

Transcriptome assembly

And quality assessment. And annotation

Alexander Tkachenko, 26.10.19

Transcriptome assembly

>30 tools in listed on

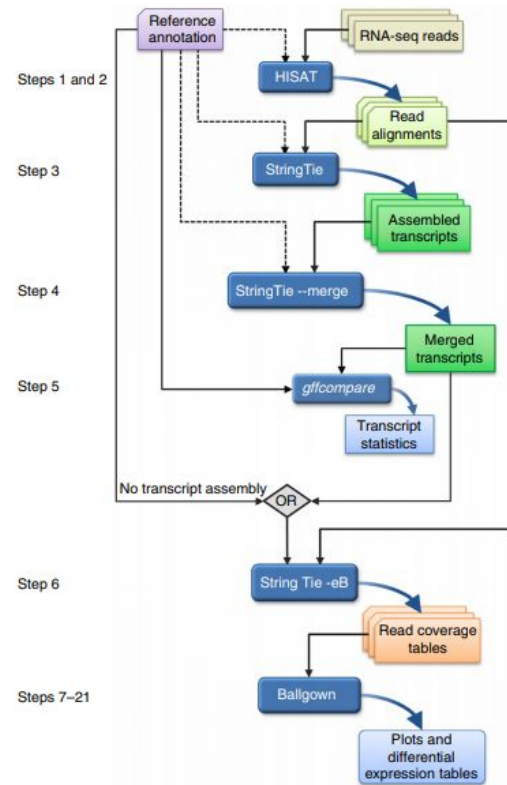
https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools#Transcriptome_assemblers

Two main approaches: mapping-first
and de novo

StringTie

Alignment-first assembler

Can also de novo assemble separate loci



Trinity

De novo and reference-based assembly
of transcriptomes

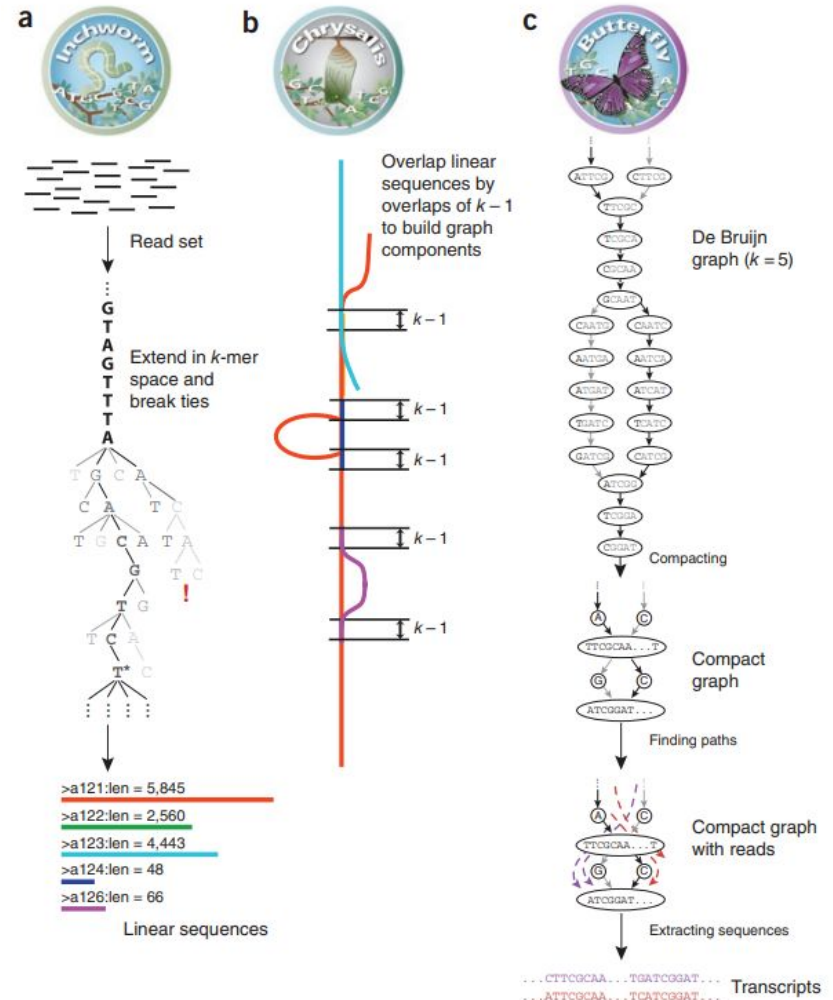


Trinity workflow

Inchworm greedily constructs contigs

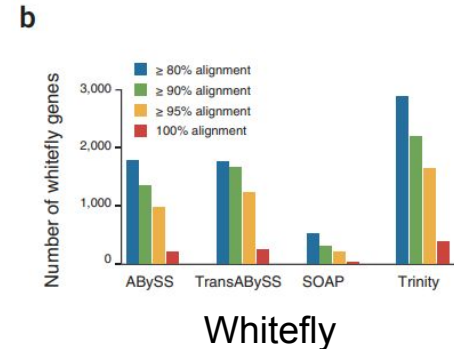
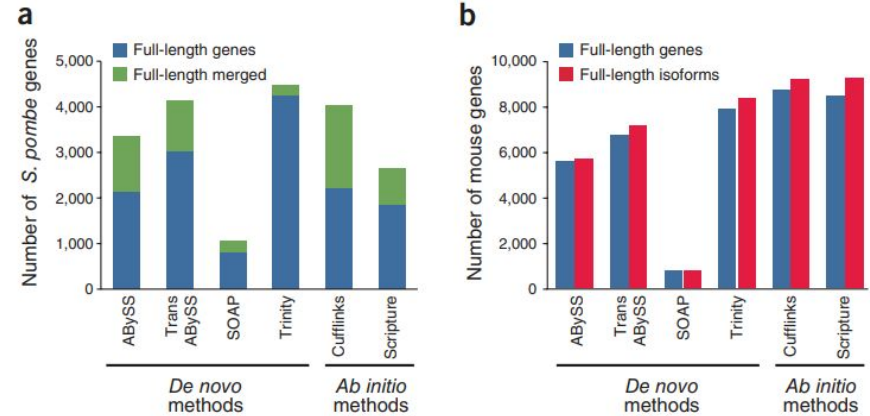
Chrysalis combines contigs

Butterfly build De Bruijn graph, compacts linear paths and reconciles paths with reads



Trinity comparison to other tools

Trinity compares better to other de novo assemblers across multiple parameter sets and tested species

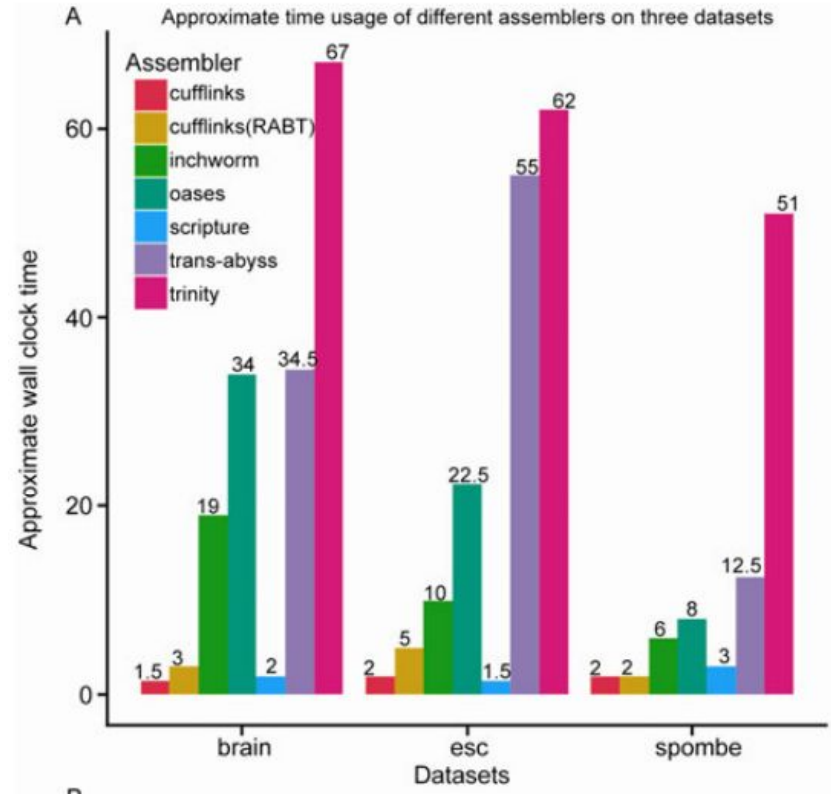


Other Trinity advantages

Regularly maintained (last update - 14.10.2019)

Contains a huge pipeline for all kinds of downstream analyses

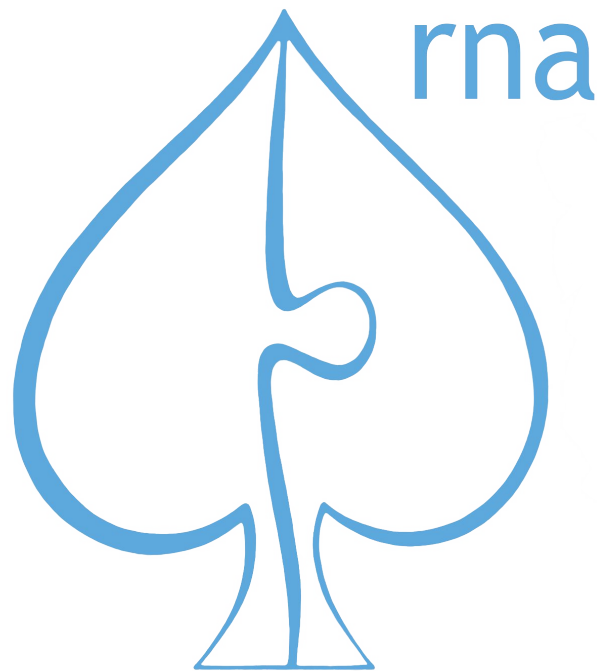
Regularly wins in comparisons with other tools



rnaSPAdes

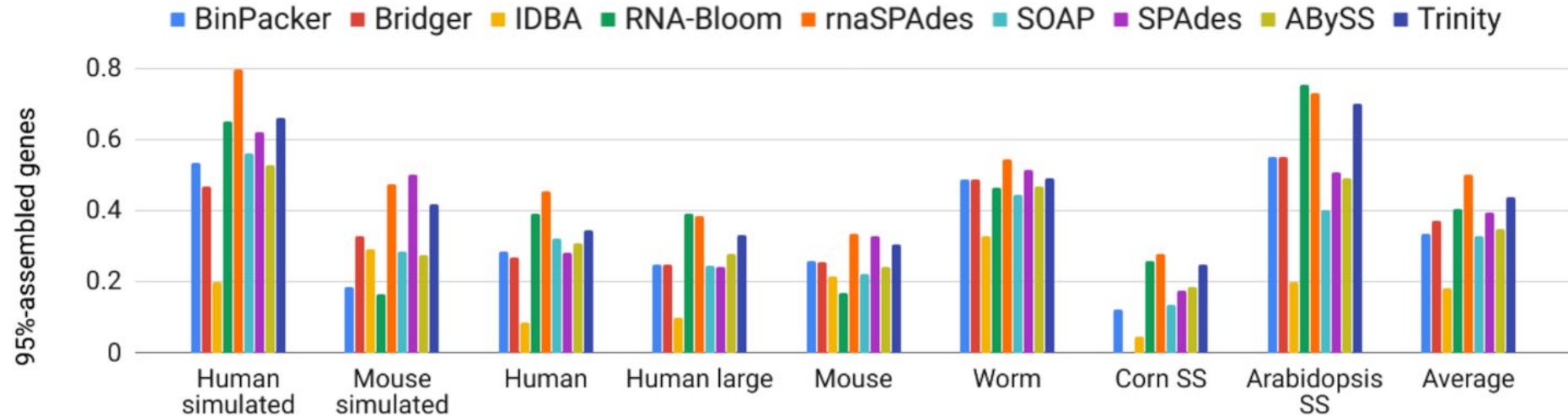
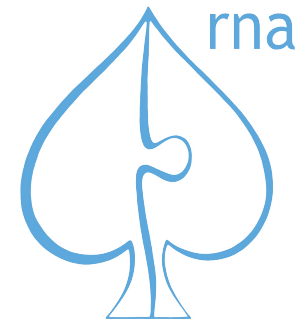
De novo assembler

<http://cab.spbu.ru/software/rnaspades/>



rnaSPAdes and other assemblers

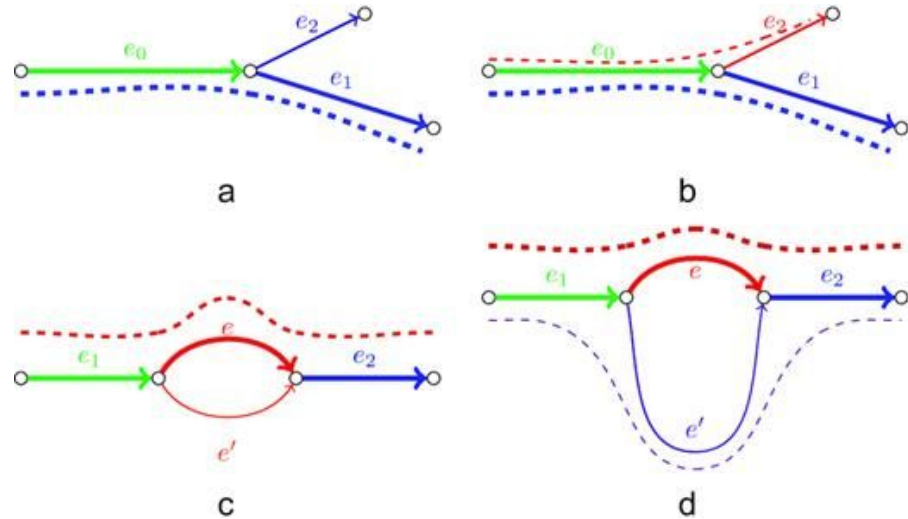
rnaSPAdes is consistently good but there is no clear winner



RNAseq data peculiarities

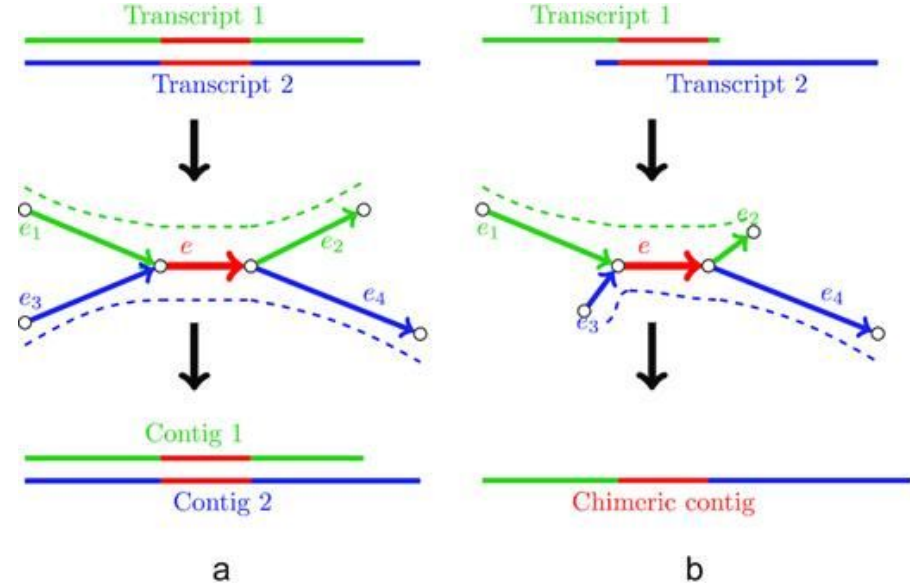
Graph tips are not necessarily errors

Graph bulges can correspond to different isoforms



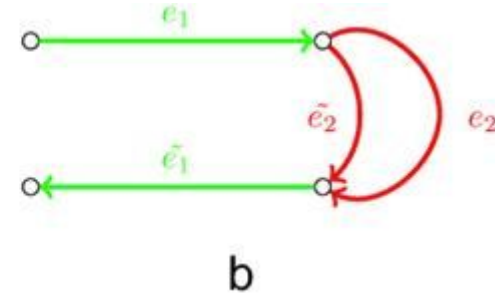
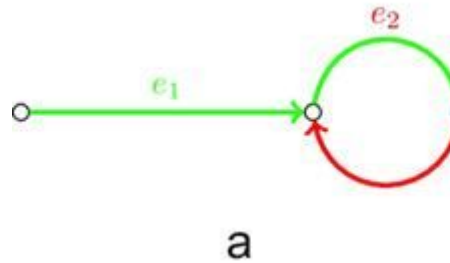
RNAseq data peculiarities

Isoforms may look like repeats



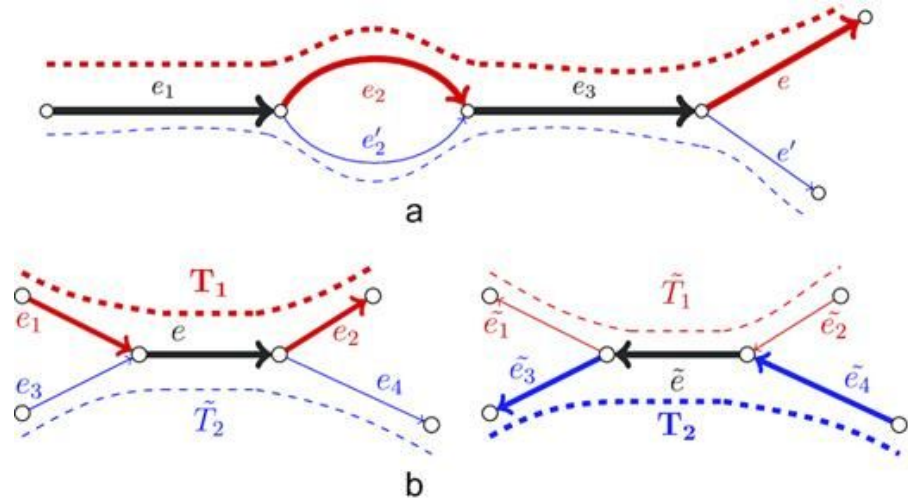
RNAseq data peculiarities

RNAseq data has specific chimeric read structures



RNAseq data peculiarities

Isoforms can be resolved with coverage and strand-specific data



Other notable tools

Velvet, Trans-ABYSS, IDBA-Tran

More benchmarks and guidelines

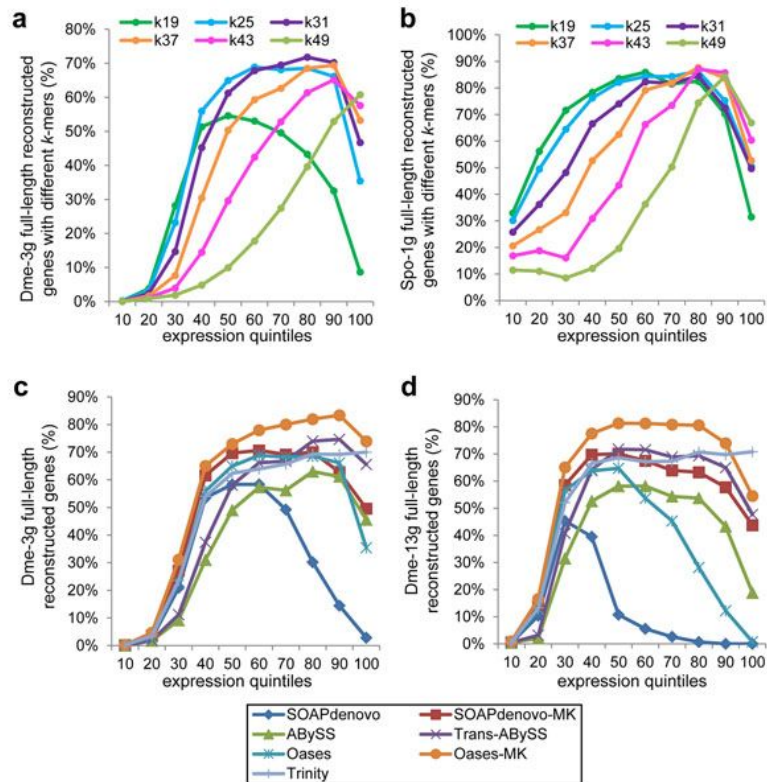
Multikmer assembly is generally better

Trinity is the best single kmer assembler

Oases-MK and Trans-ABySS produce the most diverse long transcripts

SOAP is the fastest and most memory-efficient but produces short transcripts

100x coverage is recommended for de novo

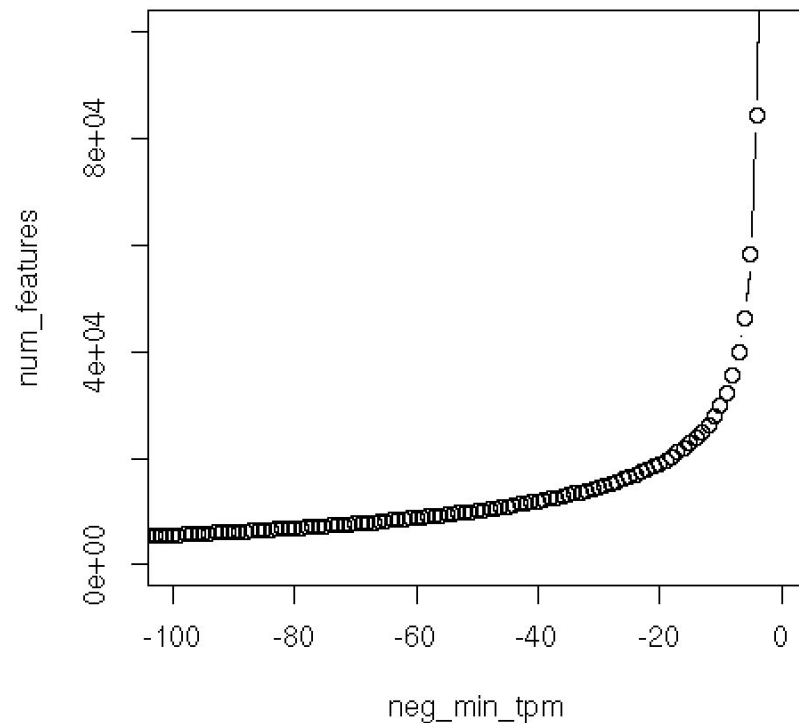


Reducing number of transcripts

TransPS

CD-HIT-EST

Coverage filtering

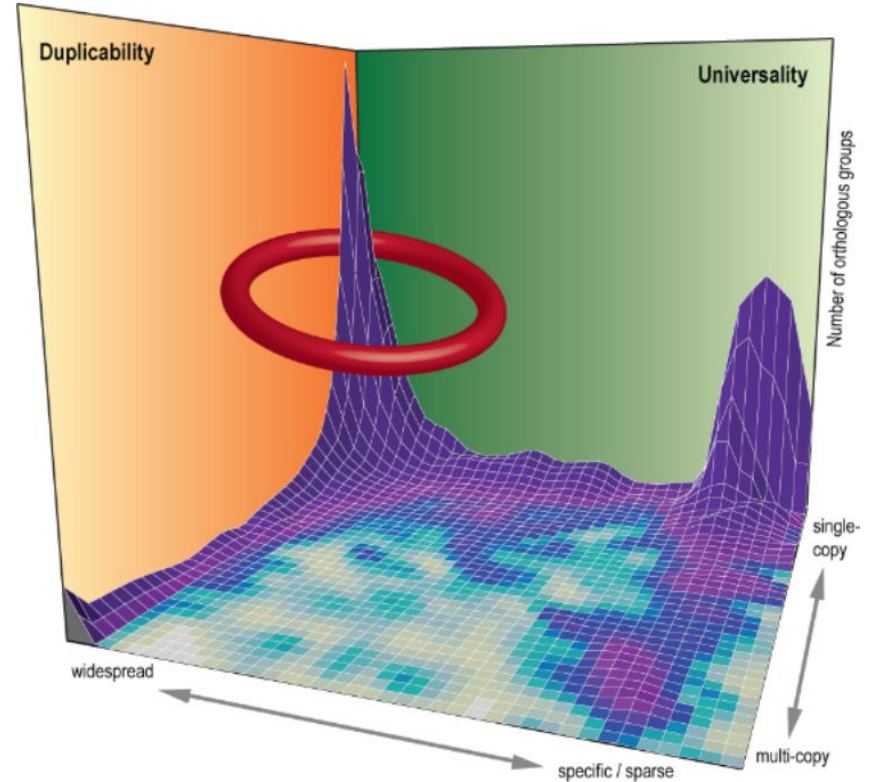


Assembly quality control

How to check which assembly is the best?

BUSCO

Benchmarking Universal Single-Copy
Orthologs

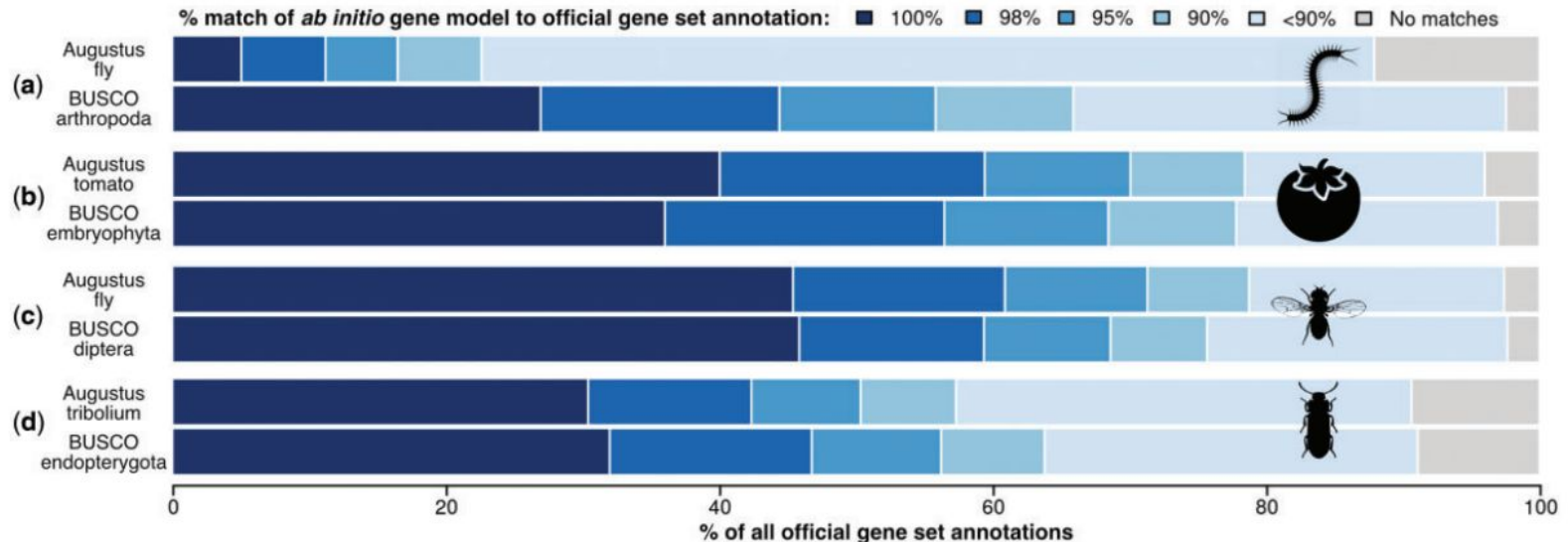


BUSCO genes can be used to construct sets of orthologous genes for phylogeny



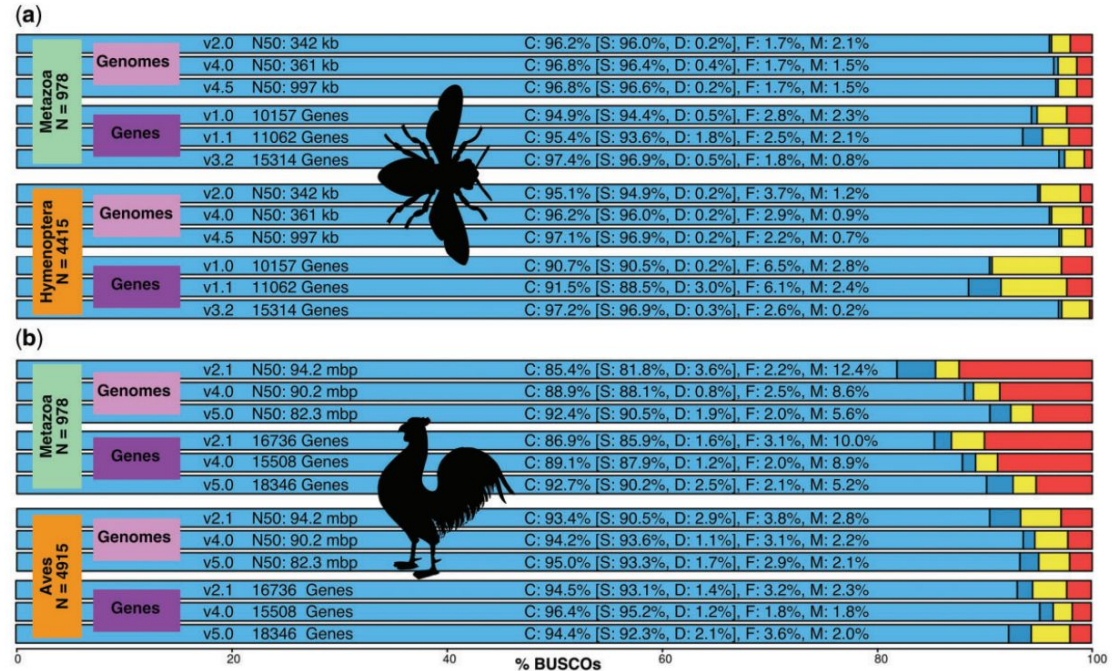
BUSCO-trained gene predictions

Gene annotations trained on BUSCO sets sometimes work better than *ab initio*



BUSCO for genomics quality control

BUSCO scores are in good concordance with increased quality of genome assemblies



BUSCO datasets

Several reference datasets based on OrthoDB

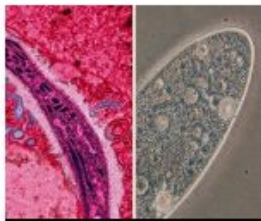
Datasets



Bacteria sets



Eukaryota sets



Protists sets



Metazoa sets



Fungi sets



Plants set

OrthoDB

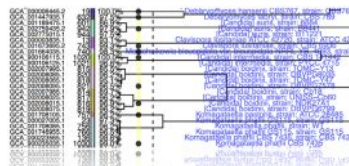
Database of orthologs
across multiple species

<https://www.orthodb.org/>

1. collect genomes



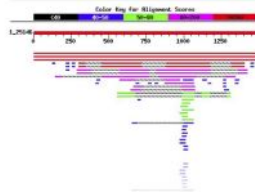
2. select representatives



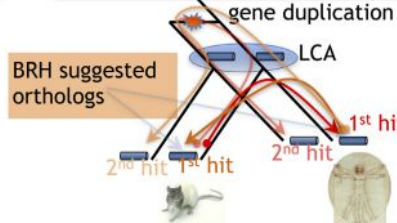
3. collate gene annotations



4. find all-to-all homologs



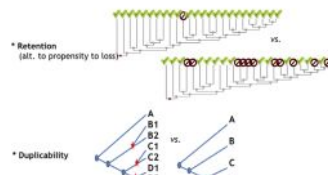
5. filter Best Reciprocal Hits



6. cluster BRHs and homologs



7. sample evolutionary traits



8. summarise OG annotation



9. make the data available

OrthoDB Data Download

Application Programming Interface - API

This is the recommended way to download data if the data set is not too large. See note on documentation and examples are found [here](#).

Flat files

OrthoDB data is also available for download from [here](#). This is recommended if the user intends to process large parts of the data or /flat (2000).

Your SPARQL query

Query Text

BUSCO workflow

For transcriptome assessment BUSCO only uses BLAST and HMMER

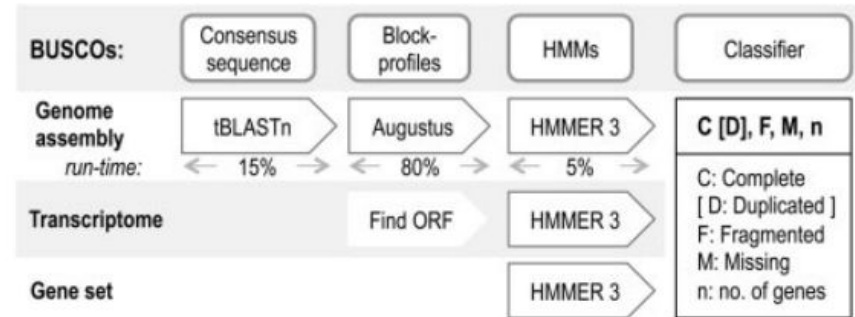
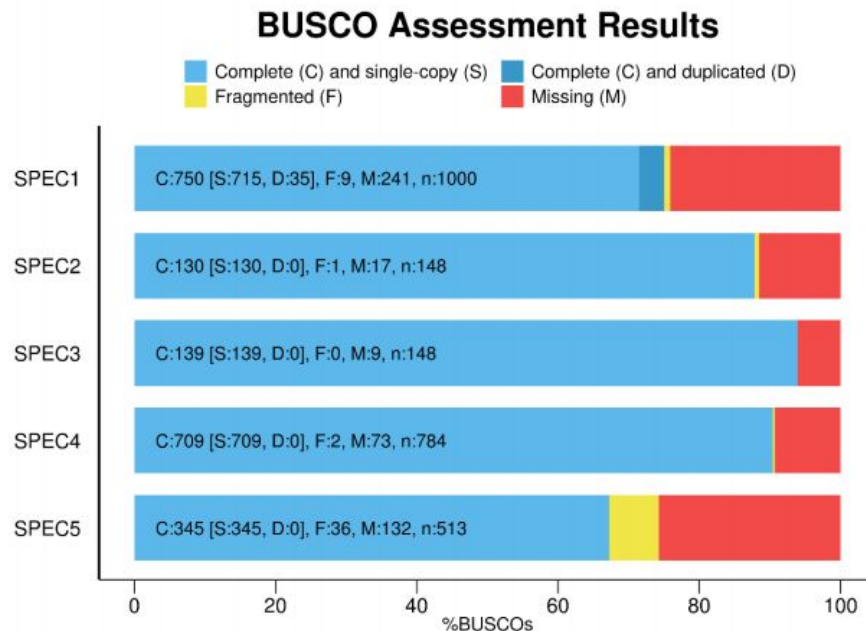


Fig. 1. BUSCO assessment workflow and relative run-times

BUSCO results example

BUSCO builds a plot with complete (single-copy or duplicated), fragmented and missing gene models



rnaQUAST

Tool for transcriptome assembly
assessment

Works for both reference-based and de
novo assemblies

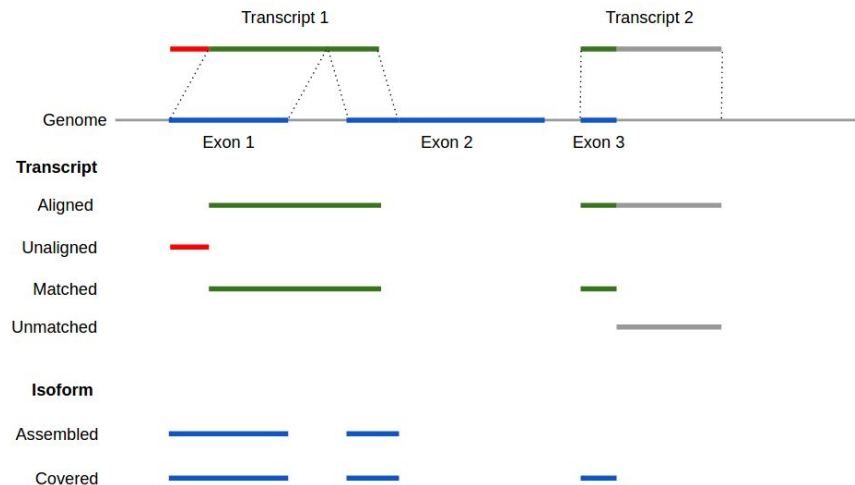
<http://cab.spbu.ru/software/rnaquast/>



rnaQUAST

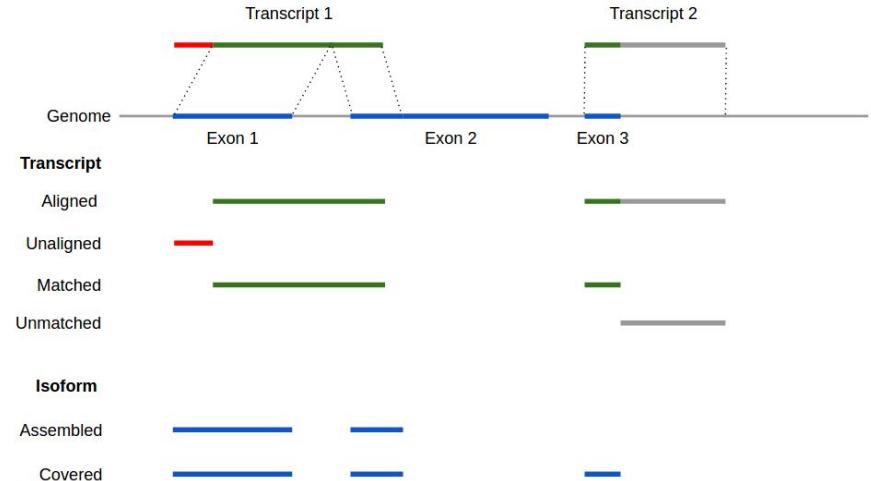
For reference-based assemblies
rnaQUAST aligns transcripts to genome
using BLAT or GMAP

It also estimates coverage with STAR
alignments



rnaQUAST

For de novo assemblies rnaQUAST
runs BUSCO and GeneMarkS-T



rnaQUAST metrics

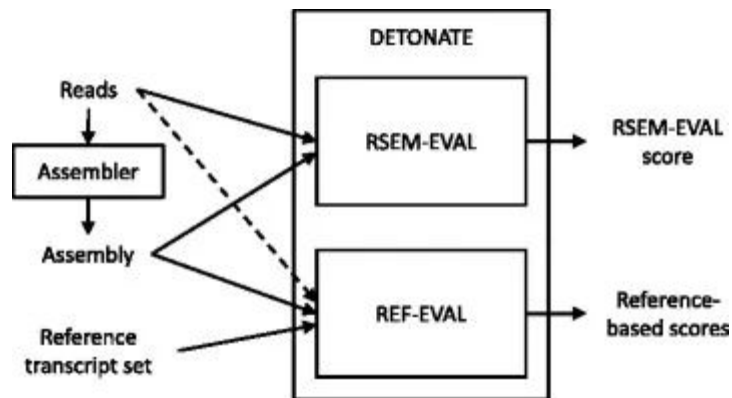
rnaQUAST reports a lot of basic and alignment metrics

Assembler	ABYSS	IDBA	SOAP	SPAdes	Trinity
<i>k</i> -mer size	32	default	31	default	default
rnaQUAST metrics					
Transcripts	107202	38294	69331	48706	51245
Transcripts \geq 500 bp	17882	17542	16021	17512	21994
Aligned	95884	38198	68591	48027	51112
Uniquely aligned	94681	37288	67878	45091	49846
Unaligned	11318	96	740	679	133
50%-matched	66744	32574	54581	37447	43039
95%-matched	61633	29429	50876	32565	35239
Unannotated	26678	3905	12252	7102	5740
Database coverage	18.5	16.9	17.2	17.6	18.1
50%-assembled isoforms	7061	6777	6241	6887	7020
95%-assembled isoforms	1907	1611	1397	2292	2053
99%-assembled isoforms	432	431	347	754	710
Misassemblies	267	471	26	942	465
Mismatches per transcript	0.50	1.04	0.58	1.13	1.28
REF-EVAL scores					
Nucleotide precision	0.69	0.86	0.84	0.81	0.69
Nucleotide recall	0.76	0.75	0.75	0.79	0.78
Nucleotide F_1	0.73	0.80	0.79	0.80	0.73
Contig precision	0.095	0.17	0.14	0.14	0.14
Contig recall	0.096	0.063	0.089	0.066	0.068
Contig F_1	0.095	0.092	0.11	0.090	0.092
<i>k</i> -mer recall	0.84	0.34	0.76	0.67	0.90
KC score	0.80	0.31	0.73	0.64	0.86
RSEM-EVAL score ($\times 10^9$)	-1.42	-2.31	-1.40	-1.48	-0.98

DETONATE

DE novo Transcriptome rNa-seq
Assembly with or without the Truth
Evaluation

<http://deweylab.biostat.wisc.edu/detonate/vignette.html>



RSEM-EVAL

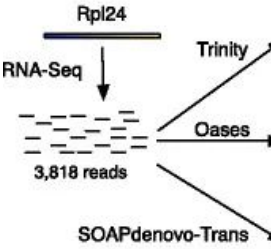
Used for assessment of de novo assemblies



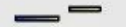
Score is probability of assembly A given reads D

$$\text{score}_{\text{RSEM-EVAL}}(A) = \log P(A, D)$$

RSEM-EVAL for Rpl24 transcript

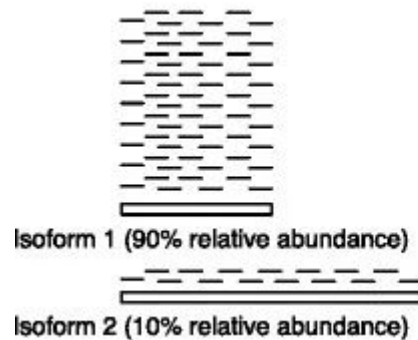
Trinity demonstrates the best result



Score Assembly	Likelihood score	Prior score	BIC score	RSEM-EVAL score
	-86542	-876	-8	-87426
	-198666	-2771	-29	-201466
	-254310	-613	-12	-254935

RSEM-EVAL for multi-isoform genes

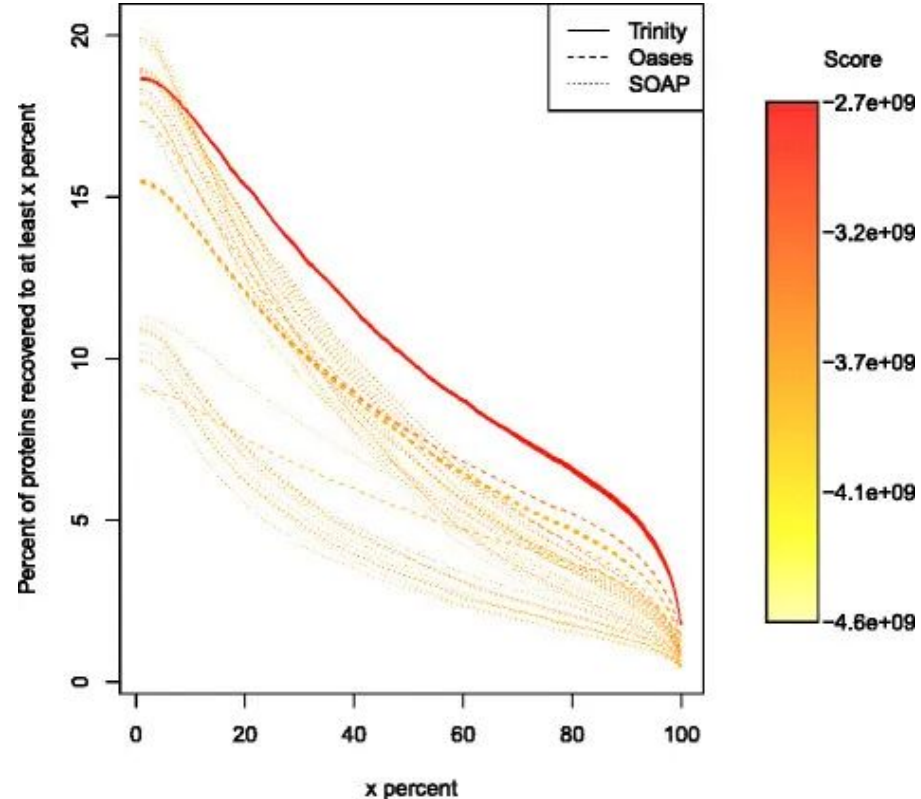
RSEM-EVAL is better at picking up truth than metagenome assessment tools



Score Assembly Truth	RSEM-EVAL	GENOVO	ALE
	-43720	-19557	-116316
Long only	-44403	-18199	-88905
Short only	-104963	-68997	-52090

RSEM-EVAL for *Xenopus* transcriptome

Trinity shows the best results






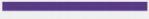




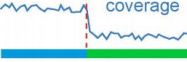
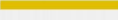
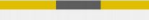










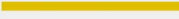


Transrate

Tool for de novo transcriptome
assembly control

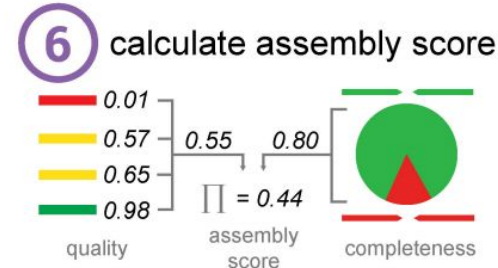
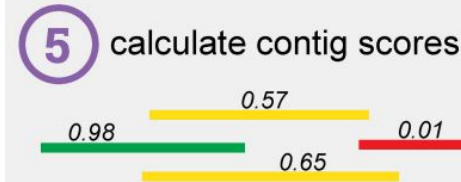
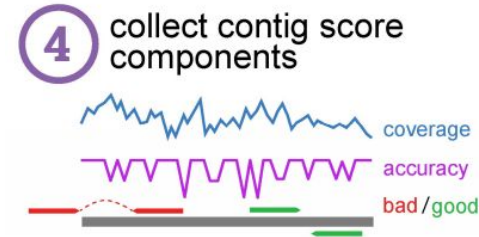


Transrate

Several ways how contigs can be wrong and how mistakes can be detected

Error type	Transcripts	Assembly	Read evidence
Family collapse	geneAA  geneAB  geneAC  n=3	 n=1	
Chimerism	 geneC  geneB n=2	 n=1	
Unsupported insertion	 n=1	 n=1	
Incompleteness	 n=1	 n=1	
Fragmentation	 n=1	 n=4	
Local misassembly	 n=1	 n=1	
Redundancy	 n=1	 n=3	

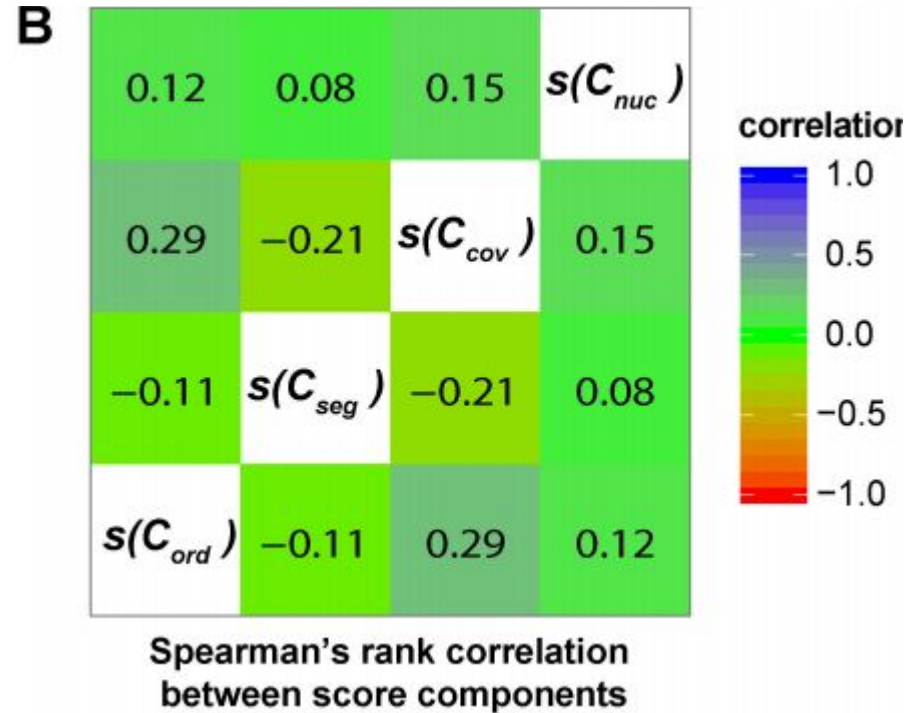
Transrate workflow



Transrate metrics

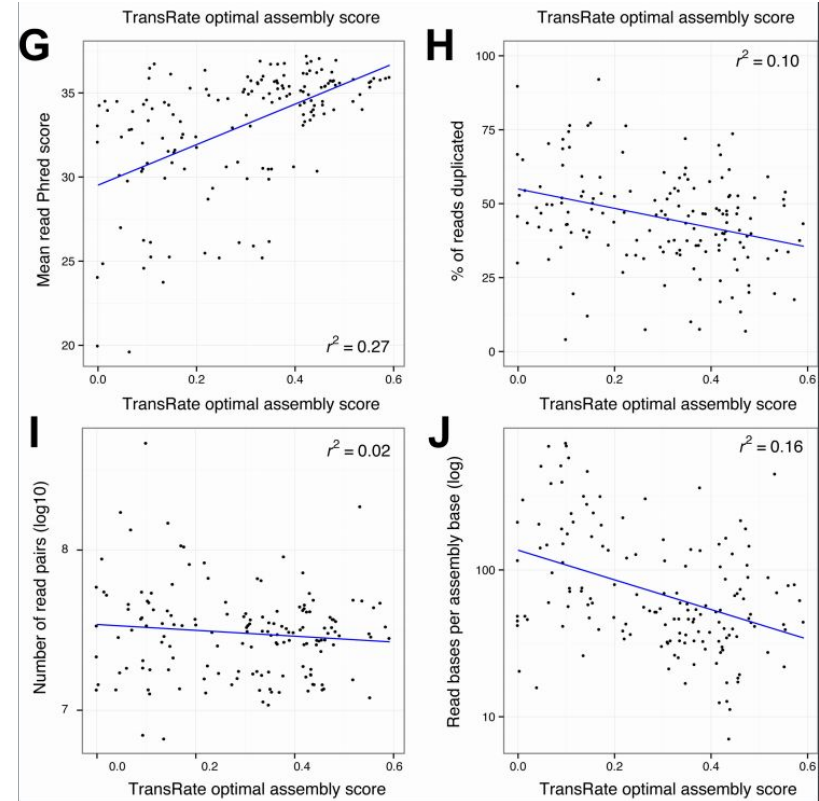
Score component	Description
$s(C_{nuc})$	The proportion of nucleotides in the mapped reads that are the same as those in the assembled contig
$s(C_{cov})$	The proportion of nucleotides in the contig that have have no supporting read data
$s(C_{ord})$	The extent to which the order of the bases in contig are correct by analyzing the pairing information in the mapped reads
$s(C_{seg})$	The probability that the coverage depth of the transcript is univariate

Transrate metrics correlations



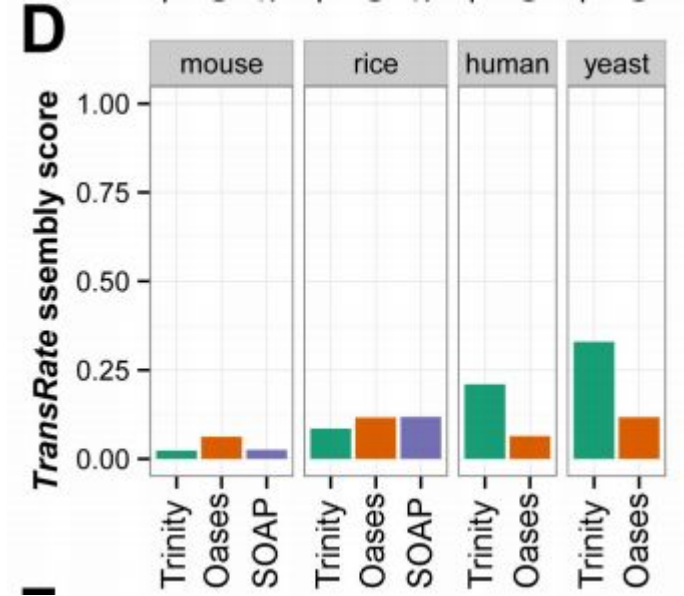
Transrate and data parameters

Assembly scores do not depend on number of read pairs and duplication of reads, but depend on quality of reads



Assemblers comparison

Assemblers' rankings depend on the organism



Assembly and QC take-home points

Selecting assembler and parameters is a non-trivial matter

Best choice is to try everything available and compare results (possibly with downsampling to use resources more efficiently)

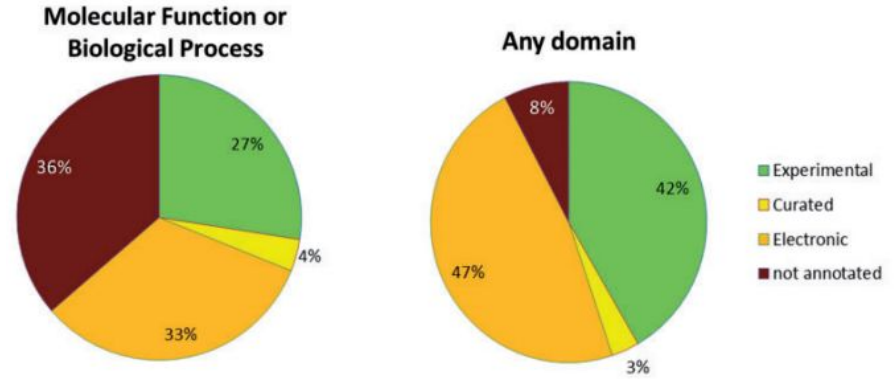
Annotation

“Annotation”

BLAST results against a relatively close annotated genome

Annotation “dark matter”

A lot of genes can not be assigned to any functional class



Protein prediction

Some pipelines require protein sequences for annotation

ORFs in transcripts can be predicted with several tools

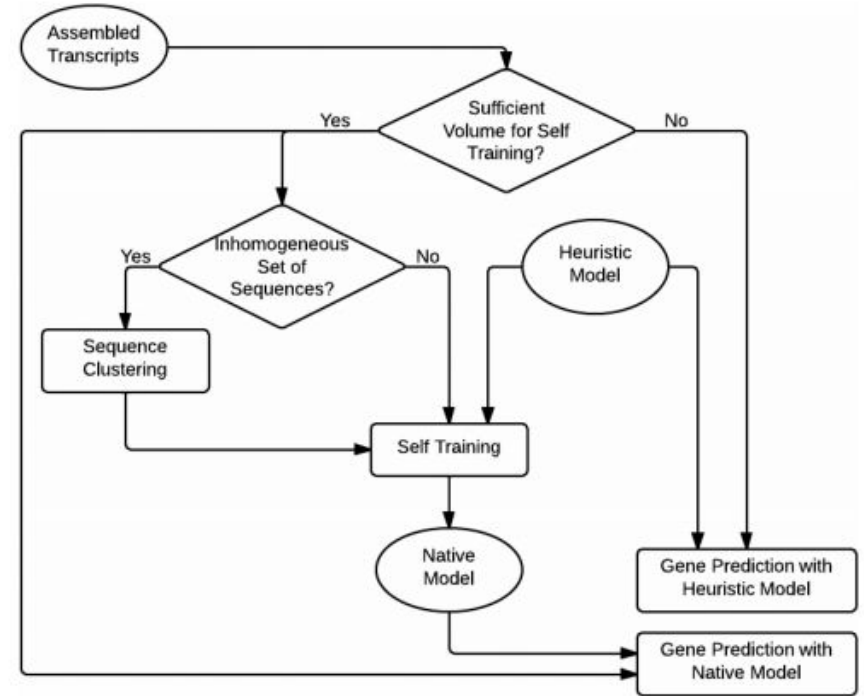


TransDecoder

Identifies coding regions within transcripts

GeneMarkS-T

Ab initio finding of proteins in eukaryotic transcripts



Protein families annotations

Resource	Version	Families	Web address	Comments
PFAM	30.0	16 306	http://pfam.xfam.org/	
TIGRFAM	15.0	4488	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	
PANTHER	11.0	13 096	http://pantherdb.org	
SMART	7.1	1312	http://smart.embl-heidelberg.de/	License necessary
EggNOG	4.5	190 648 (37 127 plants)	http://eggnogdb.embl.de/#/app/home	
INTERPROSCAN	58.0	>40 000 integrated entries	https://www.ebi.ac.uk/interpro/search/sequence-search	Meta engine including all other resources except EggNOG but not necessarily the most recent version at all times
CDD	3.15	52 411 (11 474 from CDD curation)	http://www.ncbi.nlm.nih.gov/cdd/	Uses RPS-BLAST and includes partly older versions of PFAM, SMART and TIGRFAM

Specialized protein groups annotations

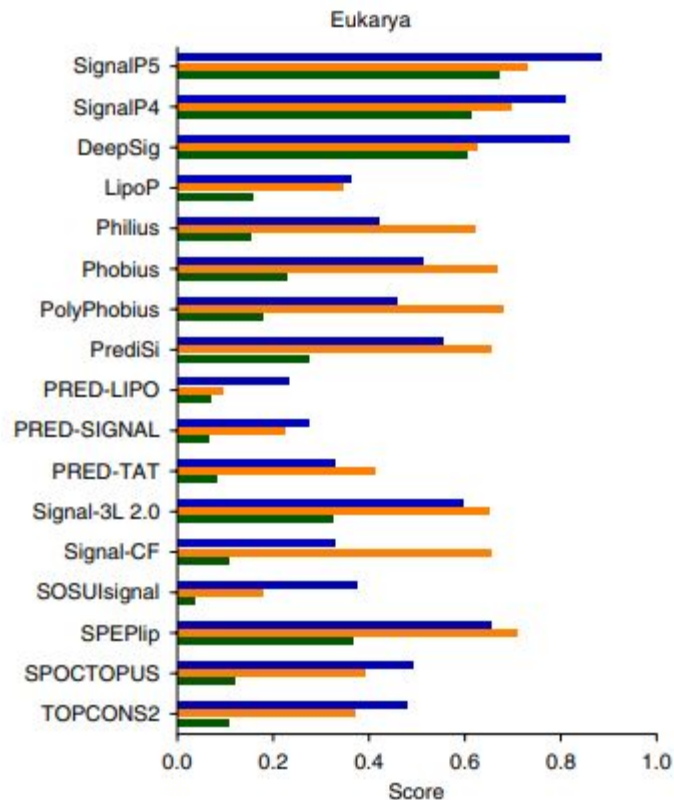
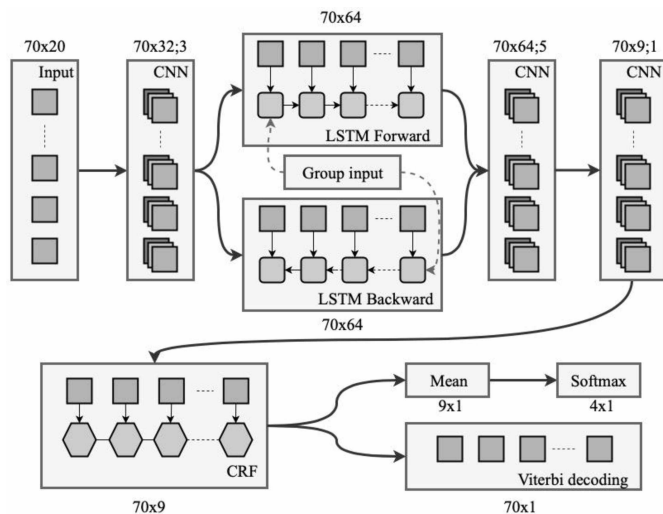
Table 4. Tools and Web sites useful in annotating large protein families

Resource	Function	Web address
CoGe	Compares genomes, find synteny	https://genomevolution.org
PlantTFDB	Plant Transcription Factor families	http://planttfdb.cbi.pku.edu.cn/
Potsdam plntfdb	Plant Transcription Factor families	http://plntfdb.bio.uni-potsdam.de/v3.0/
P450 Database	P450 protein families	http://drnelson.uthsc.edu/CytochromeP450.html
CAZy	Enzymes acting on carbohydrates	http://www.cazy.org/
Aramemnon ^a	Plant membrane proteins	http://aramemnon.uni-koeln.de/
Merops Database	Peptidases	http://merops.sanger.ac.uk
PLAZA	Generalist Plant Family database	http://bioinformatics.psb.ugent.be/plaza
GreenPhylDB	Generalist Plant Family database	www.greenphyl.org/

Note. ^aAlso lists a comprehensive set of tools for transmembrane domains, subcellular localization and lipid modifications.

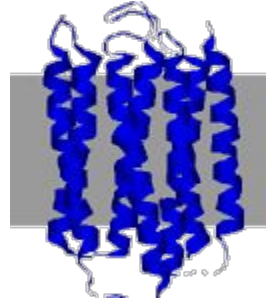
SignalP

Predicts signal peptides using neural networks



TMHMM

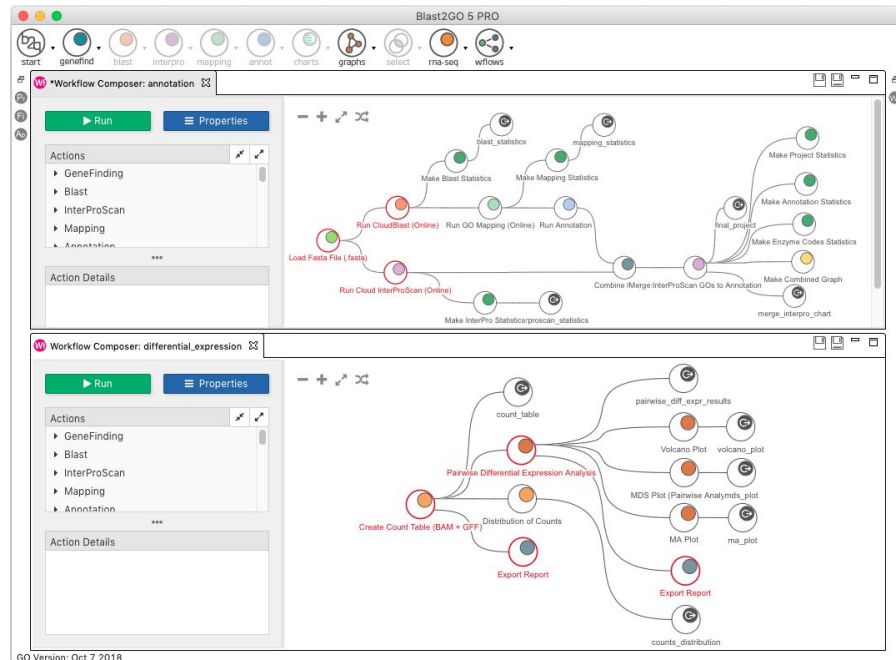
Predicts transmembrane domains



BLAST2GO

Readily available annotation pipeline

Commercial



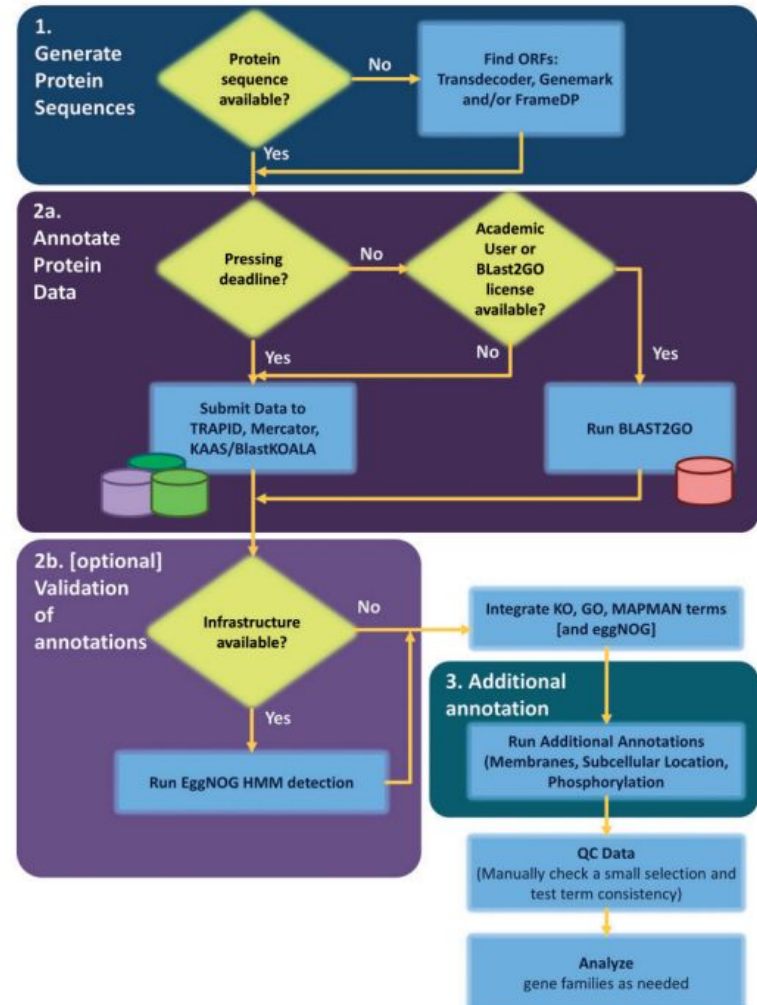
Trinotate

Annotation pipeline accompanying
Trinity

Protein prediction with TransDecoder,
annotation with Uniprot, Pfam, eggNOG



Annotation flowchart



Annotation take-home points

Annotation relies on homology and finding protein domains and protein families annotations

Annotations can be supplemented with finding additional features like signal proteins, transmembrane domains and phosphorylation sites

Annotation can be automated

Annotation should quality checked

Questions, suggestions?