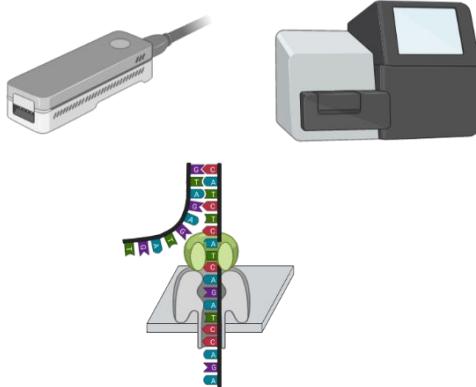
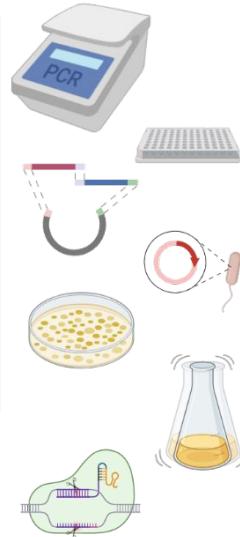
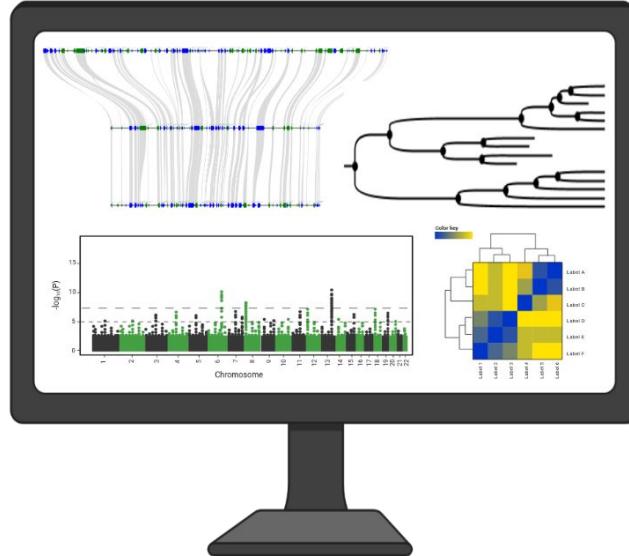




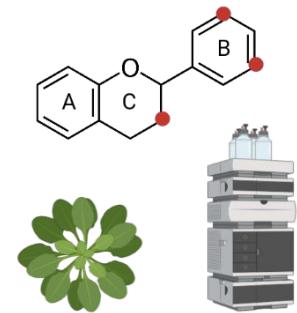
Technische
Universität
Braunschweig



Plant Biotechnology
and Bioinformatics



species biosynthesis proteins analysis different conditions
biosynthesis species activities bellman
functional variants H293-MYB
within genes site data functional Col 100% variant
duplex site data divergently expressed non-canonical
sequencer ICGM single reference multiple transcript annotation level identified
single reference structure synthesis gene annotation level identified
sites evolution pathway evolutionary
plants pigment model genome systems biology long interspecies
Keyvanda genes key against canonical conservation Arabidopsis
pigments model genome sequencing across Canophylales
plants sequencing sequencing free thaliana
flavonoid conservation read transcription synthetic accessions identification sequence
gene sequencing sequencing for conserved MYB introns residues RNA-Seq



RNA-seq (dry lab)

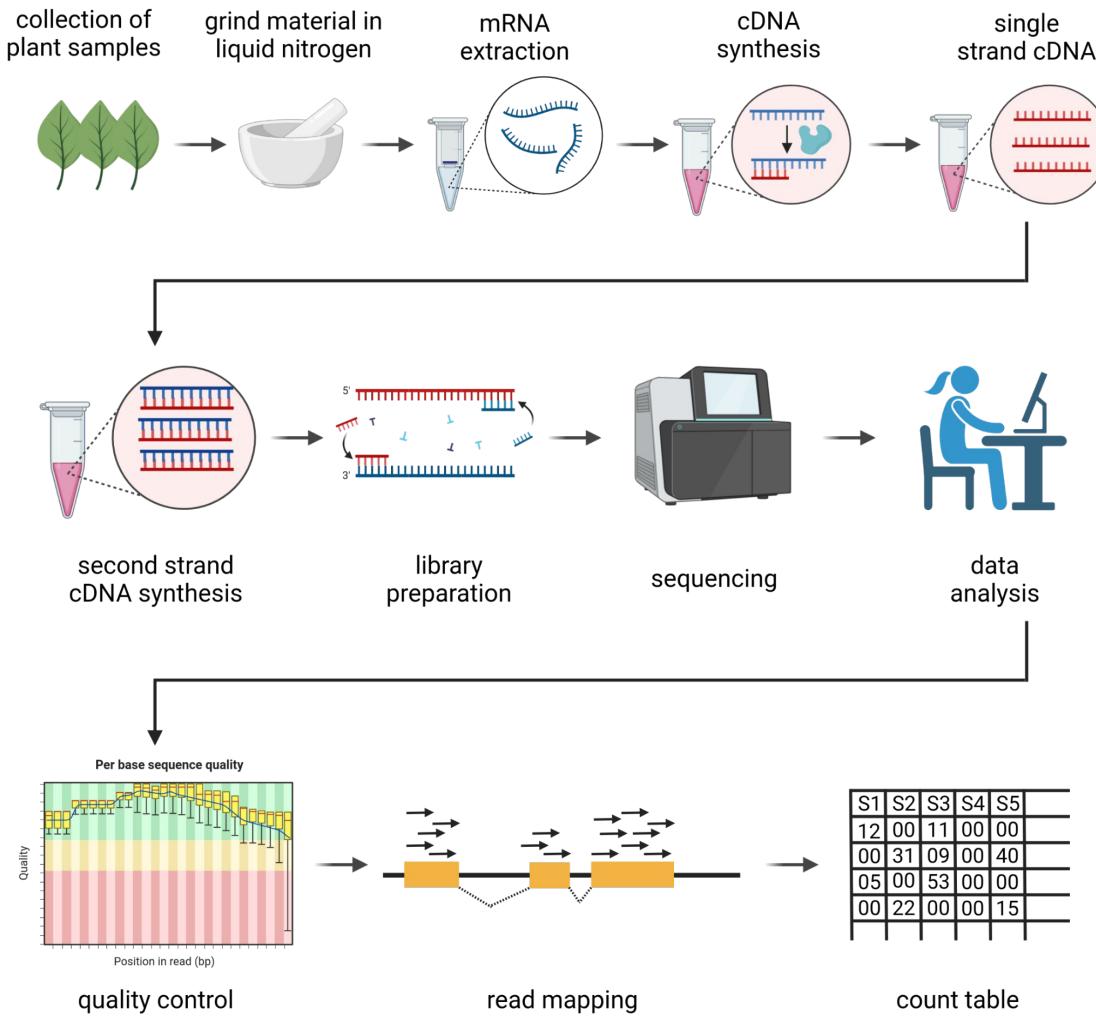
Prof. Dr. Boas Pucker
(Plant Biotechnology and Bioinformatics)

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

My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

RNA-seq workflow (reminder)



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
GNCCCGTTCTCGGTATGCCGGTAGCTATGCTGCATGTGAGATAGGTTAGGGGGGAAGCAACACTGTCTACGATCAAGTTAACATATCTTGTAC
+
F#FFFFFF:FFF,FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF
@A00125:59:H3W7MDSXX:4:1101:8260:1016 1:N:0:ACCATCAC+TGATCTCC
CNGAACATCAAGGTGCCGAGGGACGACGATGGACACGGTAGTCATACATTAAACACAGCAGGAGGTTCTGAAGTTGCAGGAGCTAGCCTTTGGTTAGCT
+
F#FF:FF:FFF:FFFF
@A00125:59:H3W7MDSXX:4:1101:10773:1016 1:N:0:ACCATCAC+TGATCTCC
CNCCTCTCTGCAACTAGCTACTGCCAATGACTAACTCTAATCTTCTTCTTCAACAAAACCTTCTTACTTTCTATTATTTTACCAAGT

ASCII code

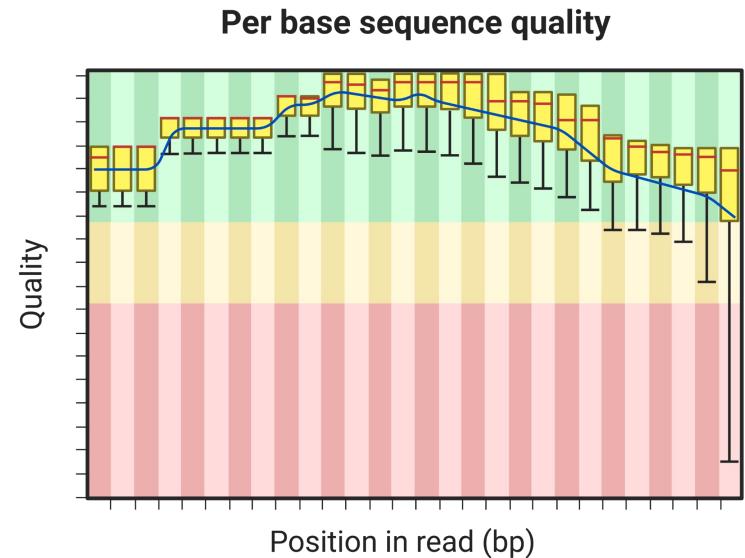
- Phred score encoded with ASCII
- Phred32, phred64 (different offset values to avoid special characters)

Binary	Oct	Dec	Hex	Glyph		
				1963	1965	1967
010 0000	040	32	20	space		
010 0001	041	33	21	!		
010 0010	042	34	22	"		
010 0011	043	35	23	#		
010 0100	044	36	24	\$		
010 0101	045	37	25	%		
010 0110	046	38	26	&		
010 0111	047	39	27	'		
010 1000	050	40	28	(
010 1001	051	41	29)		
010 1010	052	42	2A	*		
010 1011	053	43	2B	+		
010 1100	054	44	2C	,		
010 1101	055	45	2D	-		
010 1110	056	46	2E	.		
010 1111	057	47	2F	/		
011 0000	060	48	30	0		
011 0001	061	49	31	1		
011 0010	062	50	32	2		
011 0011	063	51	33	3		
011 0100	064	52	34	4		
011 0101	065	53	35	5		
011 0110	066	54	36	6		
011 0111	067	55	37	7		
011 1000	070	56	38	8		
011 1001	071	57	39	9		
011 1010	072	58	3A	:		
011 1011	073	59	3B	;		
011 1100	074	60	3C	<		
011 1101	075	61	3D	=		
011 1110	076	62	3E	>		
011 1111	077	63	3F	?		
100 0000	100	64	40	@	'	@

100 0000	100	64	40	@	'	@
100 0001	101	65	41		A	
100 0010	102	66	42		B	
100 0011	103	67	43		C	
100 0100	104	68	44		D	
100 0101	105	69	45		E	
100 0110	106	70	46		F	
100 0111	107	71	47		G	
100 1000	110	72	48		H	
100 1001	111	73	49		I	
100 1010	112	74	4A		J	
100 1011	113	75	4B		K	
100 1100	114	76	4C		L	
100 1101	115	77	4D		M	
100 1110	116	78	4E		N	
100 1111	117	79	4F		O	
101 0000	120	80	50		P	
101 0001	121	81	51		Q	
101 0010	122	82	52		R	
101 0011	123	83	53		S	
101 0100	124	84	54		T	
101 0101	125	85	55		U	
101 0110	126	86	56		V	
101 0111	127	87	57		W	
101 1000	130	88	58		X	
101 1001	131	89	59		Y	
101 1010	132	90	5A		Z	
101 1011	133	91	5B		[
101 1100	134	92	5C	\	-	\
101 1101	135	93	5D]	
101 1110	136	94	5E	↑	^	
101 1111	137	95	5F	←	-	
110 0000	140	96	60		@	'
110 0001	141	97	61		a	
110 0010	142	98	62		b	
110 0011	143	99	63		c	
110 0100	144	100	64		d	
110 0101	145	101	65		e	
110 0110	146	102	66		f	
110 0111	147	103	67		g	
110 1000	150	104	68		h	

Quality checks

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

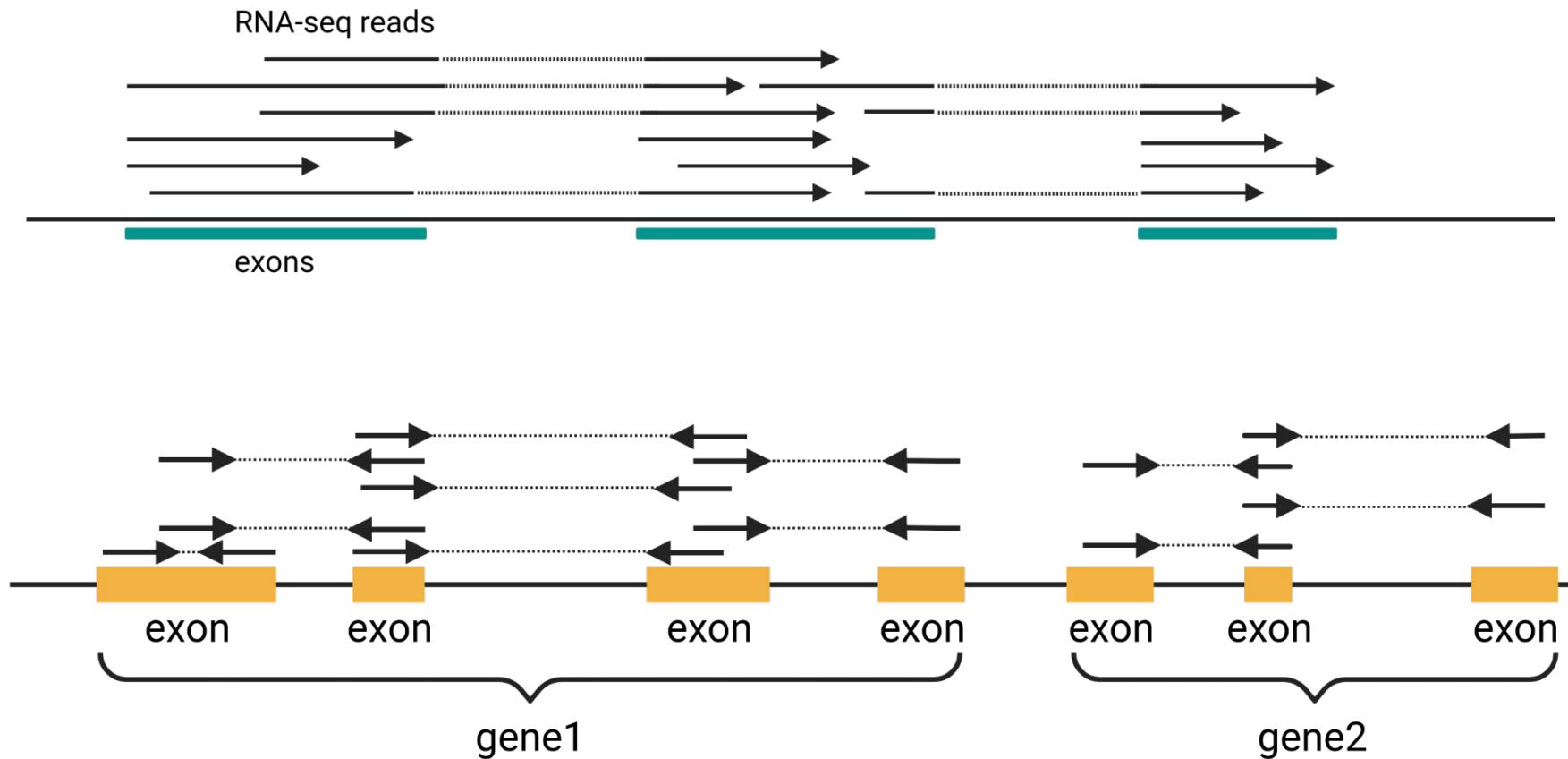


Phred score

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

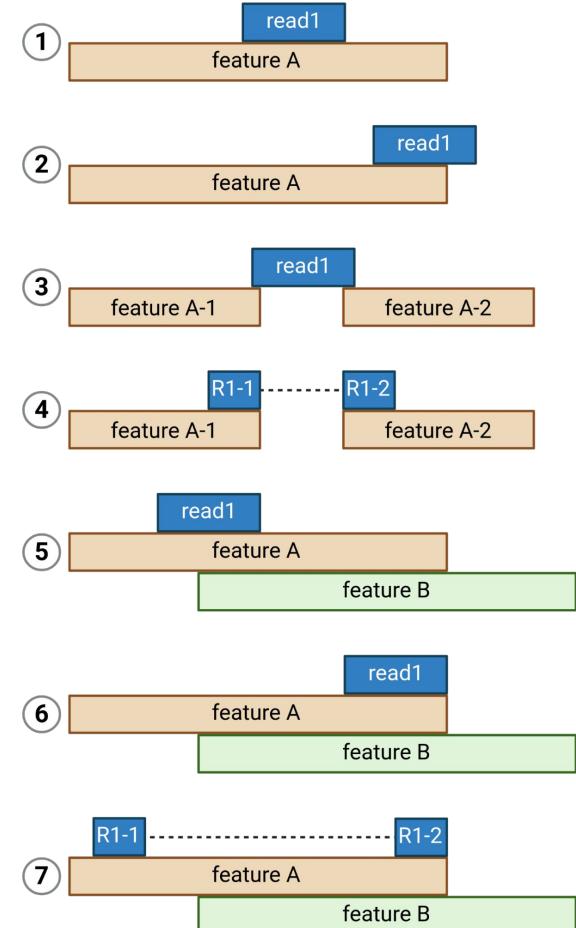
Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Read mapping



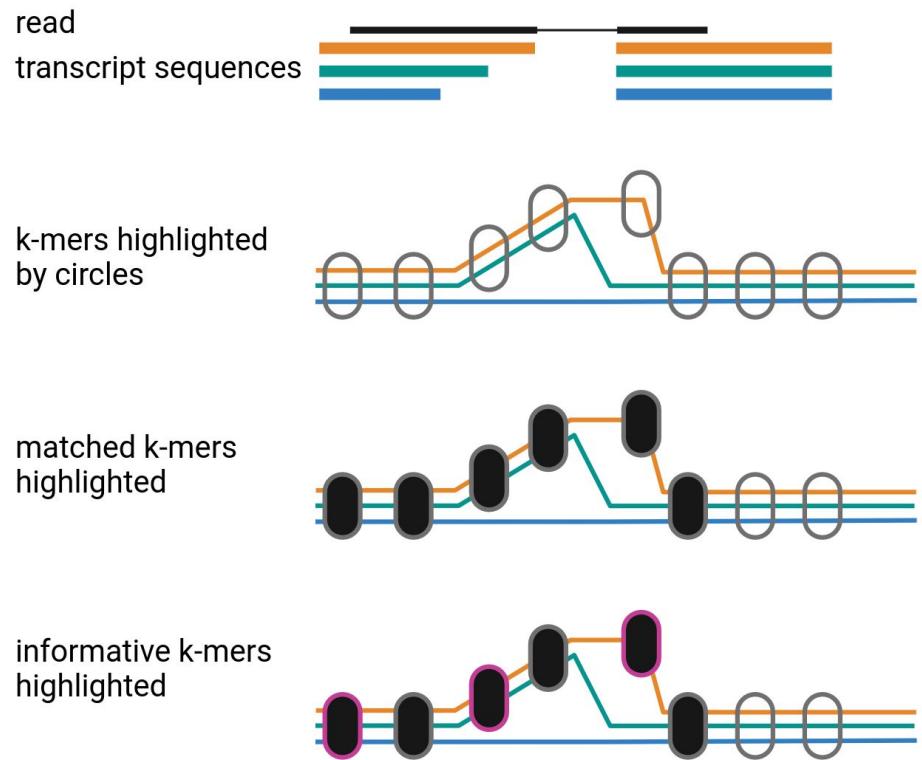
Counting mapped reads

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



Mapping-free quantification: kallisto

- Much faster than proper read mappings with STAR/HISAT2
- Context of k-mers is considered and not only individual k-mers



Count tables

- Gene/transcript IDs in first column
- Sample IDs in first row
- Expression values in all fields of the table

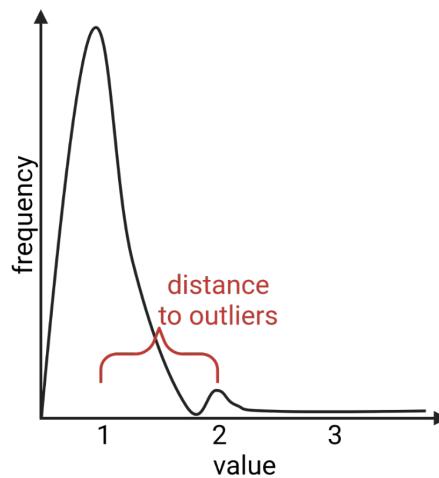
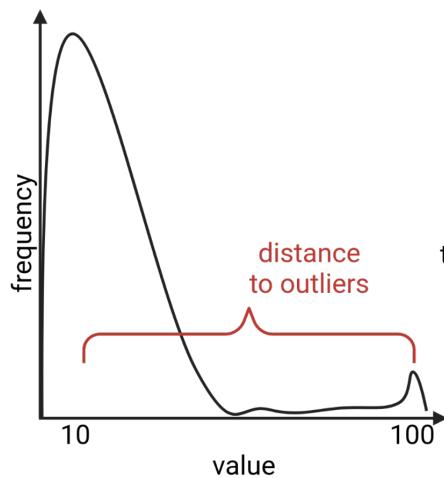
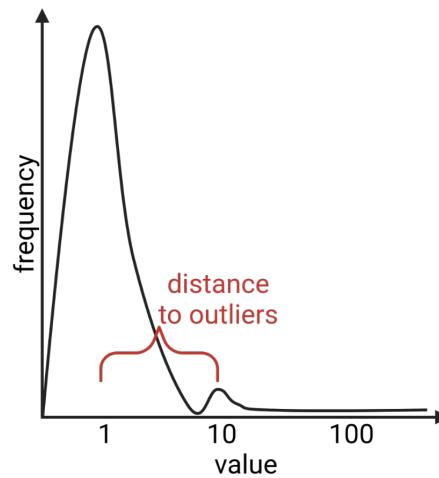
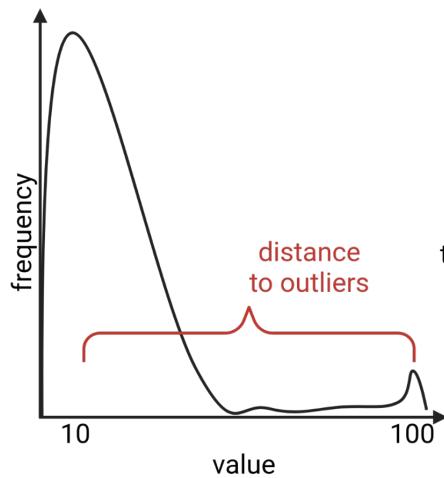
gene	SRR11603178	SRR11603179	SRR11603180	SRR11603181	SRR11603182	SRR11603183	SRR11603184	SRR11603185
TRINITY_DN10003_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10005_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
TRINITY_DN10005_c1_g1_i1	0.0	0.0	0.0	0.0	1.0	2.0	0.0	2.0
TRINITY_DN10008_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10014_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10016_c0_g1_i1	0.0	2.0	0.0	0.0	1.0	0.0	2.0	0.0
TRINITY_DN10016_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10018_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10019_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10029_c0_g1_i1	12.0	4.0	3.0	1.0	9.0	2.0	63.0	1.0
TRINITY_DN1002_c0_g1_i1_0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0
TRINITY_DN10032_c0_g1_i1	3.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TRINITY_DN10032_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
TRINITY_DN10034_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10037_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
TRINITY_DN10038_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10040_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10041_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10043_c0_g1_i1	0.0	0.0	0.0	0.0	2.0	3.0	1.0	0.0
TRINITY_DN10049_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0
TRINITY_DN10050_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10052_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10053_c0_g1_i1	2.0	0.0	0.0	1.0	0.0	0.0	4.0	0.0
TRINITY_DN10054_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TRINITY_DN10054_c0_g2_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10056_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
TRINITY_DN10057_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
TRINITY_DN10058_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10058_c0_g1_i2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TRINITY_DN10059_c0_g1_i1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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

Normalized gene expression (units of gene expression)

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

Other normalization methods



DEG identification

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

DESeq2

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

DESeq2

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

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

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.16)

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

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

Maintainer: Michael Love < michaelisaiahlove at gmail.com >

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

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

edgeR

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

edgeR

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

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

Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.16)

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

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

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

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

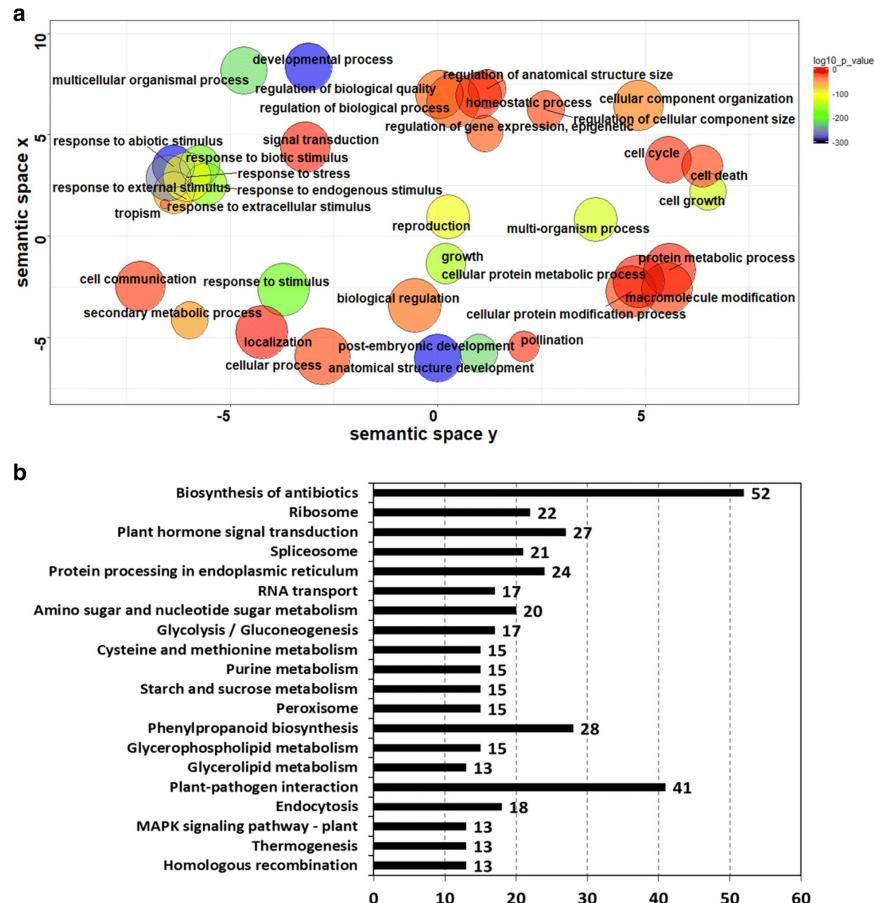
Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140. doi: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616).

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

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

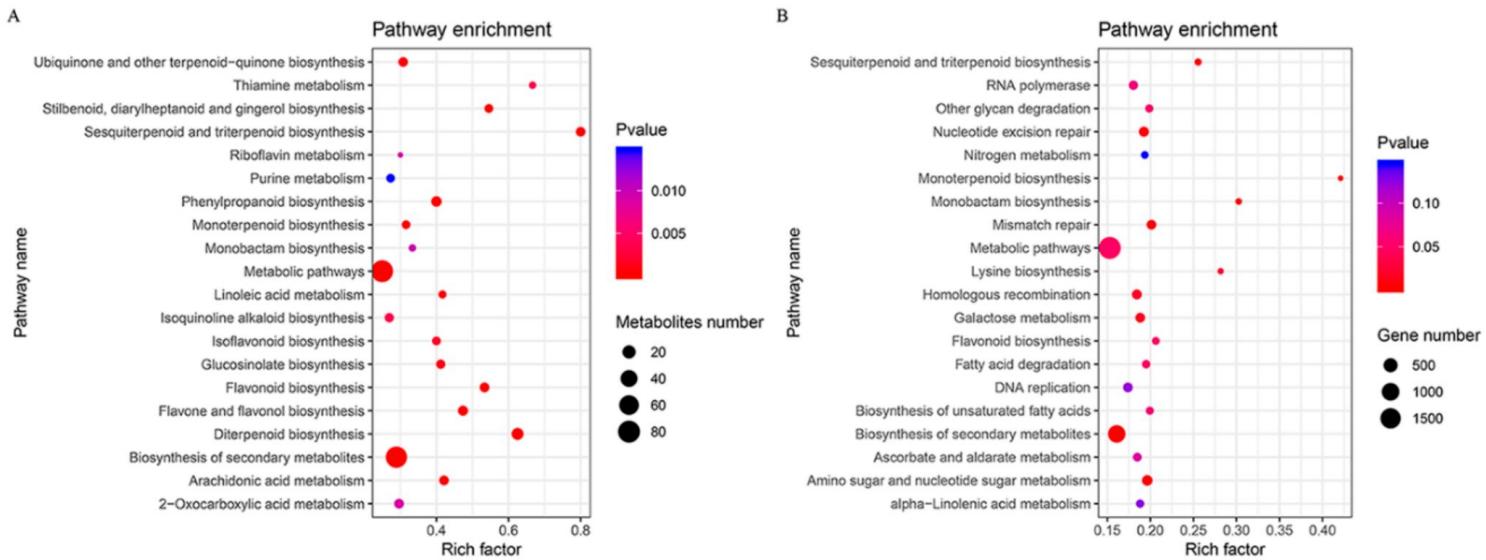
GO enrichment

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



KEGG enrichment

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



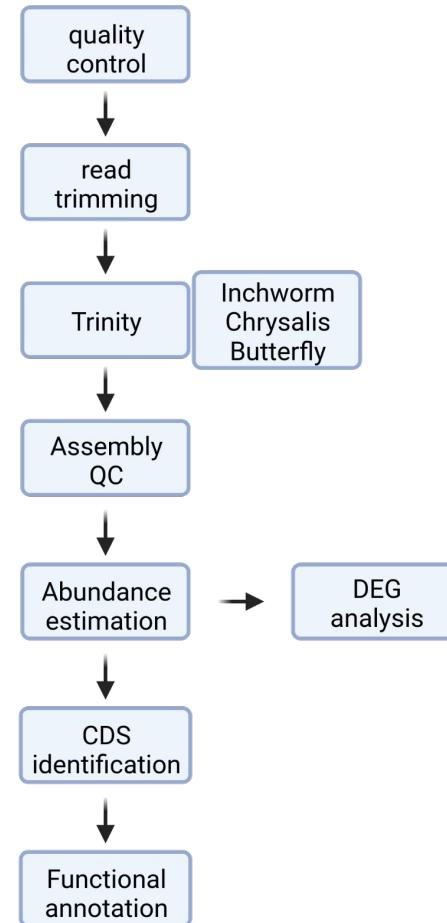
<https://doi.org/10.3390/genes11020187>

De novo transcriptome assemblers

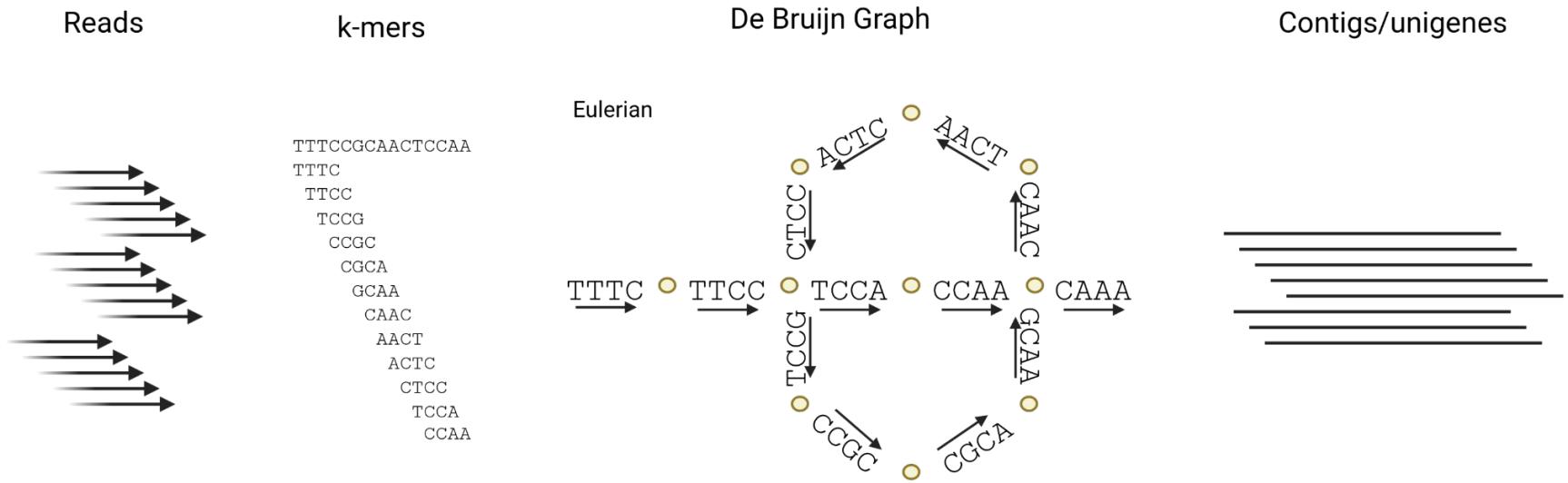
- Trinity: assembly process comprises three steps
 - Most frequently deployed transcriptome assembler
 - <https://github.com/trinityrnaseq/trinityrnaseq/wiki>
- rnaSPAdes:
 - Recent assembler with support for long reads
 - <https://cab.spbu.ru/software/rnaspades/>

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

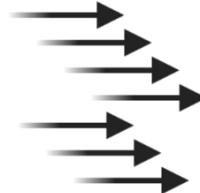


De Bruijn Graph

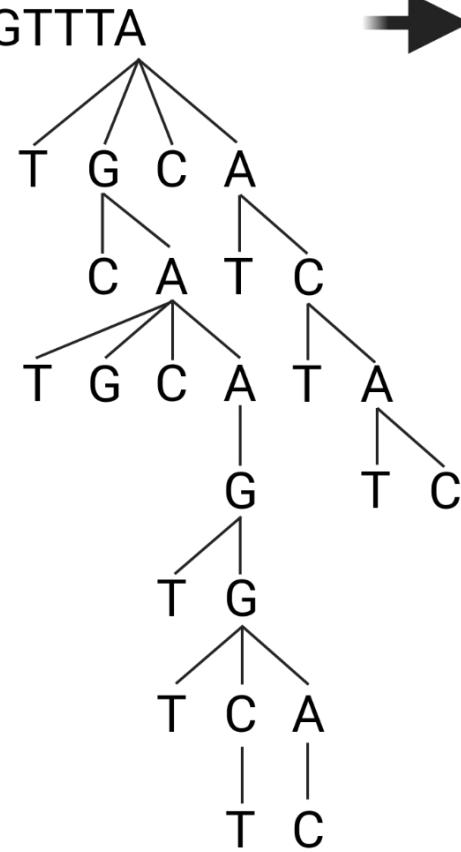


Stages of Trinity: Inchworm

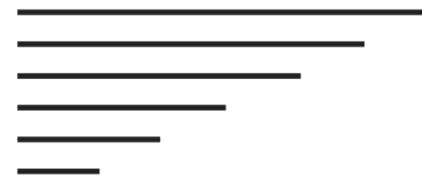
Read set



Extension in k-mer space

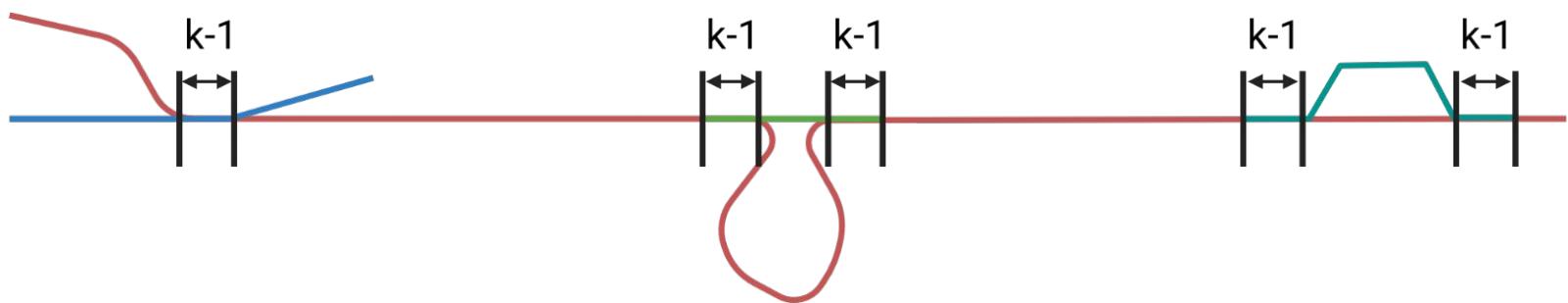


Linear sequences



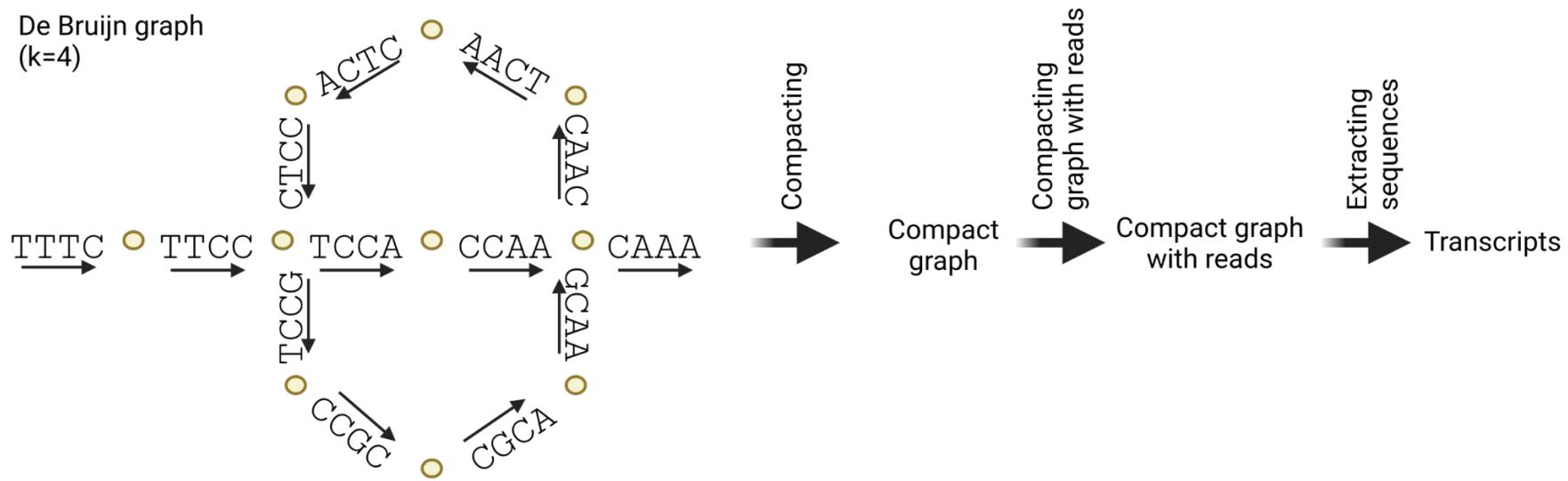
Stages of Trinity: Chrysalis

- 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

De novo transcriptome assembly

Read mapping to assembly

Cluster transcripts into genes

Calculate the counts per gene

DEG analysis

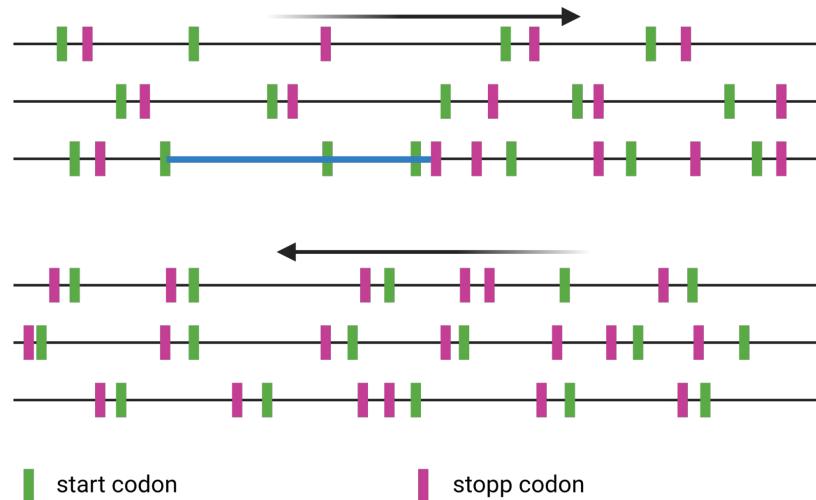
Corset

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



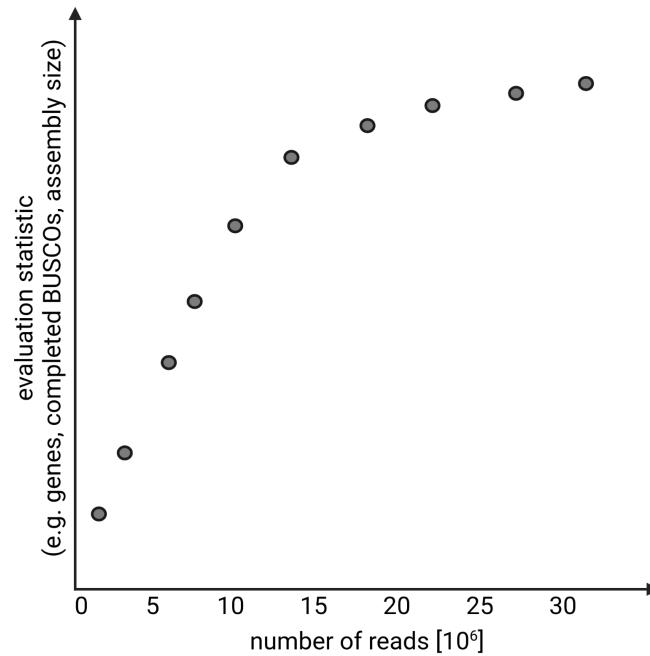
<https://github.com/bpucker/PBBtools>

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)

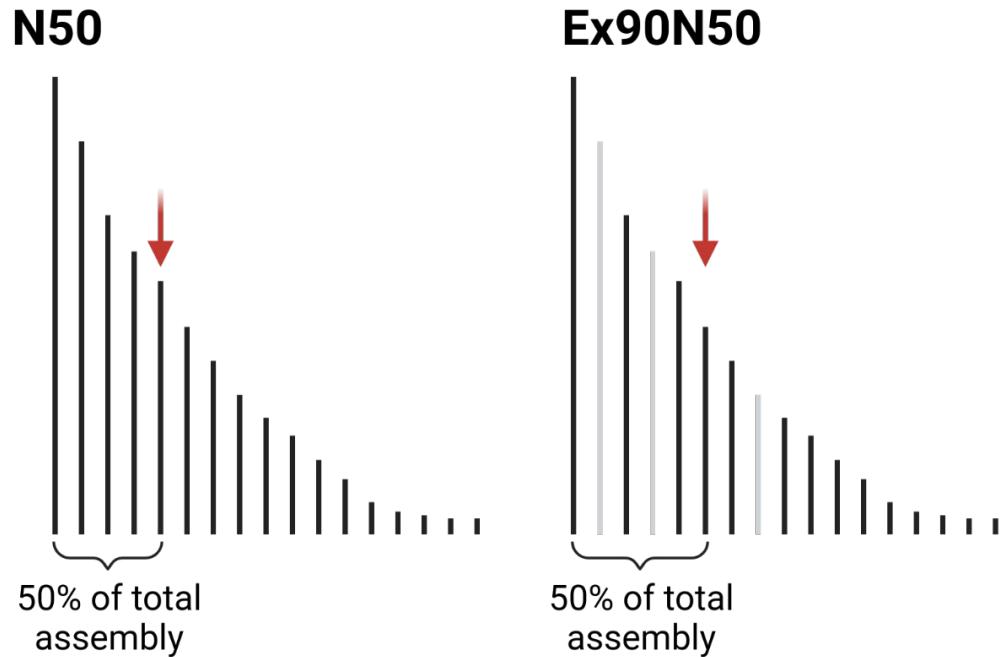
- Example: 20 million fragments; 1+2+4 hours (20 CPUs)
- Example: 200 million fragments: 1+120+7 hours (20 CPUs)
- Trinity is restartable if the process is interrupted
- Restart failed Butterfly jobs (see log file) with reduced CPU number
 - Optional settings: --bflyCPU 20 --bflyHeapSpaceMax 4G

Assembly evaluation

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

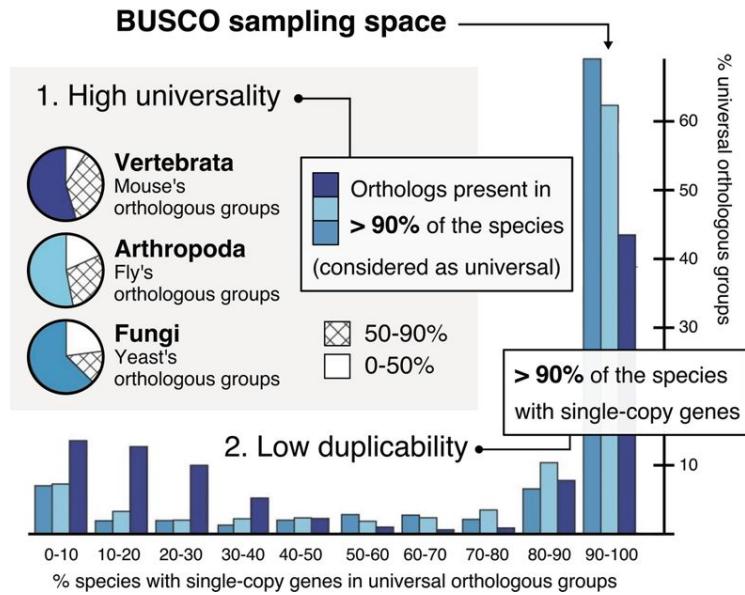
N50 & Ex90N50

- 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



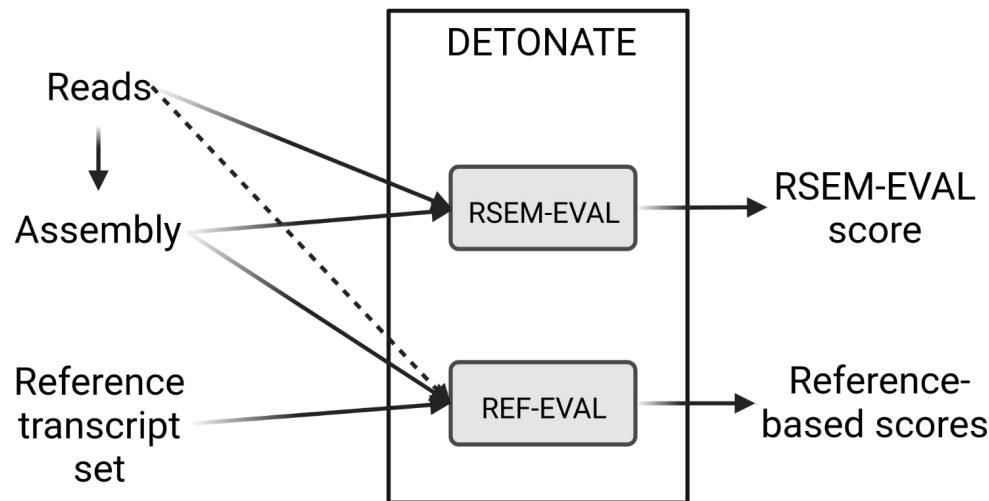
BUSCO

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



DETONATE

- **D**E novo **T**ranscript**O**me **r**Na-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

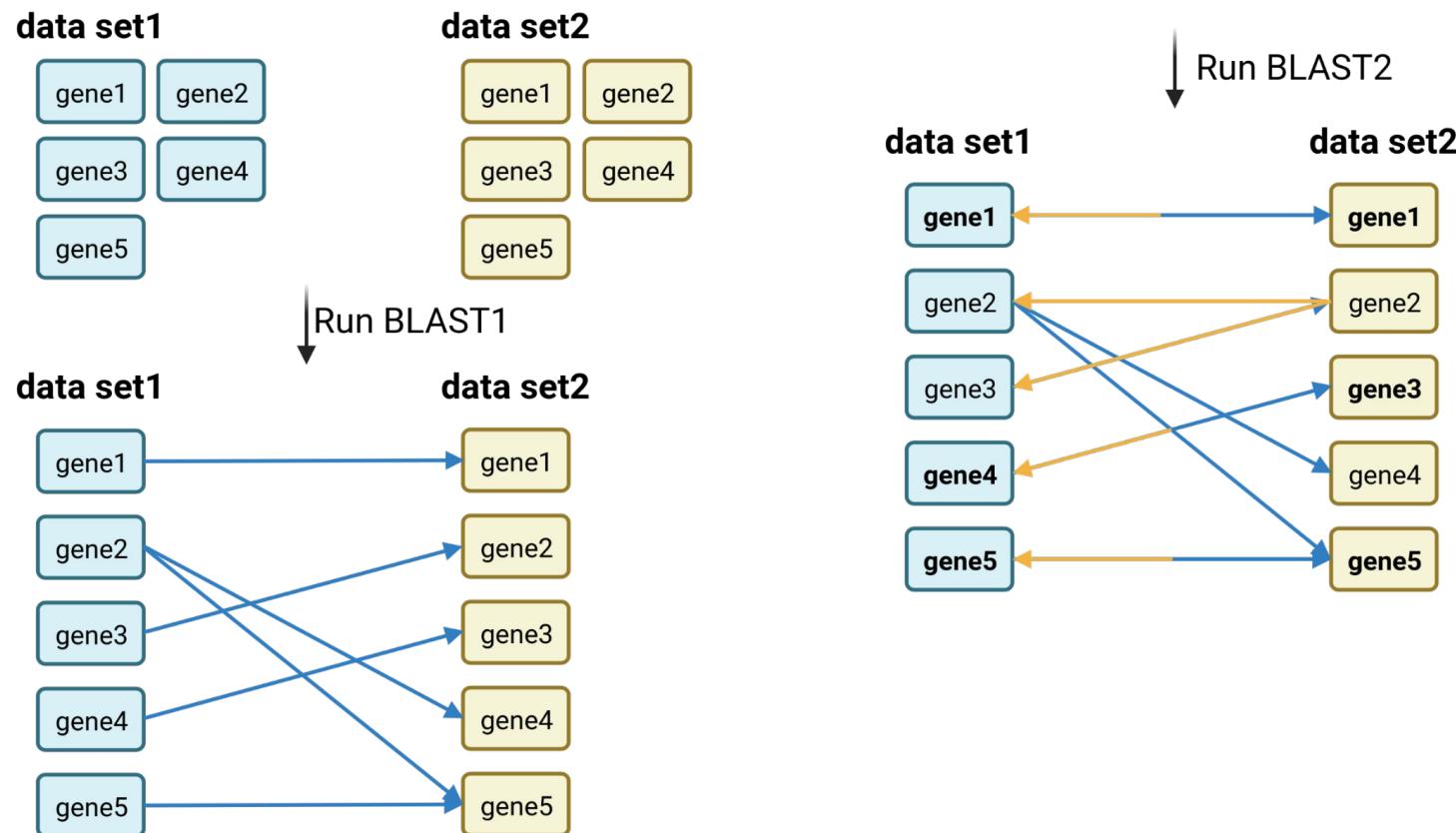


<https://github.com/deweylab/detonate>
<https://doi.org/10.1186/s13059-014-0553-5>

Functional annotation

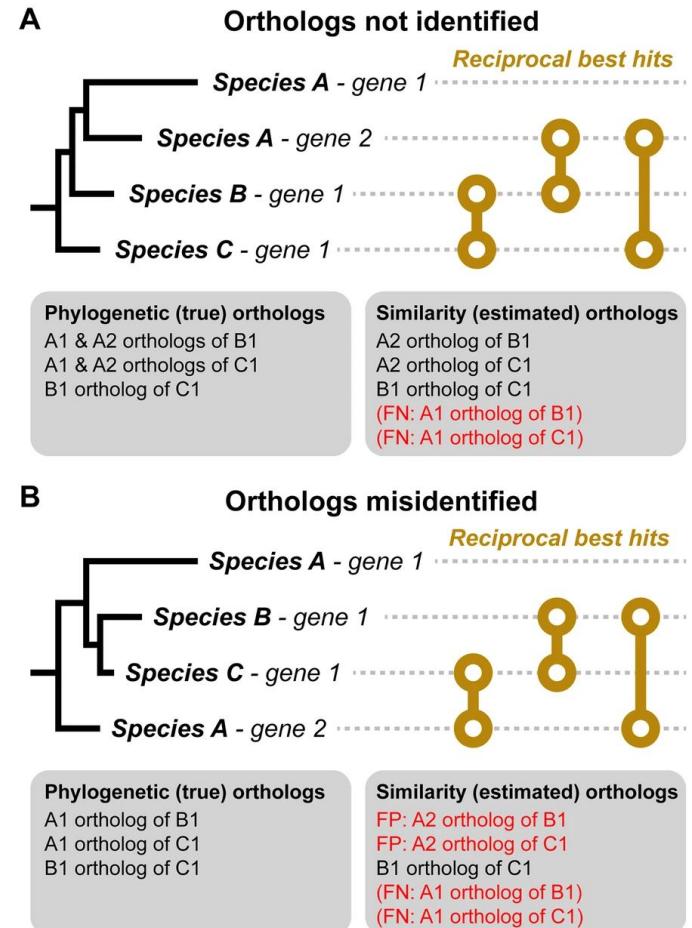
- Reciprocal Best BLAST Hits (RBHs)
- OrthoFinder2
- Knowledge-based Identification of Pathway Enzymes (KIPES)
- Mercator
- InterProScan5

Reciprocal Best BLAST Hits (RBHs)



OrthoFinder2

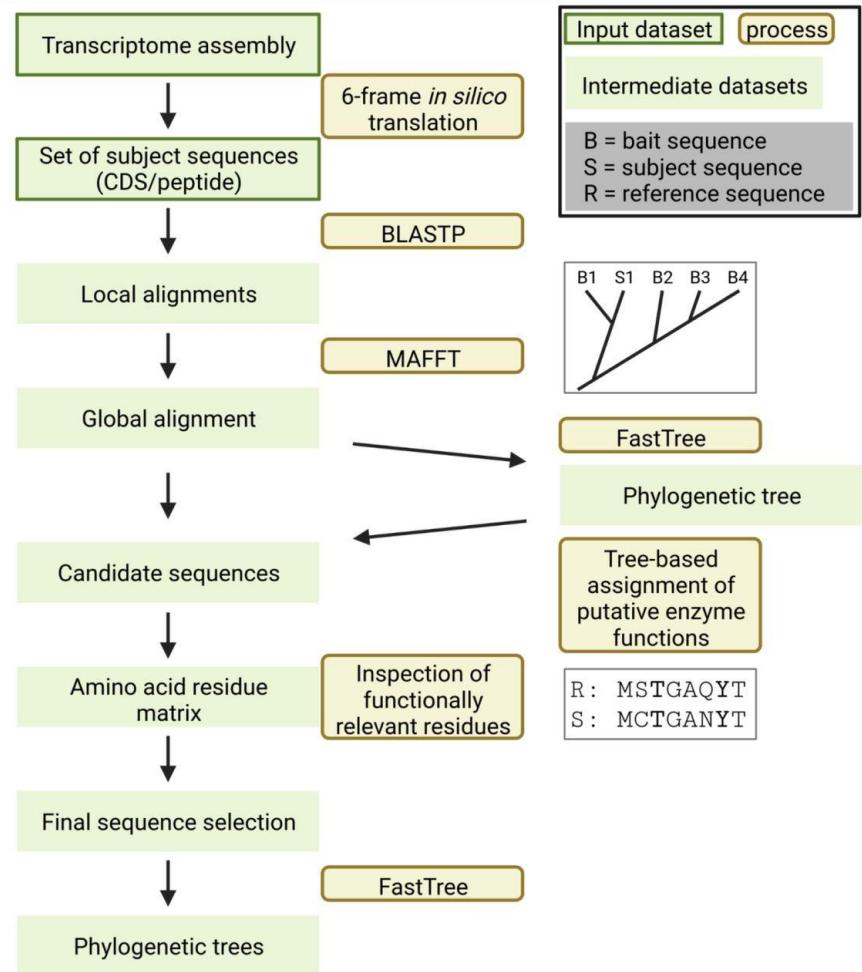
- Identification of orthologs often not possible
- Orthogroups are better reflection of reality
- Assumption: orthologs have the same function



<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



Mercator

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

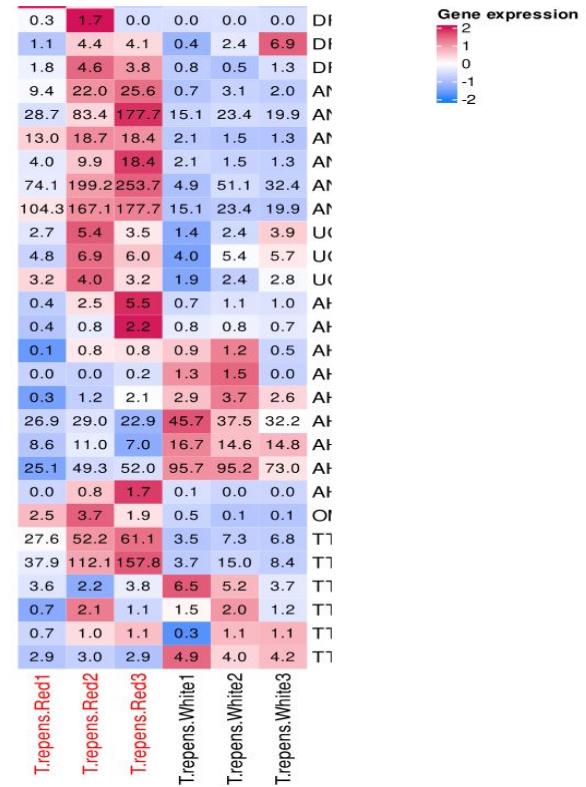


InterProScan

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

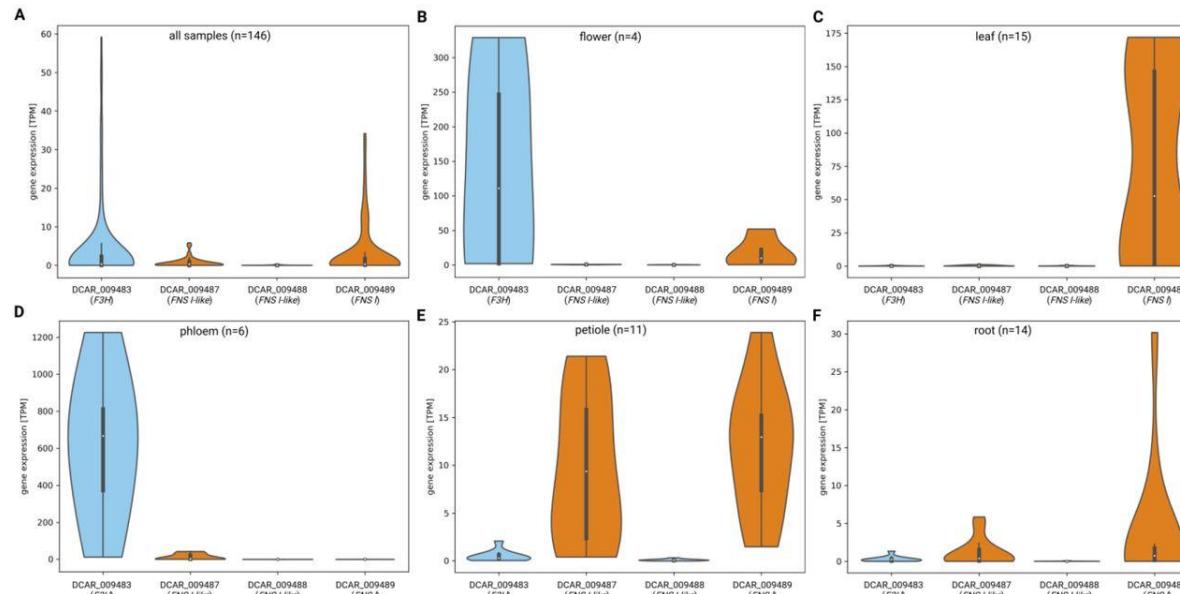
Heatmaps

- 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

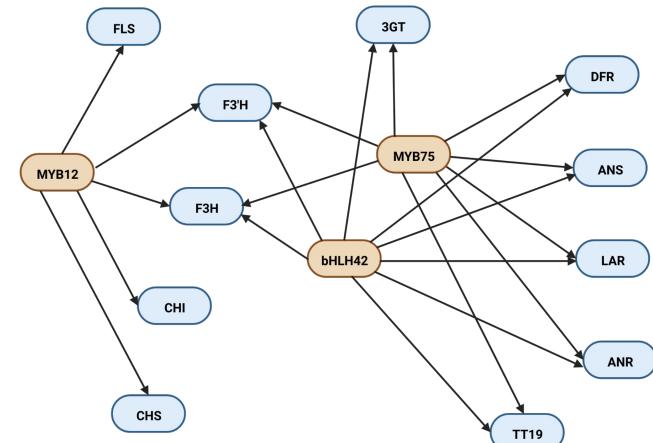


<https://doi.org/10.1371/journal.pone.0280155>



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

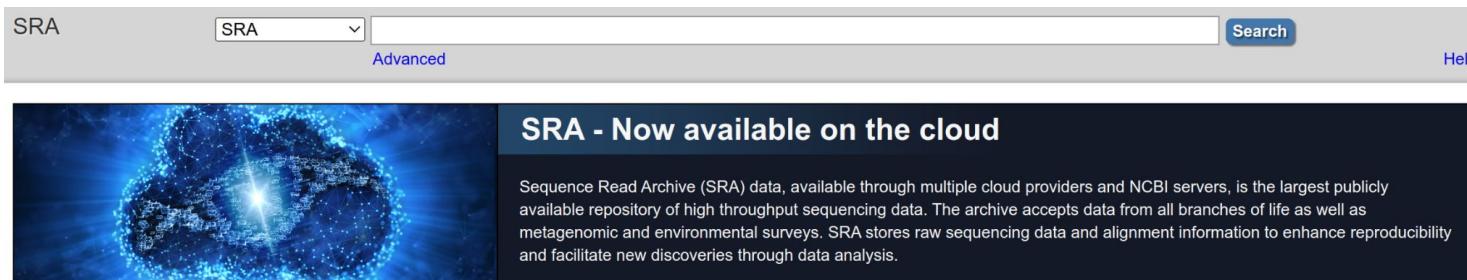
- Gene Expression Omnibus (GEO)

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.



- Sequence Read Archive (SRA)



SRA SRA Advanced Search Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

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

```
bp@bp-:~/Downloads$ md5sum SRA.png
18a5a74af965752b7b716162202f745 SRA.png
bp@bp-:~/Downloads$ █
```

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?