGGGGAAGCAACACTGTCTACGATCAAGTTGAACATATCTTGCTAC
+
F#FFFFFFFF:FFF,FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF
@A00125:59:H3W7MDSXX:4:1101:8260:1016 1:N:0:ACCATCAC+TGATCTCC
CNGAATCAAGGTCGCCGAGGGACGACGATGGACACGGTAGTCATACATTAACAACAGCAGGAGGTTCTGAAGTTGCAGGAGCTAGCCTCTTTGGTTTAGCT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF: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

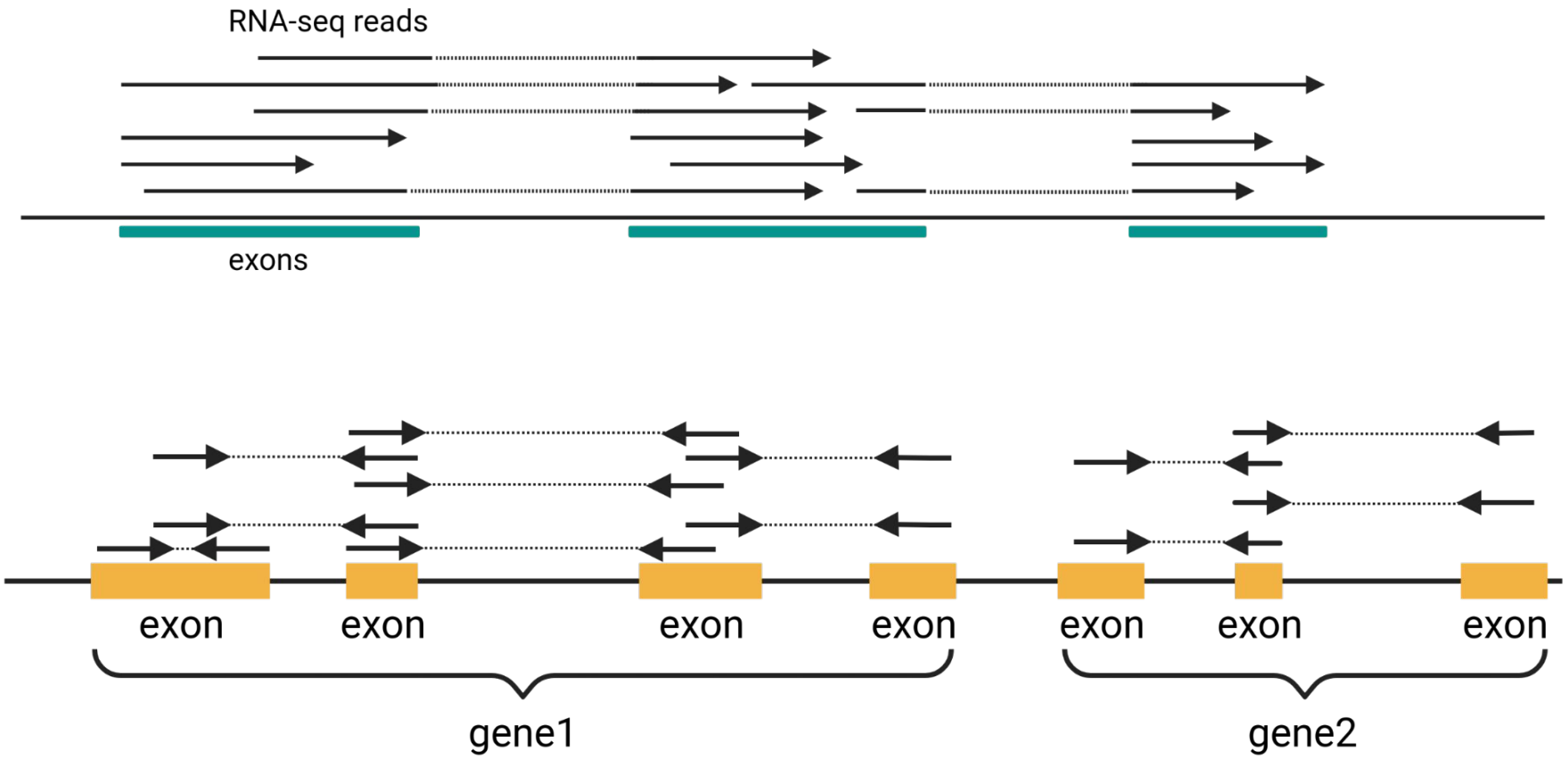


Phred score

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

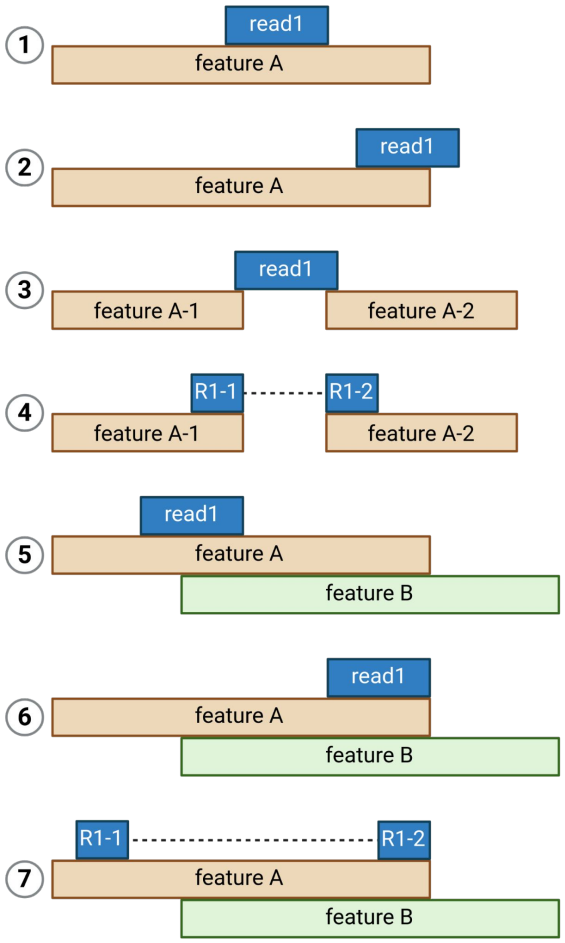
Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Read mapping



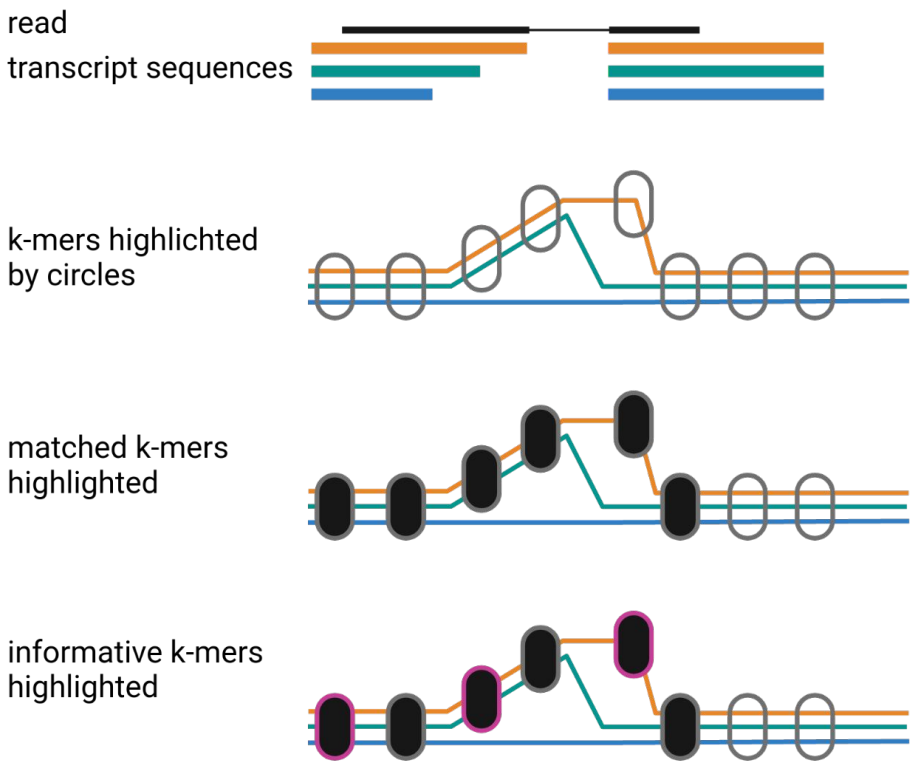
Counting mapped reads

- 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

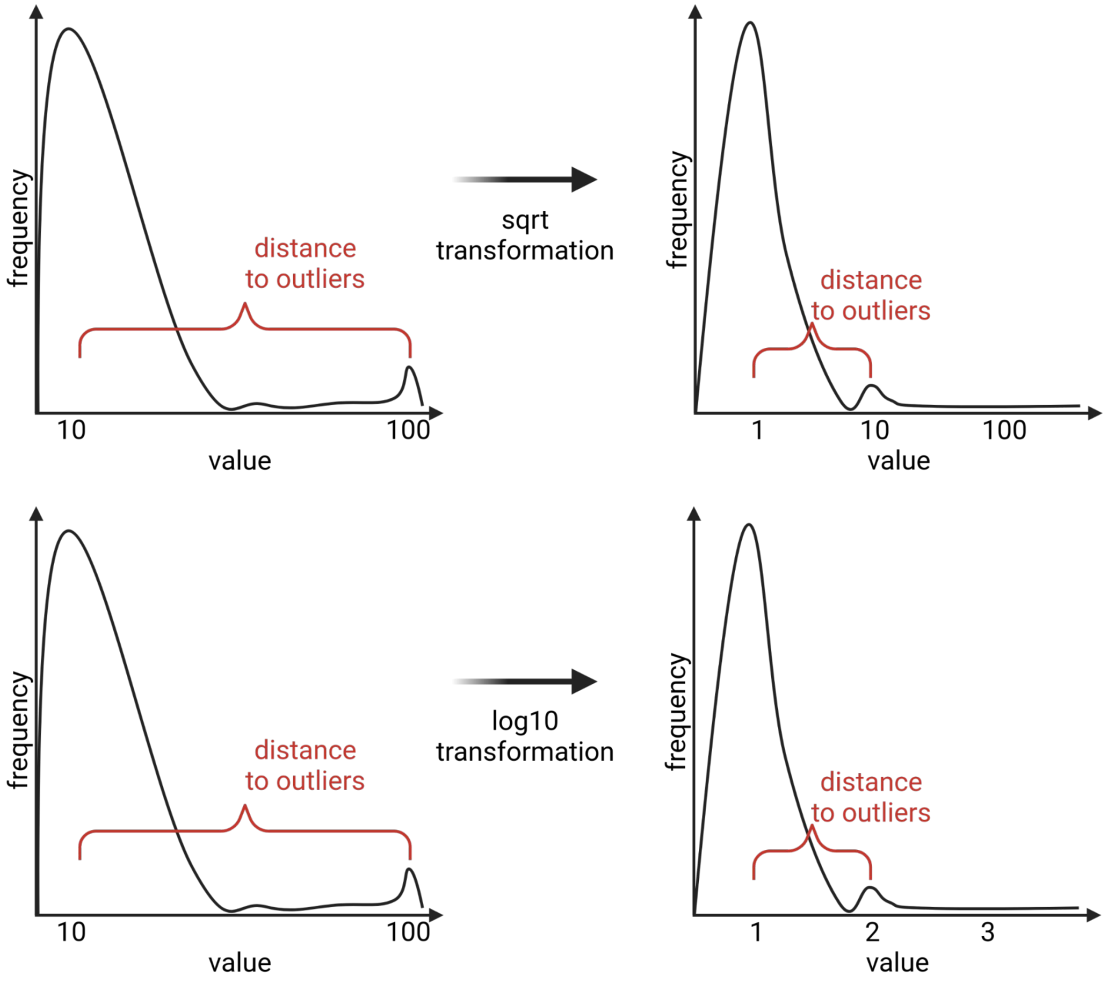
gene	SRR11603178	SRR11603179	SRR11603180	SRR11603181	SRR11603182	SRR11603183	SRR11603184	SRR11603185
TRINITY_DN10003_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10005_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	2.0
TRINITY_DN10005_c1_g1_i1	0.0	0.0	0.0	0.0	1.0	1.0	2.0	0.0	0.0	1.0	1.0	0.0	0.0
TRINITY_DN10008_c0_g1_i1	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	1.0	3.0	1.0	1.0	2.0	0.0
TRINITY_DN10032_c0_g1_i1	3.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	7.05039	11.2251
TRINITY_DN10032_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.94961	1.77489
TRINITY_DN10034_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0
TRINITY_DN10037_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	5.0
TRINITY_DN10038_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	2.0
TRINITY_DN10040_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0
TRINITY_DN10041_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	2.0
TRINITY_DN10043_c0_g1_i1	0.0	0.0	0.0	0.0	2.0	3.0	1.0	0.0	0.0	1.0	1.0	5.0	4.0
TRINITY_DN10049_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0
TRINITY_DN10050_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0
TRINITY_DN10052_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
TRINITY_DN10053_c0_g1_i1	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.0	0.0
TRINITY_DN10054_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
TRINITY_DN10054_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
TRINITY_DN10056_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0
TRINITY_DN10057_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	3.0	0.0	1.0
TRINITY_DN10058_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10058_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10059_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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

Normalized gene expression (units of gene expression)

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

Other normalization methods



DEG identification

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

DESeq2

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

edgeR

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

edgeR

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

DOI: 10.18129/B9.bioc.edgeR

Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.16)

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

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

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

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

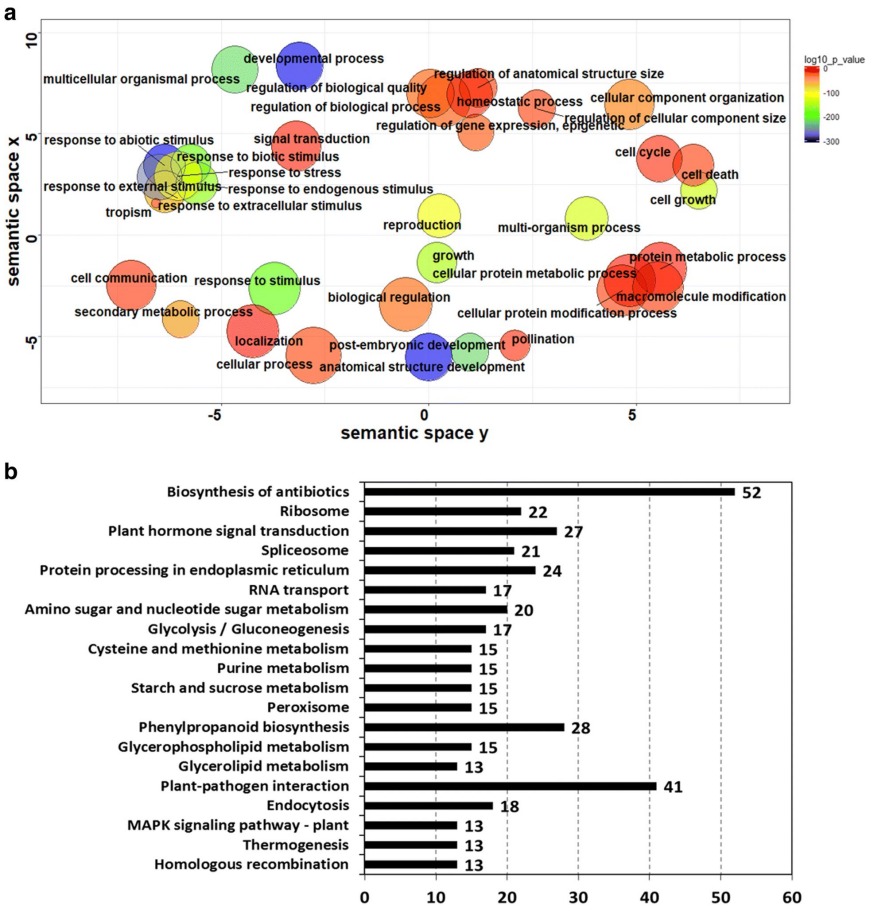
Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140. doi: 10.1093/bioinformatics/btp616.

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

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

GO enrichment

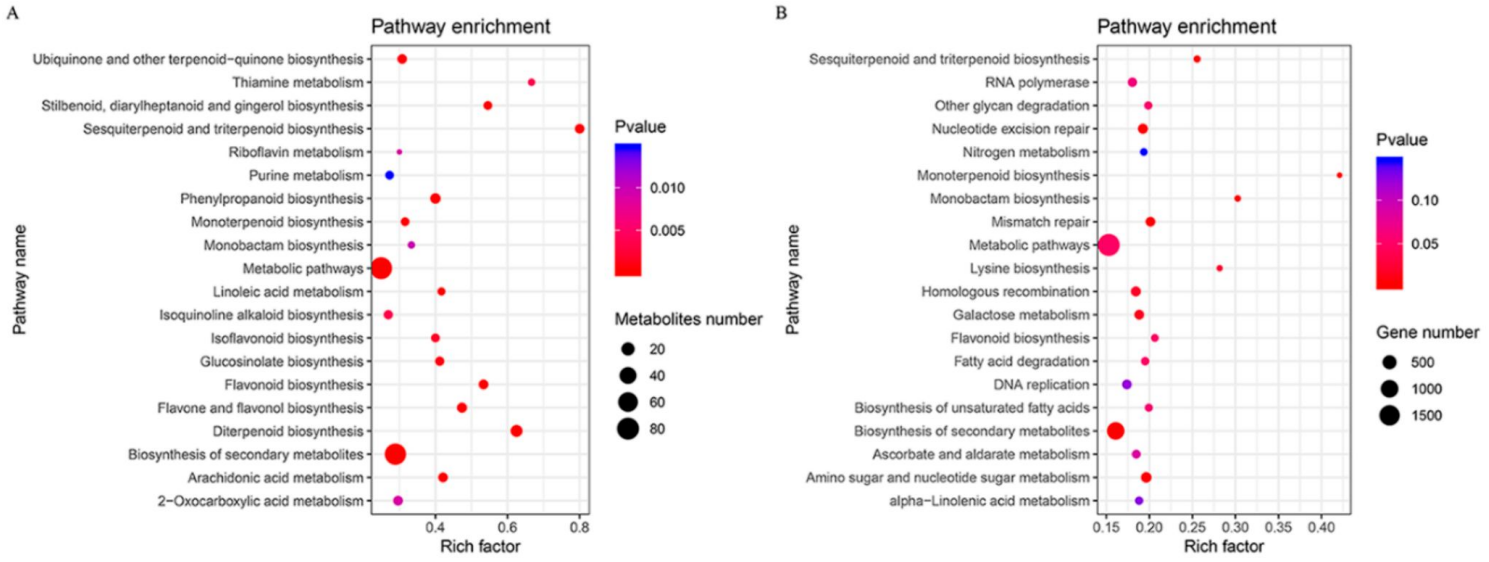
- GO = Gene Ontology
- GO term can be assigned to multiple genes (e.g. ‘flavonoid biosynthesis’)
- Up-regulated GO = larger proportion of genes with this GO term are up-regulated than expected by chance
- Tools: AmiGO, GOrilla, ShinyGO, GOnet, g:Profiler, …



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

KEGG enrichment

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



https://doi.org/10.3390/genes11020187

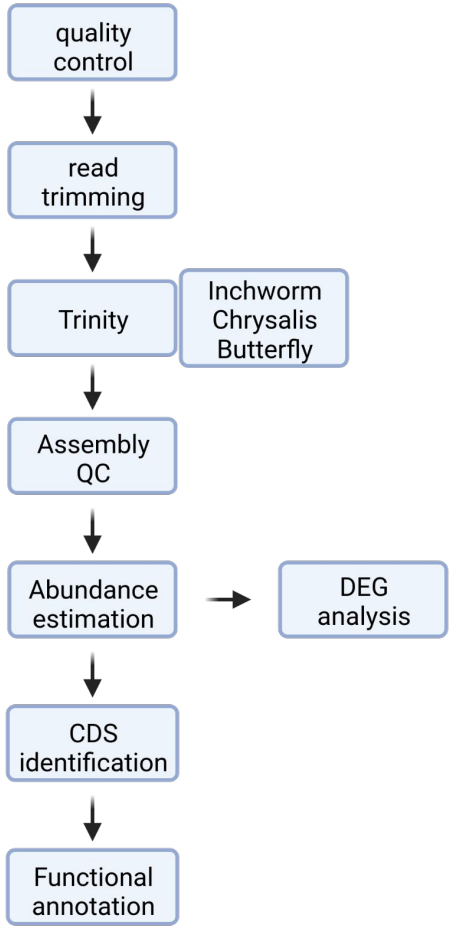
De novo transcriptome assemblers

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

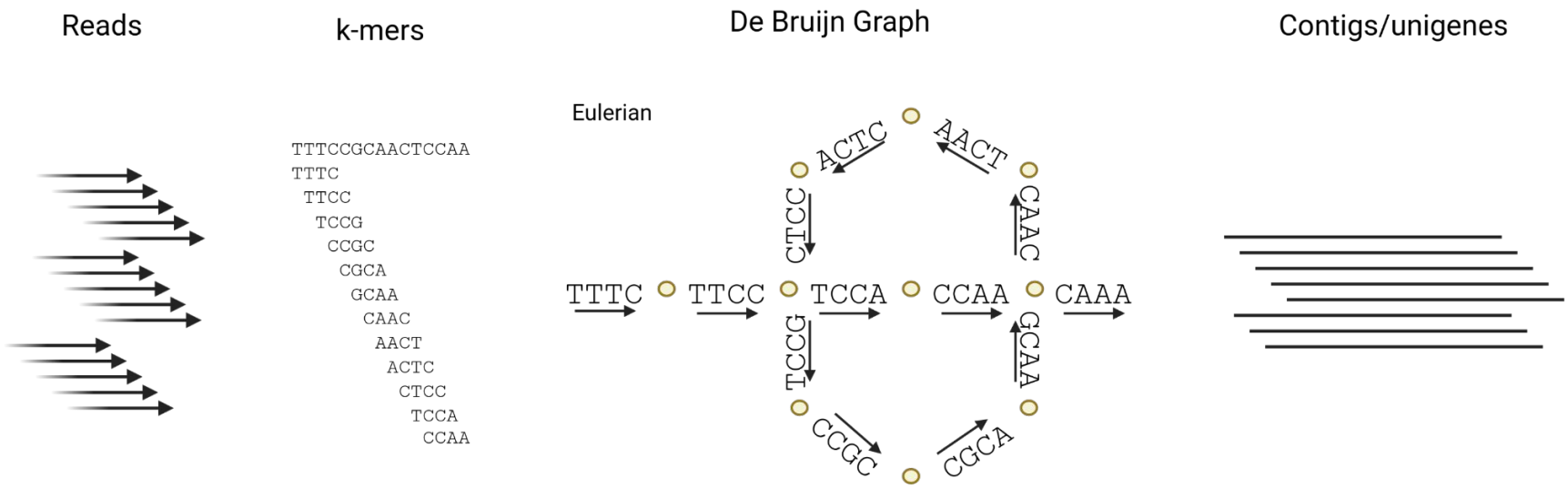
- rnaSPAdes:
  - Recent assembler with support for long reads
  - https://cab.spbu.ru/software/rnaspades/

Trinity *de novo* transcriptome assembly - overview

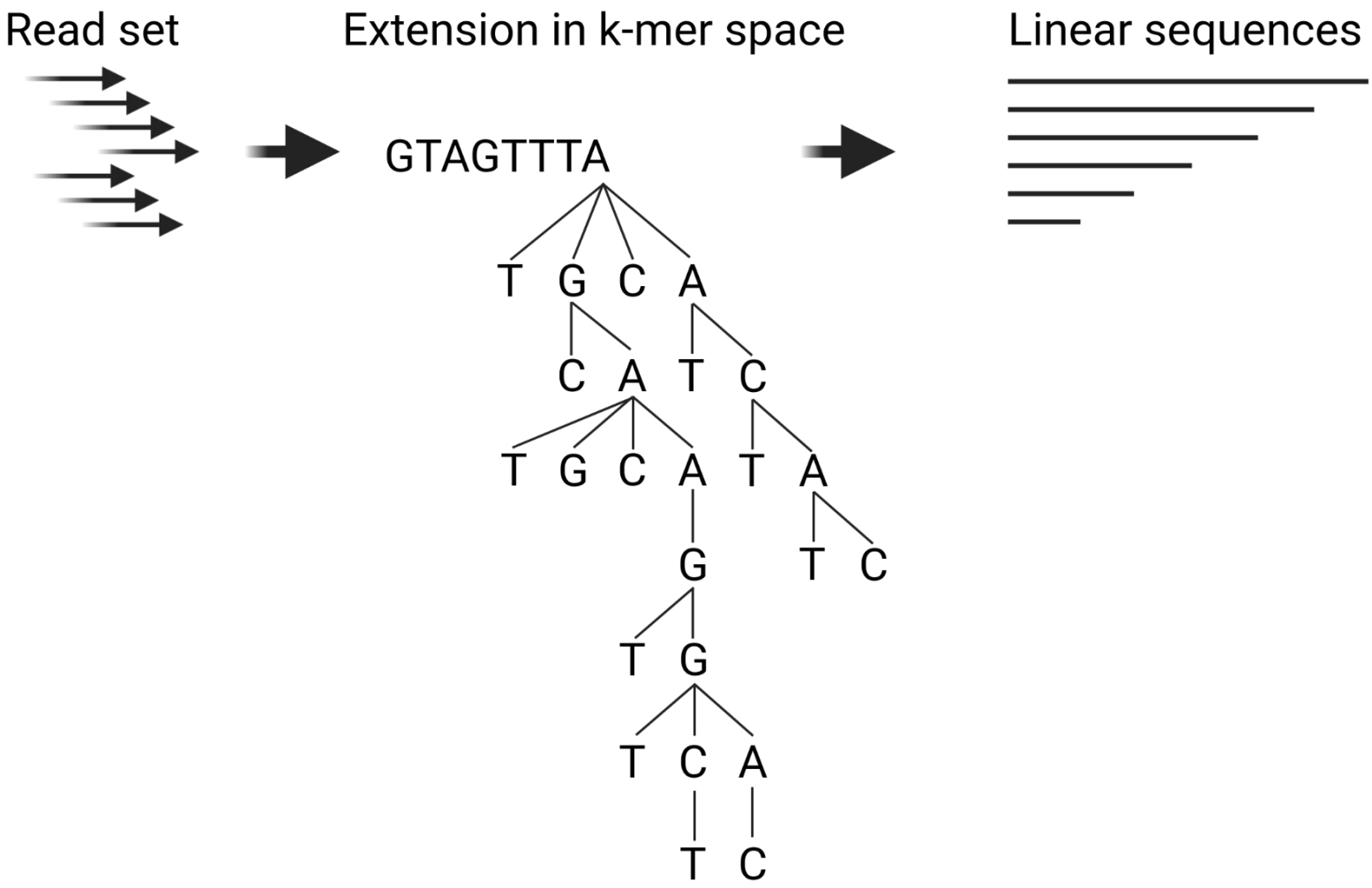
- Quality of reads should be checked prior to the assembly process
- Adapters and low quality parts need to be removed from the reads
- (Read name adjustment)
- Normalization of reads prior to assembly
- Assembly comprises multiple internal steps
- Many downstream analyses are possible depending on the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



De Bruijn Graph

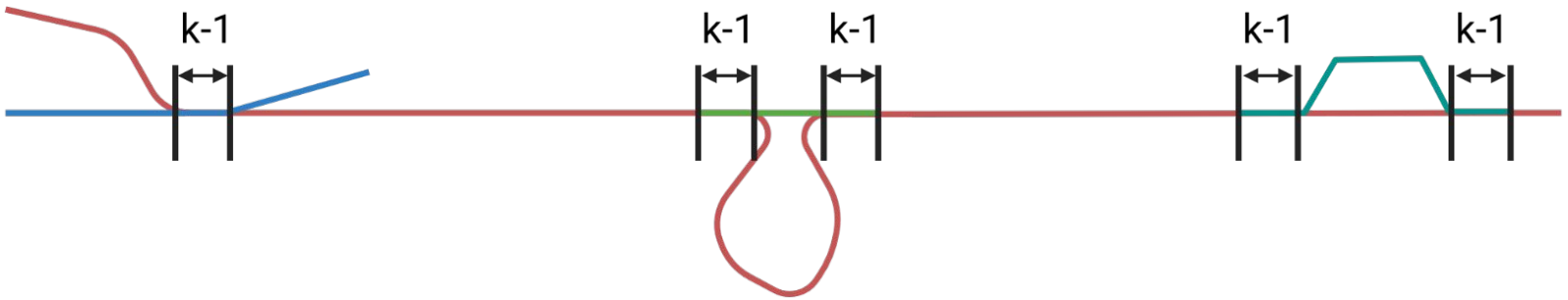


Stages of Trinity: Inchworm



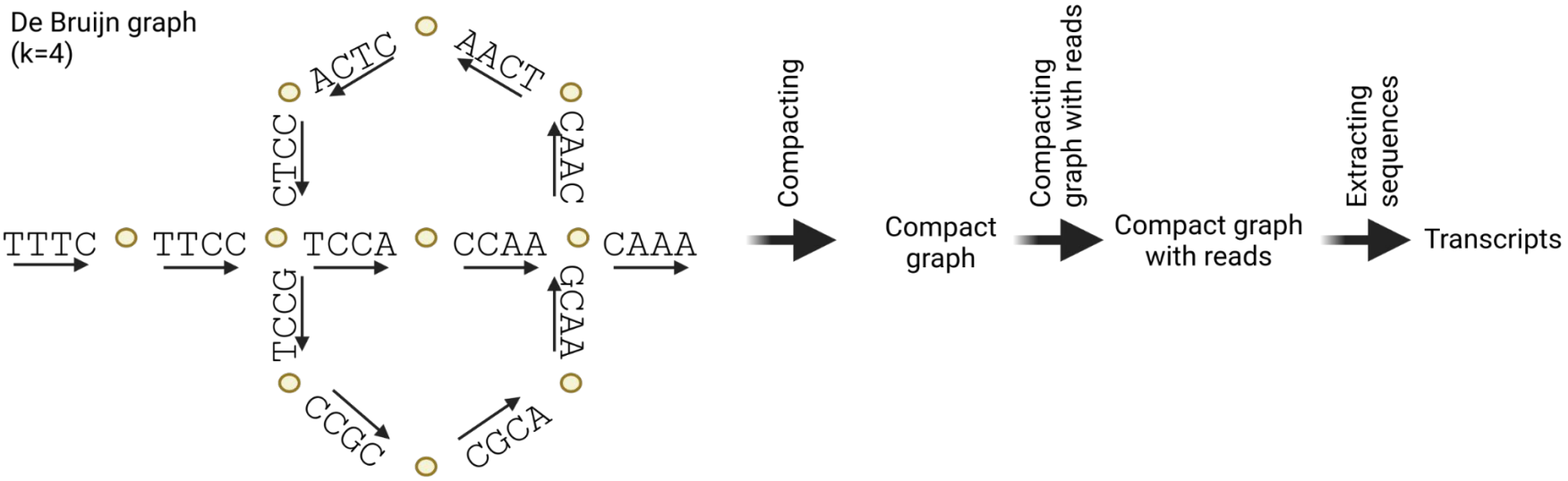
Stages of Trinity: Chrysalis

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



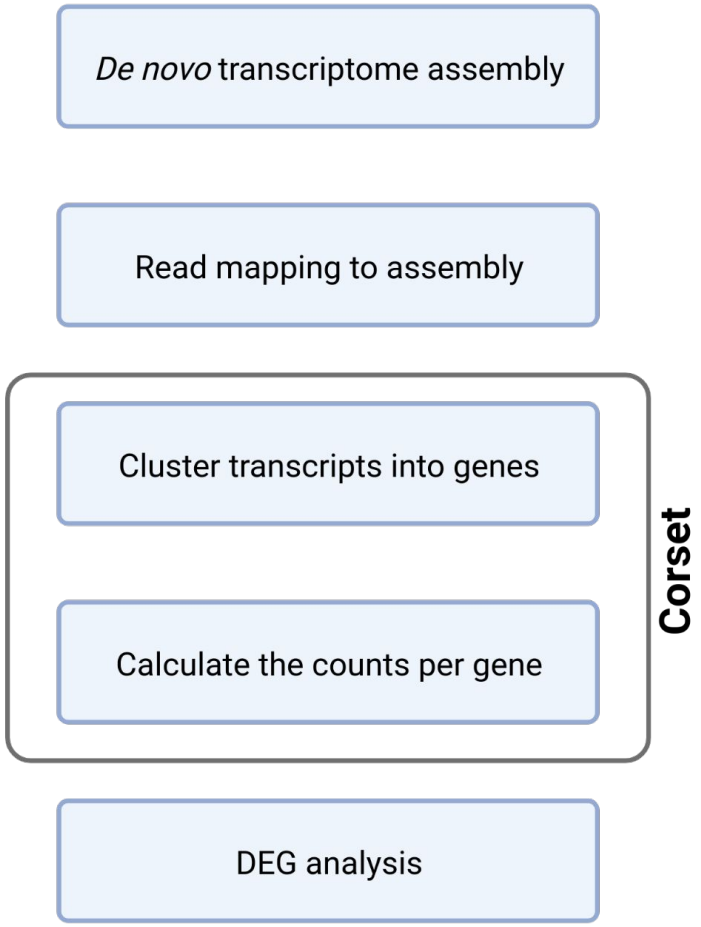
Stages of Trinity: Butterfly

- 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



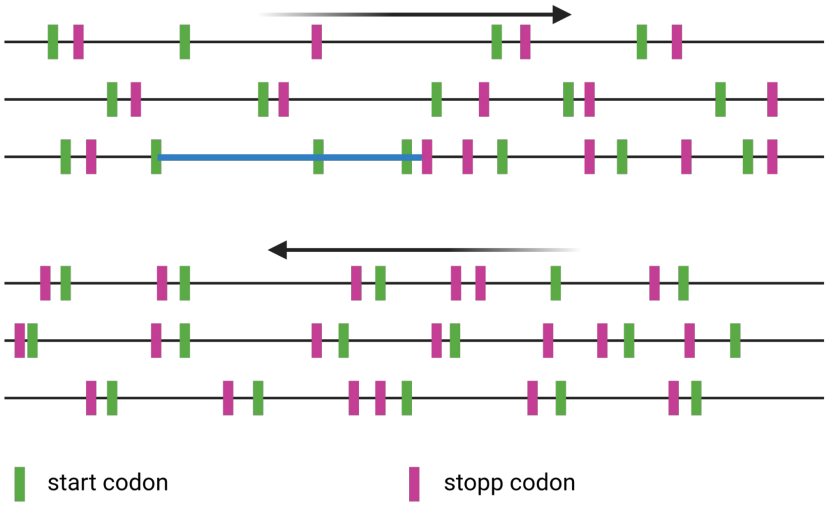
https://github.com/Oshlack/Corset
https://doi.org/10.1186/s13059-014-0410-6

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

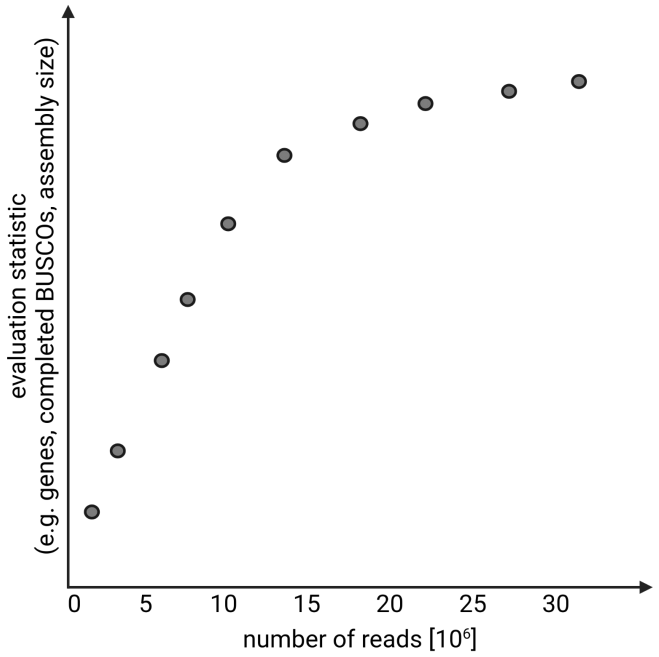
- Normalization of reads: exclude reads of highly abundant transcripts
  - Minimal and maximal k-mer coverage used as filter

- 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)

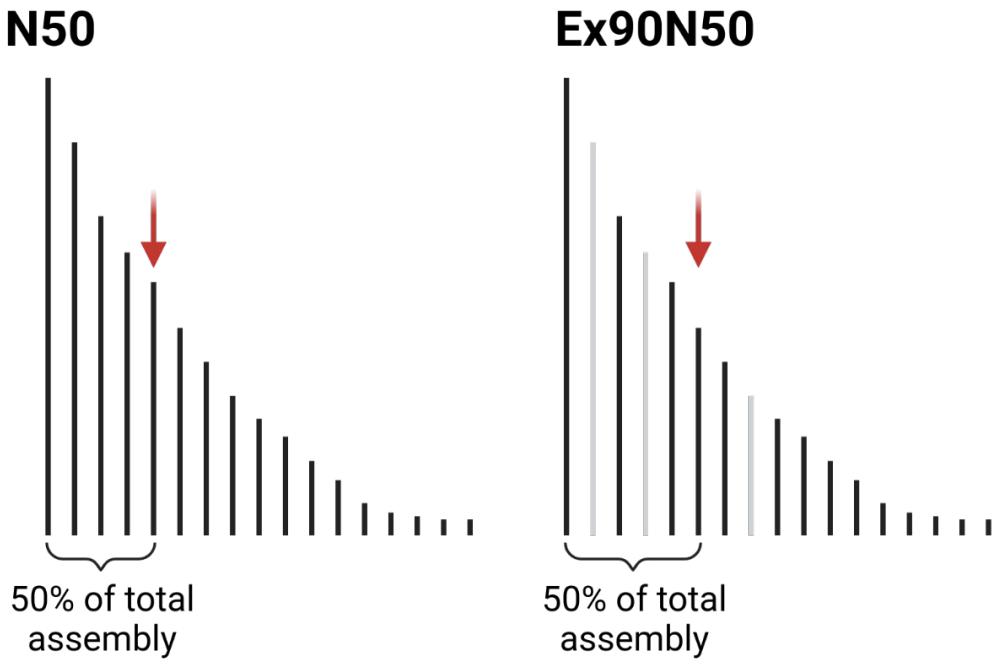
- Example: 20 million fragments; 1+2+4 hours (20 CPUs)
- Example: 200 million fragments: 1+120+7 hours (20 CPUs)
- Trinity is restartable if the process is interrupted
- Restart failed Butterfly jobs (see log file) with reduced CPU number
  - Optional settings: --bflyCPU 20 --bflyHeapSpaceMax 4G

Assembly evaluation

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

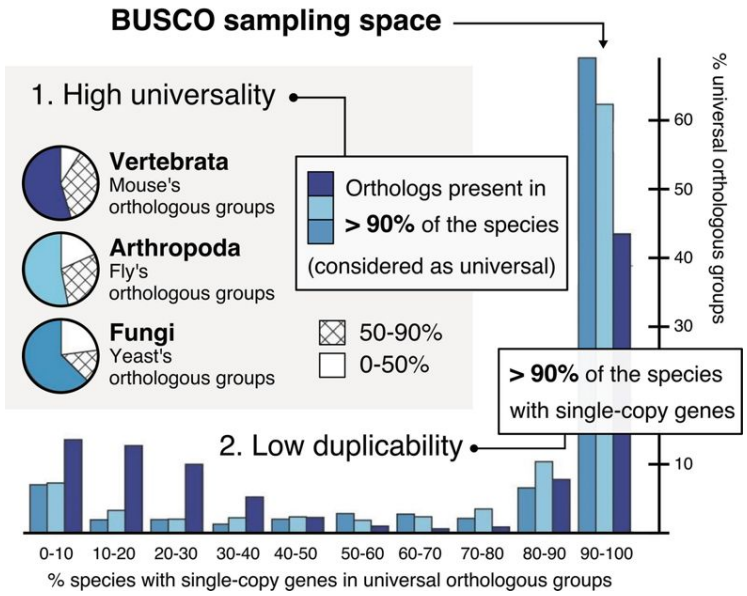
N50 & Ex90N50

- 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

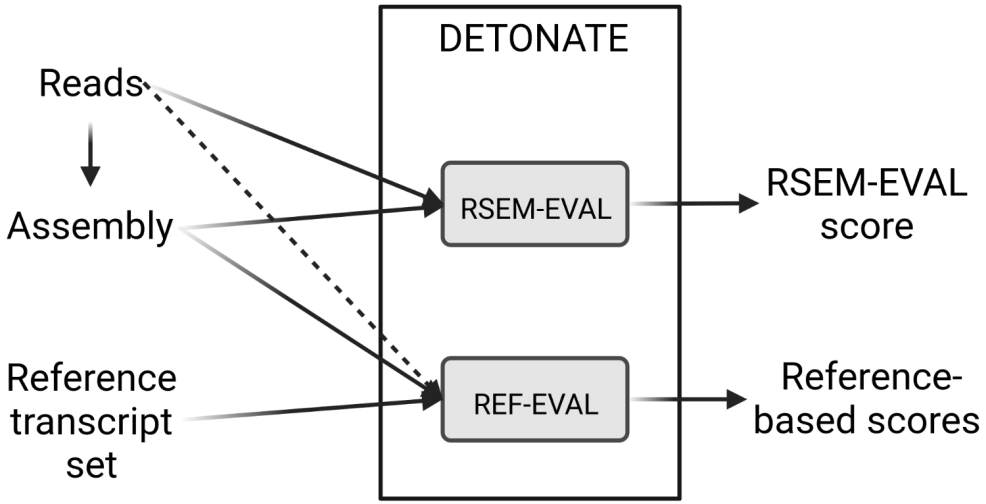
- BUSCO = Benchmarking Universal Single-Copy Orthologs
- Measuring assembly completeness based on presence of conserved genes
- Assessment of annotation quality



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

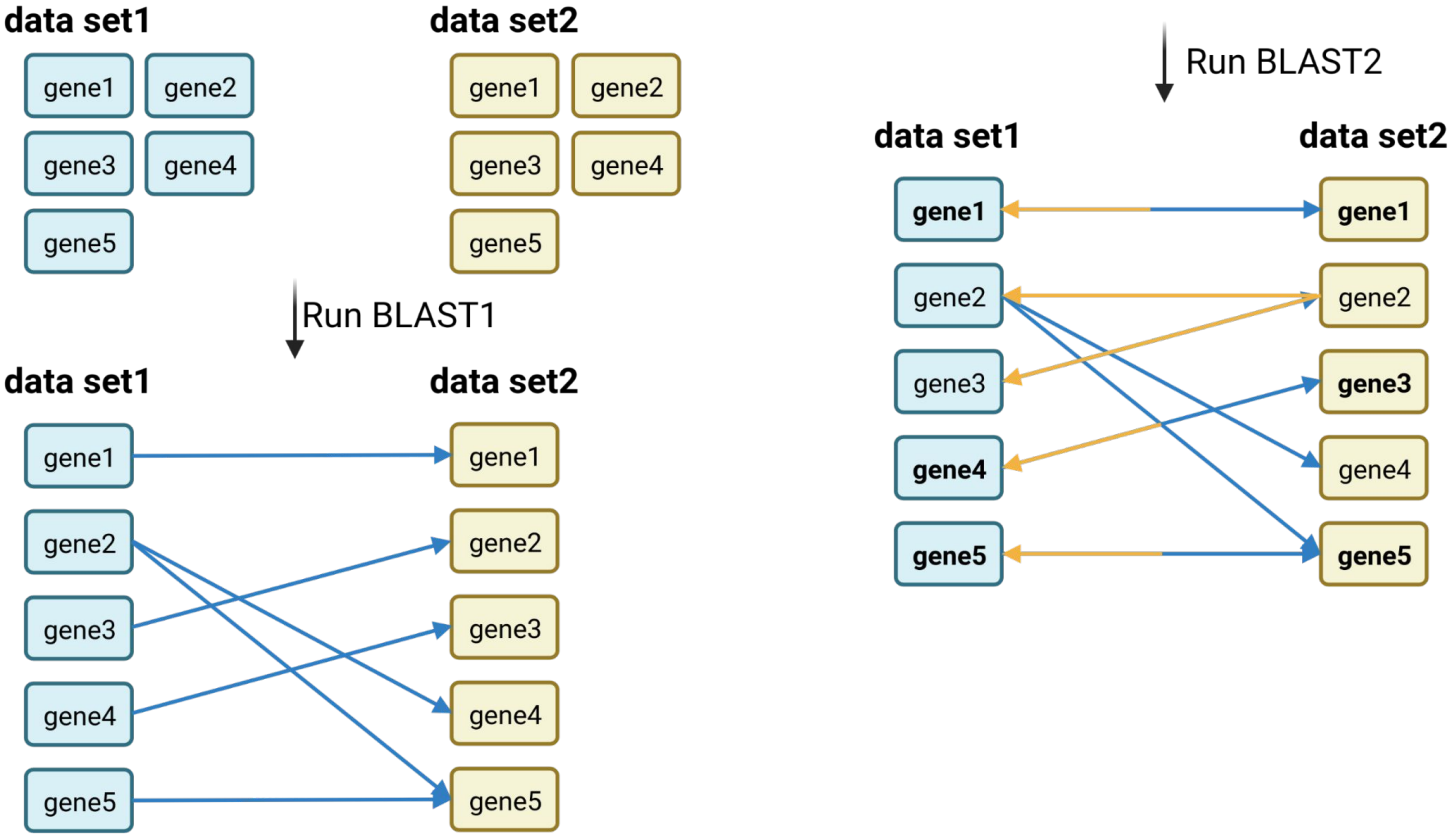


https://github.com/deweylab/detonate
https://doi.org/10.1186/s13059-014-0553-5

Functional annotation

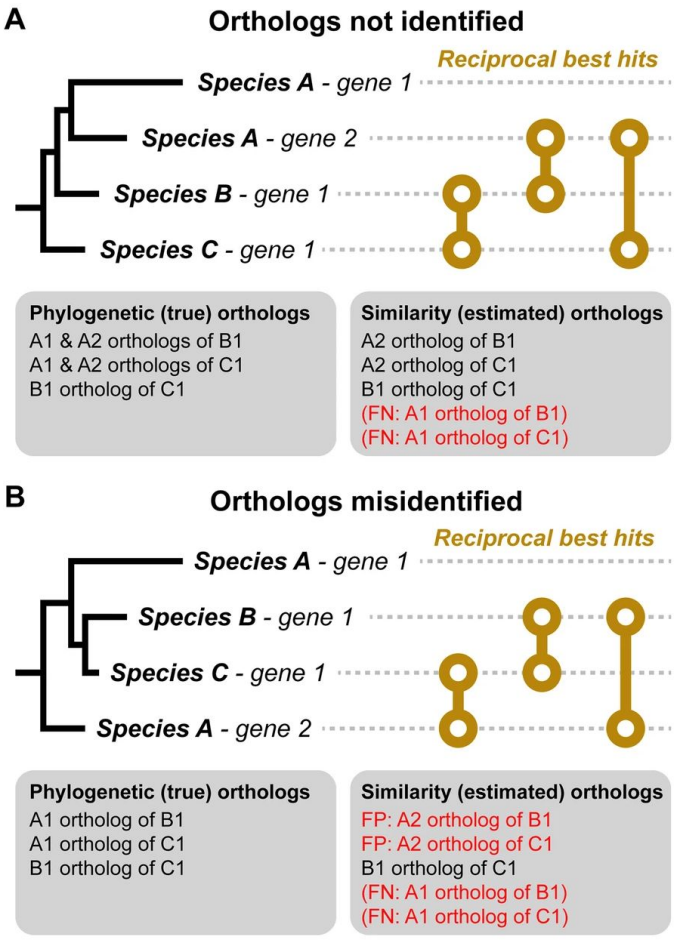
- Reciprocal Best BLAST Hits (RBHs)
- OrthoFinder2
- Knowledge-based Identification of Pathway Enzymes (KIPEs)
- Mercator
- InterProScan5

Reciprocal Best BLAST Hits (RBHs)



OrthoFinder2

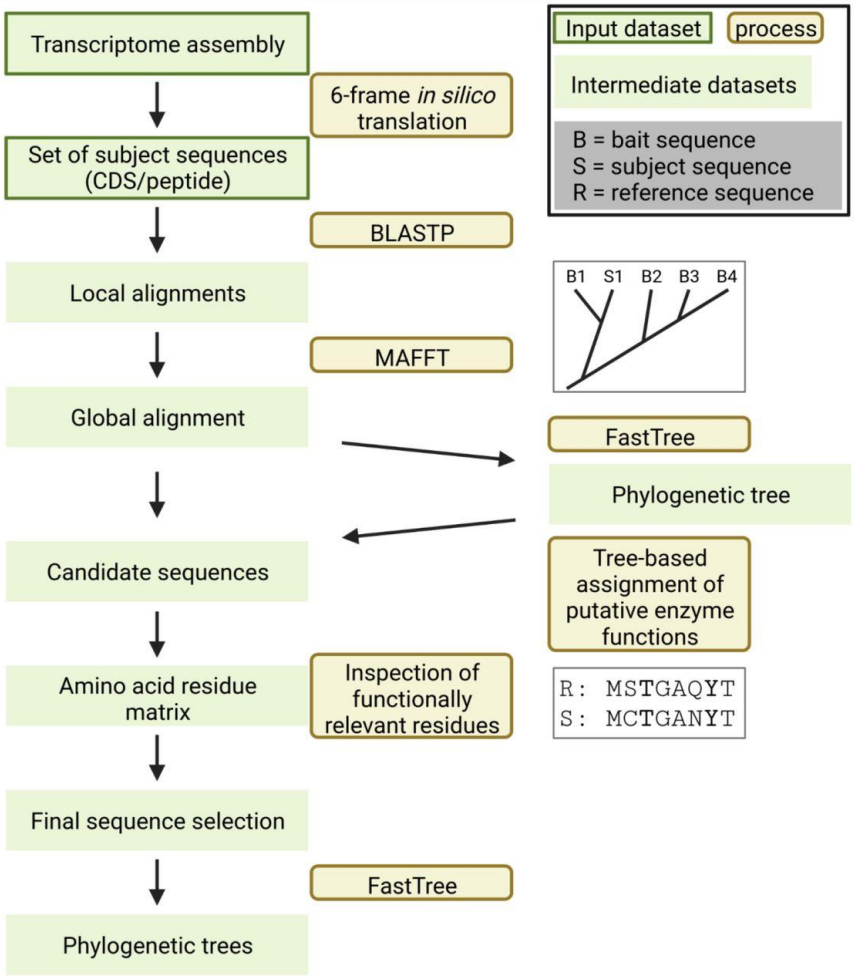
- Identification of orthologs often not possible
- Orthogroups are better reflection of reality
- Assumption: orthologs have the same function



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

KIPEs

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



https://doi.org/10.1101/2022.06.30.498365

Mercator

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

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

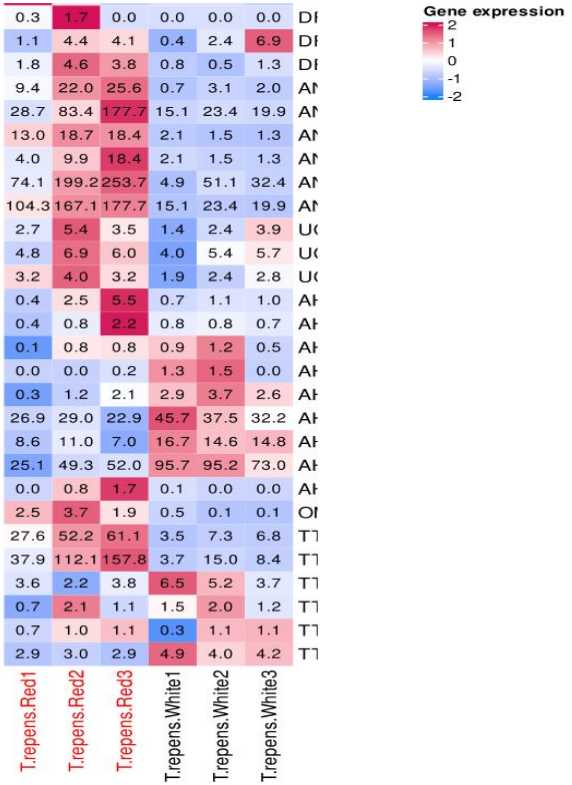
InterProScan

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

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

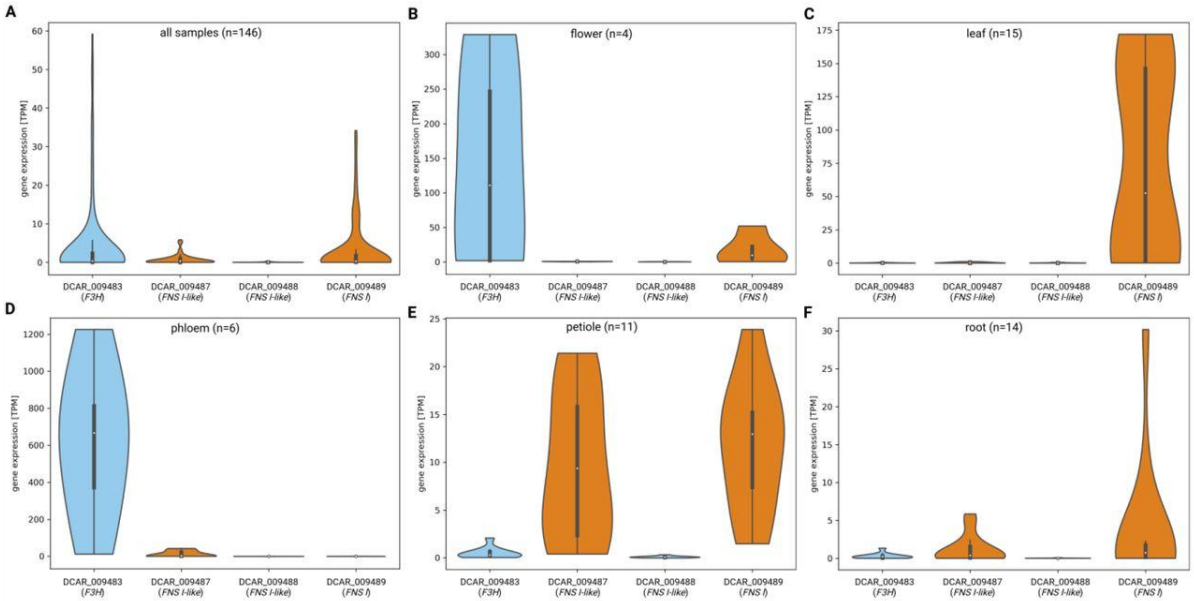
Heatmaps

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



Violin plots

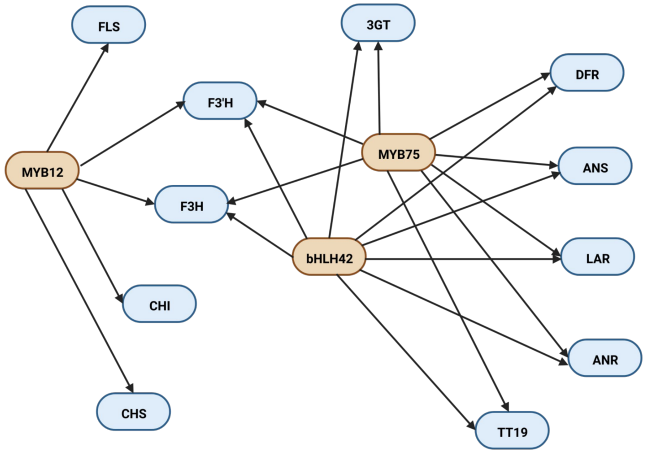
- Summarization of multiple data points
- Indication of data point distribution
- Mean and median can be displayed



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

Gene expression networks

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

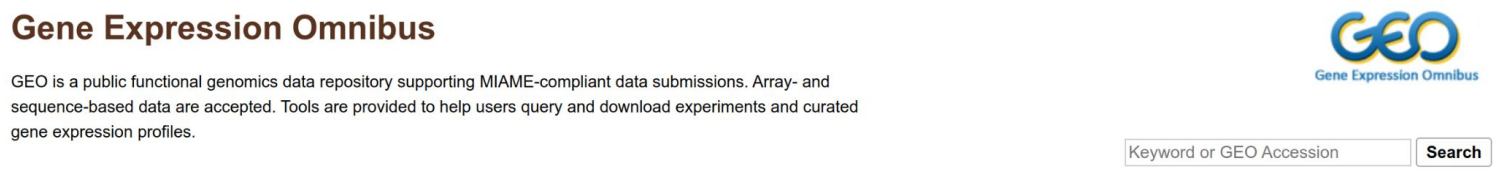


Data submission

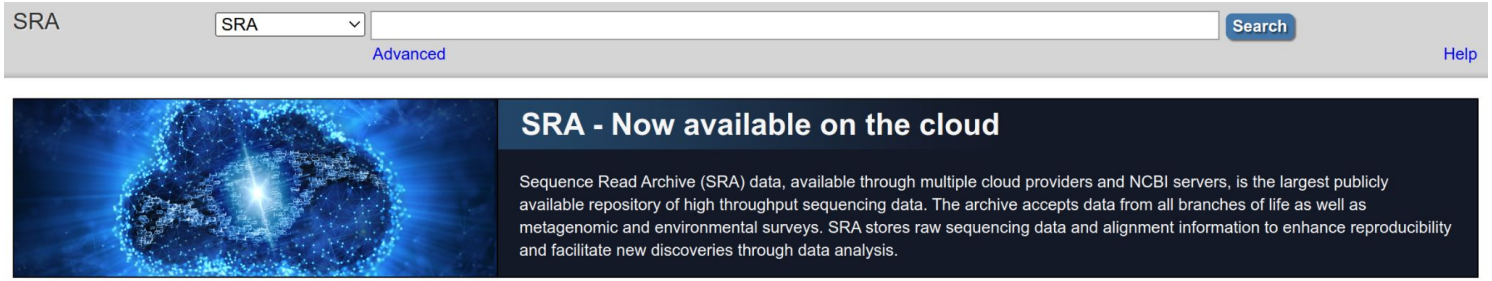
- Databases: GEO/SRA
- Hashes for submission check: md5sum, sha256
- Filezilla for transfer
- Specify IDs in data availability statement
- Sharing metadata is crucial for efficient reuse

RNA-seq databases

- Gene Expression Omnibus (GEO)



- Sequence Read Archive (SRA)



Hashes

- Hash = digital fingerprint of a file
- Hashes are useful to check completeness of file transfer
- Md5sum: frequently used hash to validate file completeness
- Sha256sum: more secure hash to exclude file manipulation

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

Time for questions!

Questions

1. What is the structure of a FASTQ file?
2. Which parameters can be analyzed to check the RNA-seq data quality?
3. What is the Phred score?
4. What are challenges in the read counting to quantify gene expression?
5. What does a count table look like?
6. What are the different units of gene expression?
7. What are DEGs and how are they identified?
8. Which tools can be used to functionally annotated unigenes?
9. Which approaches can be used to visualize gene expression?
10. What are important metadata elements that should be shared with data sets?