

de-novo Transcriptome Assembly

Ecole EB3I n1 - 2025

Erwan CORRE

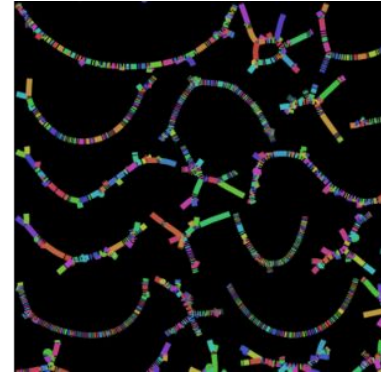
<https://orcid.org/0000-0001-6354-2278>

ABiMS – Station Biologique Roscoff

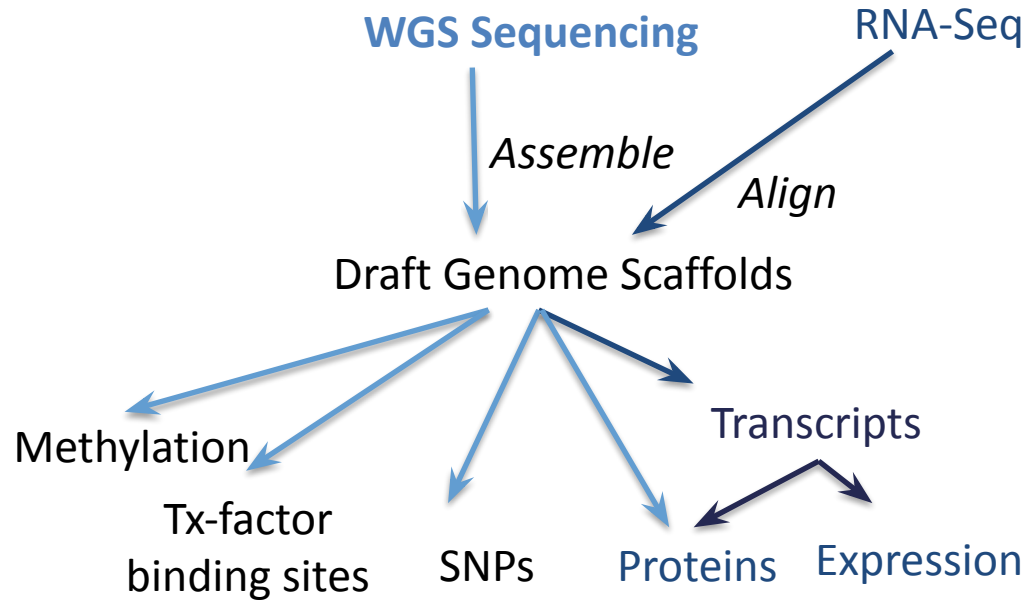
Mathieu GENETE

<https://orcid.org/0000-0002-9640-8793>

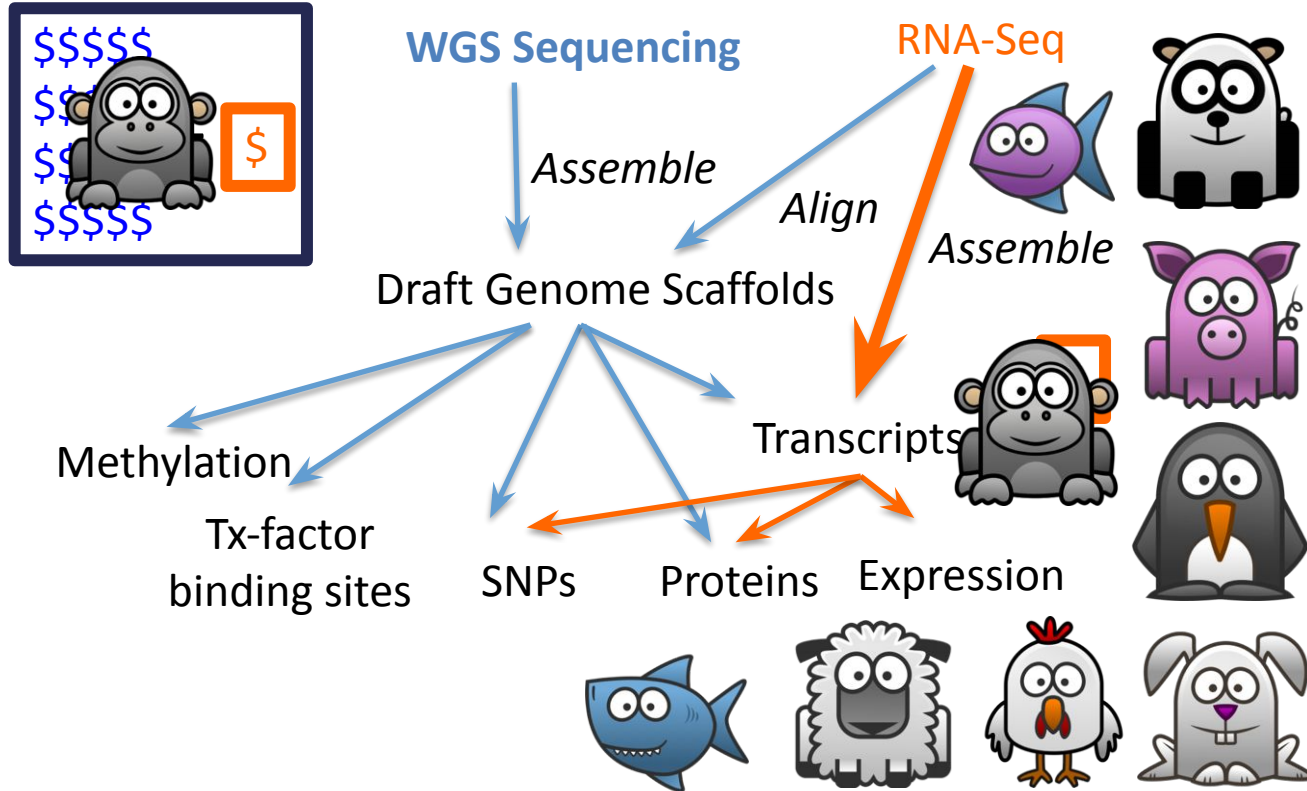
Unité EEP – UMR 8198 Lille



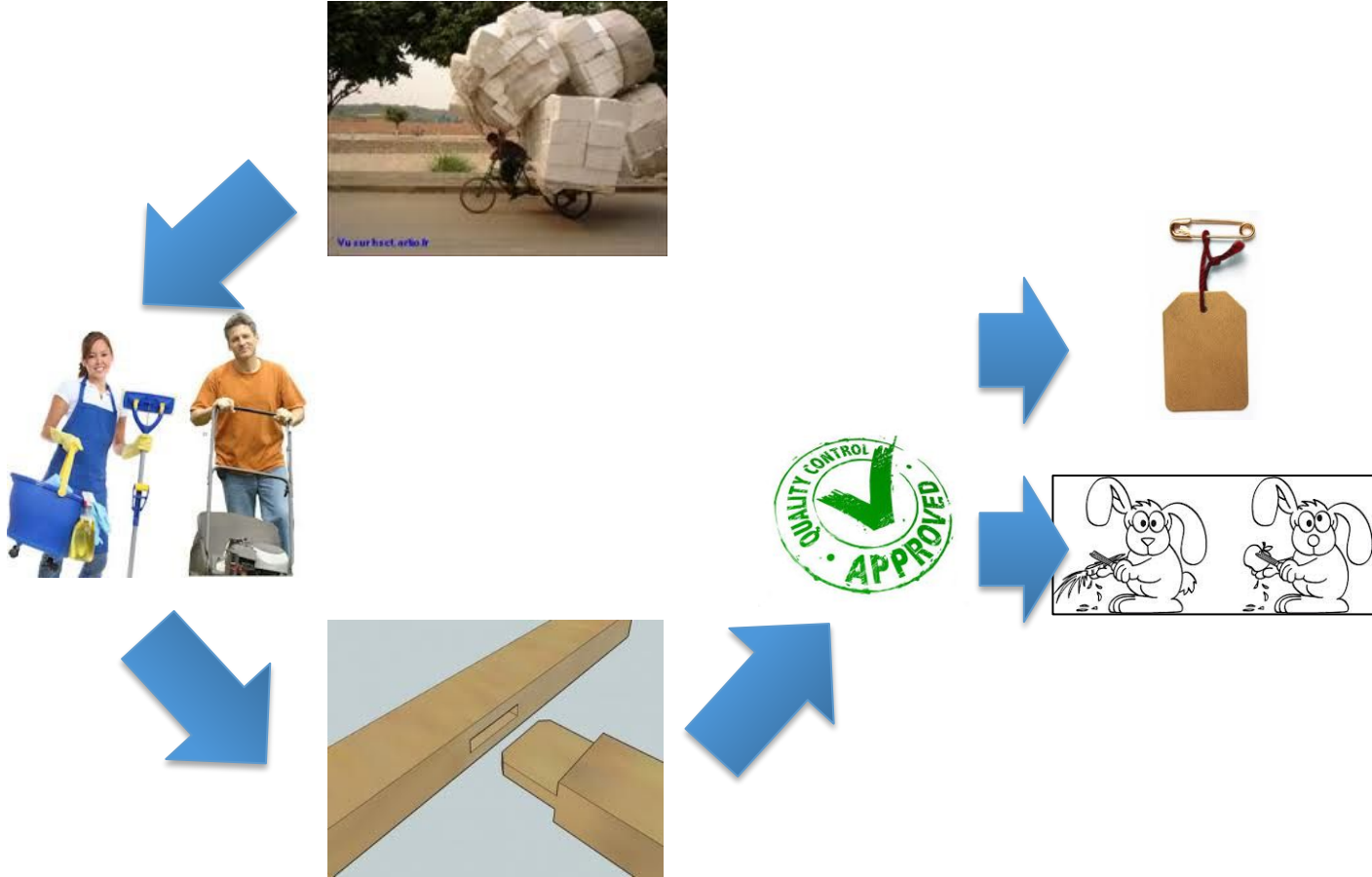
A Paradigm for Genomic Research



A Maturing Paradigm for Transcriptome Research



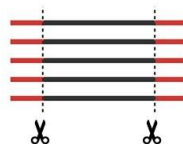
RNA Seq de novo analysis workflow



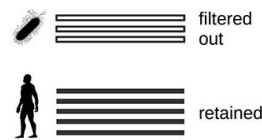
Data cleaning



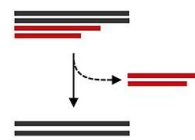
Adapter trimming



Contaminant removal



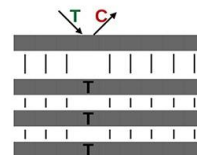
Discarding short reads



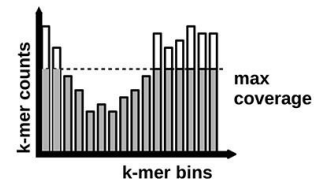
Erroneous read removal



Read correction



In silico read normalization

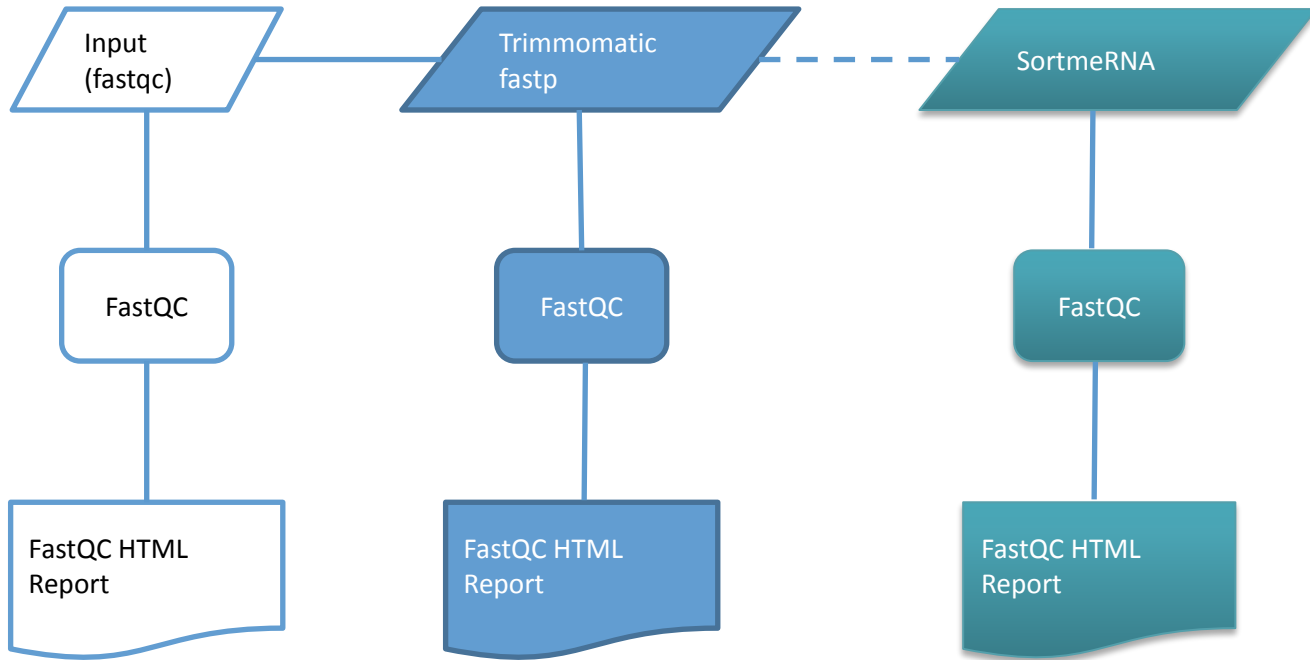


<https://doi.org/10.1093/bib/bbab563>

Bias should be corrected in reverse order of their generation

1. Sequencing biases (bad quality, unknowns)
2. Library preparation
 - Adaptors and primers sequences
 - Poly A/T tails
3. Biological sample (low complexity, rRNA, contaminants)

Data cleaning



Alternative : Sickle, TrimGalore

Alternative : phyloFlash, BBMap

Contaminations



Euphausia superba (Uwe Kils. 2011)



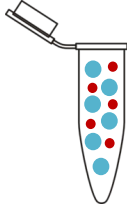
Contaminations



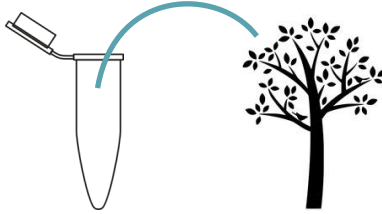
Euphausia superba (Uwe Kils. 2011)



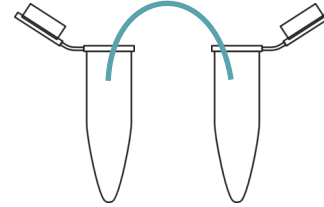
Contaminations



in-contamination
for ex. rRNA



third-party contamination
for ex. food - parasite



cross-contamination
for ex. experiment

- Most of (all) Illumina sequencing dataset are somewhat contaminated
- Illumina sequencing is especially susceptible to contamination due to the coverage depth
- It seems inherent to the method
- “Index misassignment between multiplexed libraries is a known issue” (Illumina, Inc., 2018); it potentially can produce contaminations in the sequenced datasets

Contaminations

Genome Biology

[Home](#) | [About](#) | [Articles](#) | [Submission Guidelines](#)

[Method](#) | [Open Access](#) | [Published: 12 May 2020](#)

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

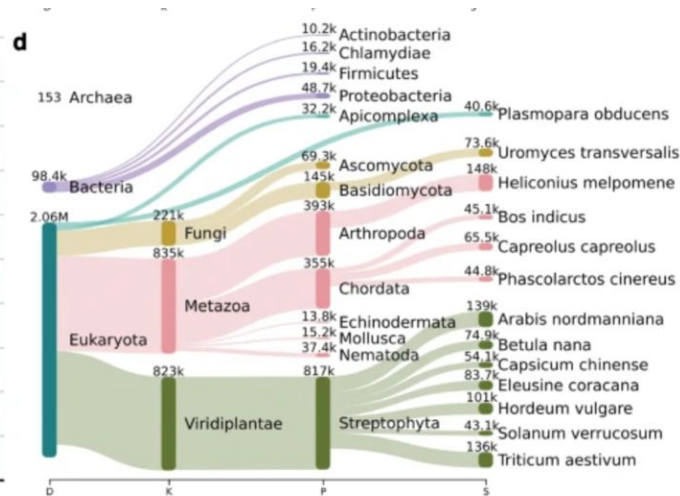
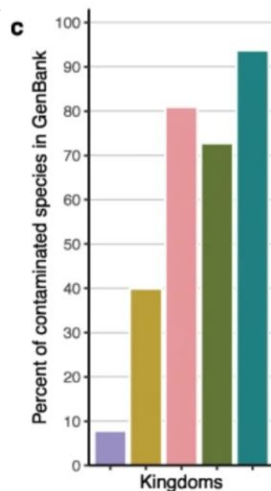
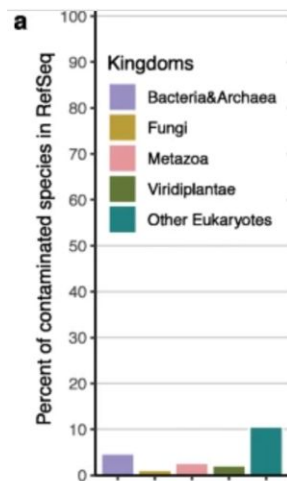
[Martin Steinegger](#) & [Steven L. Salzberg](#)

[Genome Biology](#) 21, Article number: 115 (2020) | [Cite this article](#)

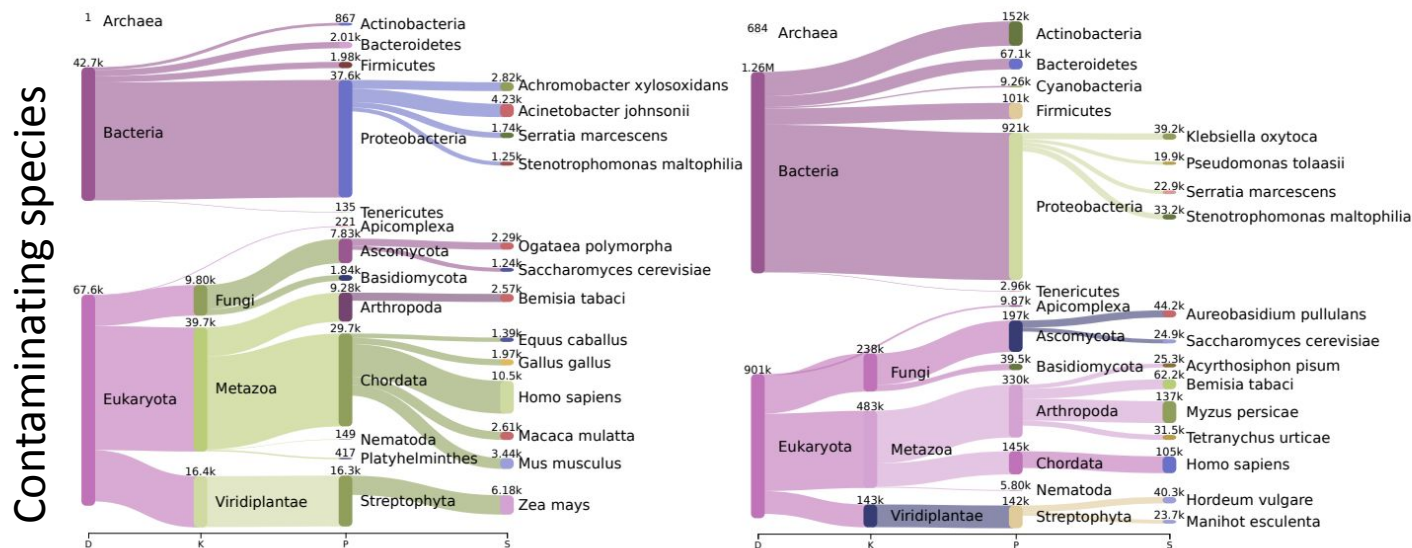
6825 Accesses | 32 Citations | 82 Altmetric | [Metrics](#)

« Conterminator reported 114,035 and 2,161,746 contaminated sequences affecting 2767 and 6795 species in RefSeq and GenBank, respectively »

Contaminated sequences



Contaminations



Supplementary Figure 1: Sanksey plot of most contaminating species in RefSeq and GenBank. left
 Sanksey plot five kingdoms: Bacteria&Archaea, Fungi, Metazoa, Viridiplantae and other Eukaryotes. right
 Distribution of contaminating species in GenBank.

Steinegger, M., Salzberg, S.L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**, 115 (2020). <https://doi.org/10.1186/s13059-020-02023-1>

rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

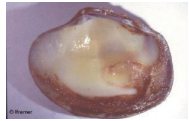
Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or Aliens

rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

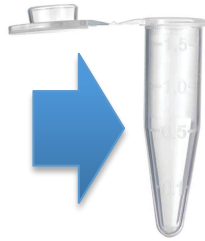
Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or Aliens



Ruditapes philippinaru



Vibrio tapetis

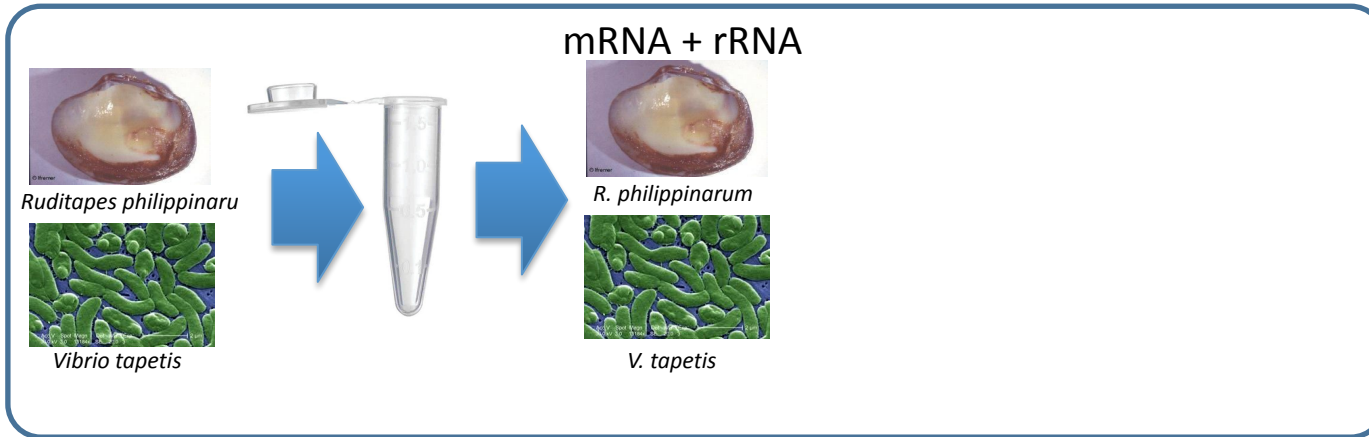


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or Aliens

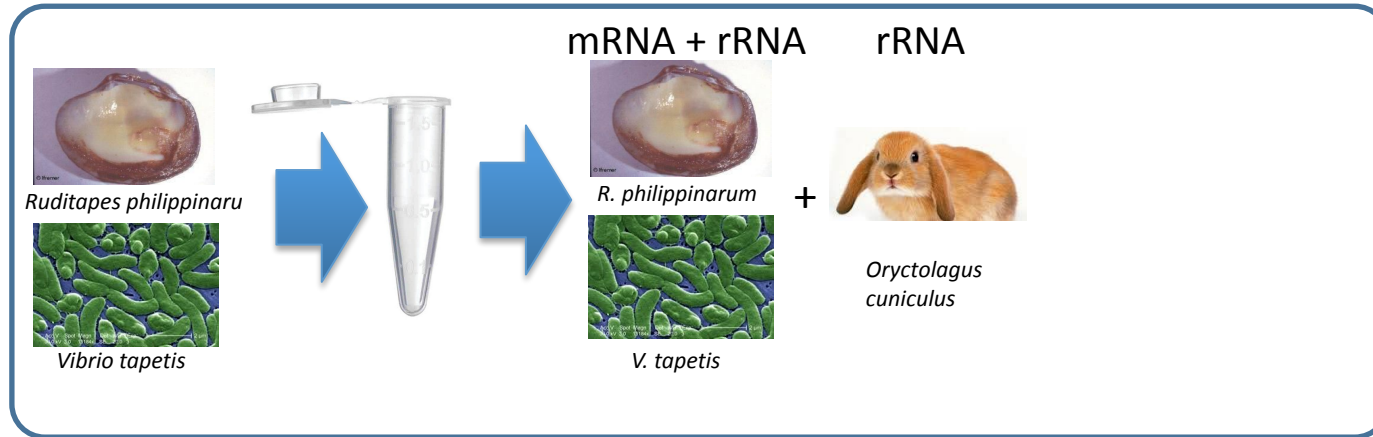


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or Aliens

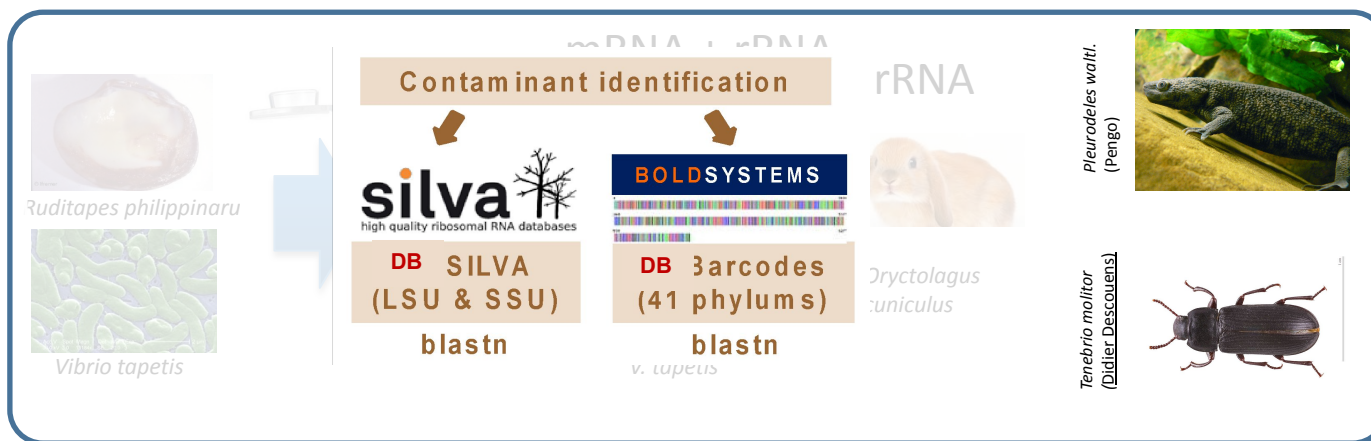


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or Aliens



Detect third-party/cross contamination

Reads that do not originate from the organism and/or RNA species of interest

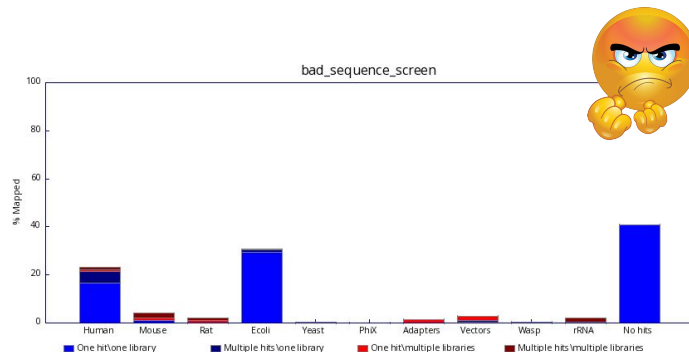
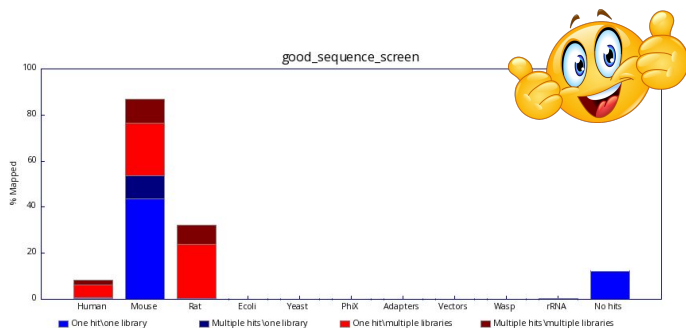
f. ex. reads originating from an endosymbiont bacterium in an eukaryote organism of interest

Tools : short-read taxonomic classifiers

With a eukaryotic read dataset, **kraken2** could be used to exclude reads classified as bacterial, archaeal, fungal or from plants.

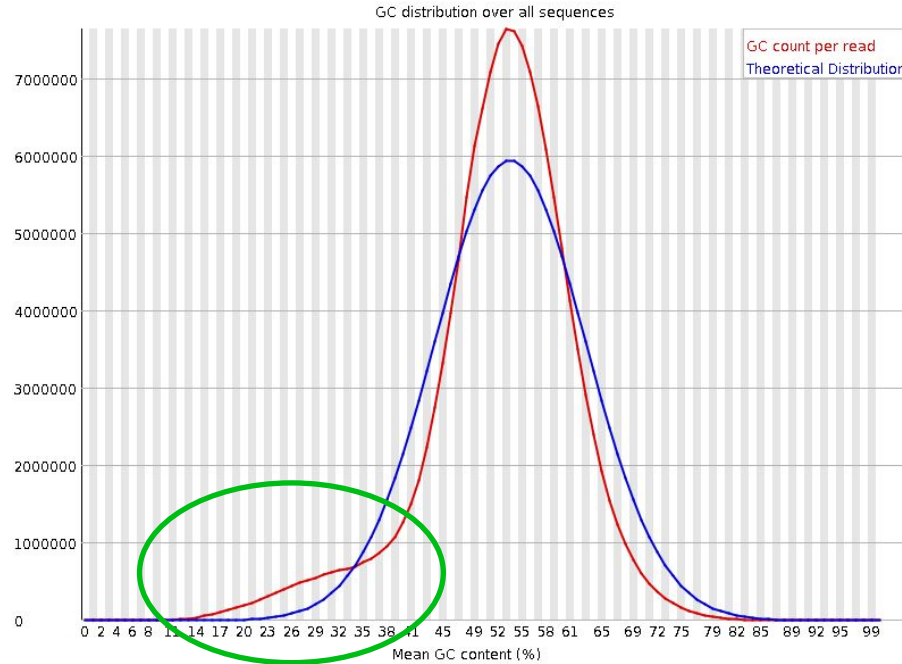
Alternative **Centrifuge** for microbial reads

screen-only alternative **FastQ Screen**



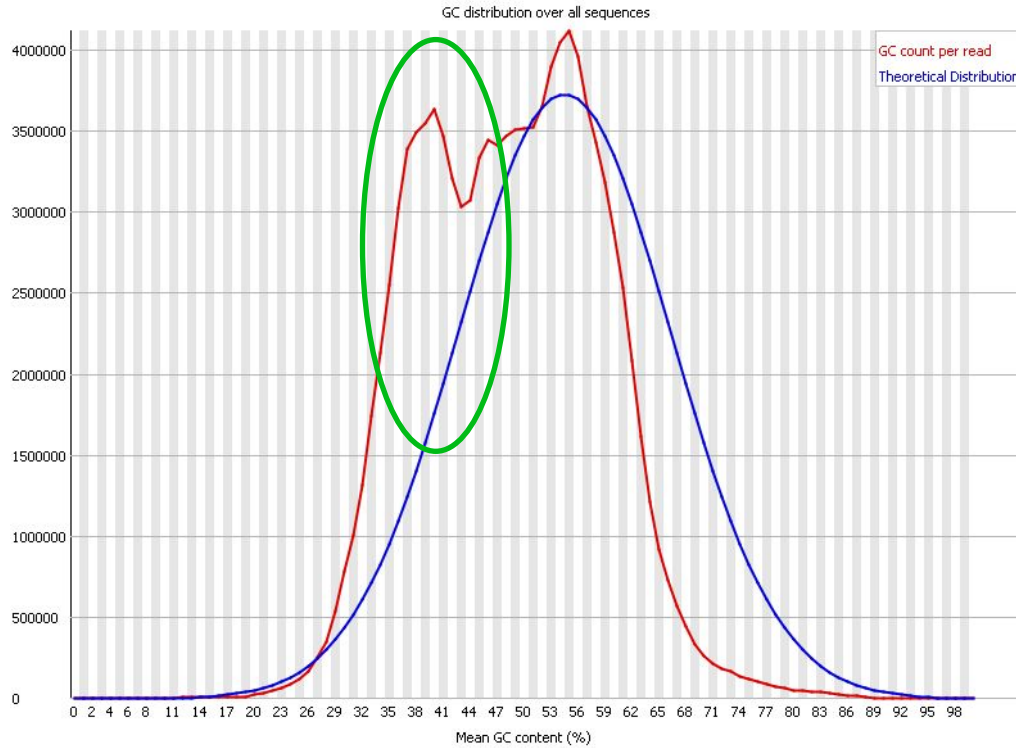
FastQC: Per sequence GC content

A contamination ?



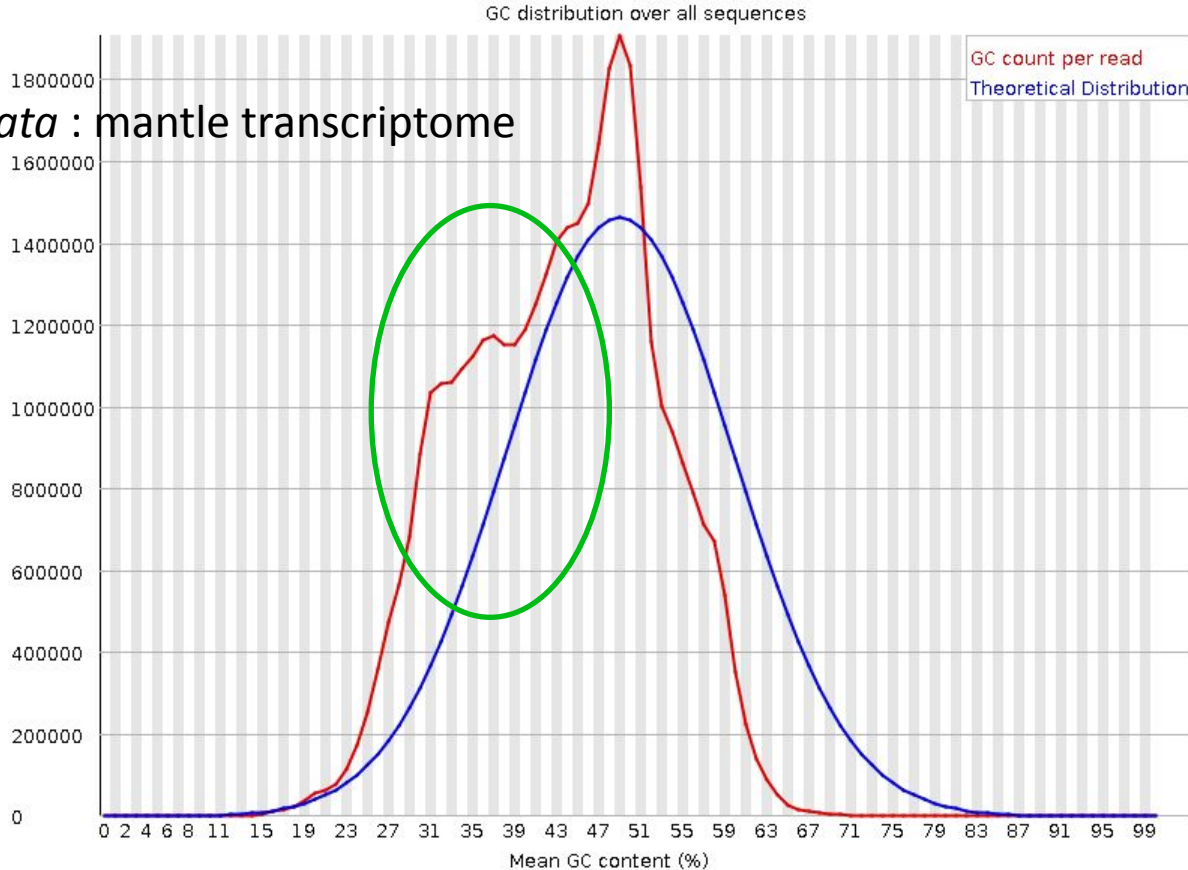
Can this be fixed ? Maybe...

FastQC: Per sequence GC content



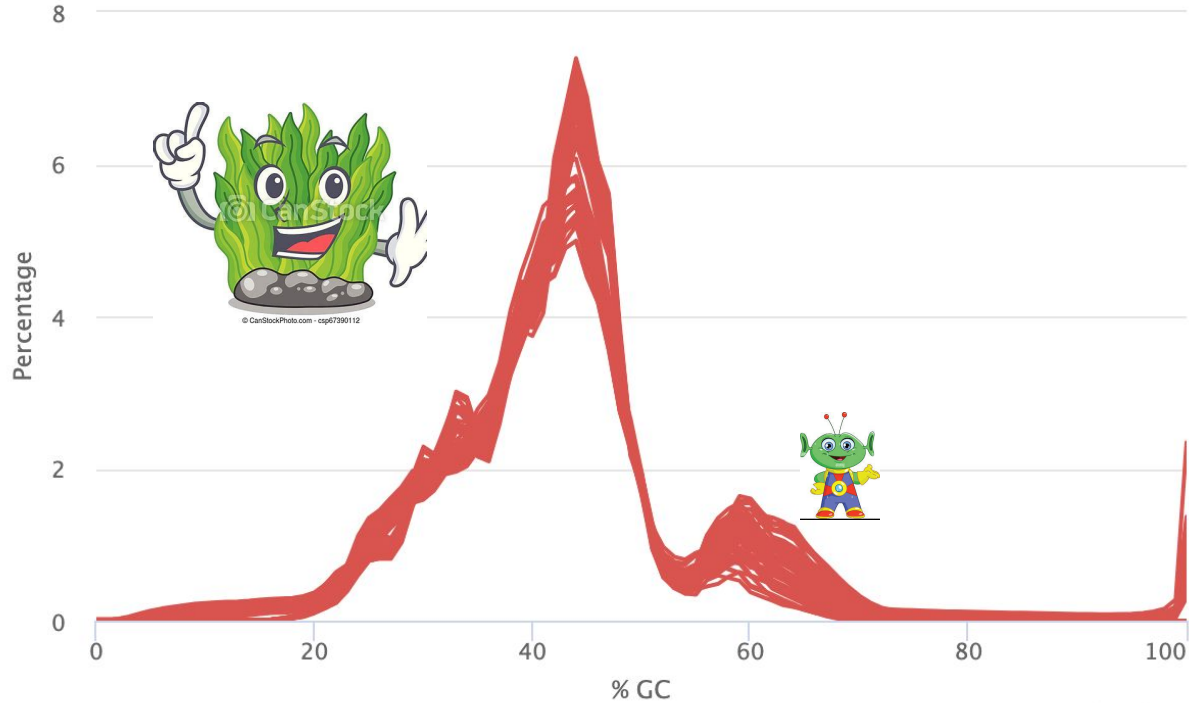
Third-party contamination : detection

Sabellaria alveolata : mantle transcriptome



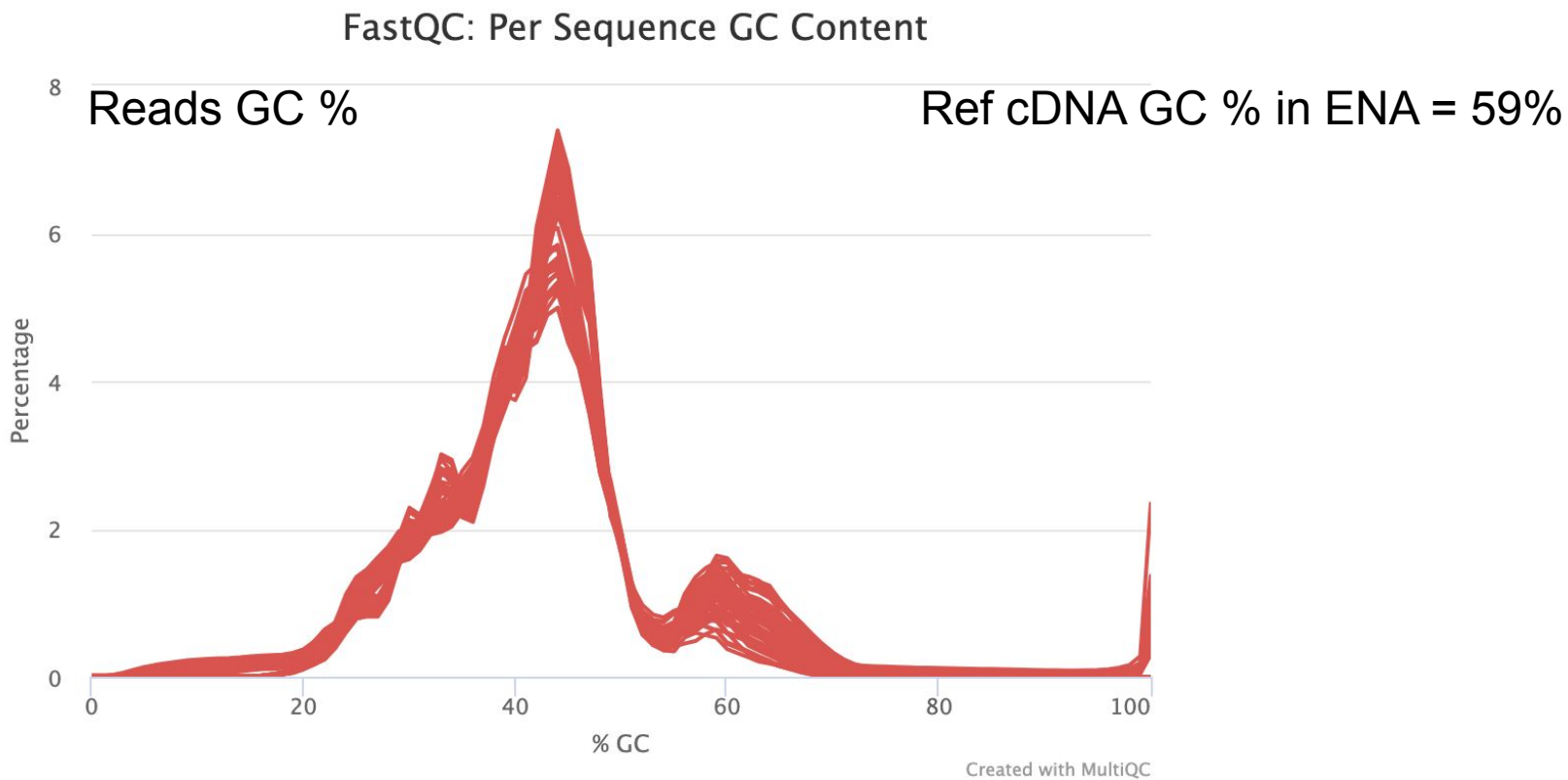
Third-party contamination : detection

FastQC: Per Sequence GC Content



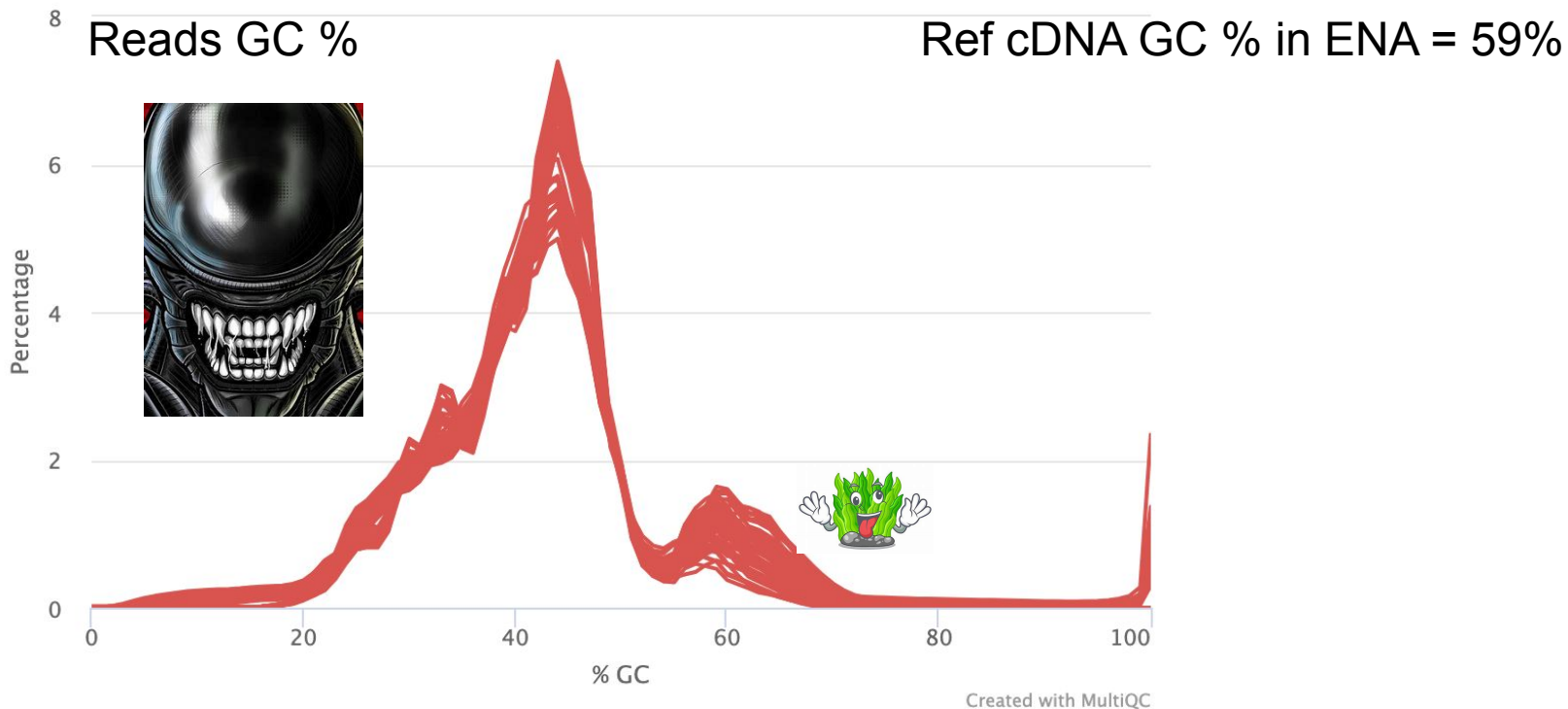
Created with MultiQC

Third-party contamination : detection



Third-party contamination : detection

FastQC: Per Sequence GC Content



Removing rRNA contamination

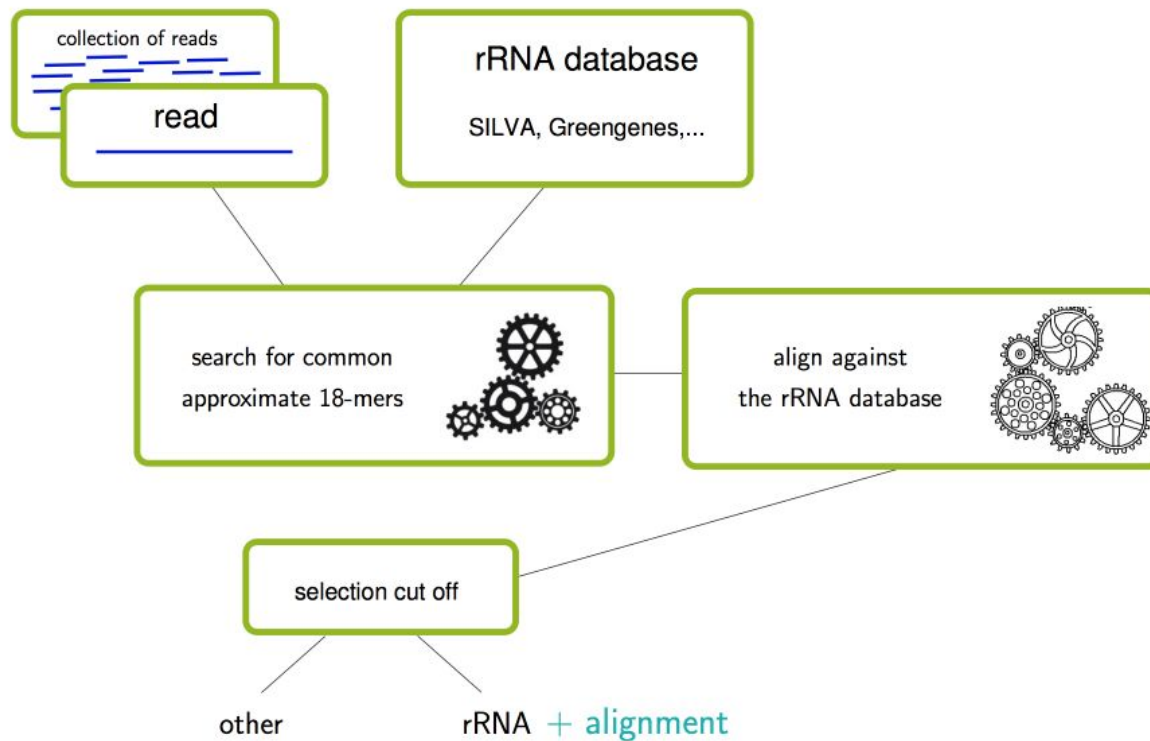
Prior to sequencing :

- Ribodepletion kits
- Selection polyA

After sequencing :

- Remove rRNA reads from raw reads (ex. SortMe RNA)
- Detect rRNA transcripts in the assembly (RNAmmer)

SortMeRNA



Selected for *de novo* assembly

<https://github.com/sortmerna>

Detect rRNA transcripts : RNAMMER

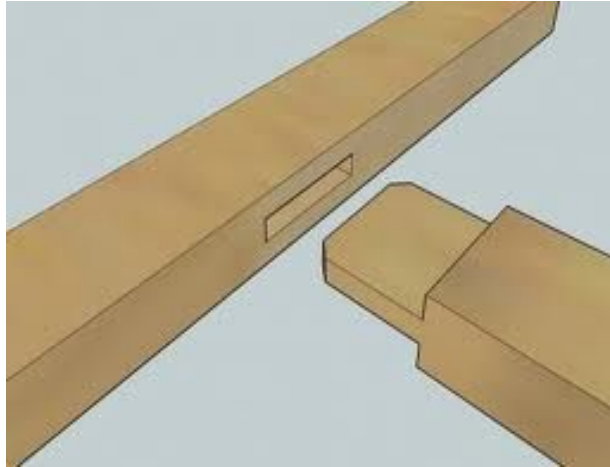


The program uses hidden Markov models trained on data from the 5S ribosomal RNA database and the European ribosomal RNA database project

Alternative Barnap : *B*ASIC *R*APID *R*IBOSOMAL RNA *P*REDICTOR

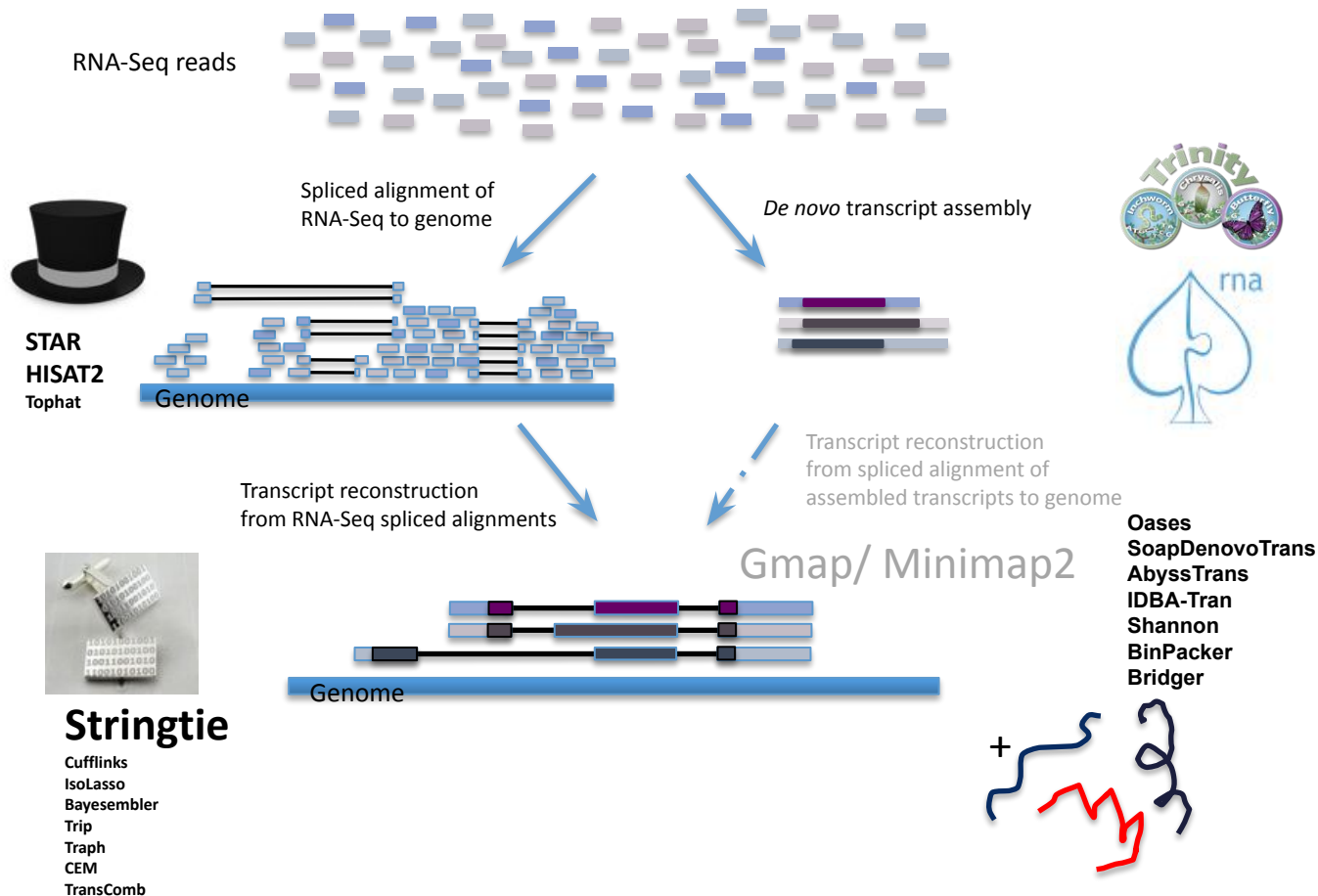
<https://github.com/tseemann/barnap>

```
# -----
##gff-version2##source-version RNAmmer-1.2##date 2009-11-16
##Type DNA
# seqname      source  feature start    end    score  +/-    frame  attribute
# -----
AE000511      RNAmmer-1.2  rRNA    448462  448577  49.2    +      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA    1473564 1473679 49.2    -      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA    1045067 1045183 40.3    +      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA    445339  448223  3056.5  +      .      23s_rRNA
AE000511      RNAmmer-1.2  rRNA    1473918 1476803 3032.8  -      .      23s_rRNA
AE000511      RNAmmer-1.2  rRNA    1207586 1209074 1801.4  -      .      16s_rRNA
AE000511      RNAmmer-1.2  rRNA    1511140 1512627 1803.6  -      .      16s_rRNA
```

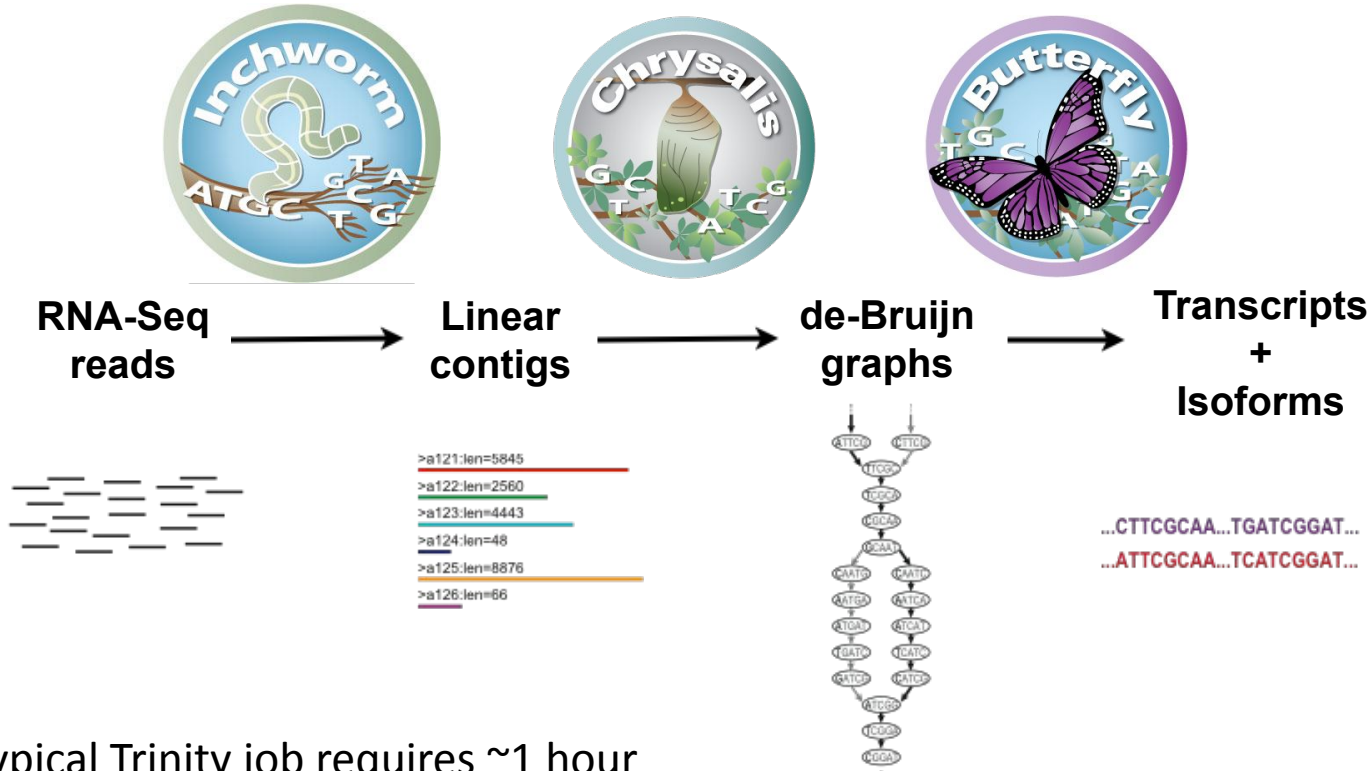


TRANSCRIPTOME ASSEMBLY STRATEGIES

Contemporary strategies for transcript reconstruction from RNA-Seq



Trinity – How it works:



Running a typical Trinity job requires ~1 hour and ~1G RAM per ~1 million PE reads.

Thousands of disjoint graphs

Results

Result: linear sequences grouped in *components*, *contigs* and sequences

```
>TRINITY_DN889_c0_g1_i1 len=259 path=[473:0-258] [-1, 473, -2]
GAACAATGTCTACACTGTCTTCAACTTGGATGACAAGGAACCTTCATTGGCTCAAGCTAA
CTACAATTCATCTCTGAAACCAGATATTGAAGAAATCAAGGATACTGTCCCTAGCGCTGT
GCTGGCTCCACAATACTACAACACATTCTCAGCTGACCCAACTGCCACTGCAGTCACTGG
TAACATCTTTGCACCAGAGGCCACTATGTCCATGGCTGCTCCAGCTAATGCTTCTAGAAA
CTCTTCATTAAACTCTCCT
```

```
>TRINITY_DN810_c0_g1_i2 len=226 path=[407:0-225] [-1, 407, -2]
GATGATATCAACAATGAGACTTGTGAACCAGGTGAAGAAAACCTCTTTCTTTGTATGCGAC
CTAGGTGAAATTGAAAGATTGTACGCTAACTGGTGGAAAGAACTACCAAGAGTTCAGCCA
TTTTACGCTGTCAAGTGTAACCCAGATTTGAAGATAATAAGAAAATTGGCTGACCTCGGA
```

TRINITY_DNW|cX_gY_iZ (until release 2.0 cX_gY_iZ previously compX_cY_seqZ)

TRINITY_DNW|cX defines the graphical component generated by Chrysalis (from clustering inchworm contigs).

Butterfly might tease subgraphs apart from each other within a single component, based on the read support data. This gives rise to subgraphs (gY): trinity genes

Each subgraph then gives rise to path sequences (iZ): trinity isoforms

(path) list of vertices in the compacted graph that represent the final transcript sequence and the range within the given assembled sequence that those nodes correspond to.

rnaSPAdes usage and results



Derived from SPAdes is a versatile toolkit designed for assembly and analysis of sequencing data. SPAdes allows to assemble genomes, metagenomes, transcriptomes, viral genomes etc.

Output

- `transcripts.fasta` – default assembly
- `hard_filtered_transcripts.fasta` – includes only long and reliable transcripts with rather high expression.
- `soft_filtered_transcripts.fasta` – includes short and low-expressed transcripts, likely to contain junk sequences.

Thus, each contig name has

- individual unique id (`NODE_XX`)
- length (`length_XXXX`)
- kmer coverage (`cov_XXXX`): (is always lower than the read (per-base) coverage)

Distinct isoforms will have different unique ids, and may have different coverage and length as well. The only thing they share is the gene id (given after `_g`). For example:

```
NODE_217_length_5488_cov_28.472321_g133_i0
```

```
NODE_260_length_5302_cov_30.767137_g133_i1
```

Trinity – rnaSPAdes : Pro - Cons



Trinity

- + old (large community of users. Brian Hass)
- + Very good documentation
- + Complete: many features.
- + Ability to assemble long and complex transcripts,
- + management of multiple isoforms.
- Slower and more fragmented assemblies



rnaSPAdes

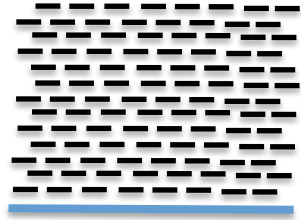
- + speed
- + precision
- + especially for data with poor coverage or heterogeneous samples.
- + wider phylogenetic range
- + include long read support
- Less rare isoforms

In silico normalization of reads

- By definition RNAseq display a wide range of expressions
Very low expressed ☐ Very highly expressed transcripts
- The information given by reads from high expression transcripts is redundant, and very high coverage also brings more sequencing errors
- De-novo assemblers do not benefit from coverage increase beyond a certain point (> 200 millions reads) , and fewer data means quicker assemblies
- ☐ How to decrease coverage of highly expressed transcripts without decreasing that of low expressed transcripts ?

In silico normalization of reads

High



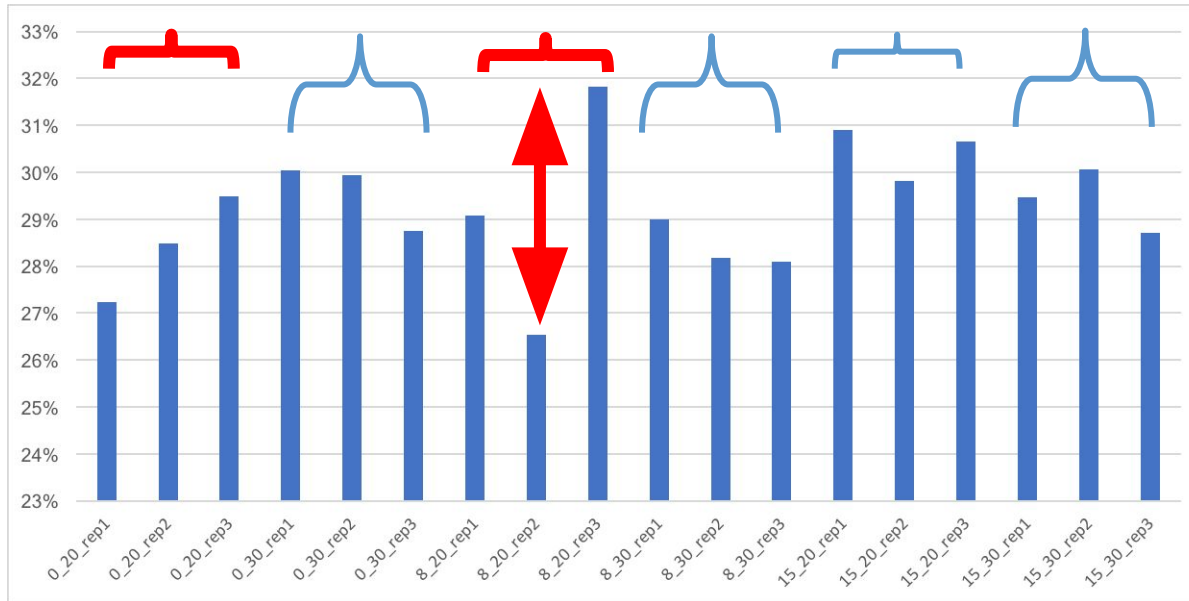
Moderate



Low



Normalisation prior to assembly : replicats congruence



Reads *in silico* normalisation

Tools that can perform *in silico* read normalization

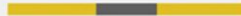
- Trinity assembler also offers in-built *in silico* normalization
- khmer – <https://github.com/dib-lab/khmer> (using the diginorm algorithm)
- Bignorm – <https://git.informatik.uni-kiel.de/axw/Bignorm>
- NeatFreq – <https://github.com/bioh4x/NeatFreq>
- ORNA – <https://github.com/SchulzLab/ORNA>



Transcriptome assembly

ASSEMBLY QUALITY ASSESSMENT AND CLEANING

De novo Transcriptome Assembly is Prone to Certain Types of Errors

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	no reads align to insertion 
Incompleteness	 n=1	 n=1	read pairs align off end of contig 
Fragmentation	 n=1	 n=4	bridging read pairs 
Local misassembly	 n=1	 n=1	read pairs in wrong orientation 
Redundancy	 n=1	 n=3	all reads assign to best contig 

Assembly quality assessment

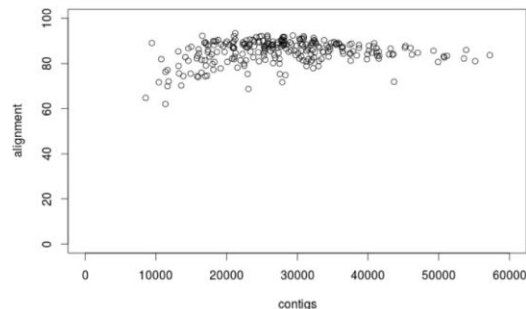
- Generating general Assembly metrics
- Comparing the assembled sequences to the reads used to generate them (reference-free)
- Comparing the assembled sequences
 - to catalogue of orthologous genes
 - to conserved gene domains
 - to transcriptomes or genomes of closely related species.

- The number of contigs in the assembly
- The size of the smallest contig
- The size of the largest contig
- The number of bases included in the assembly
- The mean length of the contigs
- The number of contigs <200 bases
- The number of contigs >1,000 bases
- The number of contigs >10,000 bases
- The number of contigs that had an open reading frame
- The mean % of the contig covered by the ORF
- NX (e.G. N50): the largest contig size at which at least X% of bases are contained in contigs at least this length
- % Of bases that are G or C
- GC skew

Realignment metrics

The realignment rate gives how much of the initial information is inside the contigs.

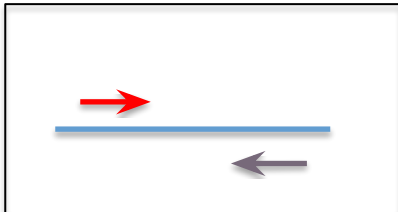
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are **properly paired**.



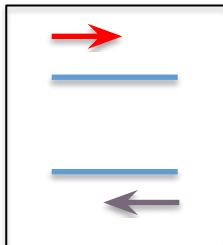
Given read pair:  

Possible mapping contexts in the Trinity assembly are reported:

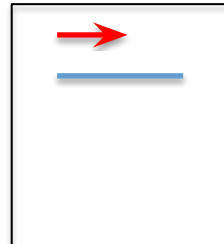
Proper pairs



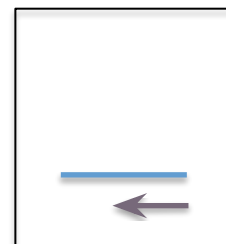
Improper pairs



Left only



Right only



Assembly evaluation : read remapping

Quality Assessment: Calculate the coverage and depth of each contig to identify transcripts that are well or poorly supported by the raw data.

Expression quantification: Estimate the level of expression (f.e. TPM) of each transcript.

Correction and Refinement: Identify regions of low coverage, sequence variations, or potential errors in the assembly.

Tools :

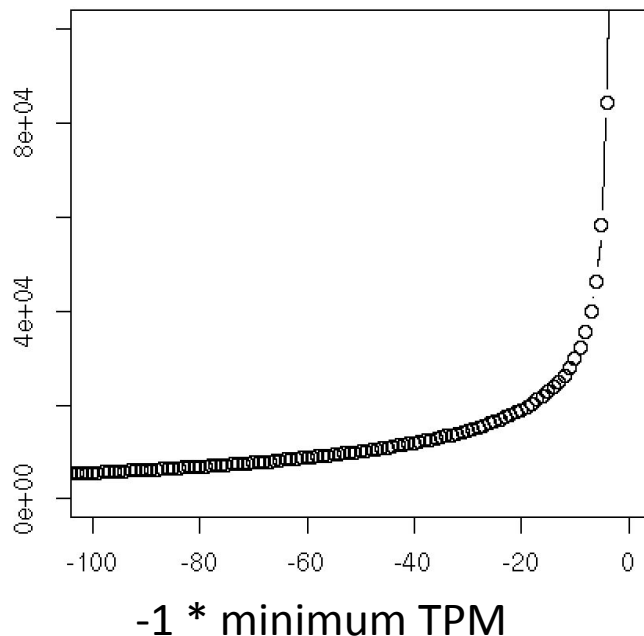
Unspliced (fast) aligners: Bowtie2 or BWA are fast and efficient for aligning short sequences to reference sequences.

K-mers-based aligners: Kallisto Or Salmon do not perform traditional base-by-base alignment, but use k-mer indexing (or near-alignments) approaches of reads on transcripts for fast and highly accurate quantification.

Alternative to N50 ?

Often, most assembled transcripts are ***very*** lowly expressed
 (How many 'transcripts & genes' are there really?)

Cumulative nb of
 Transcripts ordered
 by decreasing
 expression level



← 1.4 million Trinity transcripts

N50 ~ 500 bases

← 24k transcripts

Correspond to 90% of expression

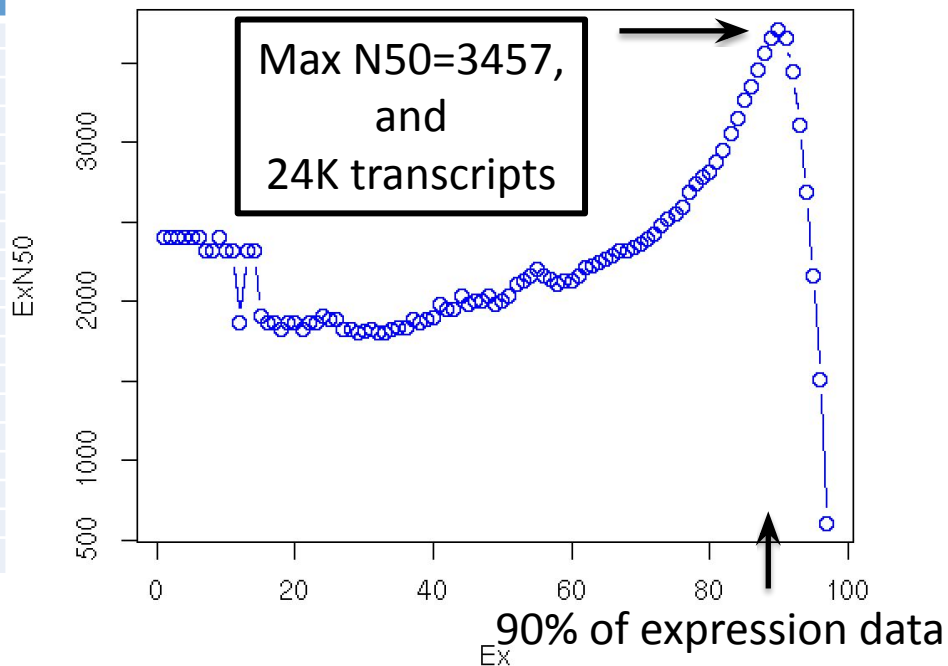
Expression

Alternative to N50 : ExN50 – E90N50

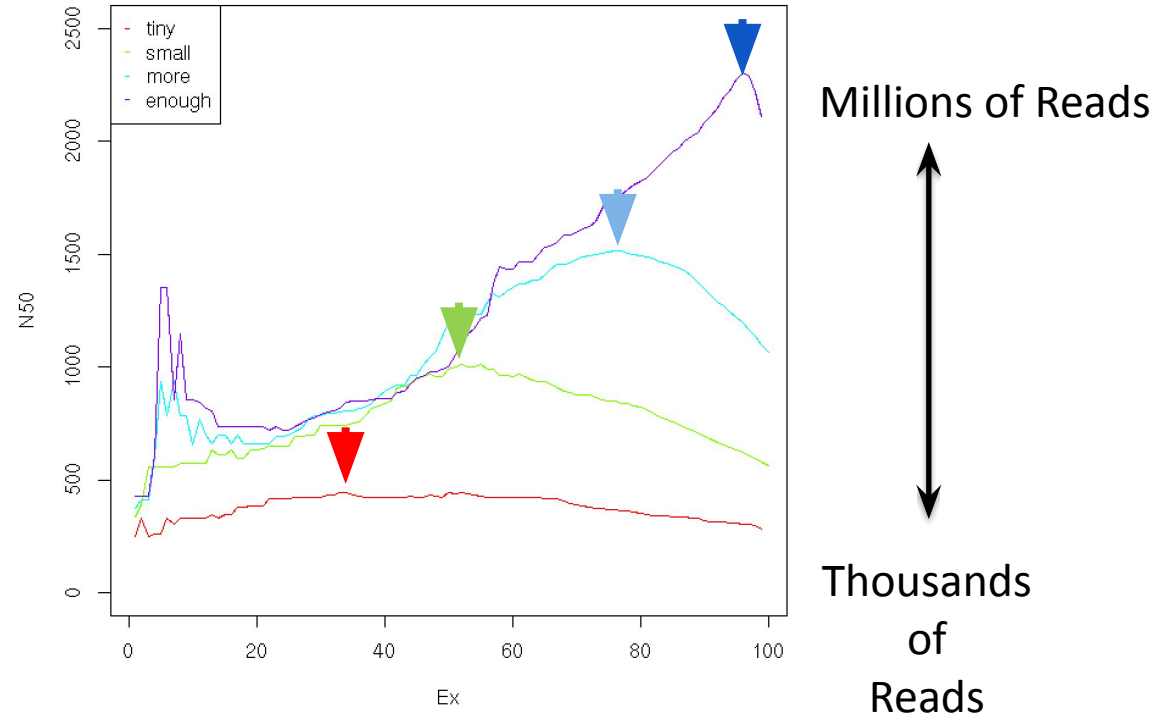
Compute N50 Based on the Top-most Highly Expressed Transcripts (ExN50)

- Sort contigs by expression value, descendingly.
- Compute N50 given minimum % total expression data thresholds => ExN50

#E	min expr	E-N50	num transcripts
E2	89129.251	2397	1
E3	89129.251	2397	2
E5	66030.692	2397	3
E6	66030.692	2397	4
E8	66030.692	2397	5
...			
E86	9.187	3056	12309
E87	7.044	3149	14261
E88	6.136	3261	16646
E89	4.538	3351	19635
E90	3.939	3457	23471
E91	3.077	3560	28583
E92	2.208	3655	35832
E93	1.287	3706	47061
...			
E97	0.235	2683	275376
E98	0.164	2163	428285
E99	0.128	1512	668589
E100	0	606	1554055



ExN50 Profiles for Different Trinity Assemblies Using Different Read Depths



Note shift in ExN50 profiles as you assemble more and more reads.

BUSCO analysis

BUSCO (<http://busco.ezlab.org/>)

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

BUSCO

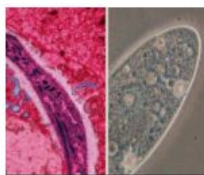
Datasets (Beta versions, updated sets and additional lineages coming soon)



Bacteria sets



Eukaryota sets



Protists sets



Metazoa sets



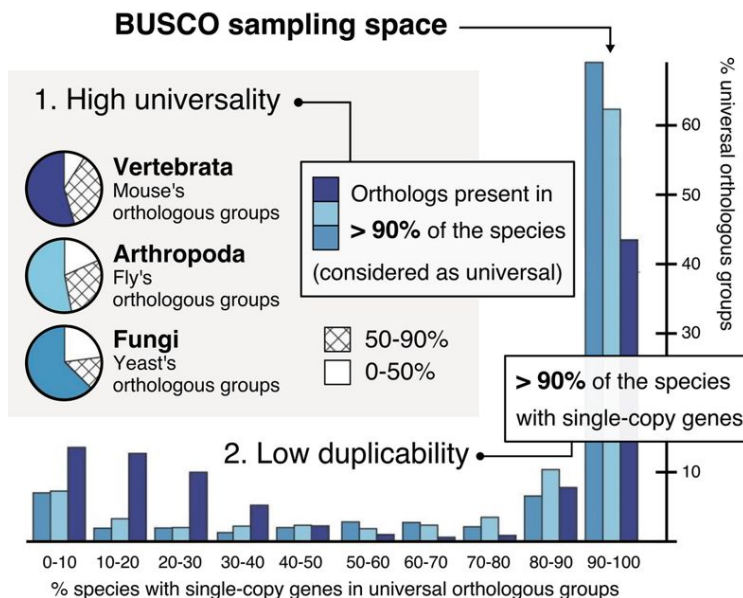
Fungi sets



Plants set

Arthropods: 
Vertebrates: 
Fungi: 
Bacteria: 
Metazoans:  &  & 
Eukaryotes:  &  &  & 

BUSCO analysis



193 lineages !!

acidobacteria_odb10.2024-01-08.tar.gz
 aconoidasida_odb10.2024-01-08.tar.gz
 actinobacteria_class_odb10.2024-01-08.tar.gz
 actinobacteria_phylum_odb10.2024-01-08.tar.gz
 actinopterygii_odb10.2024-01-08.tar.gz
 agaricales_odb10.2024-01-08.tar.gz
 agaricomycetes_odb10.2024-01-08.tar.gz
 alphabaculovirus_odb10.2024-01-08.tar.gz
 alphaherpesvirinae_odb10.2024-01-08.tar.gz
 alphaproteobacteria_odb10.2024-01-08.tar.gz
 alteromonadales_odb10.2024-01-08.tar.gz
 alveolata_odb10.2024-01-08.tar.gz
 apicomplexa_odb10.2024-01-08.tar.gz
 aquificae_odb10.2024-01-08.tar.gz
 arachnida_odb10.2024-01-08.tar.gz
 archaea_odb10.2024-01-08.tar.gz
 arthropoda_odb10.2024-01-08.tar.gz
 ascomycota_odb10.2024-01-08.tar.gz
 aves_odb10.2024-01-08.tar.gz
 aviadenovirus_odb10.2024-01-08.tar.gz
 bacillales_odb10.2024-01-08.tar.gz
 bacilli_odb10.2024-01-08.tar.gz
 bacteria_odb10.2024-01-08.tar.gz
 bacteroidales_odb10.2024-01-08.tar.gz
 bacteroidetes-chlorobi_group_odb10.2024-01-08.
 tar.gz

.....

BUSCO Results

```
# BUSCO version is: 5.3.2
# The lineage dataset is: eukaryota_odb10 (Creation date: 2020-09-10, number of genomes: 70,
number of BUSCOs: 255)
# BUSCO was run in mode: transcriptome
```

***** Results: *****

```
C:97.2%[S:57.6%,D:39.6%],F:2.4%,M:0.4%,n:255
248      Complete BUSCOs (C)
147      Complete and single-copy BUSCOs (S)
101      Complete and duplicated BUSCOs (D)
6        Fragmented BUSCOs (F)
1        Missing BUSCOs (M)
255      Total BUSCO groups searched
```

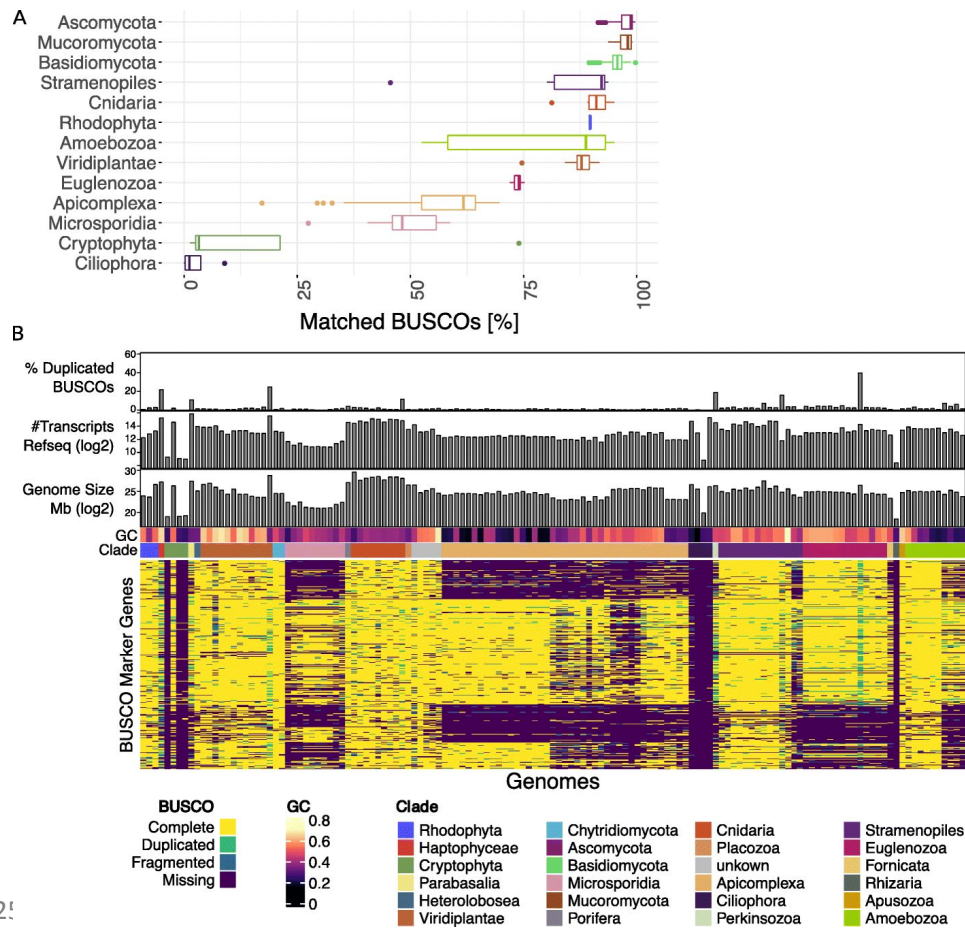
```
# BUSCO version is: 5.3.2
# The lineage dataset is: arthropoda_odb10 (Creation date: 2020-09-10, number of genomes: 90,
number of BUSCOs: 1013)
# BUSCO was run in mode: transcriptome
```

***** Results: *****

```
C:96.9%[S:56.8%,D:40.1%],F:1.8%,M:1.3%,n:1013
981      Complete BUSCOs (C)
575      Complete and single-copy BUSCOs (S)
406      Complete and duplicated BUSCOs (D)
18       Fragmented BUSCOs (F)
14       Missing BUSCOs (M)
1013     Total BUSCO groups searched
```

**BUSCO v6.0.0 is the current
stable version!**

BUSCO limitation



<https://github.com/Finn-Lab/EukCC/>

Saary, P., Mitchell, A.L. & Finn, R.D.
 Estimating the quality of eukaryotic
 genomes recovered from
 metagenomic analysis with
 EukCC. *Genome Biol* **21**, 244 (2020).
<https://doi.org/10.1186/s13059-020-2155-4>

BUSCO Alternative

Instead of scoring on the basis of conserved genes (BUSCO) , completeness is assessed on the basis of conserved protein domains (DOGMA).

The screenshot shows the DOGMA web interface. At the top, there's a navigation bar with links: DomainWorld, DOGMA webserver, Impressum & Disclaimer, Data protection, and Help. Below this is a header with the DOGMA logo and the title "Domain-based transcriptome and proteome quality assessment".

The main content area is divided into two sections:

- User Input:**
 - Upload your own file to be analyzed:
 - ☒ Choisir un fichier (Aucun fichier choisi)
 - or use one of the examples for a test:
 - ☐ Select example data (Show example)
 - ☐ translate (search for longest ORF in all six frames)
- Parameters:**
 - DOGMA mode: transcriptome (Use isoform free data for the proteome mode)
 - Plam version: Plam 35
 - Core set:

The core set defines the set of domains to search for. It is recommended to use the best fitting clade for best results.

eukaryotes
☒
vertebrates
☐
mammals
☐
arthropods
☐
insects
☐
plants
☐
eudicots
☐
monocots
☐
fungi
☐
bacteria
☐
archaea
☐

At the bottom, there is a green "Submit Job" button and a note: "The results will be available for 48 hours under the given webpage address." Below this, it says "How to cite:" followed by the citation: "Dohmen E, Kremer LPM, Bornberg-Bauer E and Kemena C, DOGMA: Domain-based transcriptome and proteome quality assessment, Bioinformatics 2016" and the URL "https://academic.oup.com/bioinformatics/article/32/17/2577/2450731".

<https://domainworld.uni-muenster.de/>

de novo assembled contigs include transcriptional artifacts,

- pre-mRNA and ncRNA in addition to the protein-coding transcripts
- alternative splicing which manifests as transcript isoforms.

It may not always be necessary to retain all such sequences

Exclude transcripts that can be considered as being lowly expressed on the basis of abundance metrics such as TPM (e.g. TPM <1.00)

Clustering tool with combination of sequence identity and sequence coverage thresholds

- CD-HIT and MMSeqs2
 - the longest sequence in each cluster or the sequence with the most commonality
 - the longest isoform is not necessarily the most expressed (and vice versa).

Clustering + shared read support

Corset, Grouper or Compacta

Creation of SuperTranscripts : stitches all unique exons from the isoforms into a single, linear sequence.

New de novo transcriptome assemblers

- IDBA-Tran (Peng et al., Bioinf., 2014)
- IDBA-MTP (Peng et al., RECOMB 2014)
- SOAPdenovo-Trans (Xie et al., Bioinf., 2014)
- Fu et al., ICCABS, 2014
- StringTie (Pertea et al., Nat. Biotech., 2015)
- Bermuda (Tang et al., ACM, 2015)
- Bridger (Chang et al., Gen. Biol. 2015)
- BinPacker (Liu et al. PLOS Comp Biol, 2016)
- FRAMA (Bens M et al., BMC Genomics 2016)
- rnaSPAdes (Bushmanova et al., *GigaScience* 2019)
-
- Cstone (Linheiro and Archer, PLOS Comp Biol, 2021)

BinPacker - <https://github.com/macmanes-lab/BINPACKER>

Bridger - https://github.com/fmaguire/Bridger_Assembler

inGAP-CDG - <https://sourceforge.net/projects/ingap-cdg/>

DTA-SiST - <https://github.com/jzbio/DTA-SiST>

IDBA-tran - <https://github.com/loneknightpy/idba>

IsoTree - <https://github.com/david-cortes/isotree>

Oases - <https://github.com/dzerbino/oases>

RNA-Bloom - <https://github.com/bcgsc/RNA-Bloom>

rnaSPAdes - <https://github.com/ablab/spades>

SOAPdenovo-Trans - <https://github.com/aquaskyline/SOAPdenovo-Trans>

Trans-ABYSS - <https://github.com/bcgsc/transabyss>

TransLig - <https://sourceforge.net/projects/transcriptomeassembly/>

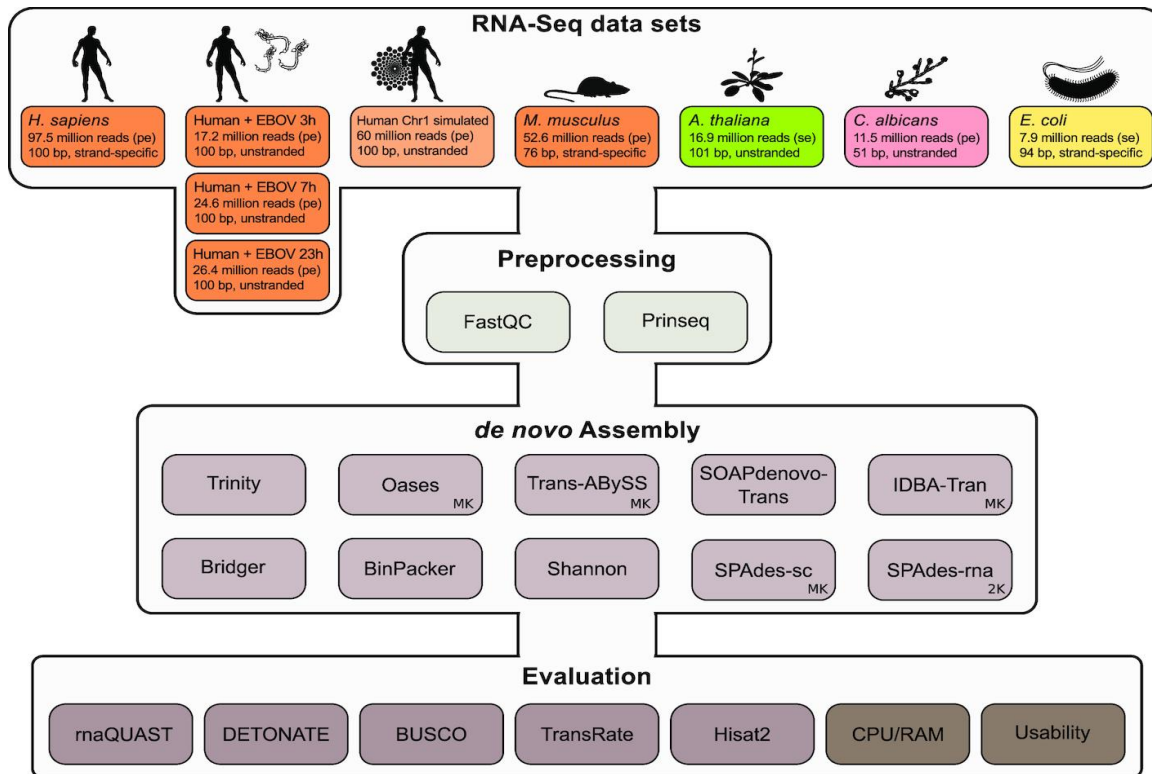
Trinity - <https://github.com/trinityrnaseq/trinityrnaseq>

Assemblers comparison



- Qiong-Yi Zhao et al., Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 2011, 12(Suppl 14):S2
- Clarke, K., Yang, Y., Marsh, R., Xie, L., & Zhang, K. K. (2013). Comparative analysis of de novo transcriptome assembly. Science China Life Sciences, 56(2), 156–162. doi:10.1007/s11427-013-4444-x
- (Vijay et al., 2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular ecology. PMID: 22998089
- (Haas et al., 2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols. PMID: 23845962
- (Lu et al., 2013) Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci China Life Sci.
- Chen, G., Yin, K., Wang, C., & Shi, T. (n.d.). De novo transcriptome assembly of RNA-Seq reads with different strategies. Science China Life Sciences, 54(12), 1129–1133. doi:10.1007/s11427-011-4256-9
- (He et al., 2015) Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. BMC genomics. DOI: 10.1186/s12864-014-1192-7
- S. B. Rana, F. J. Zadlock IV, Z. Zhang, W. R. Murphy, and C. S. Bentivegna, “Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, *Fundulus heteroclitus*,” *PLoS ONE*, vol. 11, no. 4, p. e0153104, Apr. 2016.
- (Wang and Gribskov, 2016) Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. PMID: 27694201
- M. Hölzer and M. Marz, “De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers,” *Gigascience*, vol. 8, no. 5, pp. 57–16, May 2019.
- Sadat-Hosseini et al. (2020) Combining independent *de novo* assemblies to optimize leaf transcriptome of Persian walnut. PLoS ONE 15(4): e0232005. <https://doi.org/10.1371/journal.pone.0232005>
- Jackson DJ, Cerveau N, Posnien N..De novo assembly of transcriptomes and differential gene expression analysis using short-read data from emerging model organisms - a brief guide. Front Zool. 2024 Jun 20;21(1):17. doi: 10.1186/s12983-024-00538-y.

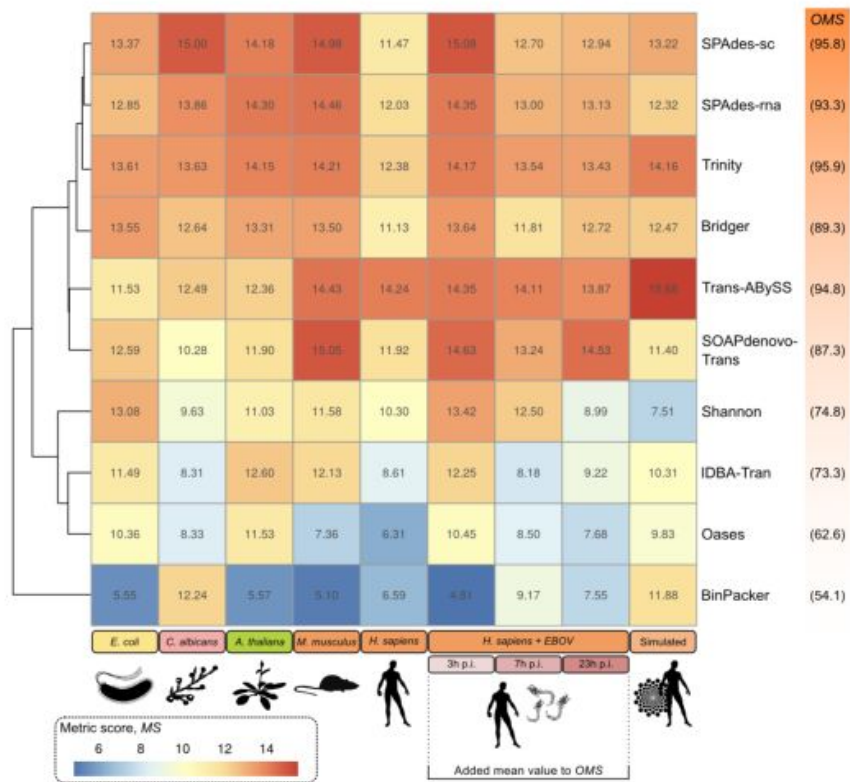
Assemblers comparison



GigaScience, Volume 8, Issue 5, May 2019, giz039, <https://doi.org/10.1093/gigascience/giz039>

The content of this slide may be subject to copyright: please see the slide notes for details.

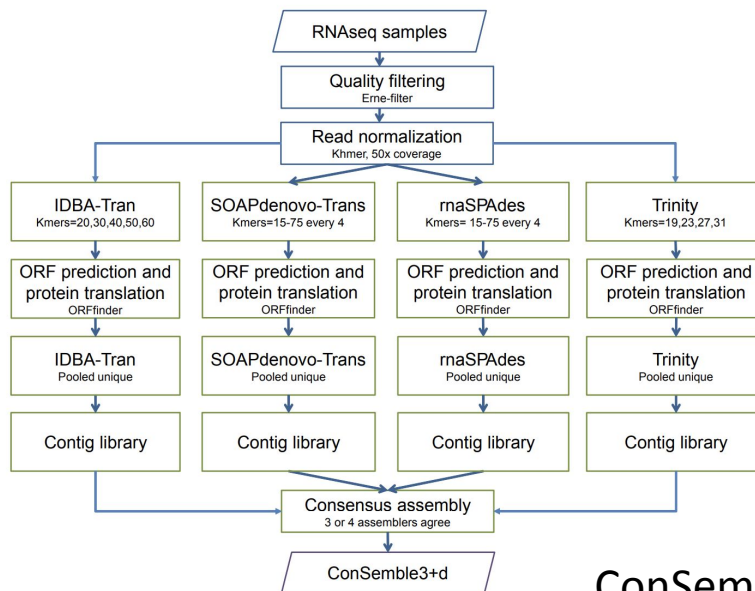
Assemblers comparison



GigaScience, Volume 8, Issue 5, May 2019, giz039, <https://doi.org/10.1093/gigascience/giz039>

The content of this slide may be subject to copyright: please see the slide notes for details.

New strategies



DRAP, EvidentialGene ,
Concatenation, ConSemble, TransPI,
Pincho.

Exploit the result of different
assemblers run in parallel and
choose the best solution

ConSemble

- Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. Compacting and correcting Trinity and Oases RNA-Seq *de novo* assemblies. PeerJ. 2017 Feb 16;5:e2988. doi: 10.7717/peerj.2988. PMID: 28224052; PMCID: PMC5316280.
- Gilbert DG. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ. 2019;7:e6374.
- Cerveau N, Jackson DJ. Combining independent *de novo* assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinform. 2016;17(1):525.
- Voshall, A., Behera, S., Li, X. *et al.* A consensus-based ensemble approach to improve transcriptome assembly. *BMC Bioinformatics* 22, 513 (2021). <https://doi.org/10.1186/s12859-021-04434-8>
- R.E. Rivera-Vicéns, C.A. Garcia-Escudero, N. Conci, M. Eitel, G. Wörheide. TransPi – a comprehensive TRanscriptome ANalysis Pipeline for *de novo* transcriptome assembly. doi: <https://doi.org/10.1101/2021.02.18.431773>