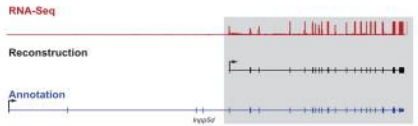
Transcriptome assembly, quality assessment and annotation

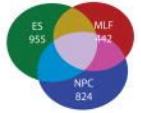
14.11.20

Why assemble anything?

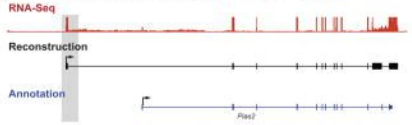
Existing annotations are not perfect

a) Internal Alternative 5' Start Sites

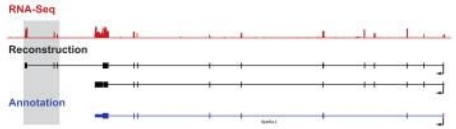


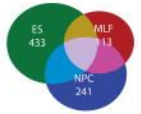


b) External Alternative 5' Start Sites

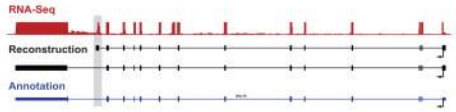


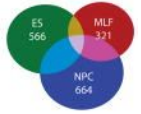
c) Alternative Downstream 3' End



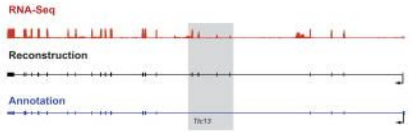


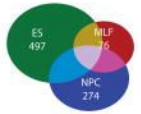
d) Alternative upstream 3' End





e) Novel Coding Exons



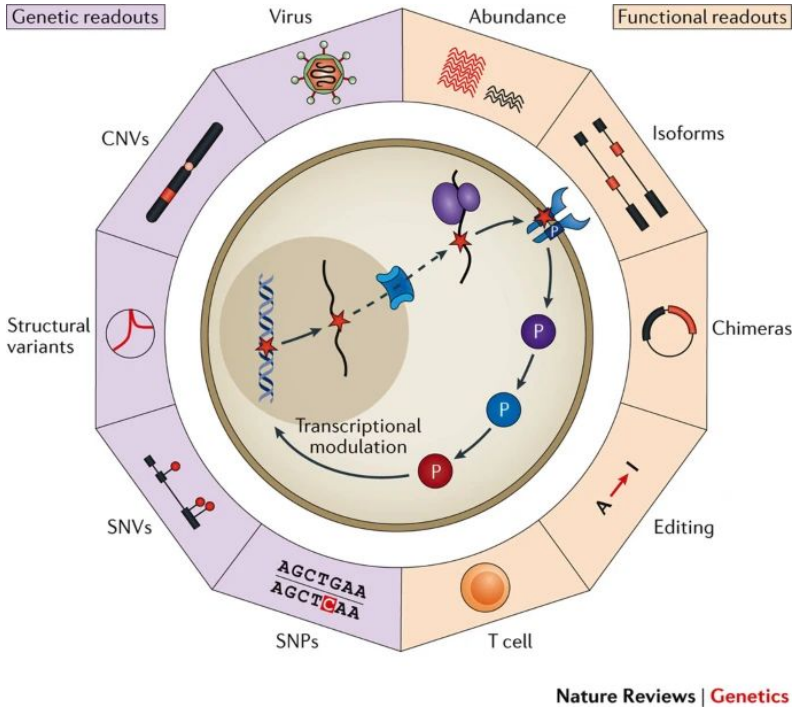


Cancer transcriptomes

Cancer transcriptomes are fundamentally different from normal ones, even if the normal reference annotation and assembly are really good

Differences account for both genetic and functional changes

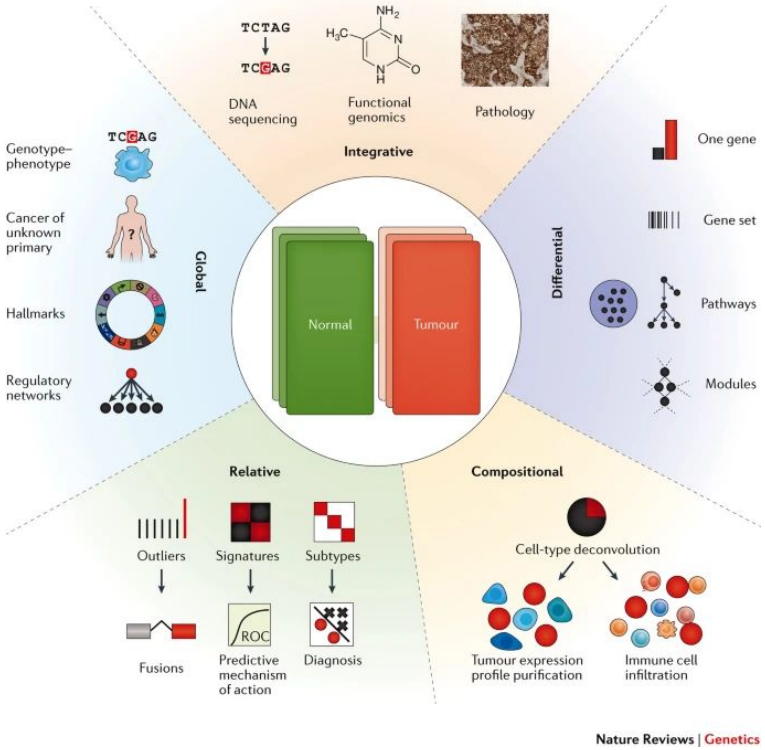
Gene fusions are the most interesting feature



10.1038/nrg.2017.96

Cancer transcriptomes

Cancer transcriptomics is not just about differential expression

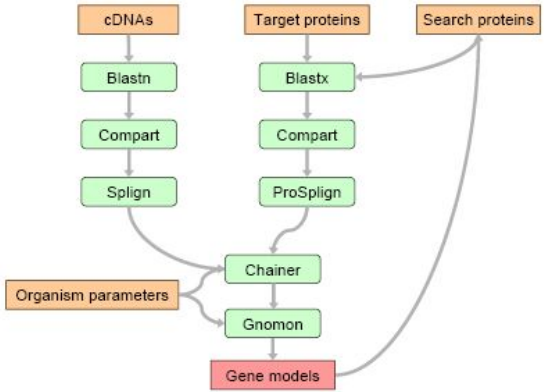


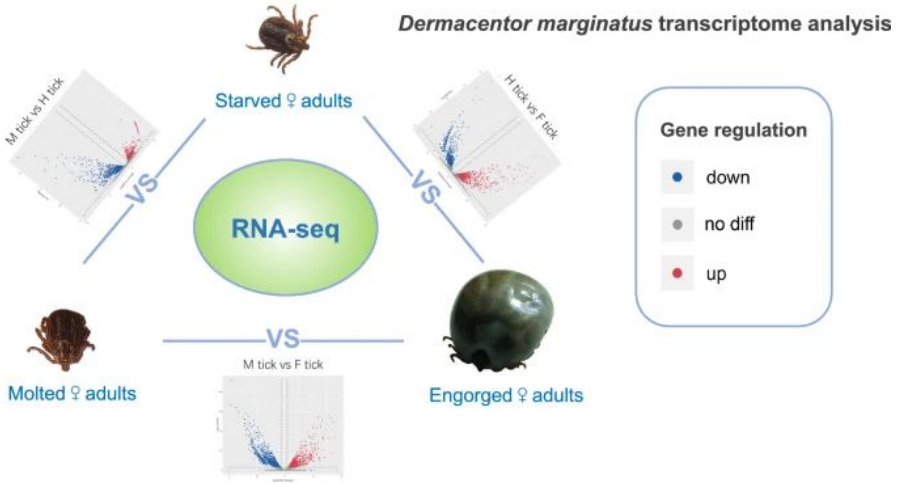
10.1038/nrg.2017.96

Novel organisms

Sometimes reference genome is not even available

RNAseq assembly improves annotation



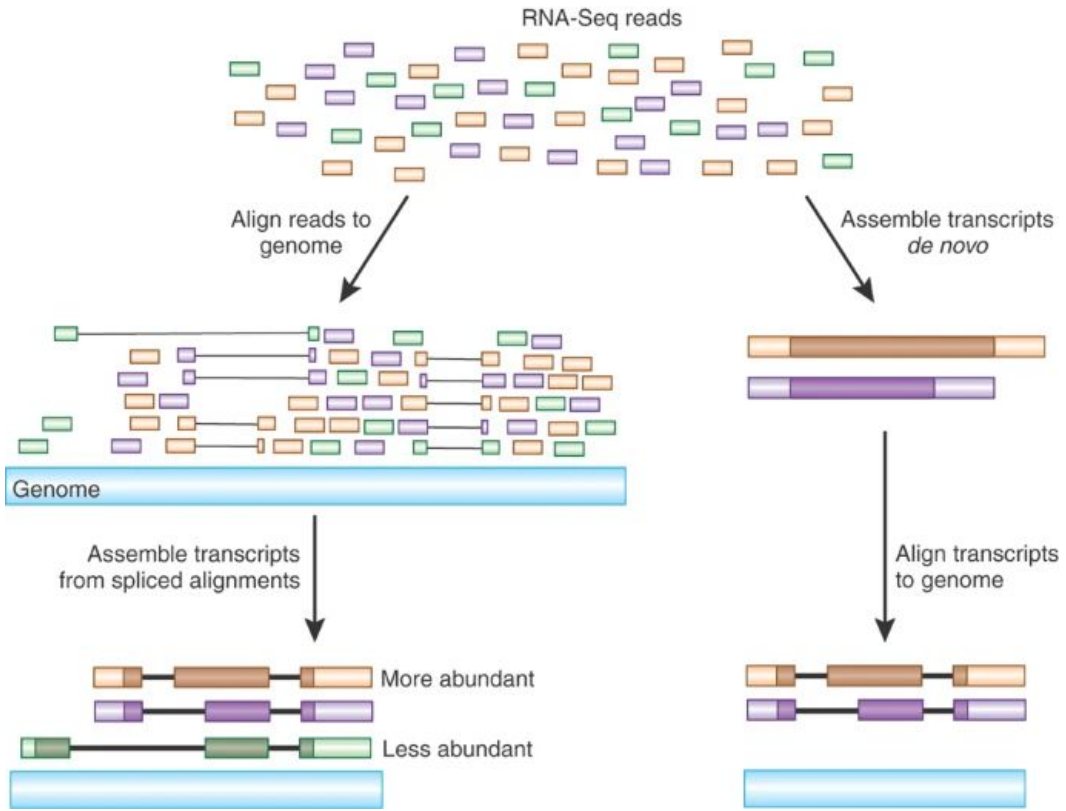


10.1186/s13071-020-04442-2

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/

Transcriptome assembly

Two main approaches: mapping-first and de novo



10.1038/nbt0510-421

Transcriptome assembly

>30 tools in listed on https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools#Transcriptome_assemblers

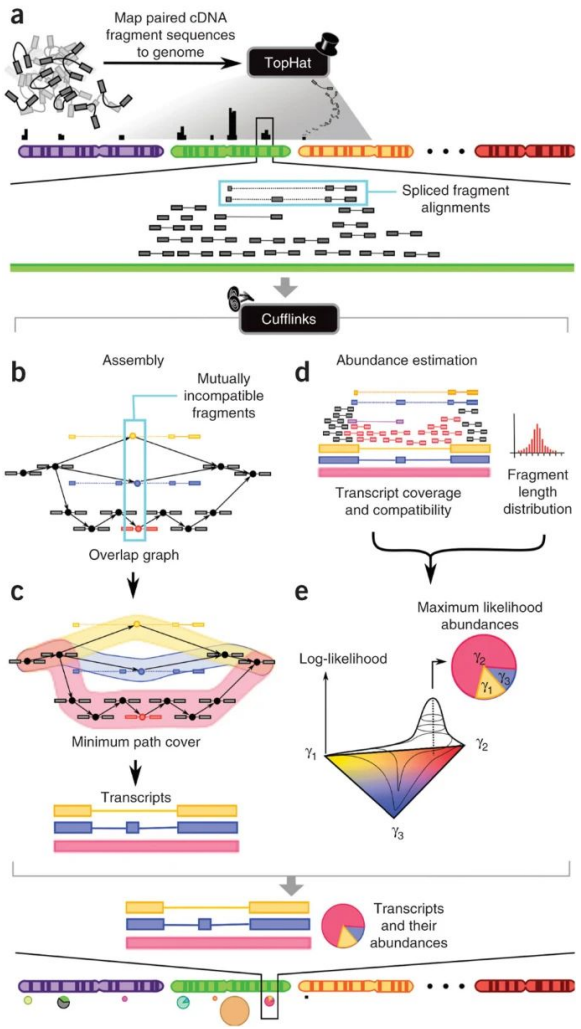
Two main approaches: mapping-first and de novo

Align-first approach

Align fragments with a splice-aware aligner

Note mutually exclusive fragments

Reconstruct isoforms and estimate their abundance

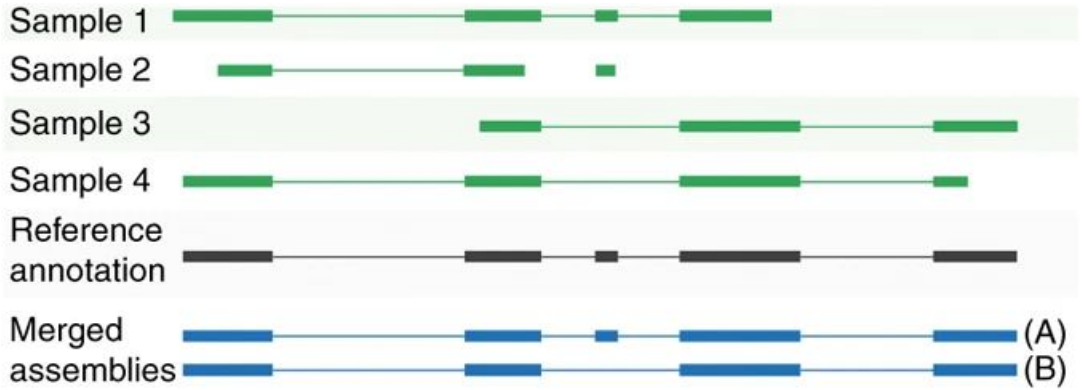


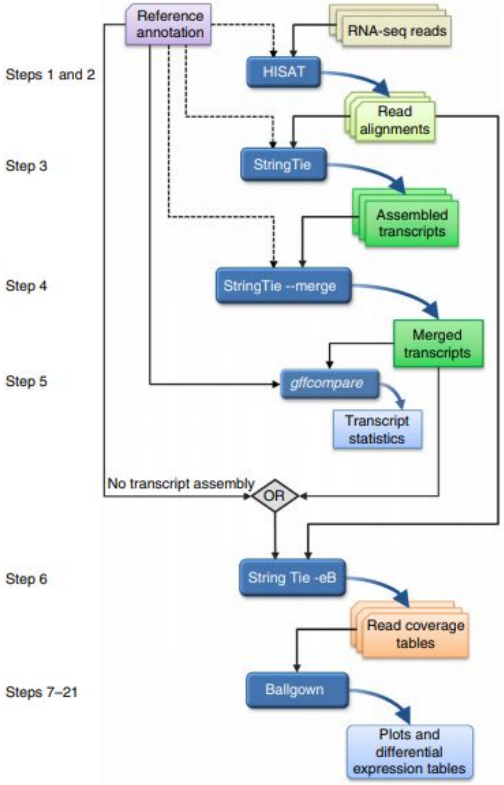
10.1038/nbt.1621

StringTie

Alignment-first assembler

Can also de novo assemble separate loci





10.1038/nprot.2016.095

Trinity

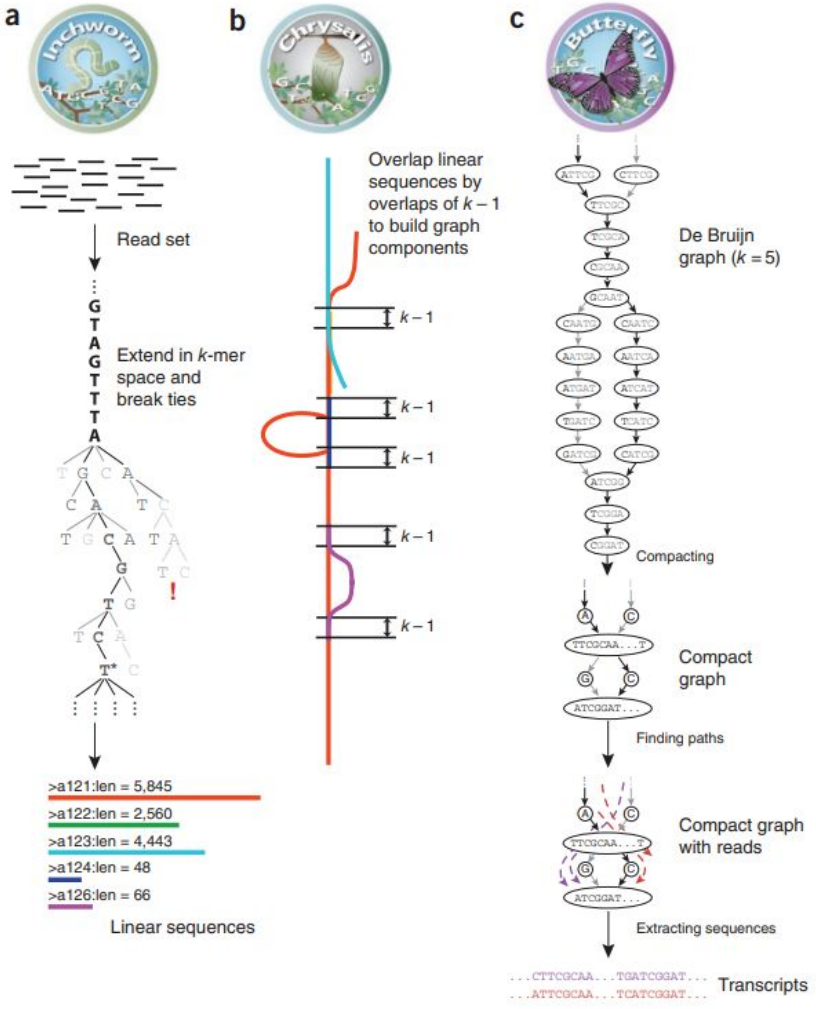
De novo and reference-based assembly of transcriptomes

Trinity workflow

Inchworm greedily constructs contigs

Chrysalis combines contigs

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

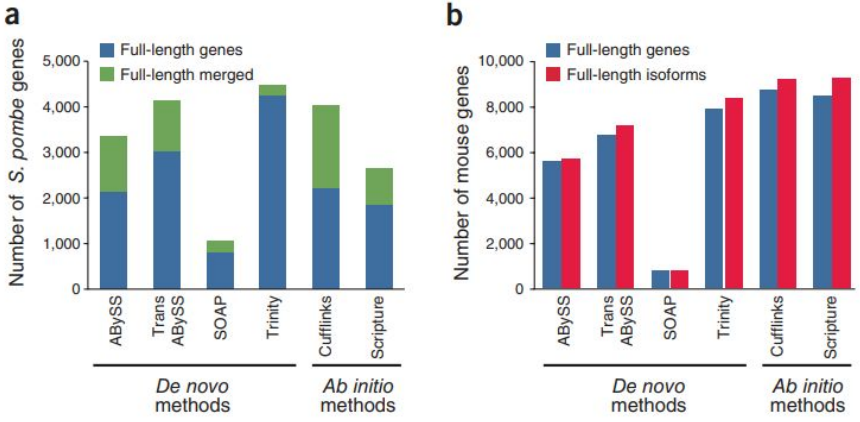


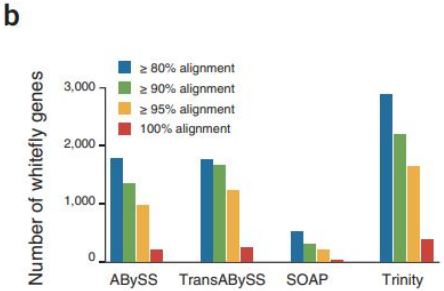
10.1038/nbt.1883

Trinity comparison to other tools

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

* according to Trinity :)





Whitefly

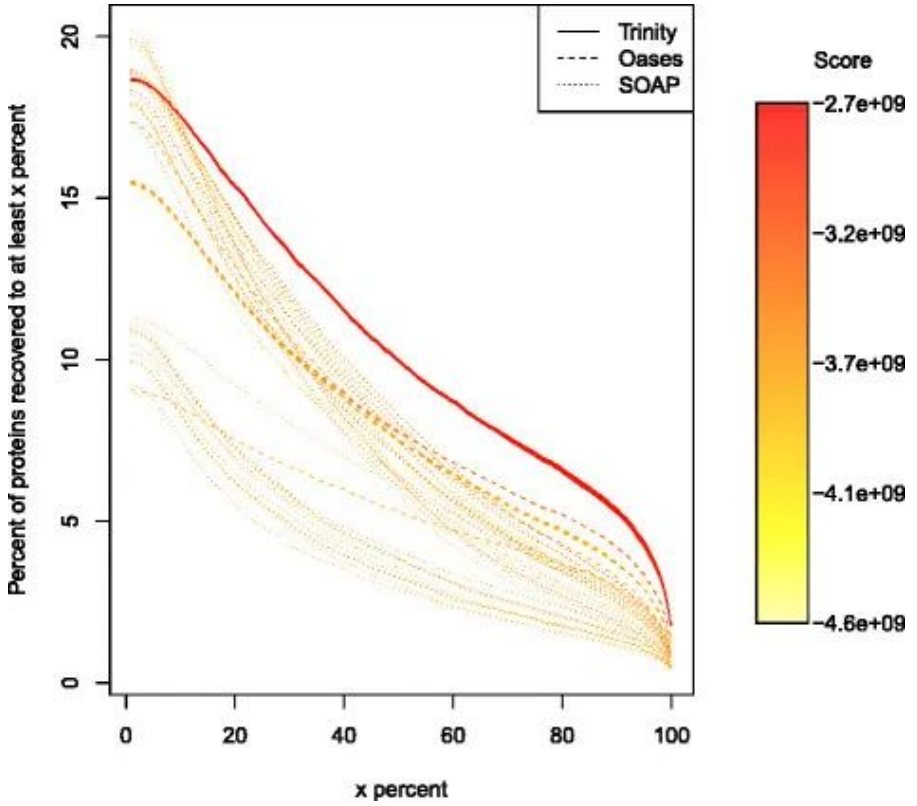
10.1038/nbt.1883

Other Trinity advantages

Regularly maintained (last update - 30.06.2020)

Contains a huge pipeline for all kinds of downstream analyses

Regularly wins in comparisons with other tools

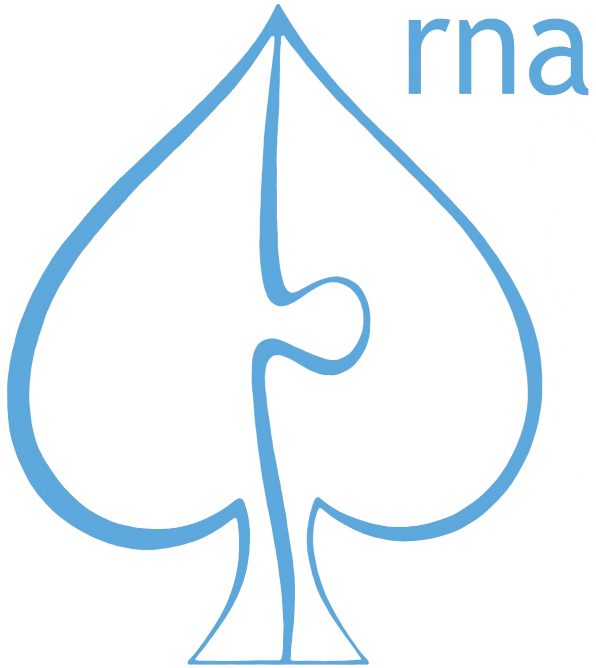


10.1186/s13059-014-0553-5

rnaSPAdes

De novo assembler

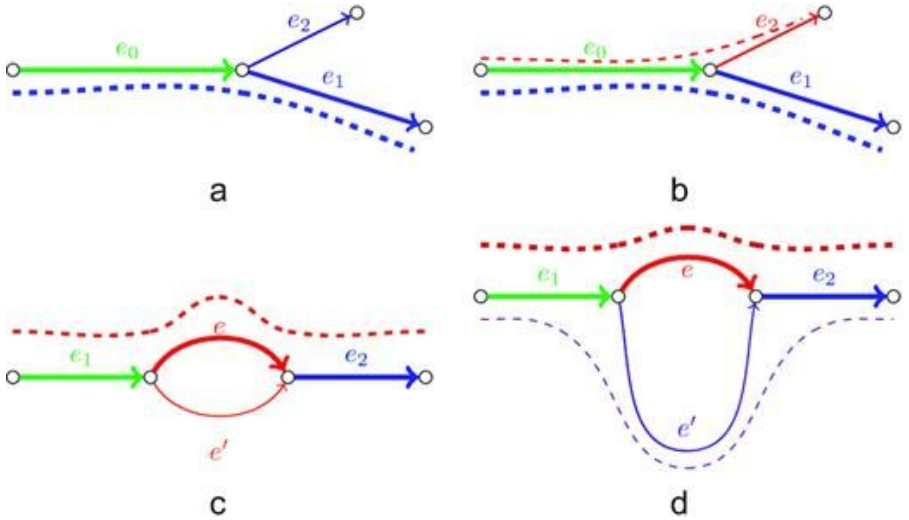
http://cab.spbu.ru/software/rnaspades/



RNAseq data peculiarities

Graph tips are not necessarily errors

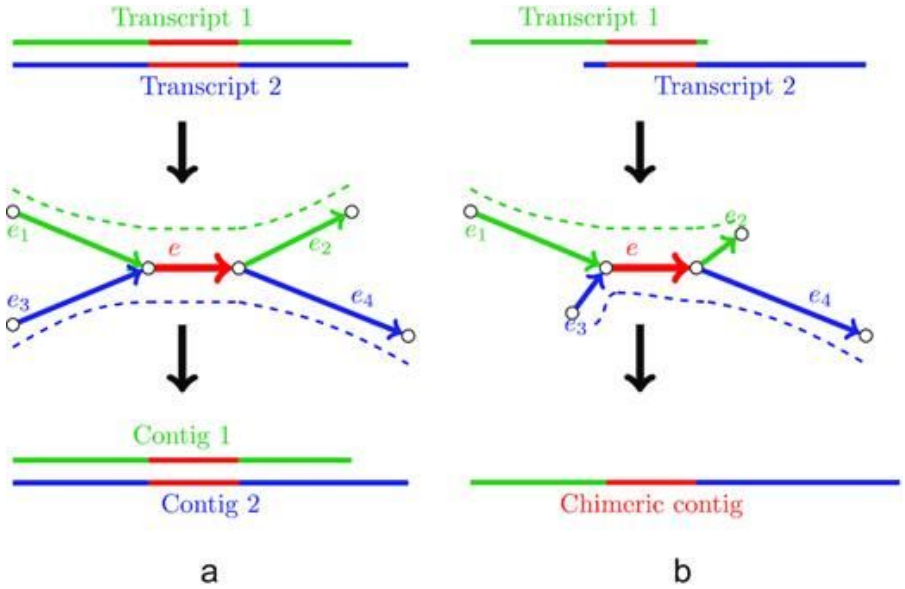
Graph bulges can correspond to different isoforms



10.1093/gigascience/giz100

RNAseq data peculiarities

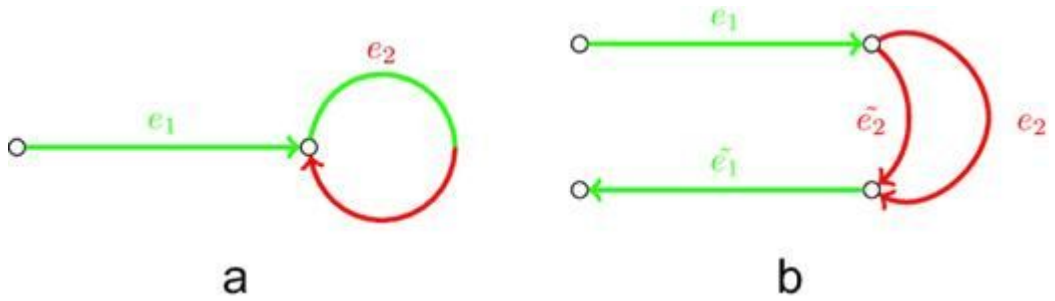
Isoforms may look like repeats



10.1093/gigascience/giz100

RNAseq data peculiarities

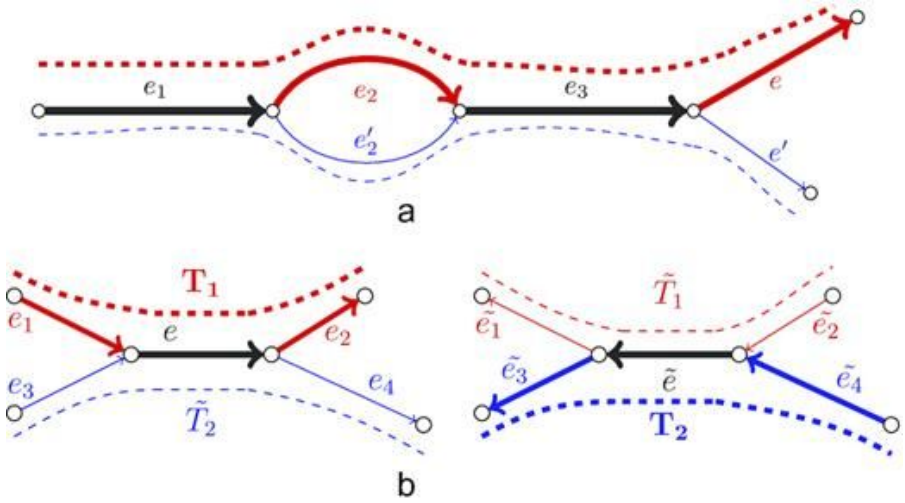
RNAseq data has specific chimeric read structures



10.1093/gigascience/giz100

RNAseq data peculiarities

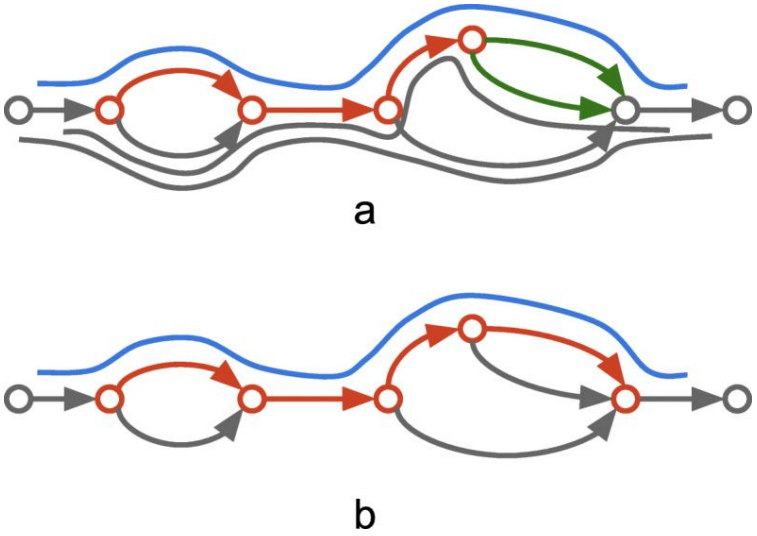
Isoforms can be resolved with coverage and strand-specific data

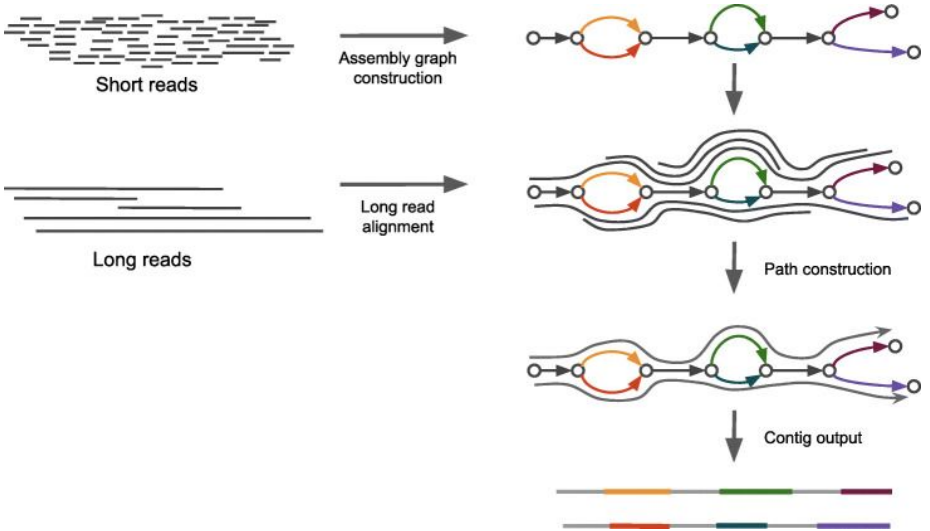


10.1093/gigascience/giz100

rnaSPAdes and long reads

rnaSPAdes can also incorporate long read data from ONT and PacBio





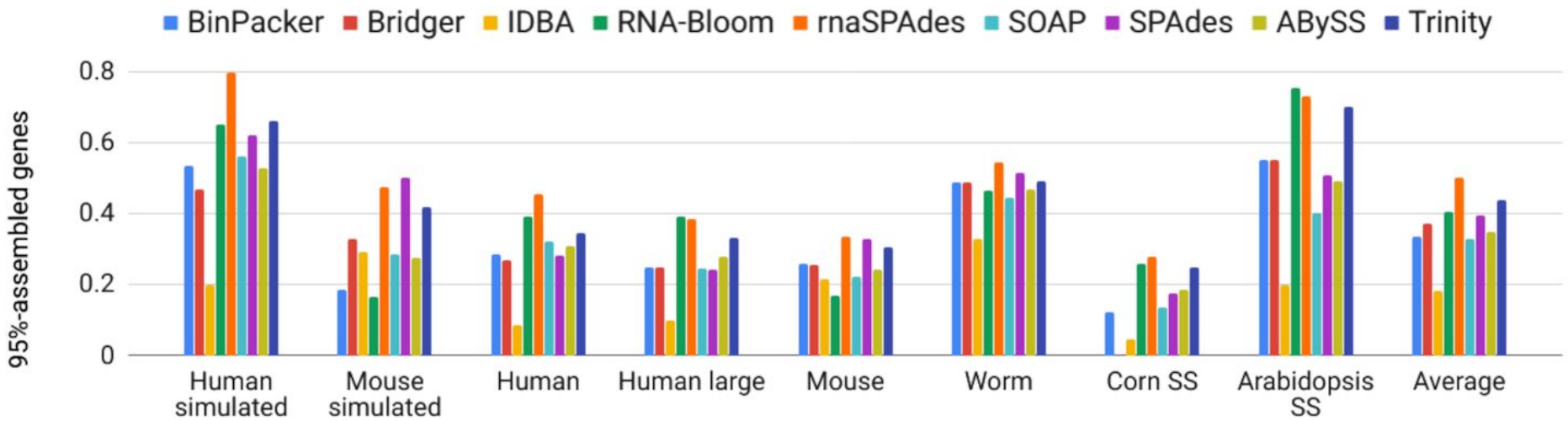
10.1186/s12859-020-03614-2

Other notable tools

Velvet, Trans-ABySS, IDBA-Tran

rnaSPAdes and other assemblers

rnaSPAdes is consistently good but there is no clear winner



10.1093/gigascience/giz100

Time and memory

Assembler	Human large		Arabidopsis SS	
	Time	RAM (GB)	Time	RAM (GB)
BinPacker	46 h 59 m	91	88 h 25 m	131
Bridger	65 h 54 m	88	49 h 58 m	126
IDBA	9 h 35 m	35	26 h 24 m	42
Bloom	37 h 52 m	38	34 h 42 m	40
rnaSPAdes	5 h 4 m	32	7 h 24 m	40
SOAP	1 h 21 m	28	1 h 58 m	20
SPAdes	11 h 39 m	39	14 h 58 m	52
ABySS	6 h 49 m	25	8 h 9 m	35
Trinity	18 h 8 m	50	8 h 30 m	123

All assemblers were launched in 16 threads on a server with 128 GB of RAM and 56 Intel Xeon 2.0 GHz cores. BinPacker, which has no options for setting the number of threads, was launched with default parameters.

10.1093/gigascience/giz100

More benchmarks and guidelines

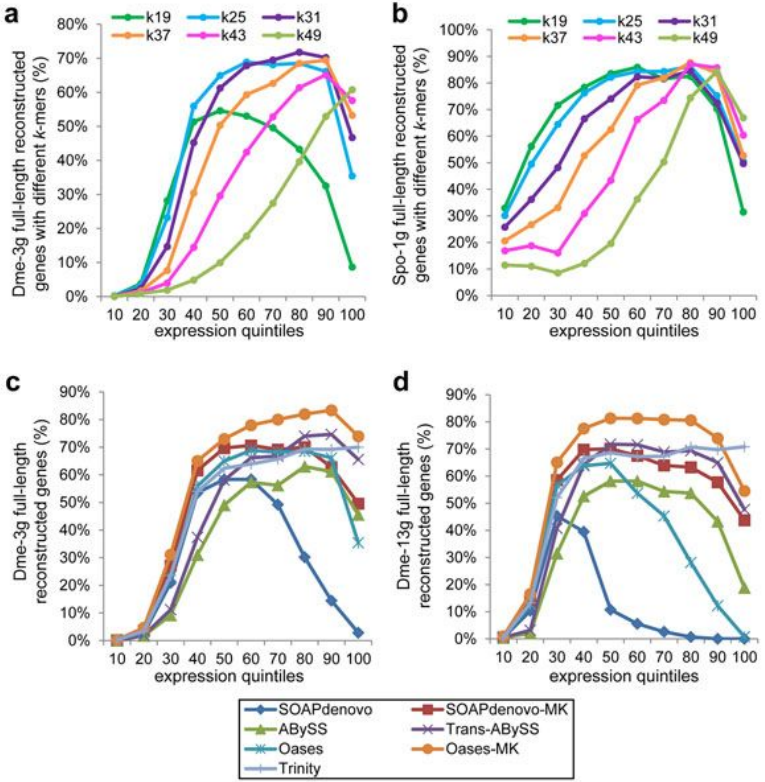
Multikmer assembly is generally better

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

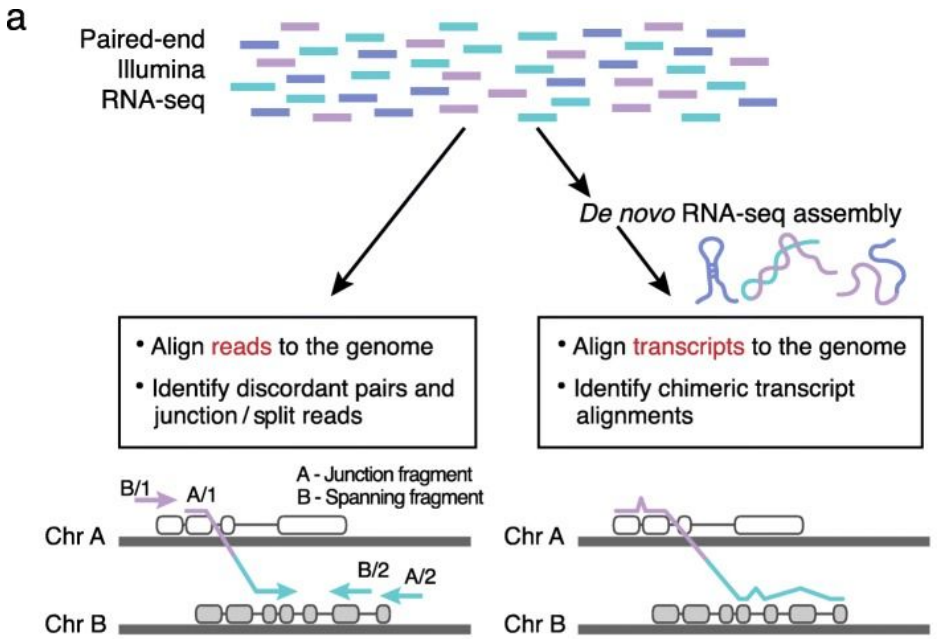


10.1186/1471-2105-12-S14-S2

Fusion detection

Recently 23 methods for fusion detection were compared

De novo methods are mostly worse in terms of accuracy, but they can find foreign transcripts



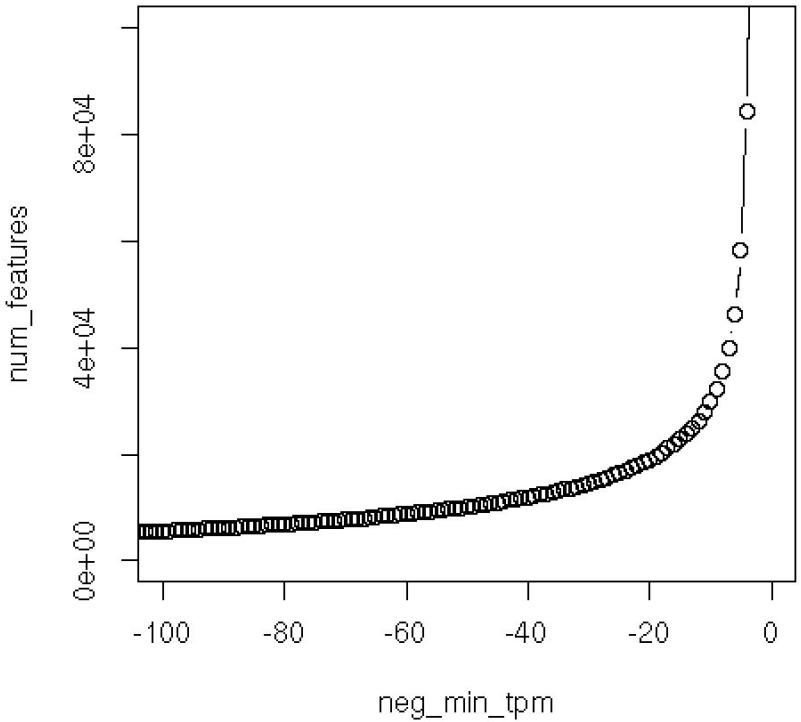
10.1186/s13059-019-1842-9

Reducing number of transcripts

TransPS

CD-HIT-EST

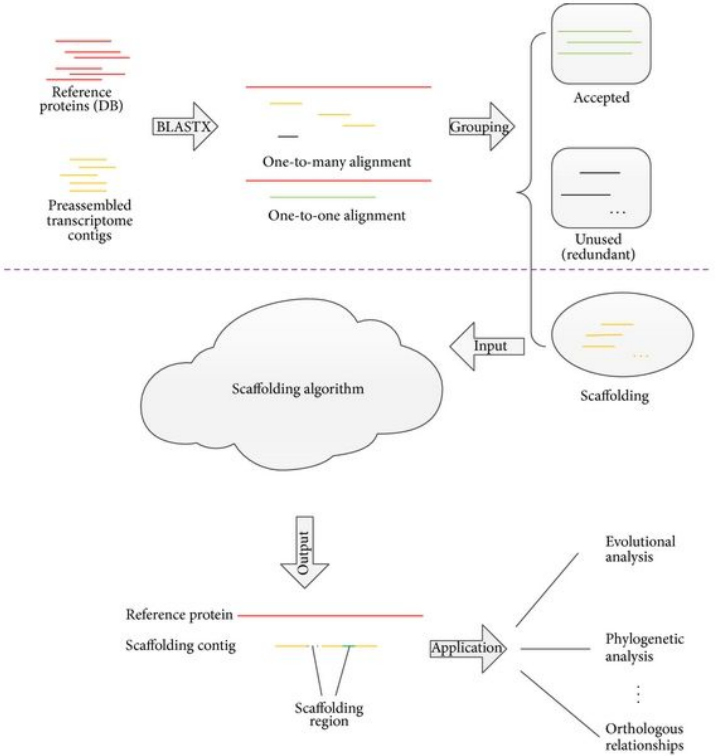
Coverage filtering



TransPS

Scaffolding transcripts based on reference proteins

Greatly reduces the number of redundant contigs and improves coverage



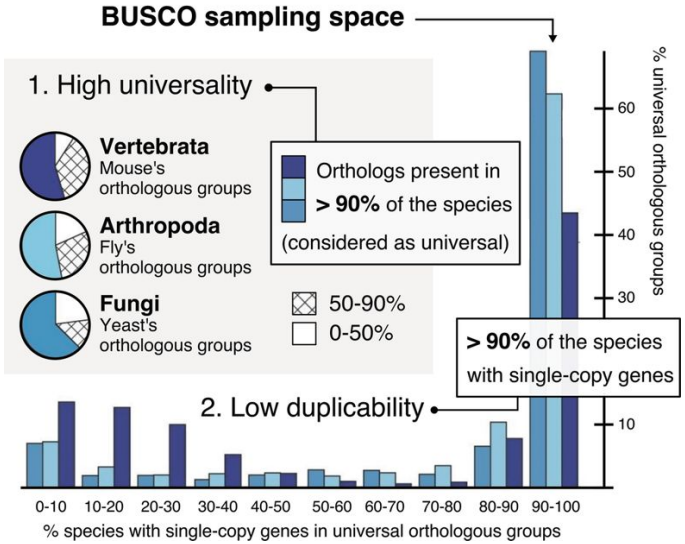
10.1155/2014/961823

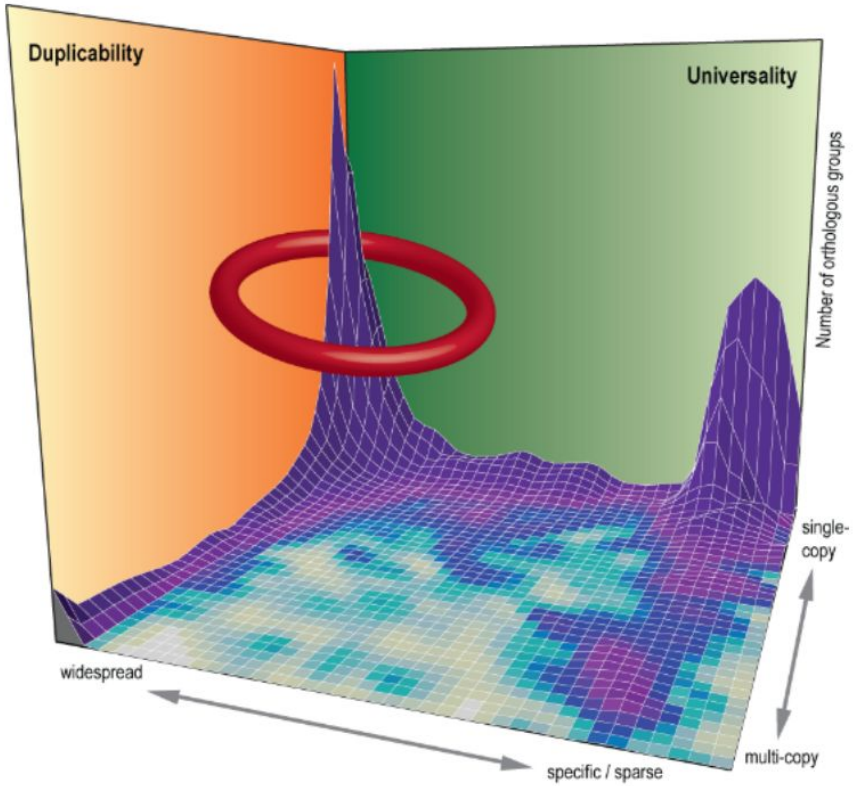
Assembly quality control

How to check which assembly is the best?

BUSCO

Benchmarking Universal Single-Copy Orthologs

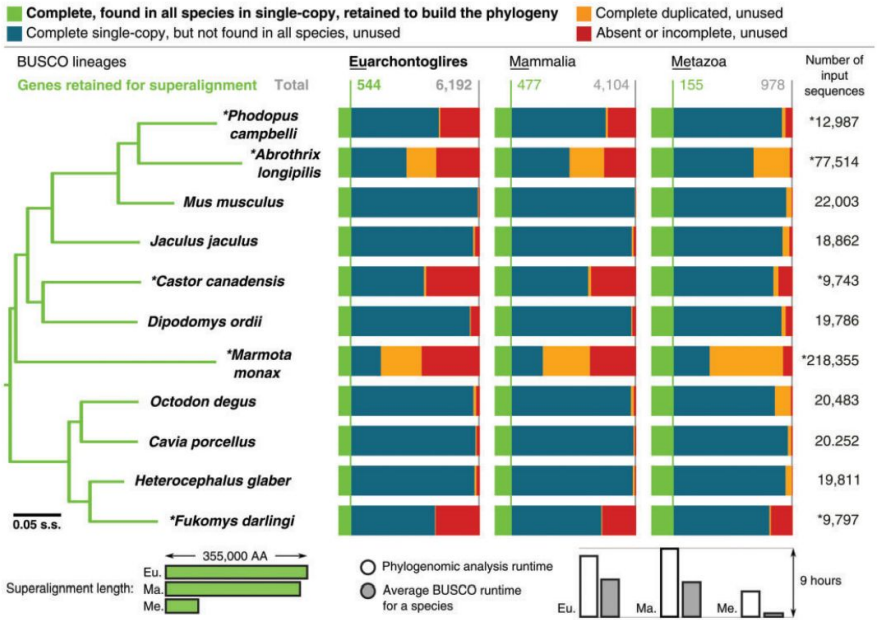




10.1093/molbev/msx319

BUSCO for phylogenetics

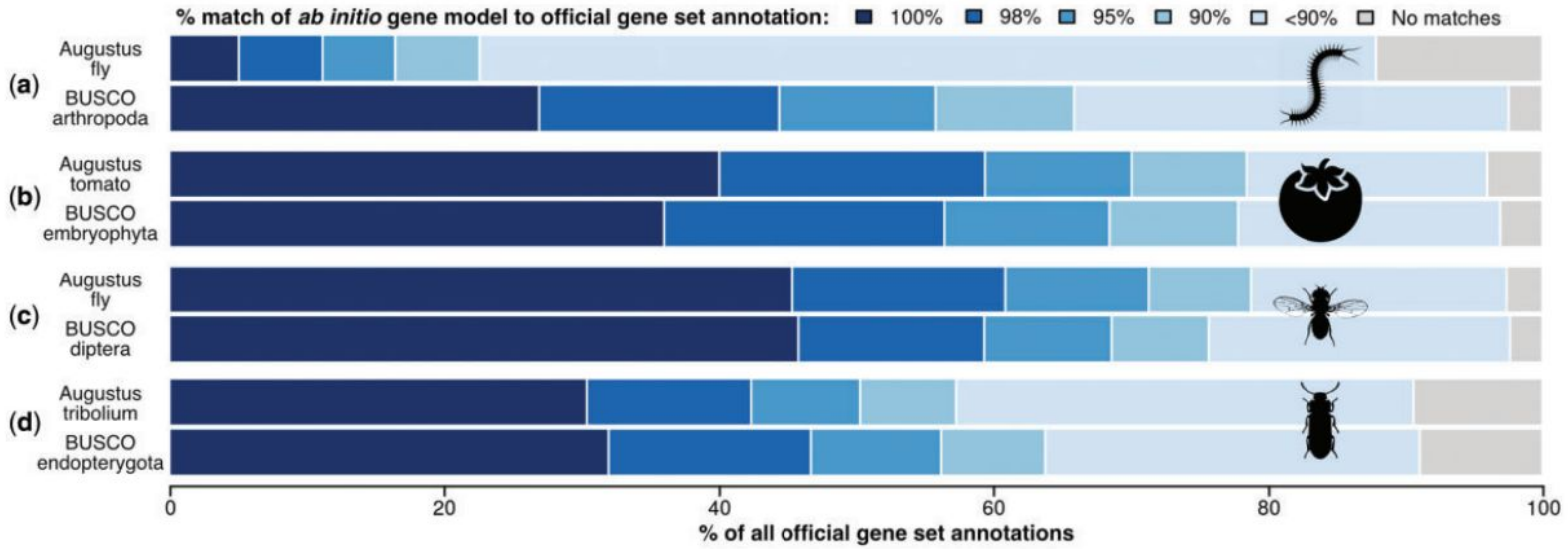
BUSCO genes can be used to construct sets of orthologous genes for phylogeny



10.1093/molbev/msx319

BUSCO-trained gene predictions

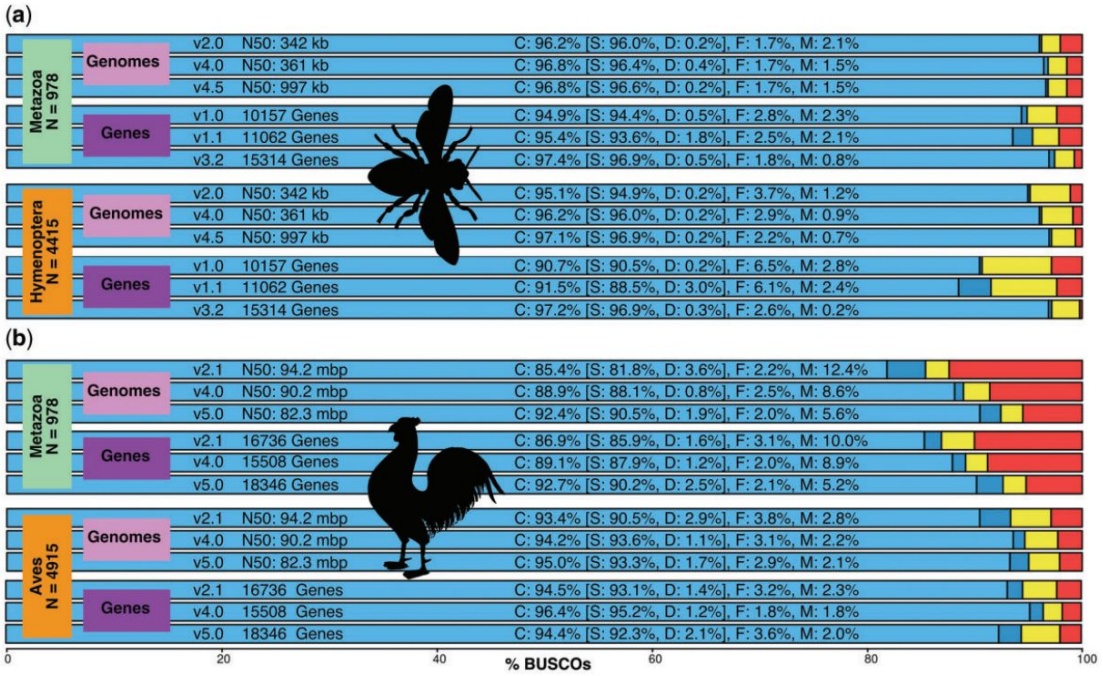
Gene annotations trained on BUSCO sets sometimes work better than ab initio



10.1093/molbev/msx319

BUSCO for genomics quality control

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



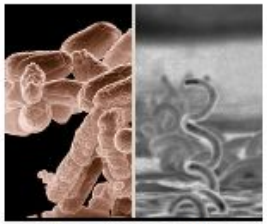
10.1093/molbev/msx319

BUSCO datasets

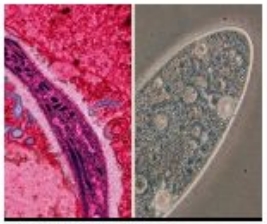
Several reference datasets based on OrthoDB

New version features automatic lineage detection

Datasets


Bacteria sets


Eukaryota sets


Protists sets


Metazoa sets

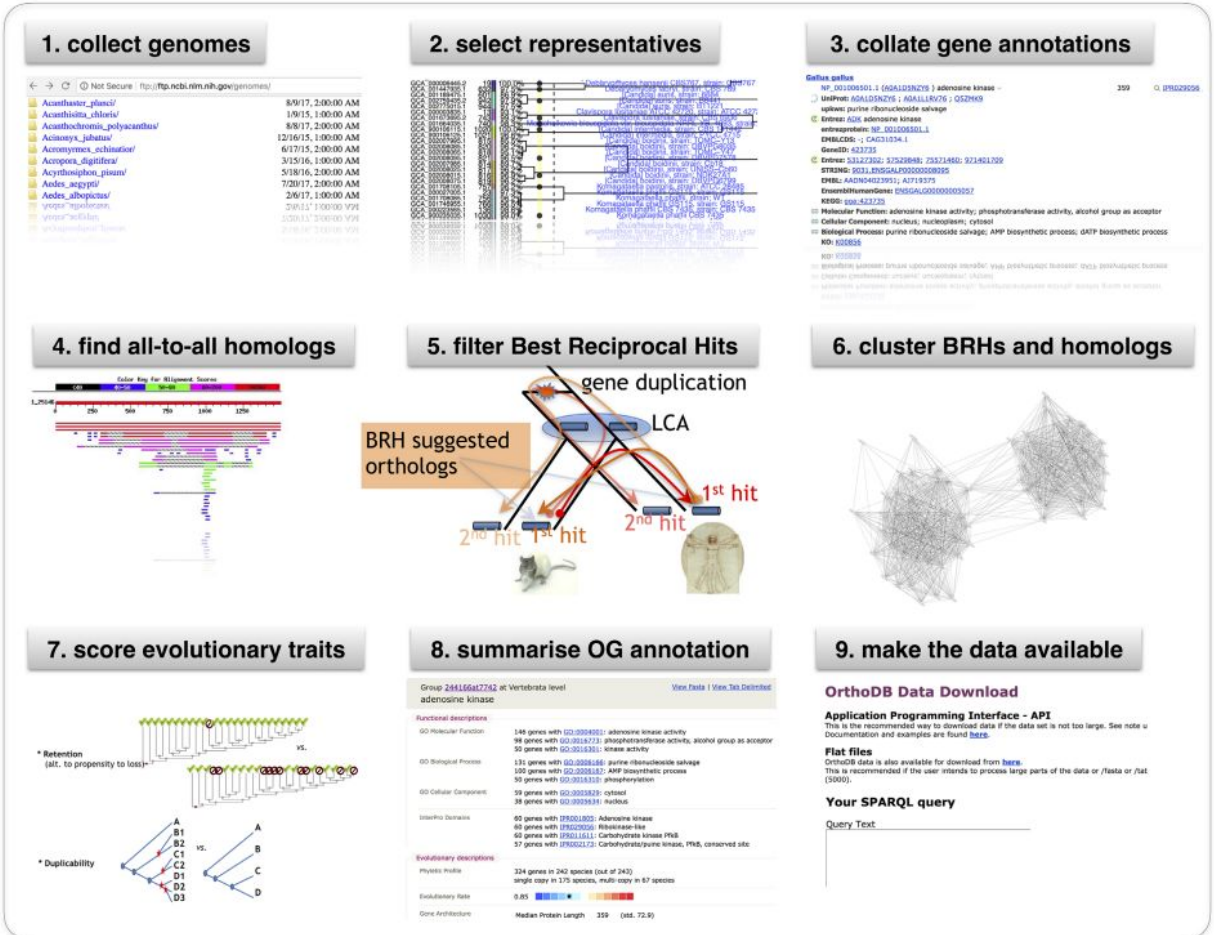

Fungi sets


Plants set

OrthoDB

Database of orthologs across multiple species

https://www.orthodb.org/



10.1093/nar/gky1053

BUSCO workflow

For transcriptome assessment BUSCO only uses BLAST and HMMER

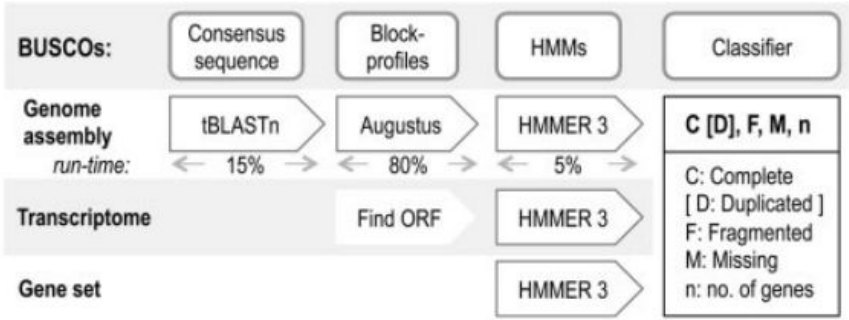
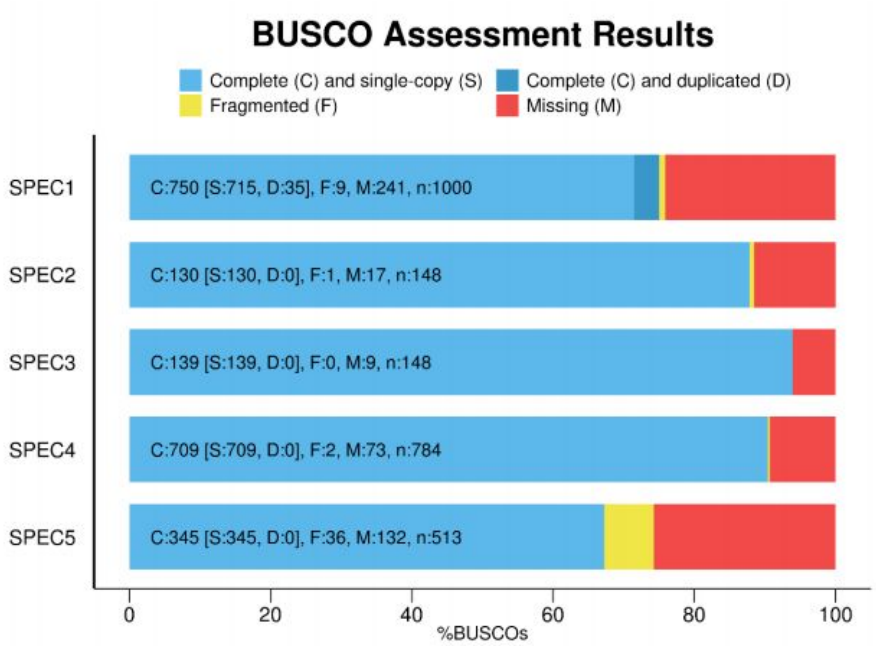


Fig. 1. BUSCO assessment workflow and relative run-times

10.1093/bioinformatics/btv351

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

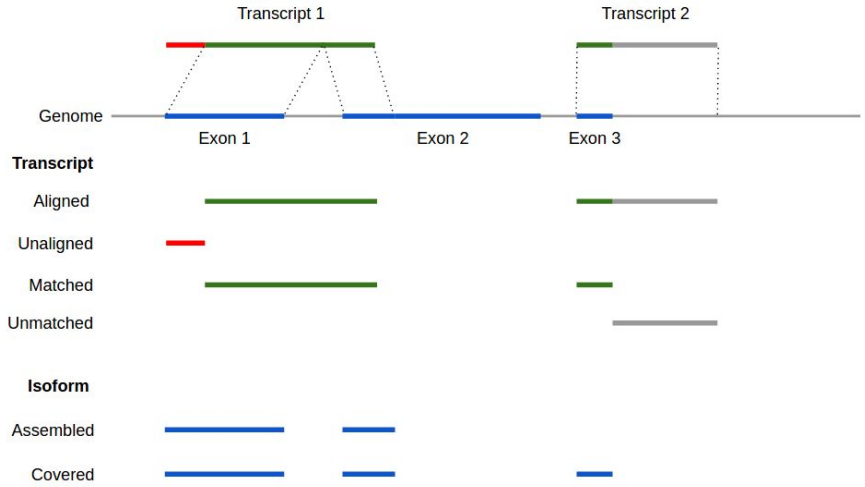
http://cab.spbu.ru/software/rnaquast/

10.1093/bioinformatics/btw218

rnaQUAST

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

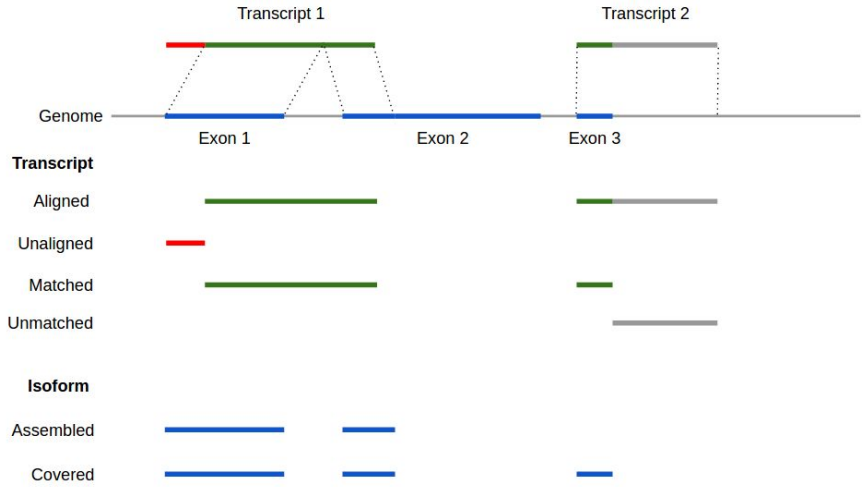
It also estimates coverage with STAR alignments



10.1093/bioinformatics/btw218

rnaQUAST

For de novo assemblies rnaQUAST runs BUSCO and GeneMarkS-T



10.1093/bioinformatics/btw218

rnaQUAST metrics

rnaQUAST reports a lot of basic and alignment metrics

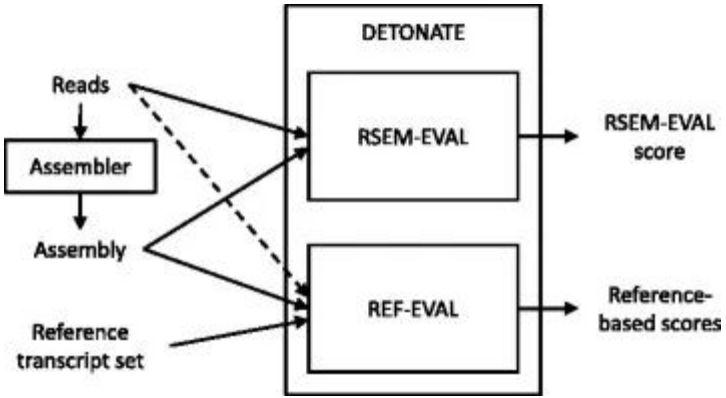
Assembler	ABySS	IDBA	SOAP	SPAdes	Trinity
k-mer size	32	default	31	default	default
rnaQUAST metrics					
Transcripts	107202	38294	69331	48706	51245
Transcripts ≥ 500 bp	17882	***17542***	16021	17512	***21994***
Aligned	95884	***38198***	68591	48027	***51112***
Uniquely aligned	94681	37288	***67878***	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
k-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***

10.1093/bioinformatics/btw218

DETONATE

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

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



10.1186/s13059-014-0553-5

RSEM-EVAL

Used for assessment of de novo assemblies

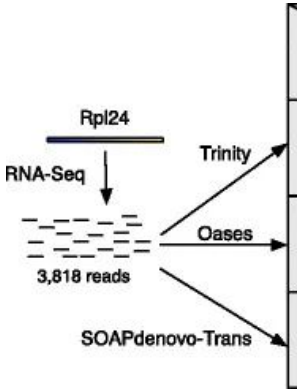
Score is the probability of the assembly A given reads D

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

10.1186/s13059-014-0553-5

RSEM-EVAL for Rpl24 transcript

Trinity demonstrates the best result

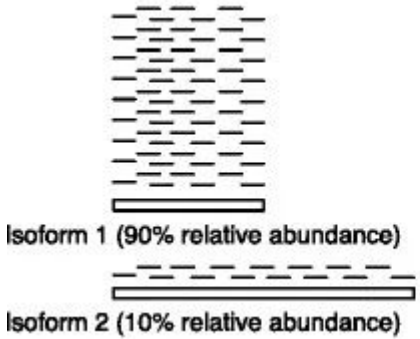


Assembly \ Score	Likelihood score	Prior score	BIC score	RSEM-EVAL score
	−86542	−876	−8	−87426
	−198666	−2771	−29	−201466
	−254310	−613	−12	−254935

10.1186/s13059-014-0553-5

RSEM-EVAL for multi-isoform genes

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

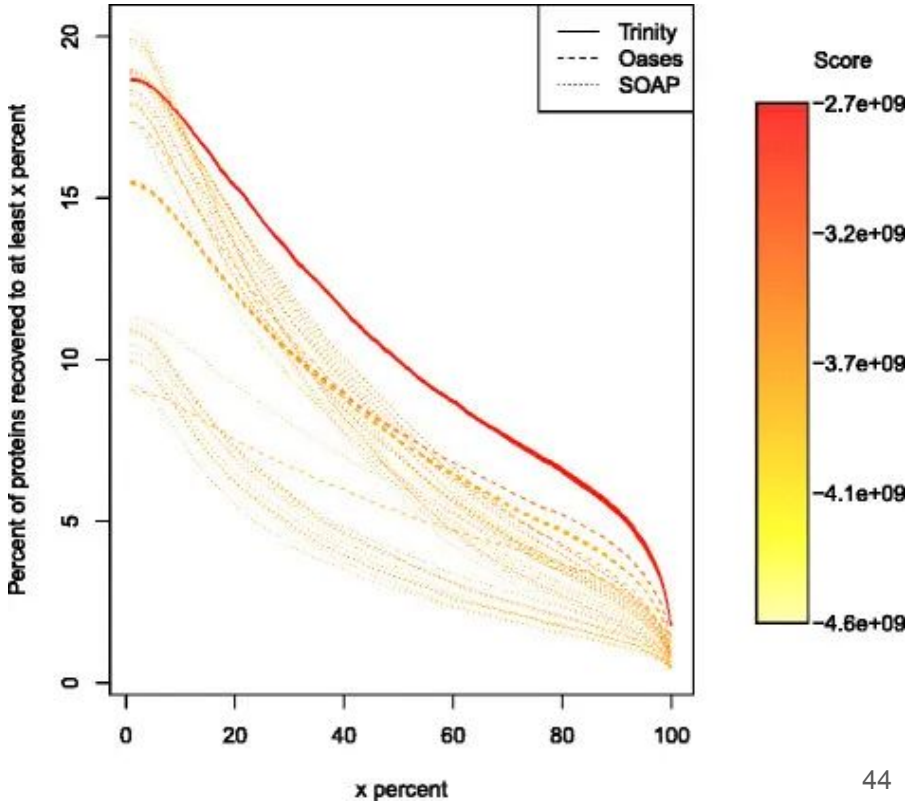


Score / Assembly	RSEM-EVAL	GENOVO	ALE
Truth	−43720	−19557	−116316
Long only	−44403	−18199	−88905
Short only	−104963	−68997	−52090

10.1186/s13059-014-0553-5

RSEM-EVAL for *Xenopus* transcriptome

Trinity shows the best results at recovering protein coding regions



10.1186/s13059-014-0553-5

Transrate

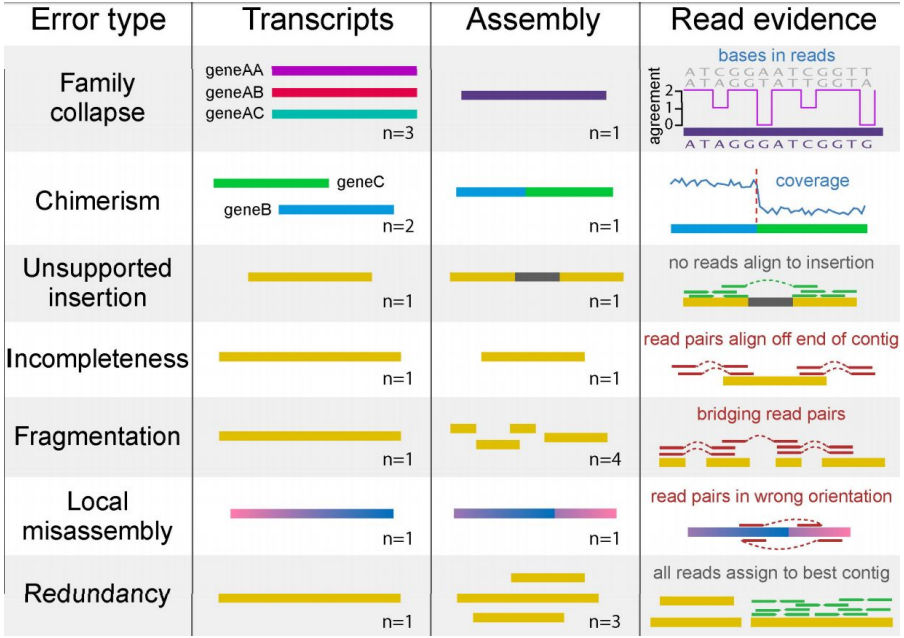
Tool for de novo transcriptome assembly control

10.1101/gr.196469.115

transrate
understand your transcriptome assembly

Transrate

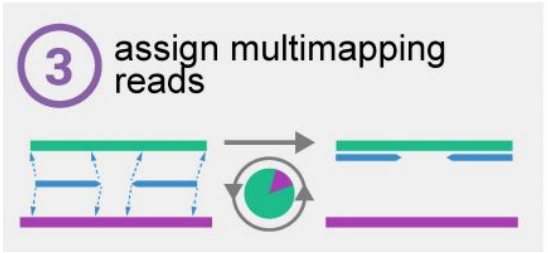
Several ways how contigs can be wrong and how mistakes can be detected

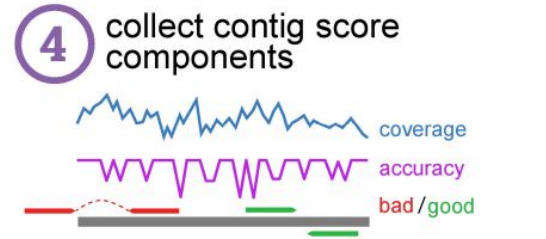


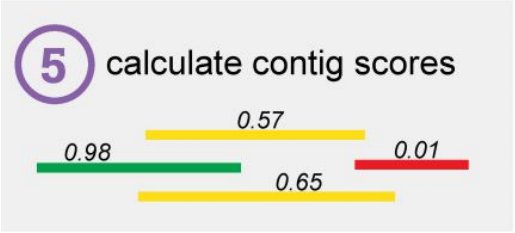
10.1101/gr.196469.115

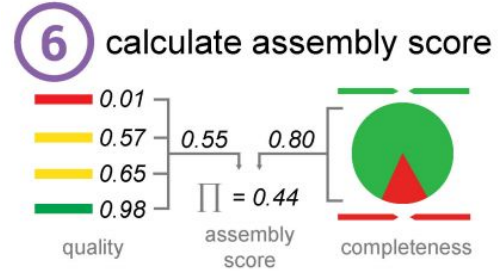
transrate
understand your transcriptome assembly

Transrate workflow









10.1101/gr.196469.115

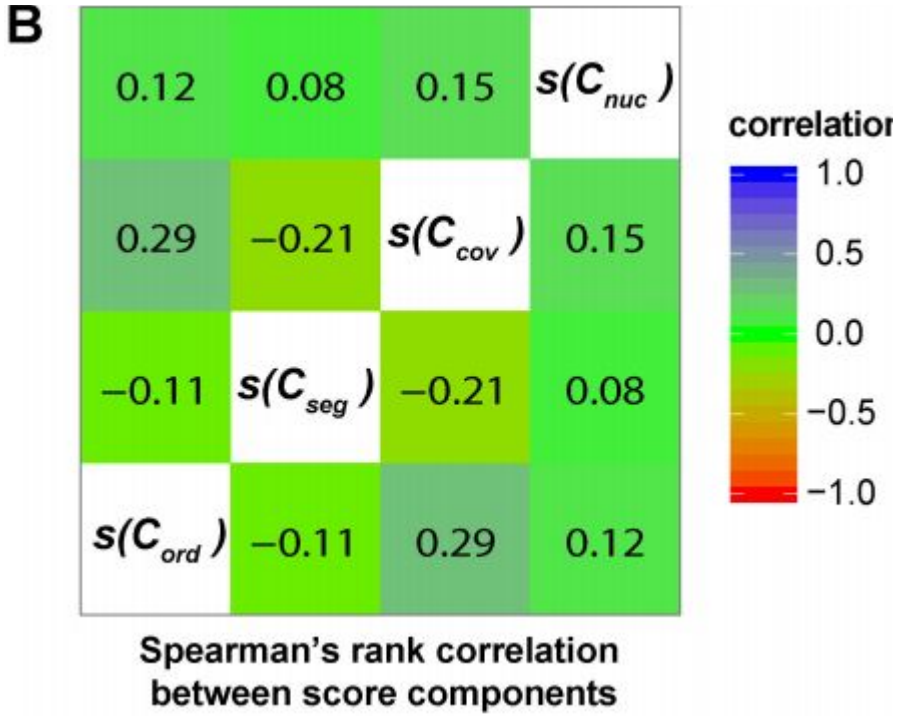
transrate
understand your transcriptome assembly

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

10.1101/gr.196469.115

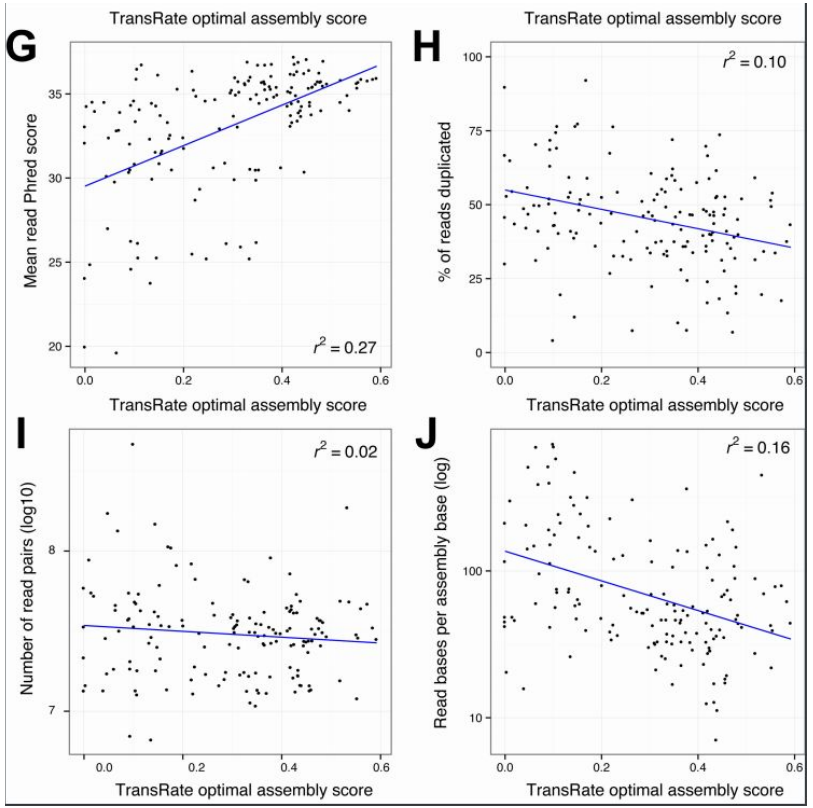
Transrate metrics correlations



10.1101/gr.196469.115

Transrate and data parameters

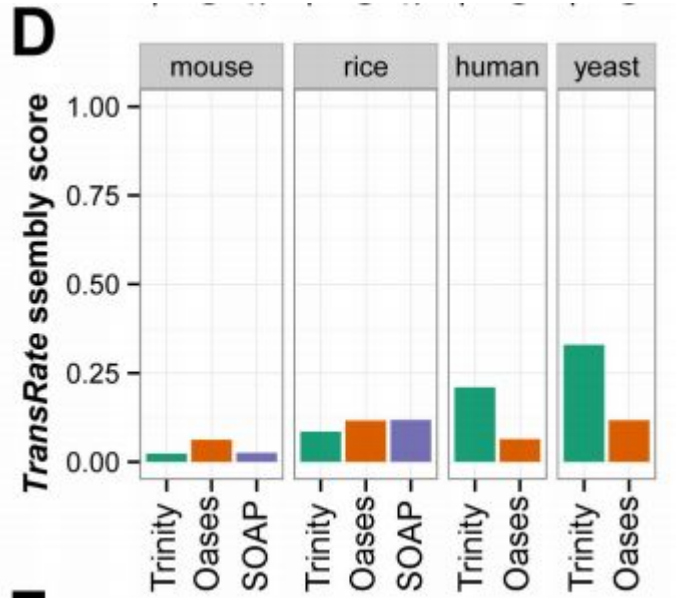
Assembly scores do not depend on number of read pairs and duplication of reads, but depend on quality of reads



10.1101/gr.196469.115

Assemblers comparison

Assemblers' rankings depend on the organism



10.1101/gr.196469.115

Assembly and QC take-home points

Selecting assembler and parameters is a non-trivial matter

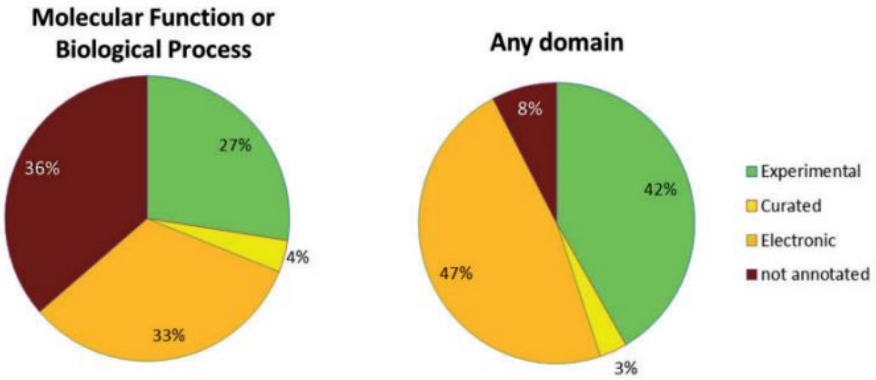
Best choice is to try everything available and compare results (possibly with downsampling to use resources more efficiently)

Annotation

BLAST results against a relatively close annotated genome

Annotation “dark matter”

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



10.1093/bib/bbw135

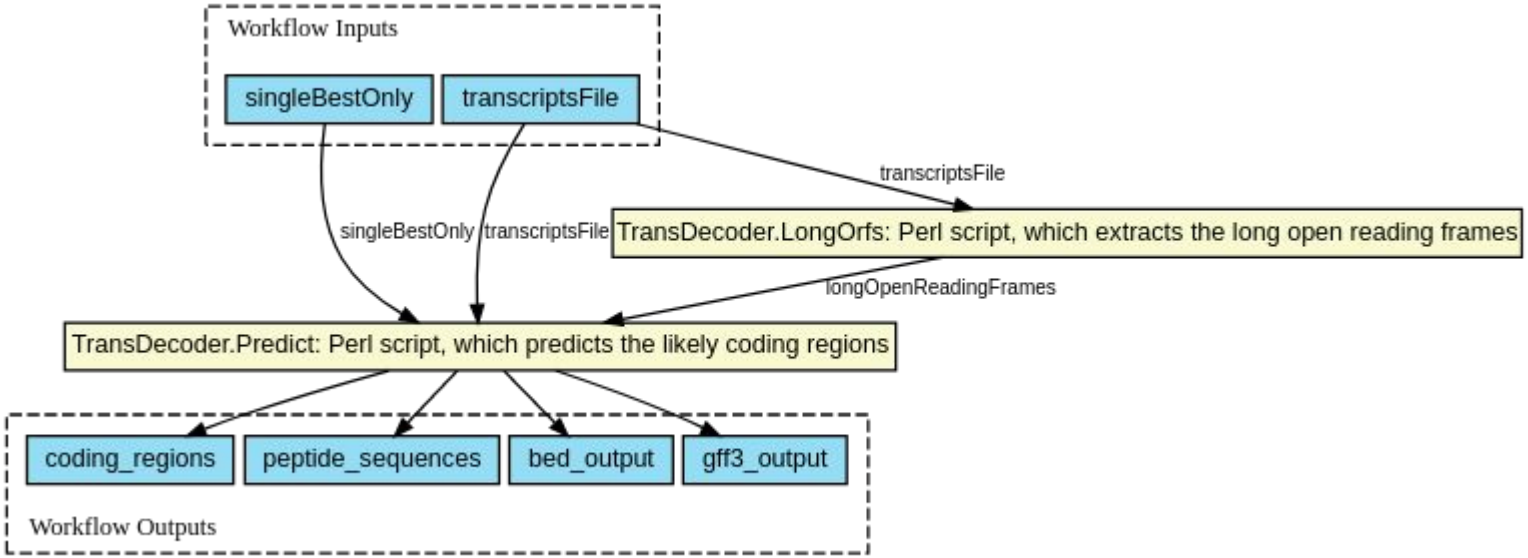
Protein prediction

Some pipelines require protein sequences for annotation

ORFs in transcripts can be predicted with several tools

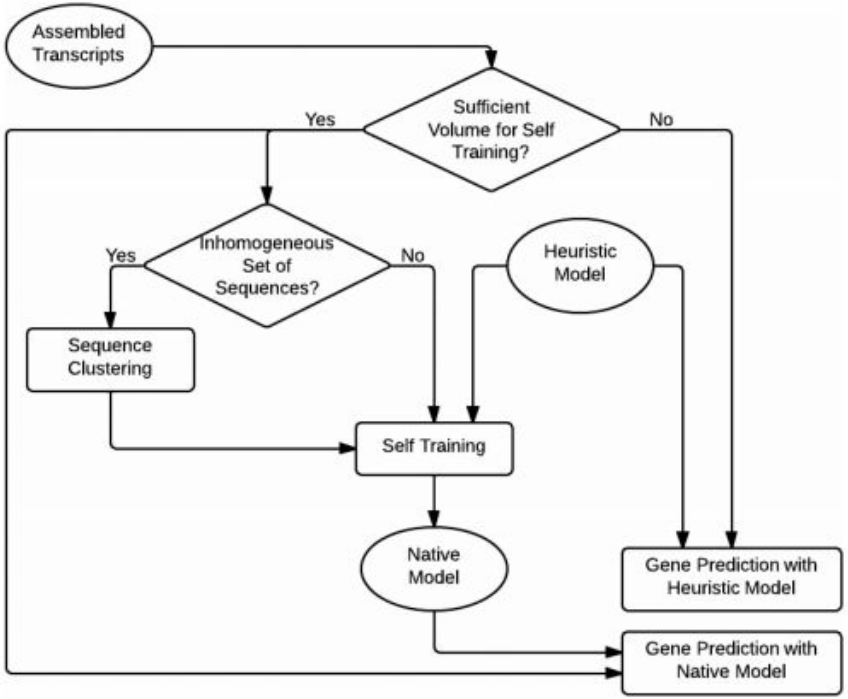
TransDecoder

Identifies coding regions within transcripts



GeneMarkS-T

Ab initio finding of proteins in eukaryotic transcripts



10.1093/nar/gkv227

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

10.1093/bib/bbw135

Specialized protein groups annotations

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

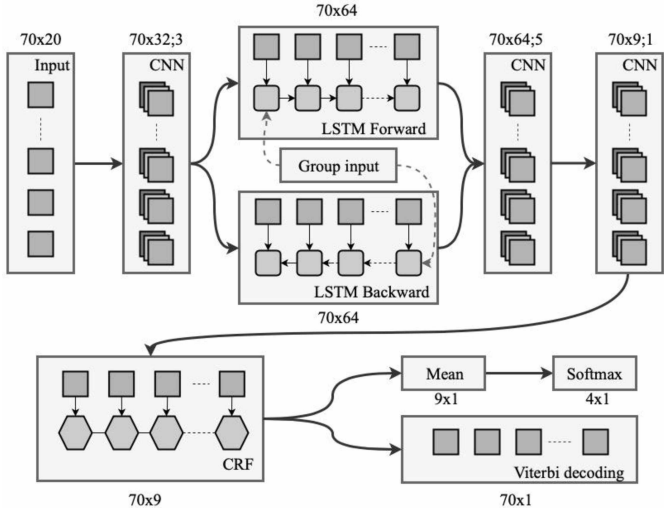
Resource	Function	Web address
CoGe	Compares genomes, find synteny	https://genomevolution.org
PlantTFDB	Plant Transcription Factor families	http://planttfdb.cbi.pku.edu.cn/
Potsdam plntfdb	Plant Transcription Factor families	http://plntfdb.bio.uni-potsdam.de/v3.0/
P450 Database	P450 protein families	http://drnelson.uthsc.edu/CytochromeP450.html
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/

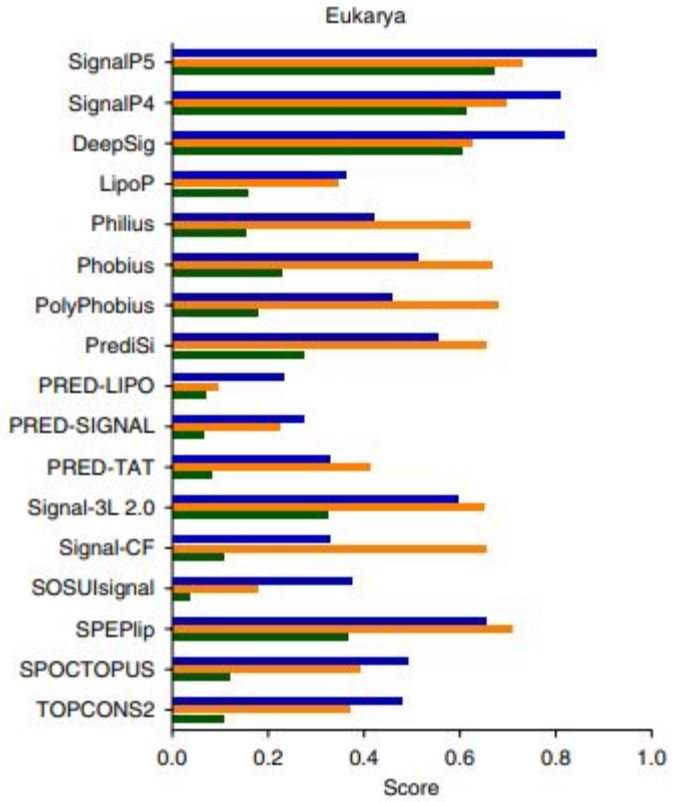
Note. [a]Also lists a comprehensive set of tools for transmembrane domains, subcellular localization and lipid modifications.

10.1093/bib/bbw135

SignalP

Predicts signal peptides using neural networks

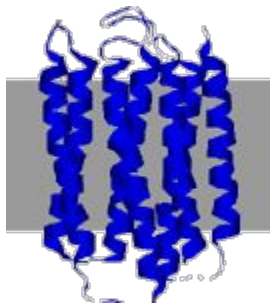




10.1038/s41587-019-0036-z

TMHMM

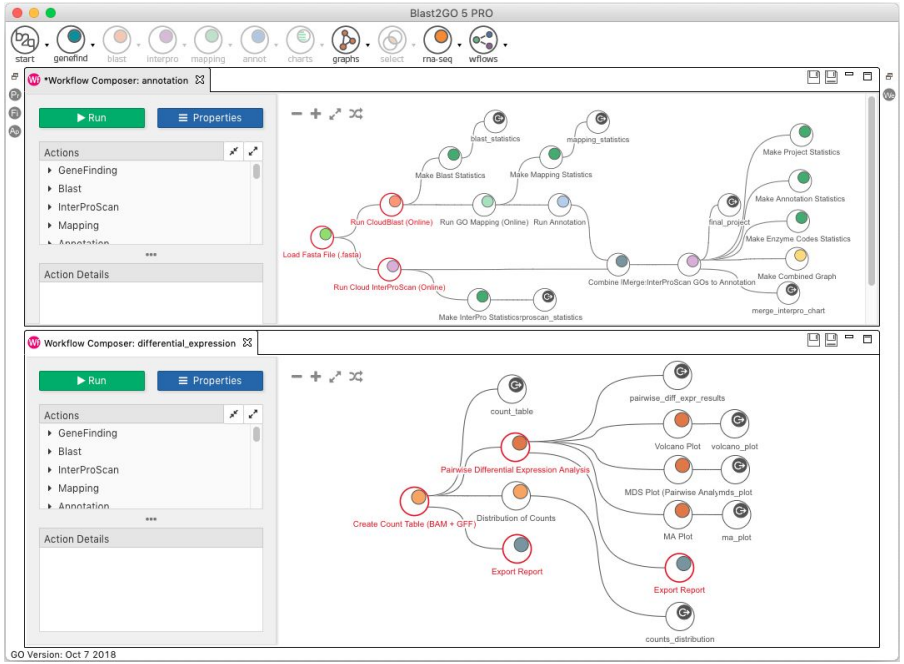
Predicts transmembrane domains



BLAST2GO

Readily available annotation pipeline

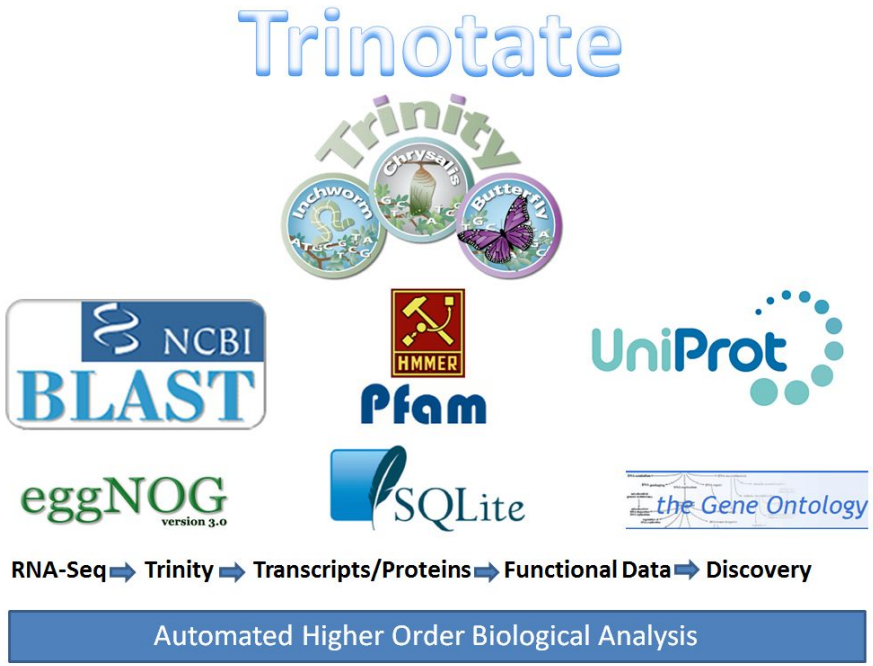
Commercial



Trinotate

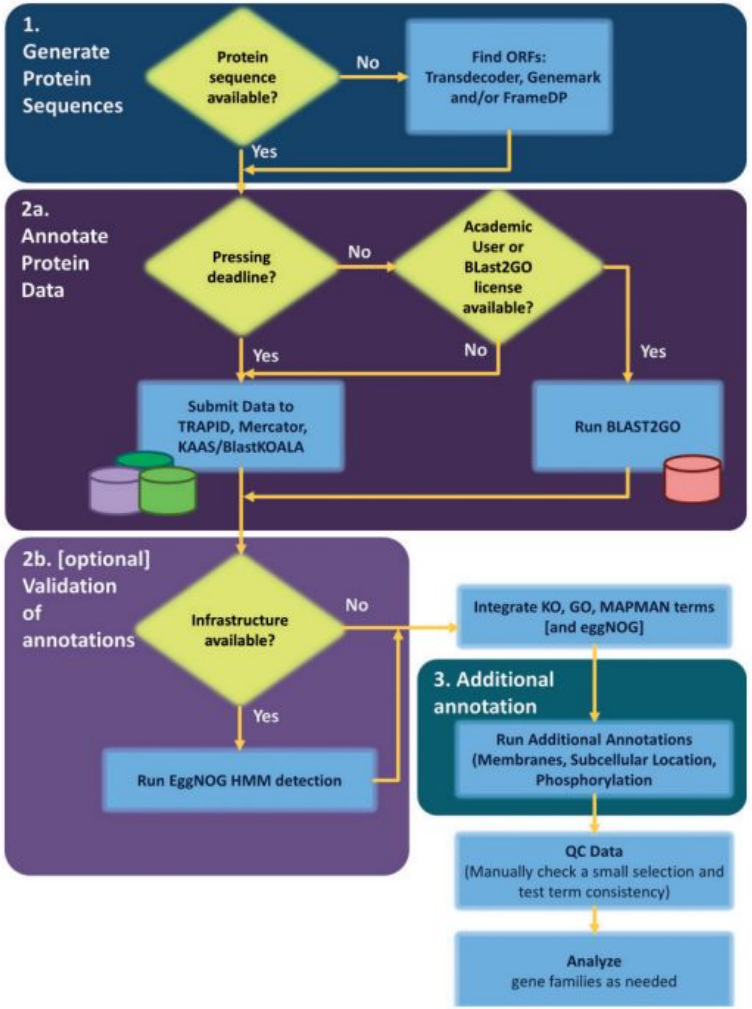
Annotation pipeline accompanying Trinity

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



Annotation flowchart

Here what typical annotation pipeline may look like



10.1093/bib/bbw135

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, and in general should be tailored to your particular needs

Annotation can be automated

Annotation should be quality checked

Questions, suggestions?