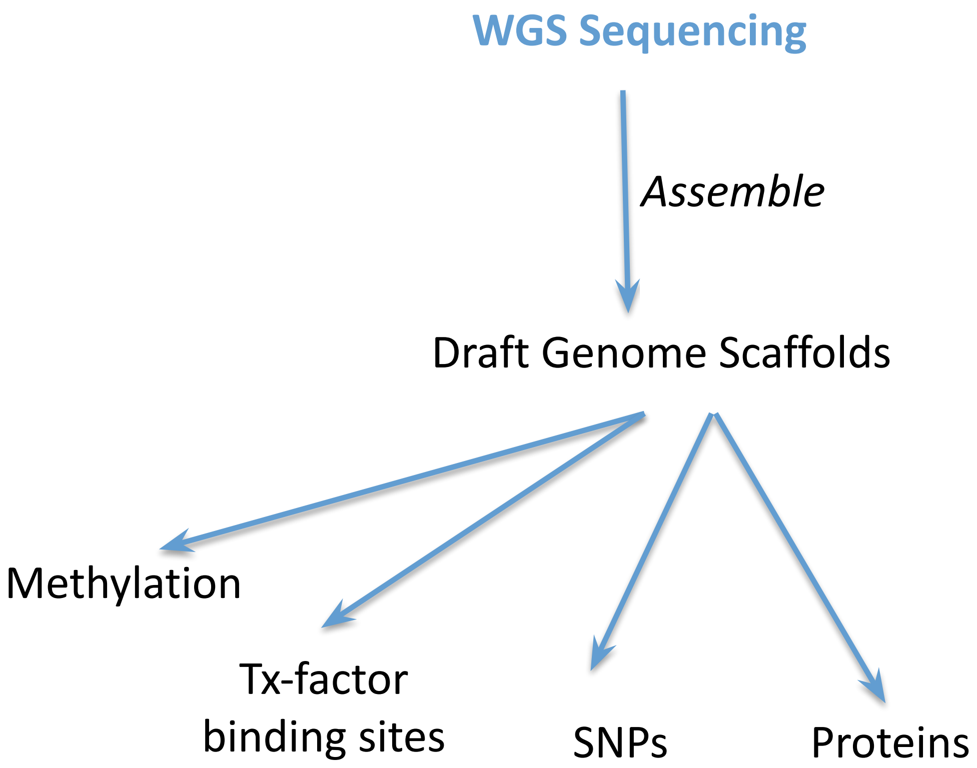
08/11/2024

Transcriptome *de-novo* Assembly

Ecole EBAII 2024

Erwan CORRE
https://orcid.org/0000-0001-6354-2278
ABiMS – Station Biologique Roscoff

A Paradigm for Genomic Research



WGS Sequencing

Assemble

Draft Genome Scaffolds

Methylation

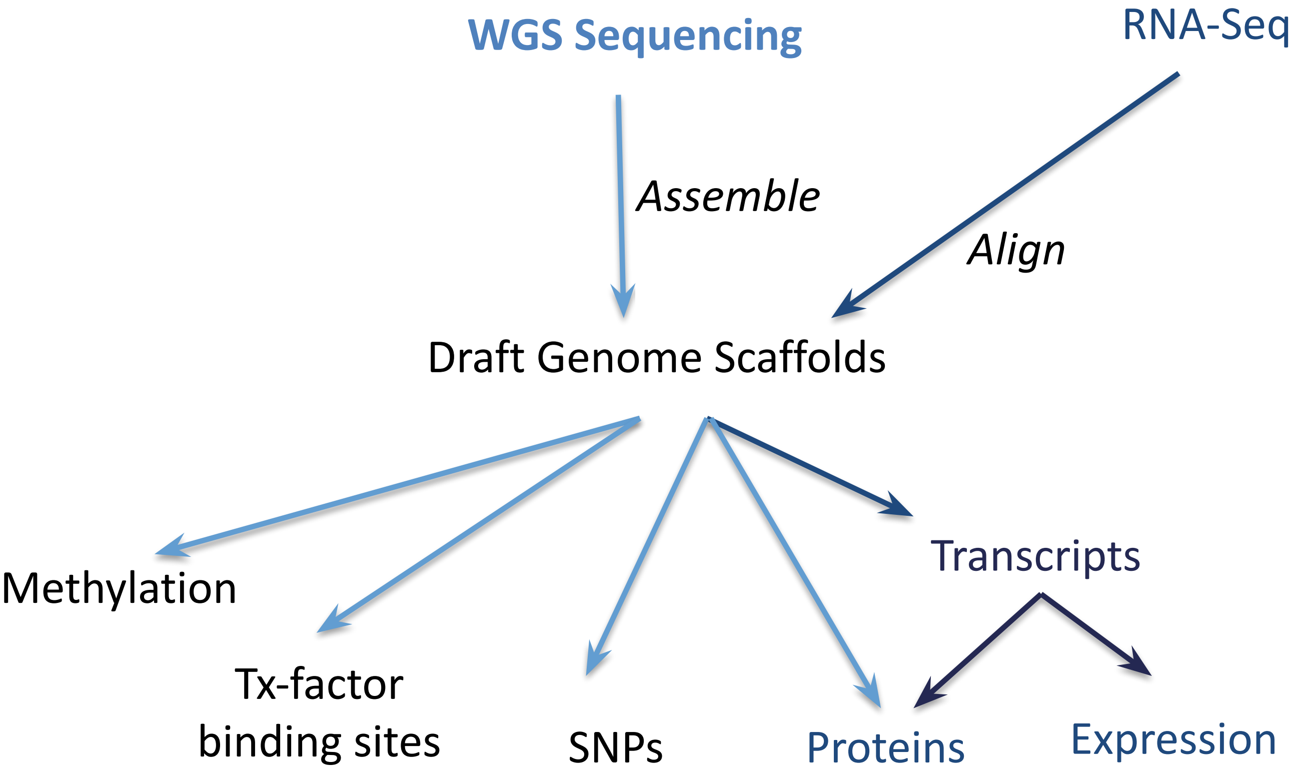
Tx-factor binding sites

SNPs

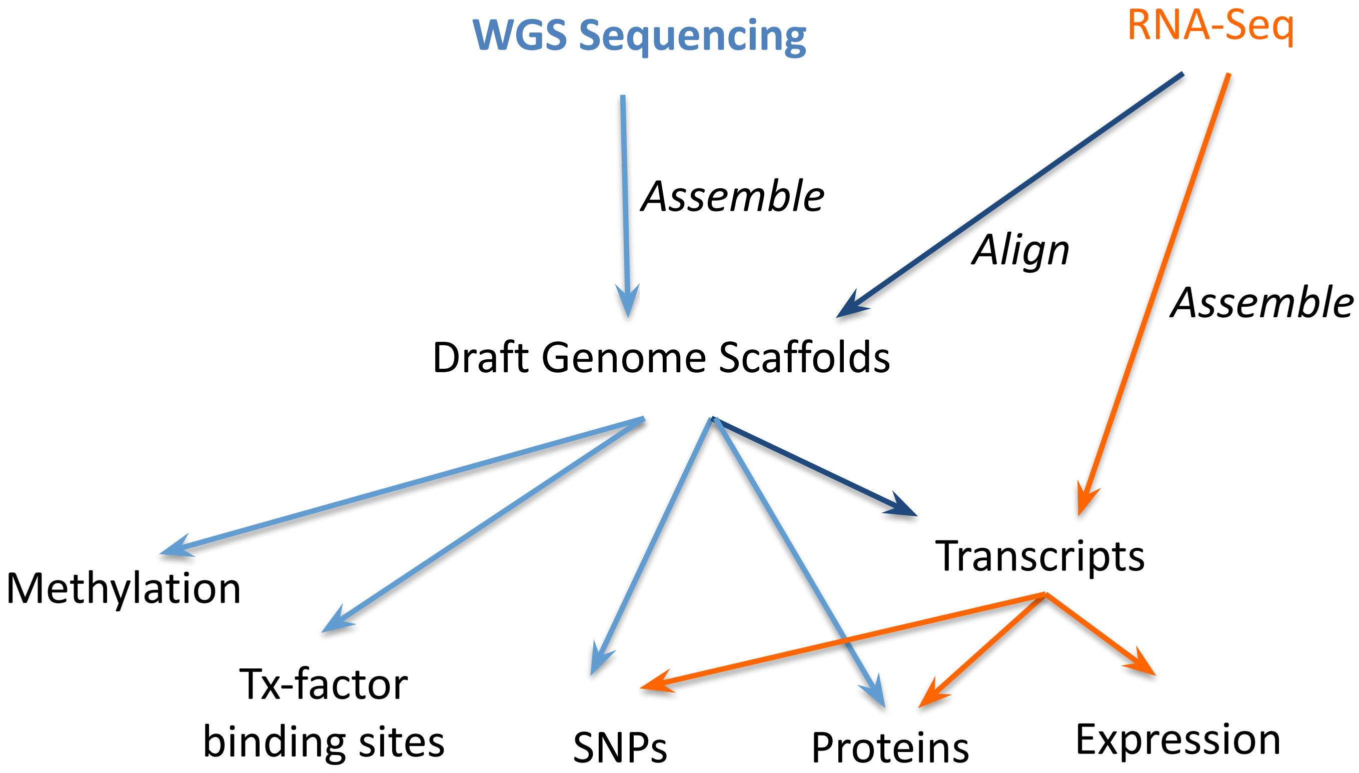
Proteins

Station Biologique de Roscoff

A Paradigm for Genomic Research



A *Maturing* Paradigm for Transcriptome Research



WGS Sequencing

RNA-Seq

Assemble

Align

Assemble

Draft Genome Scaffolds

Transcripts

Methylation

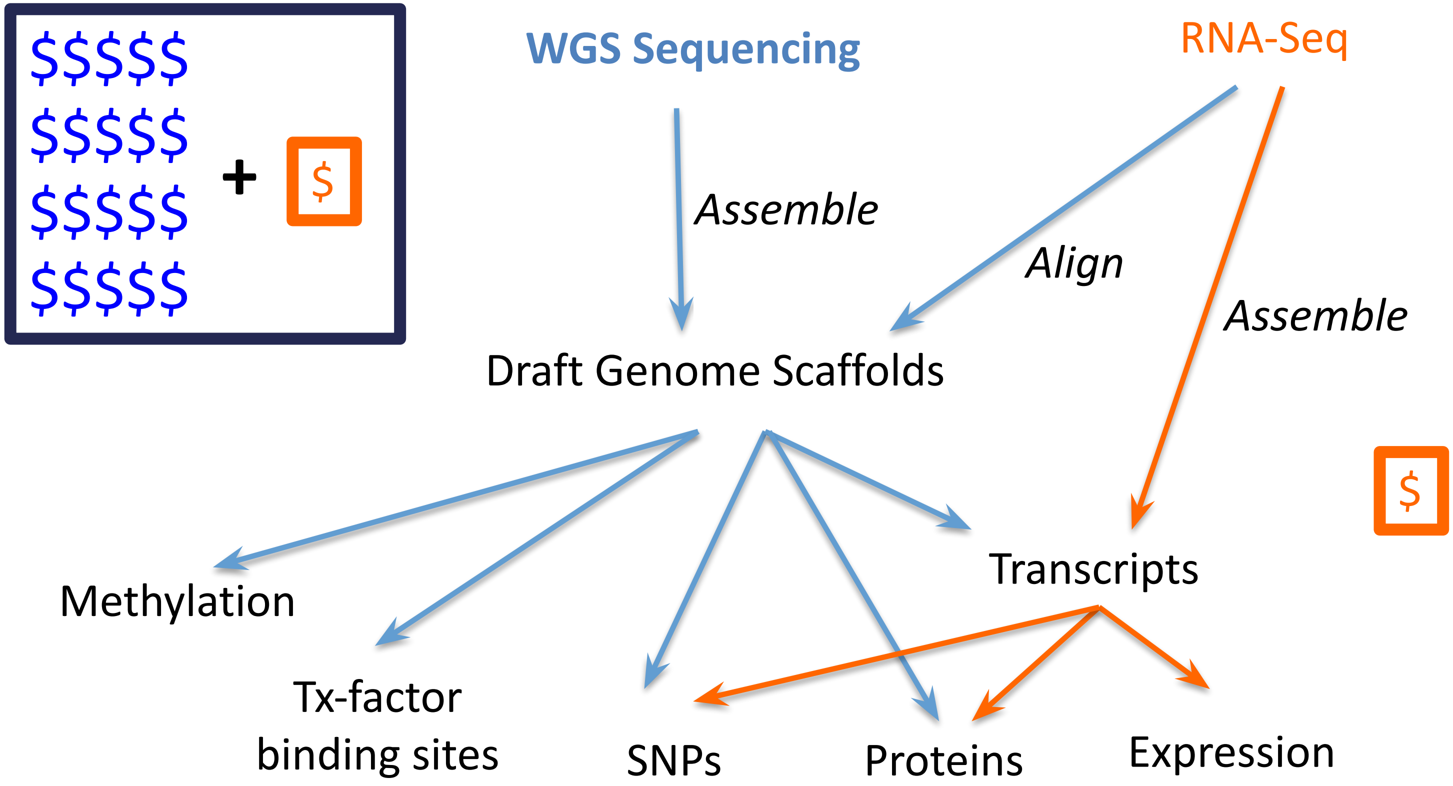
Tx-factor binding sites

SNPs

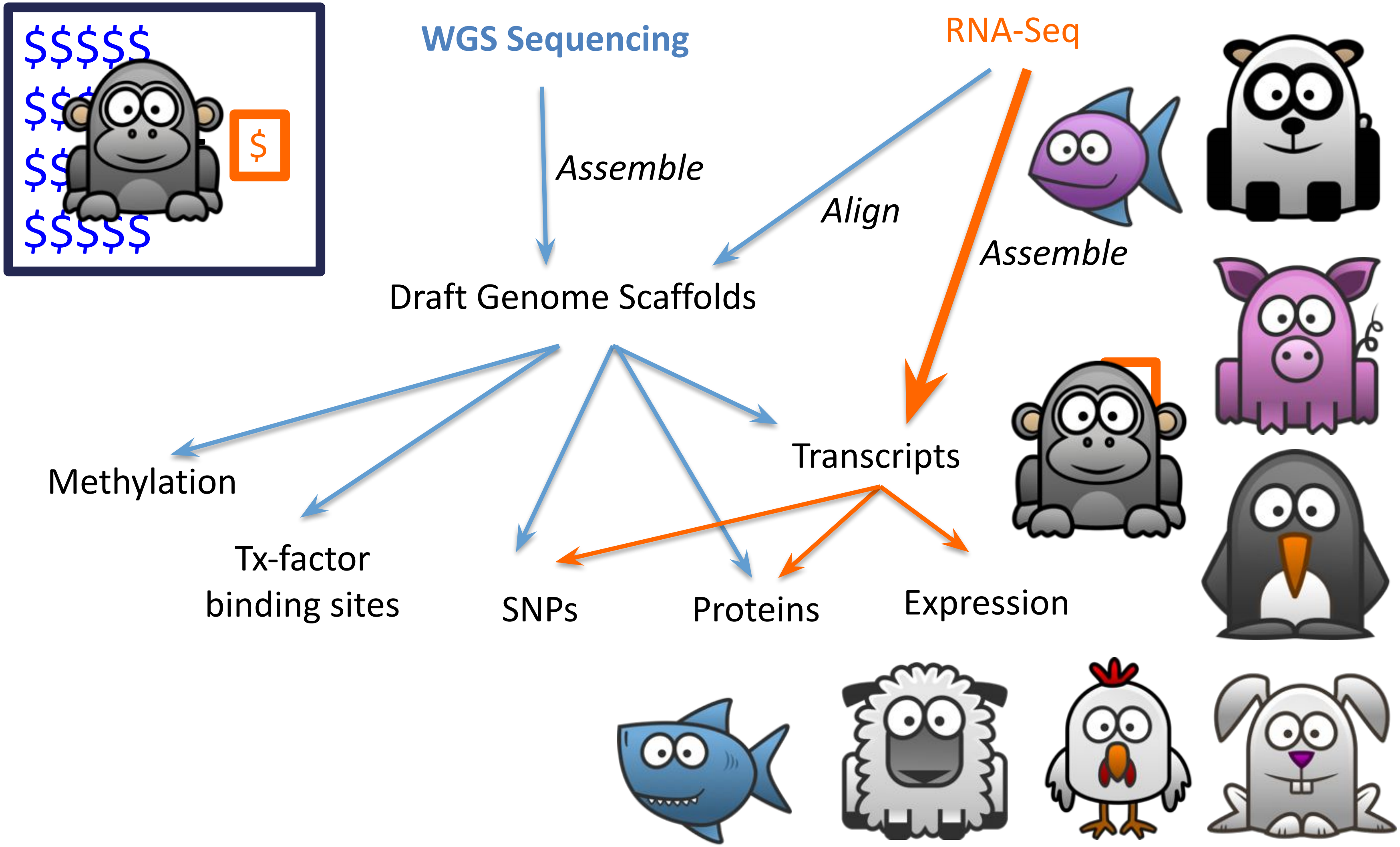
Proteins

Expression

A *Maturing* Paradigm for Transcriptome Research



A **Maturing** Paradigm for Transcriptome Research



WGS Sequencing

RNA-Seq

Assemble

Align

Assemble

Draft Genome Scaffolds

Transcripts

Methylation

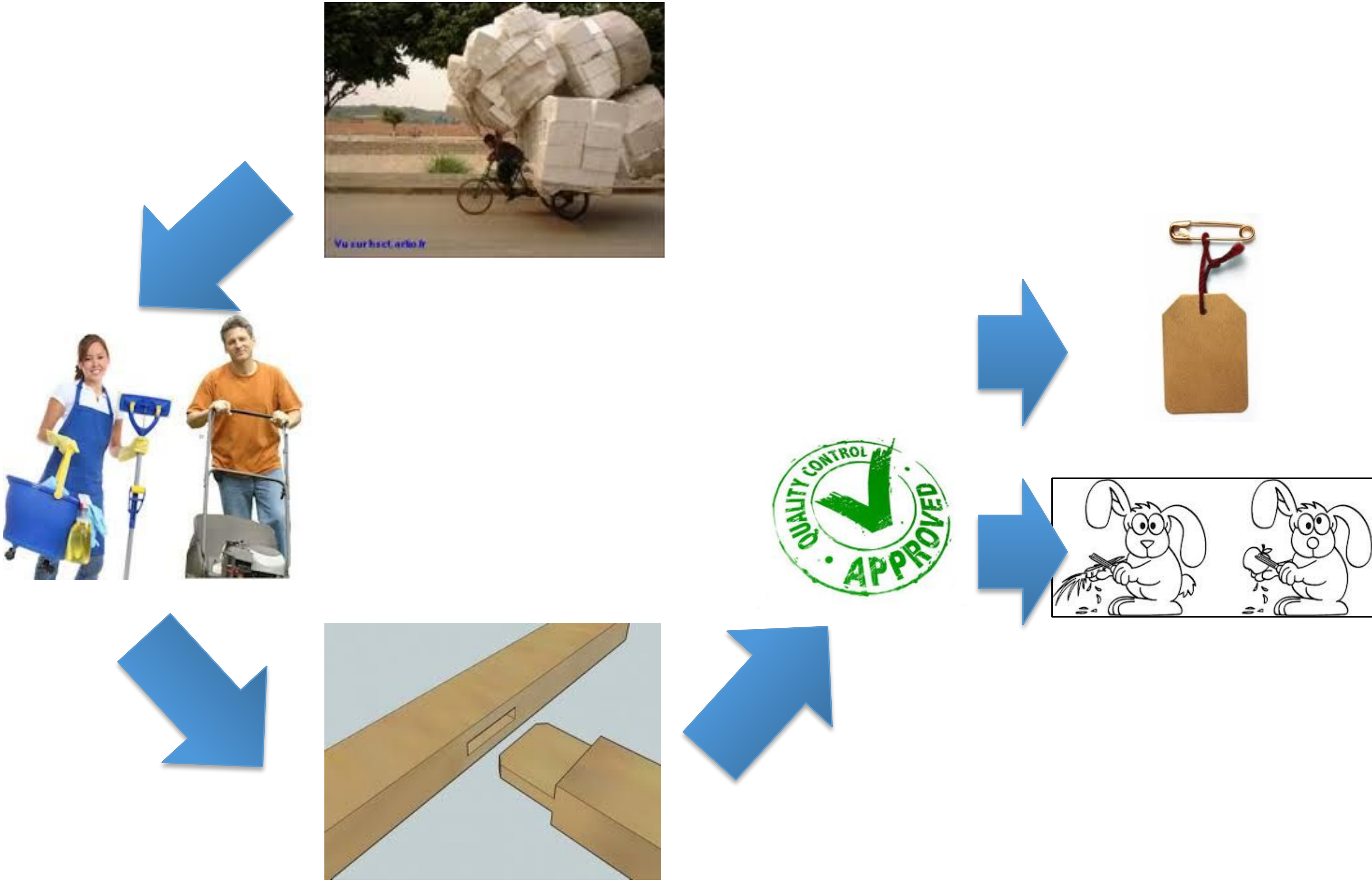
Tx-factor binding sites

SNPs

Proteins

Expression

RNA Seq de novo analysis workflow



Data Cleaning

- Unknown nucleotides
- Bad quality nucleotides
- Adaptors and primers sub-sequences
- Poly A/T tails
- Low complexity sequences
- rRNA sequences
- Contaminant sequences
- Short length sequences

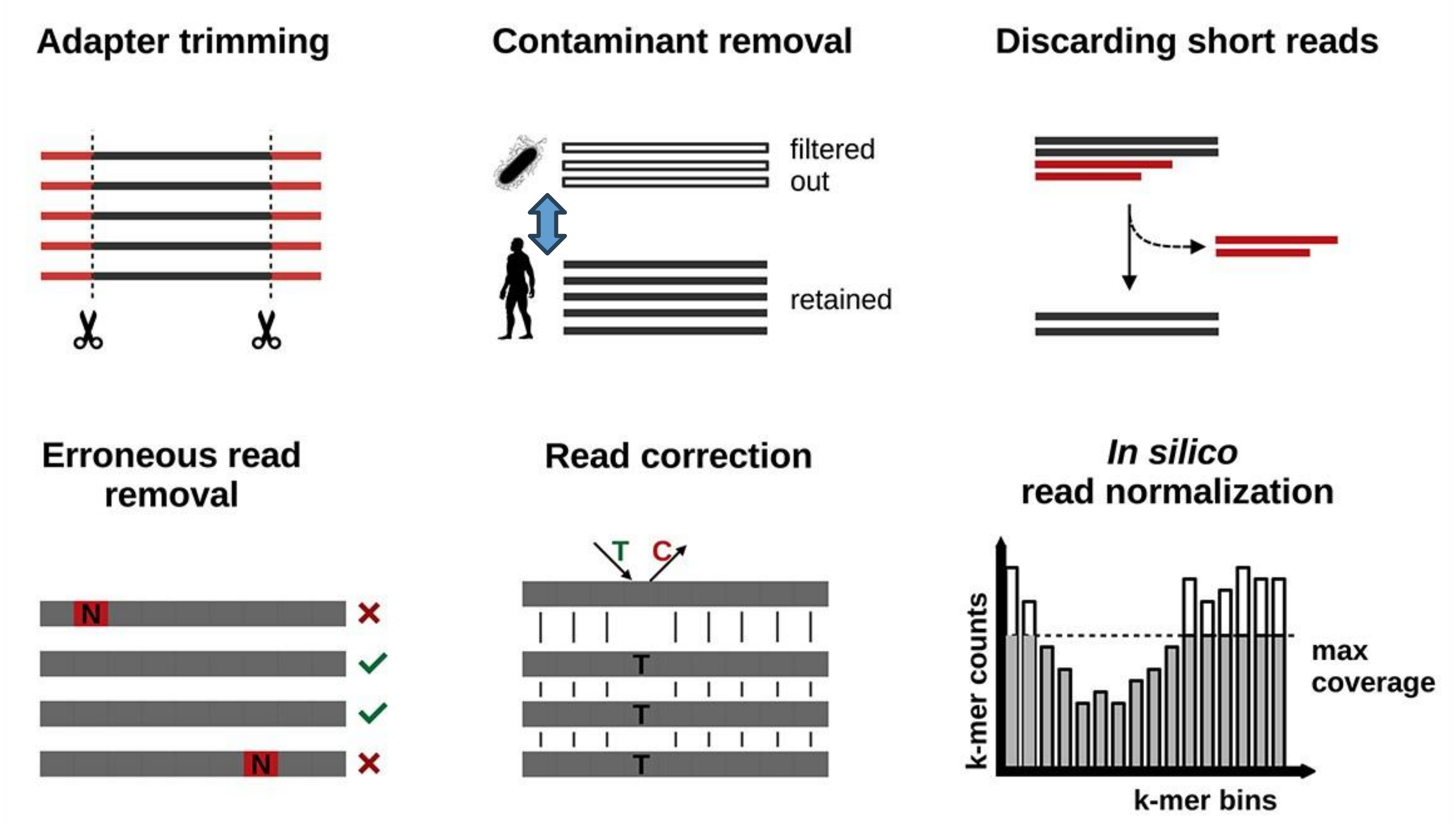
But also:

- Removing singletons
- In-silico normalization
- Sequencing errors correction
- …

Bias should be corrected in reverse order of their generation

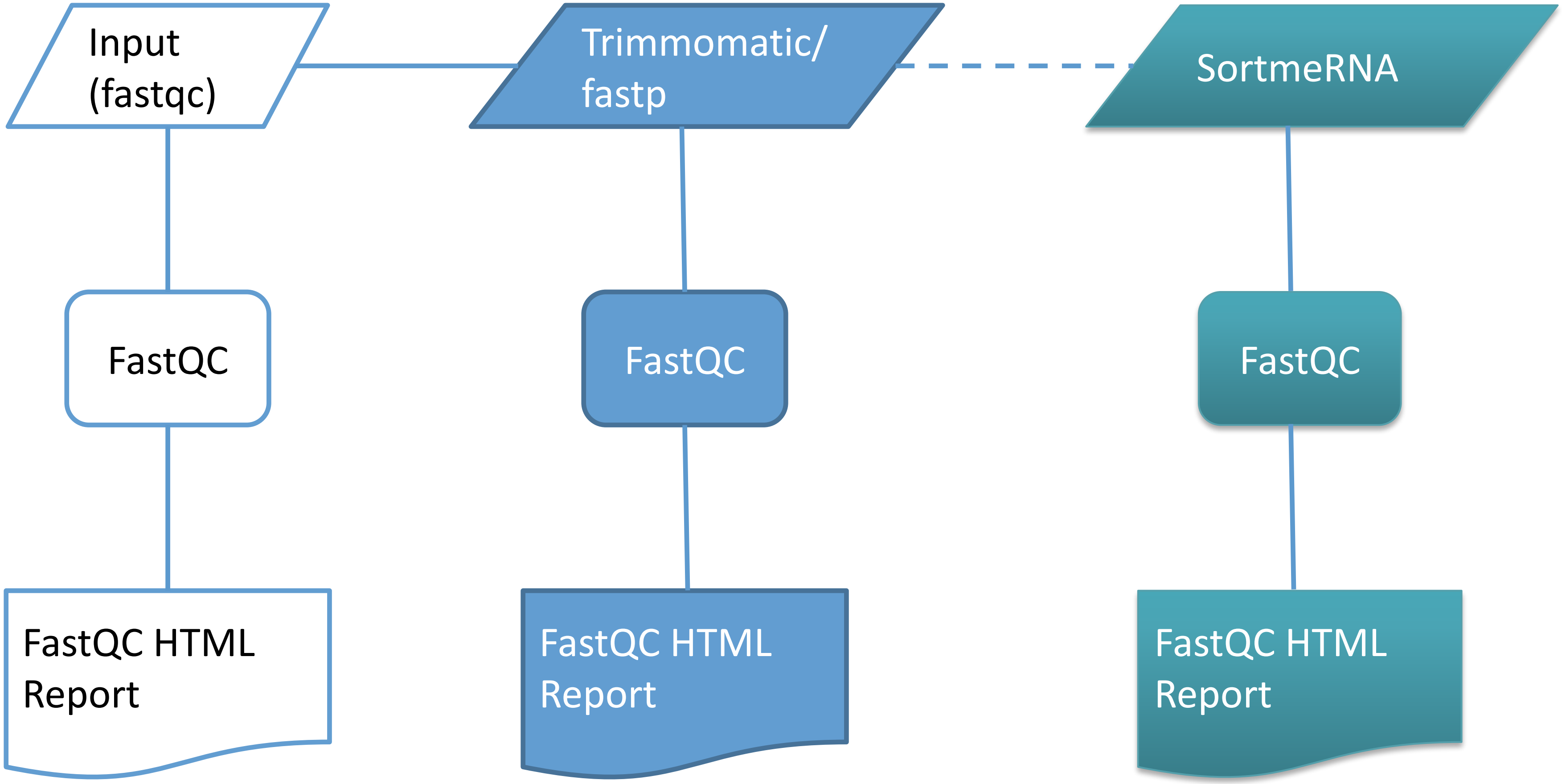
1. Sequencing biases (bad quality, unknowns)
2. Library preparation
   Adaptors and primers sequences
   Poly A/T tails
3. Biological sample (low complexity, rRNA, contaminants)

Data cleaning



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

Data cleaning



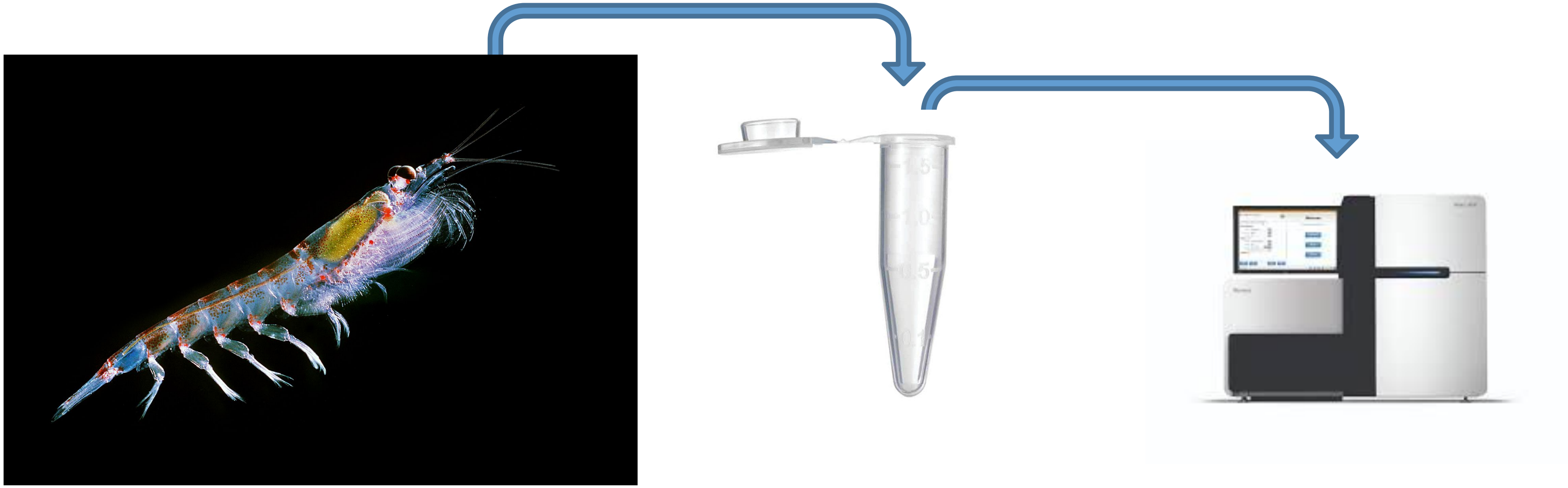
Trimmomatic command

```
java -jar trimmomatic.jar  PE -phred33
\ lib1_1.fastq lib1_2.fastq                 Raw reads
\ lib1_1.P.qtrim lib1_1.U.qtrim             Paired and unpaired reads1
\ lib1_2.P.qtrim lib1_2.U.qtrim             Paired and unpaired reads2
\ ILLUMINACLIP:illumina.fa:2:30:10          Adapters
\ SLIDINGWINDOW:4:15LEADING:5 TRAILING:5 MINLEN:25
```

```
Input Read Pairs: 2 000 000
Both Surviving: 1 879 345 (93.97%)
Forward Only Surviving: 94 153 (4.71%)
Reverse Only Surviving: 18 098 (0.90%)
Dropped: 8 404 (0.42%)

TrimmomaticPE: Completed successfully
```

Alternative : Sickle, TrimGalore, fastp

Contaminations

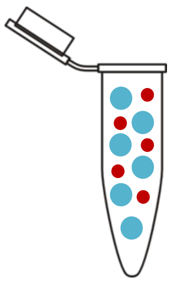


Euphausia superba (Uwe Kils. 2011)

Contaminations

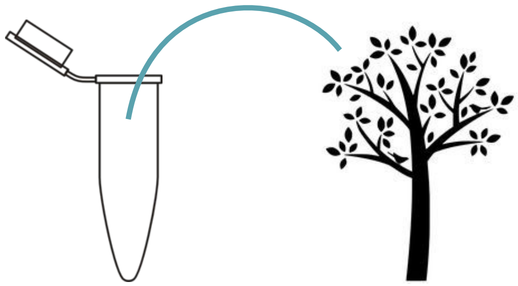
Euphausia superba (Uwe Kils. 2011)

Contaminations



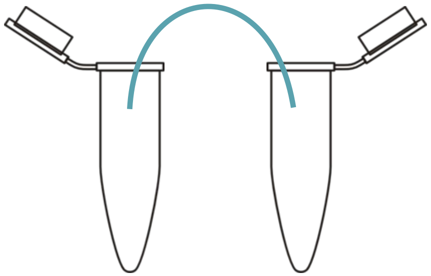
in-contamination

for ex. rRNA



third-party contamination

for ex. food - parasite



cross-contamination

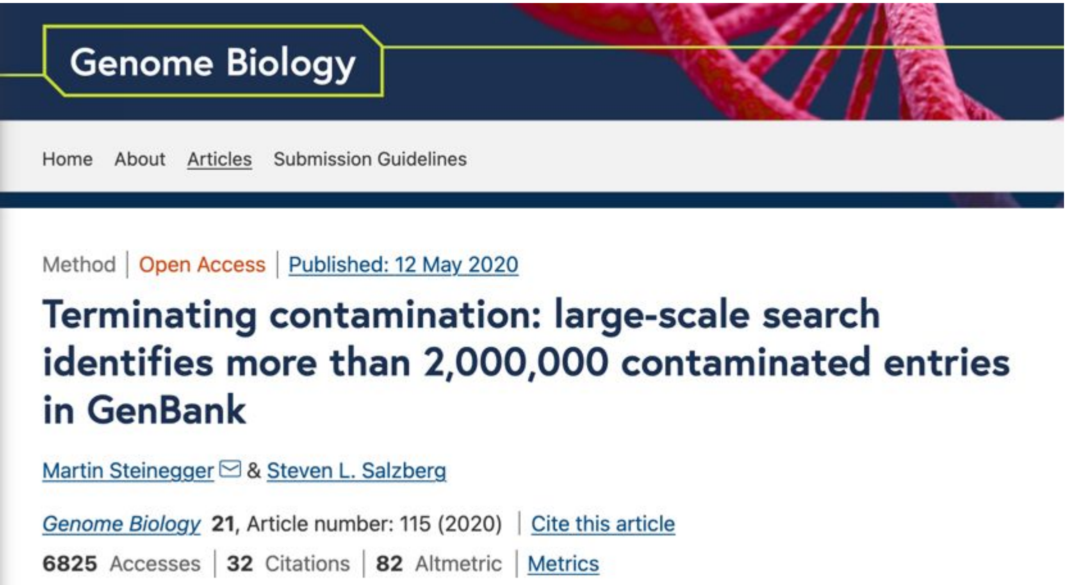
for ex. experiment

- 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

Thanks to Serafim Nenarokov

CNRS • SORBONNE UNIVERSITE
Station Biologique de Roscoff
ABiMS

Contaminations



Method | Open Access | Published: 12 May 2020

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

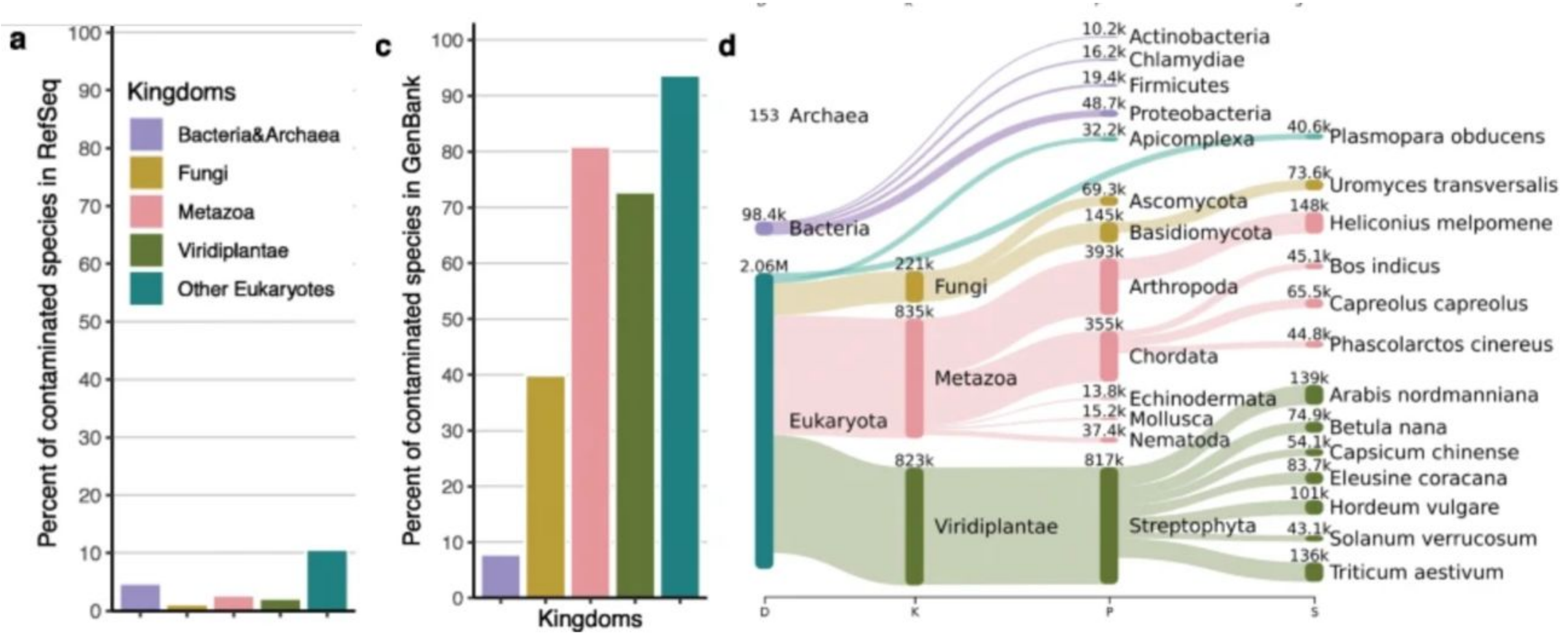
Martin Steinegger & Steven L. Salzberg

Genome Biology **21**, Article number: 115 (2020) | Cite this article

6825 Accesses | **32** Citations | **82** Altmetric | Metrics

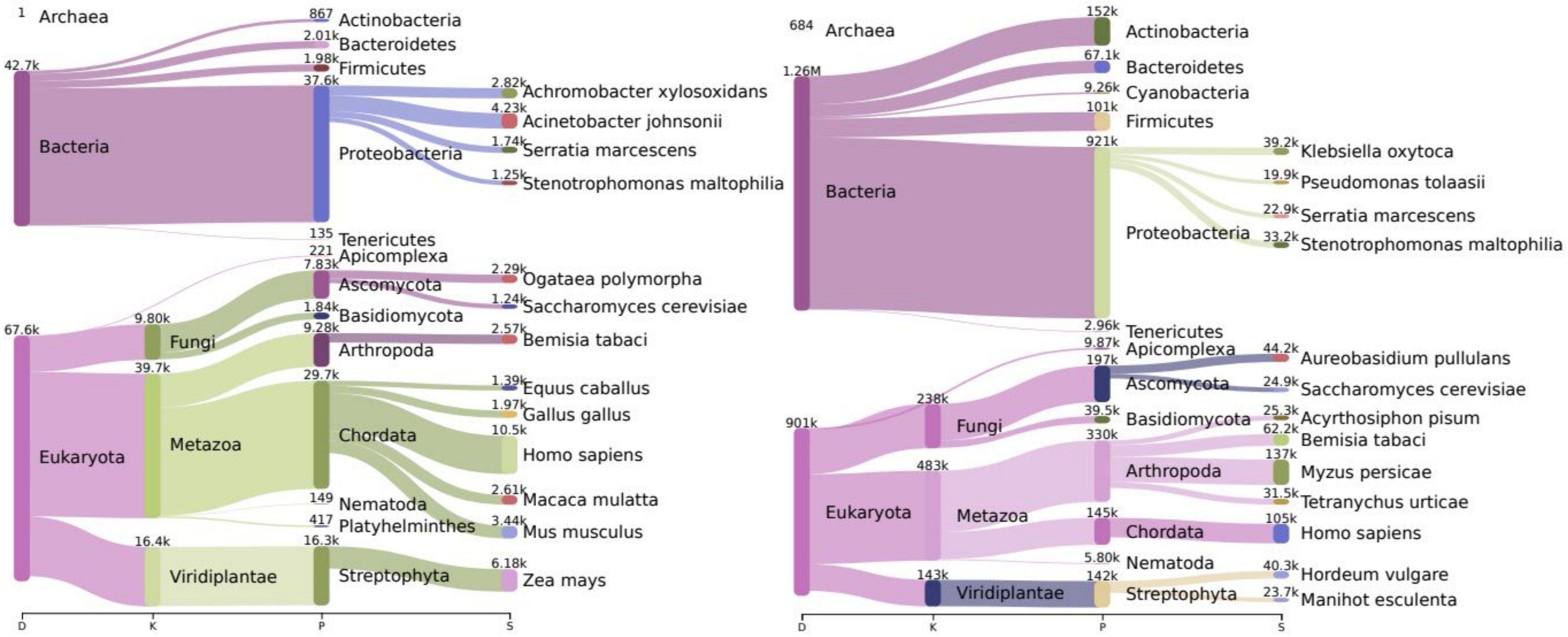
« Conterminator reported 114,035 and 2,161,746 contaminated sequences affecting 2767 and 6795 species in RefSeq and GenBank, respectively »

Contaminated sequences



Contaminations

Contaminating species



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 philippinarui

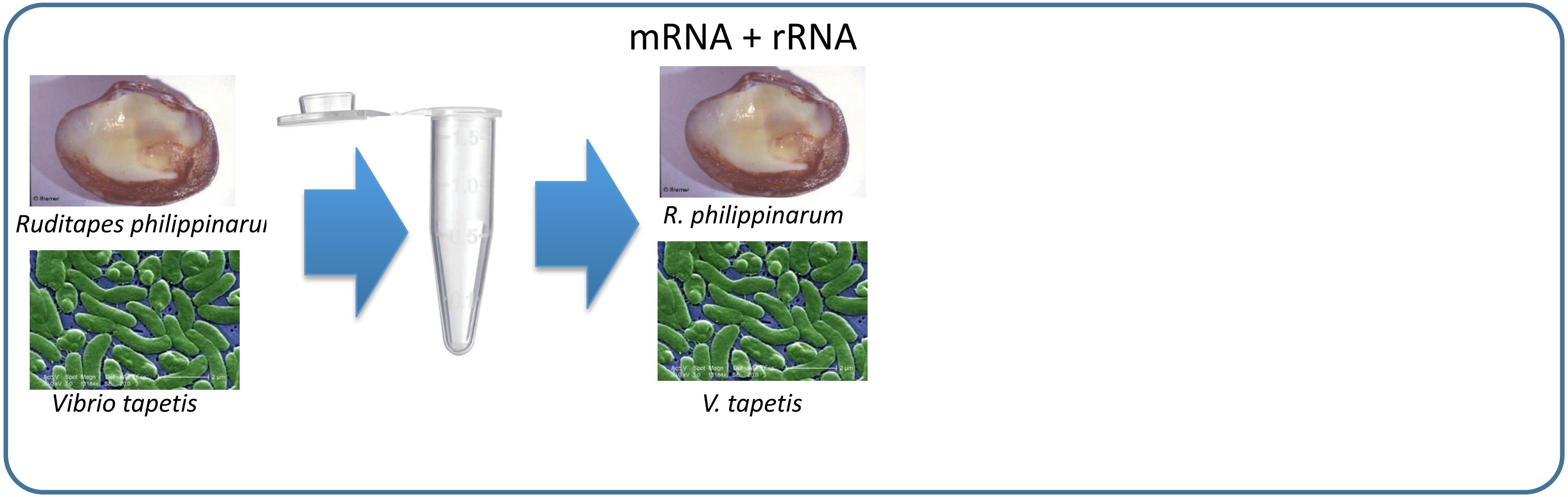
Vibrio tapetis

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



mRNA + rRNA

Ruditapes philippinarui

Vibrio tapetis

R. philippinarum

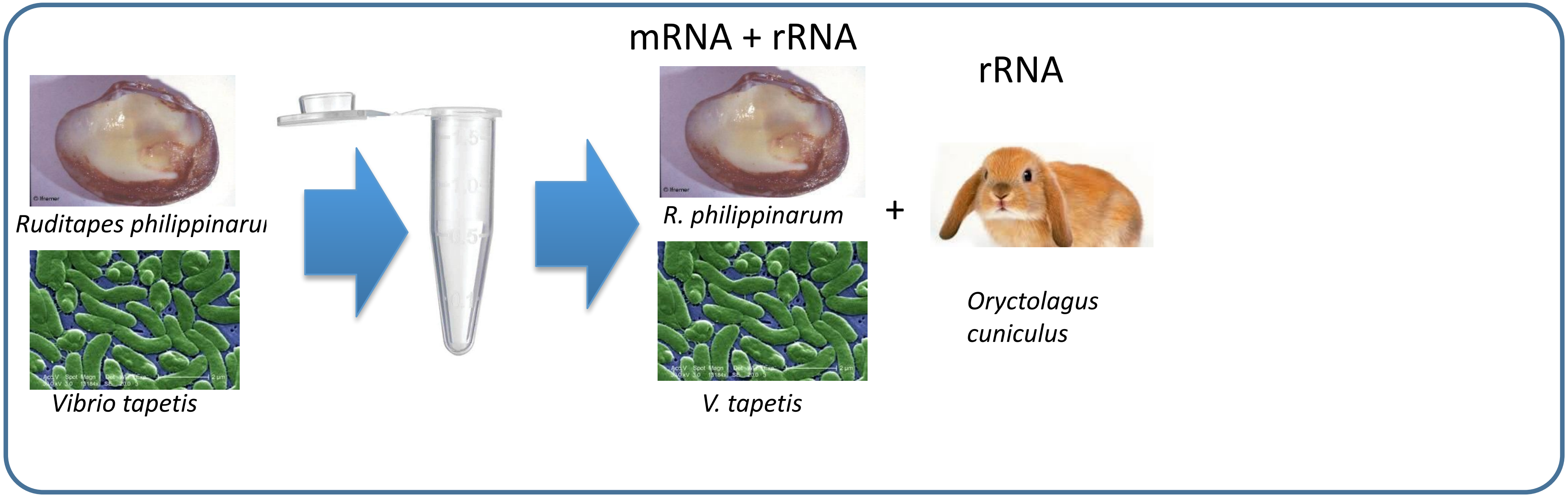
V. tapetis

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

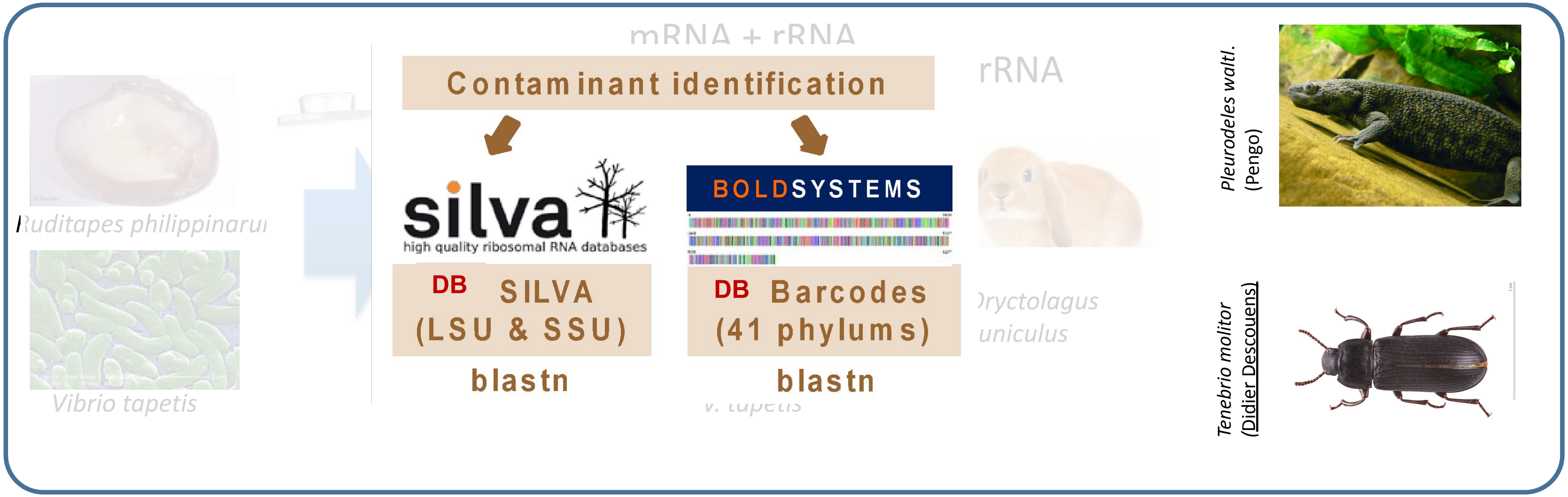


rRNA contamination

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



Contaminant identification

DB SILVA (LSU & SSU)

blastn

DB Barcodes (41 phylums)

blastn

Pleurodeles waltl. (Pengo)

Tenebrio molitor (Didier Descouens)

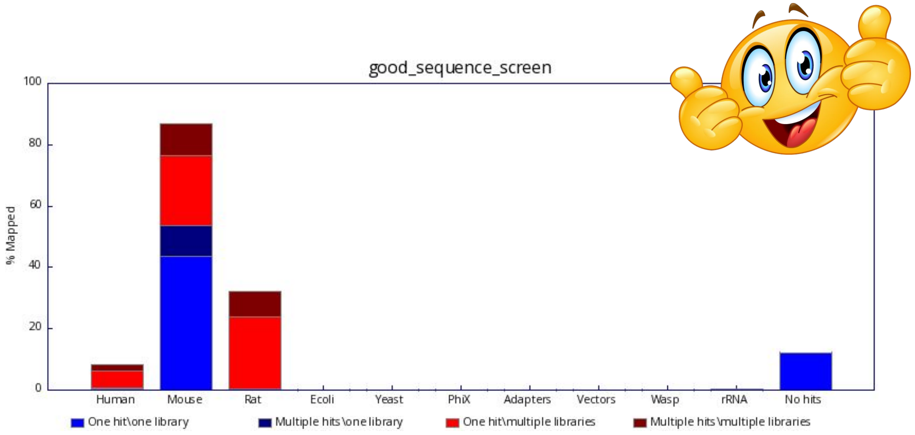
Detect third-party/cross contamination

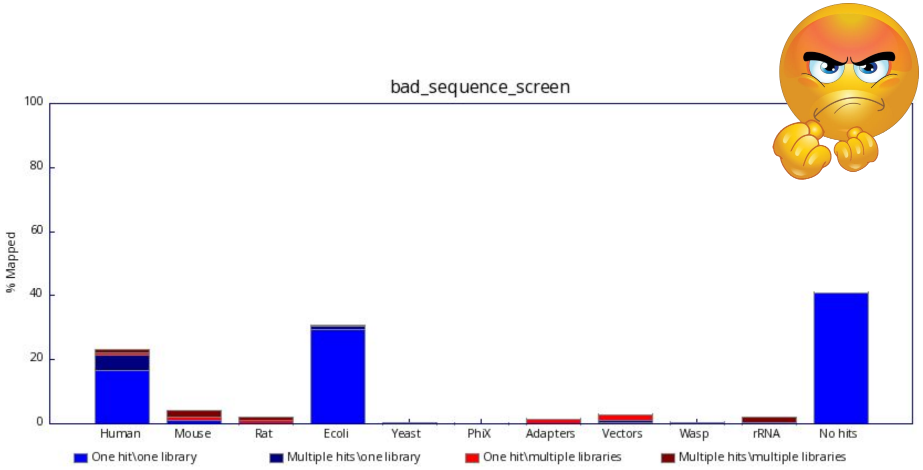
reads that do not originate from the organism and/or RNA species of interest
f. ex. reads originating from an endosymbiont bacterium in an eukaryote organism of interest

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

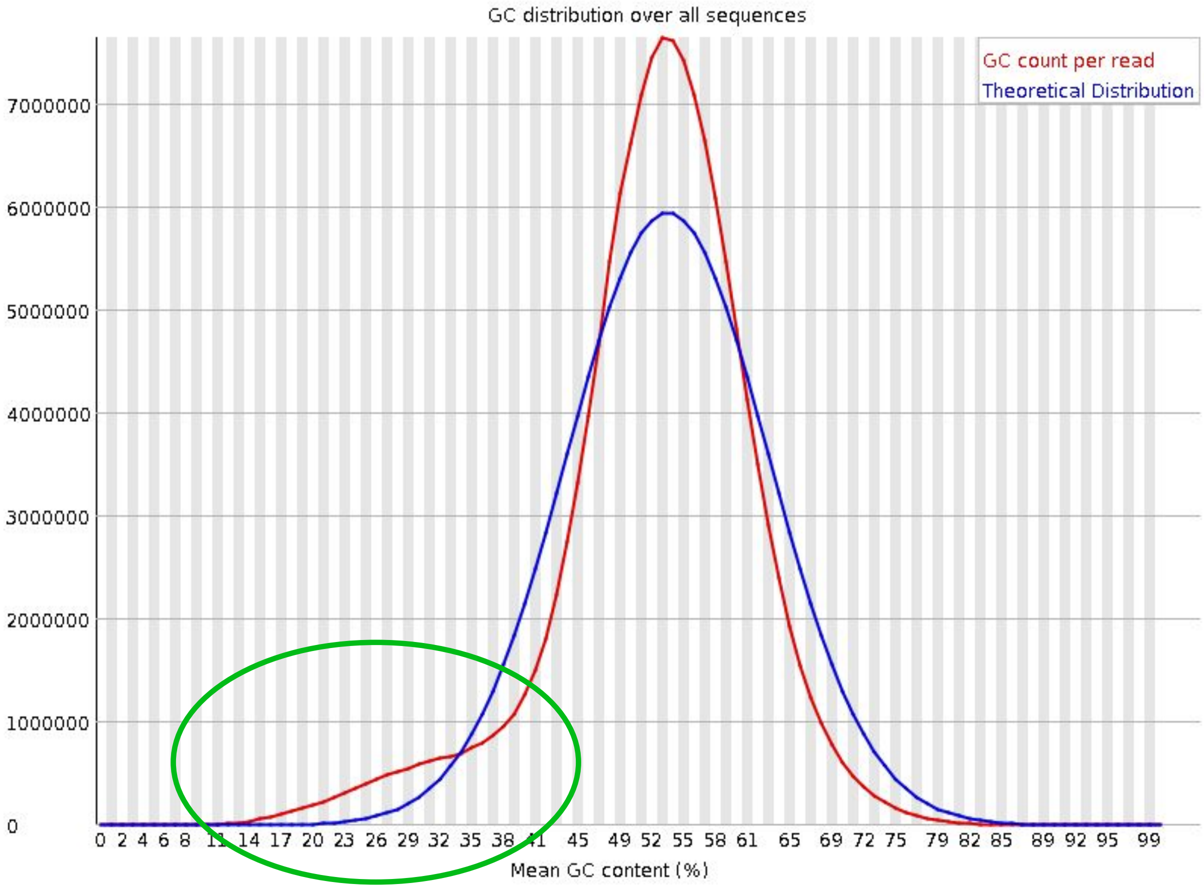
screen-only alternative **FastQ Screen**





FastQC: Per sequence GC content

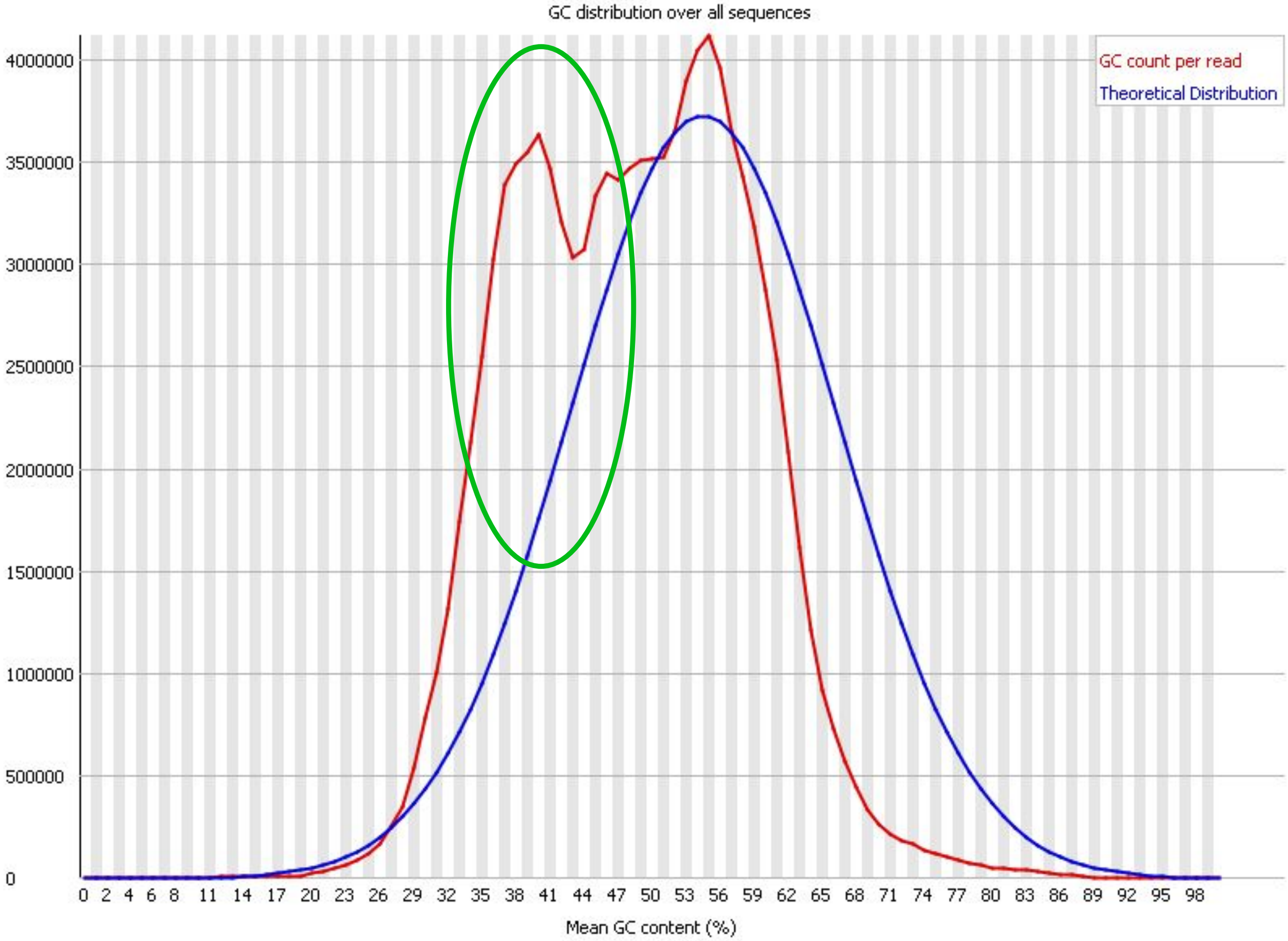
- A contamination ?



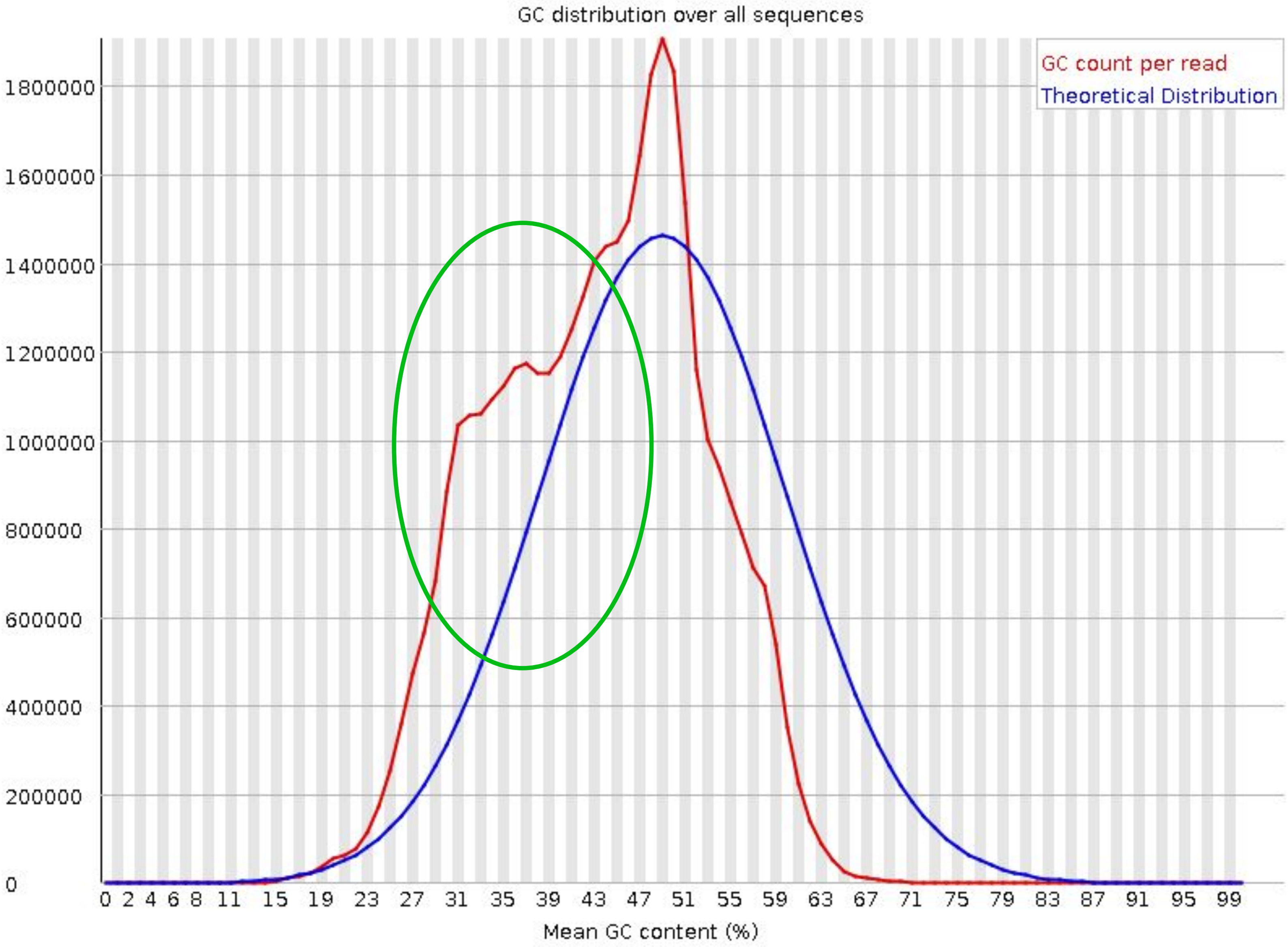
<u>Can this be fixed ?</u>

Maybe

FastQC: Per sequence GC content

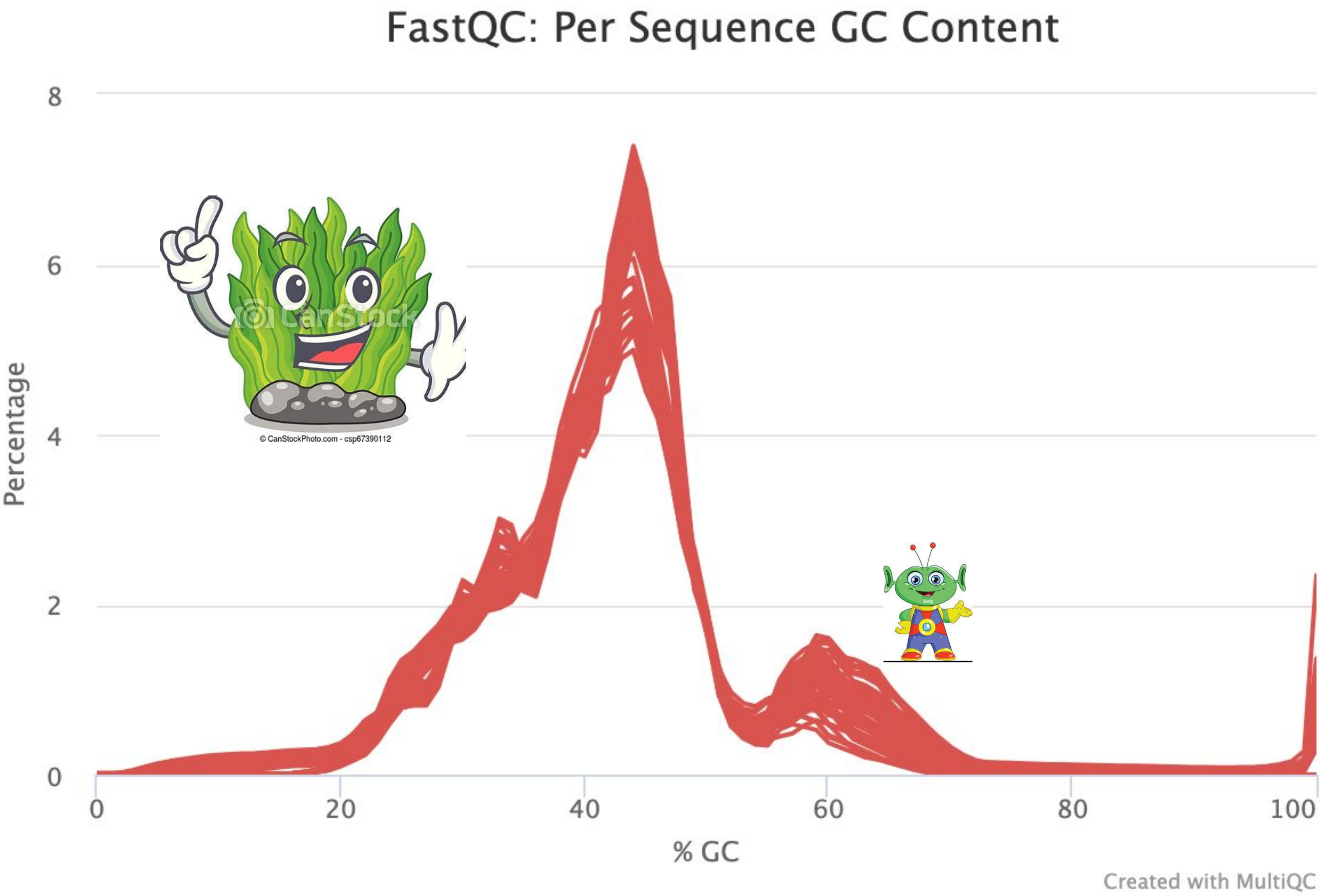


Third-party contamination : detection

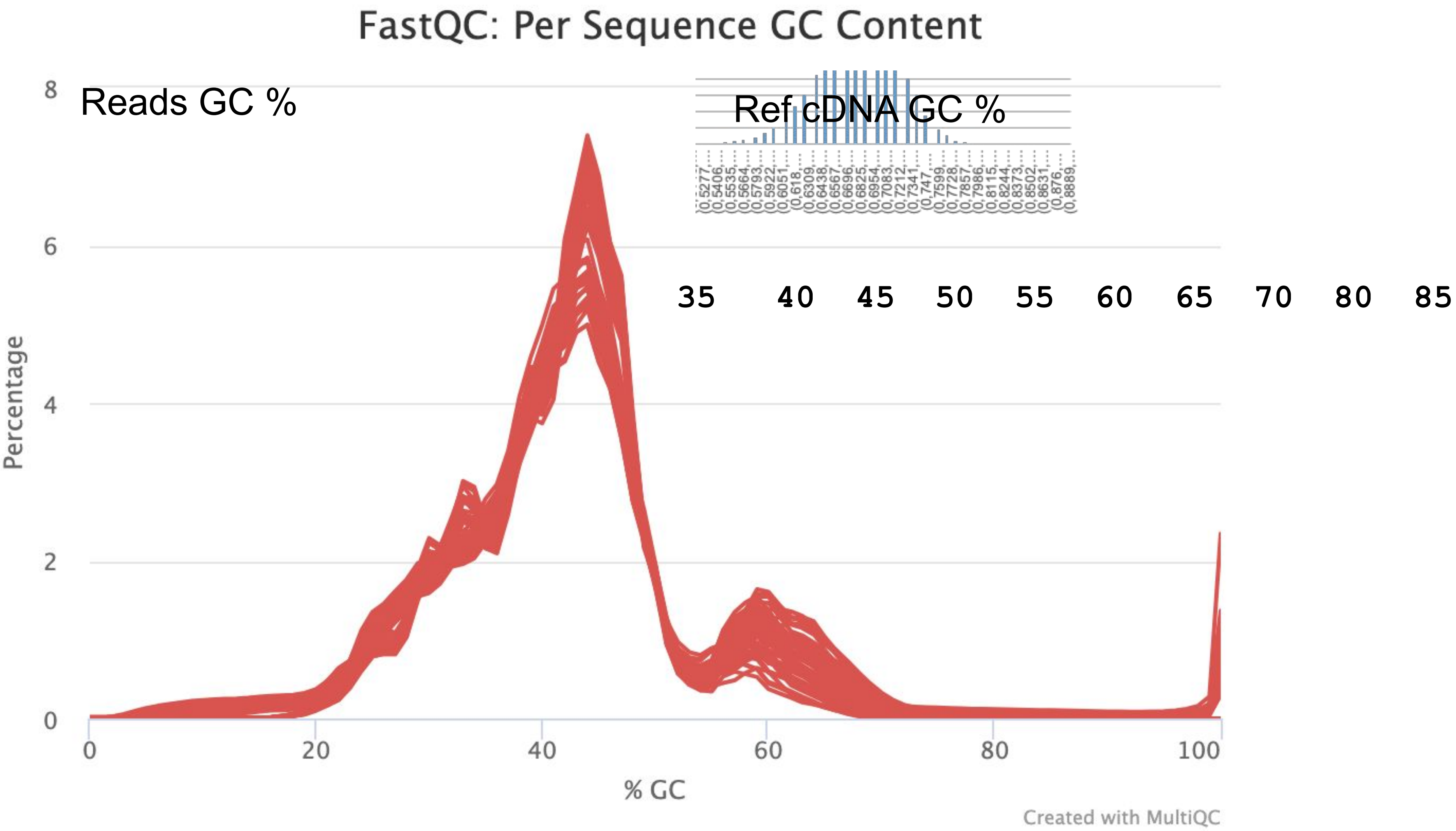


Sabellaria alveolata : mantle transcriptome

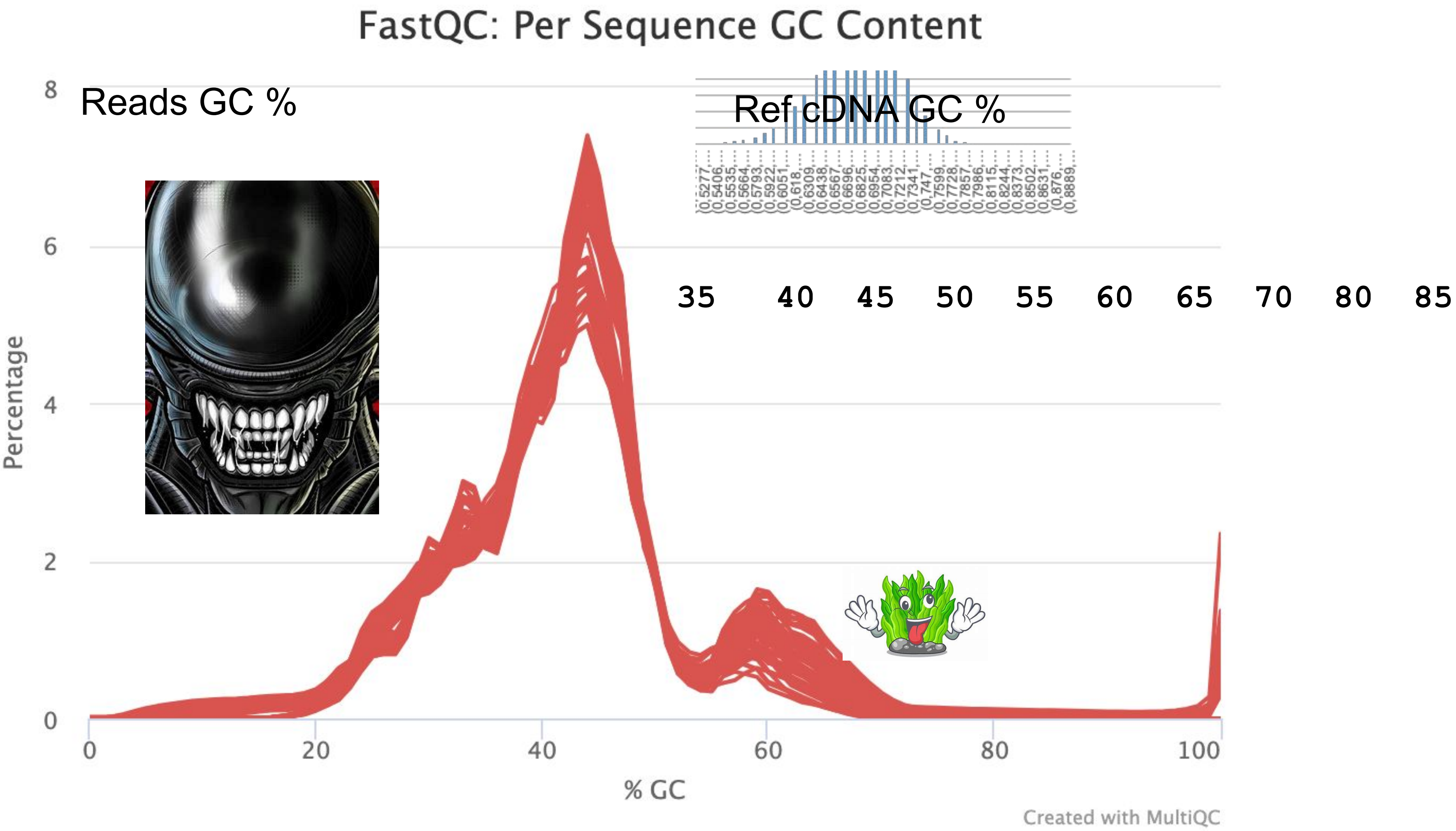
Third-party contamination : detection



Third-party contamination : detection



Third-party contamination : detection



rRNA contamination

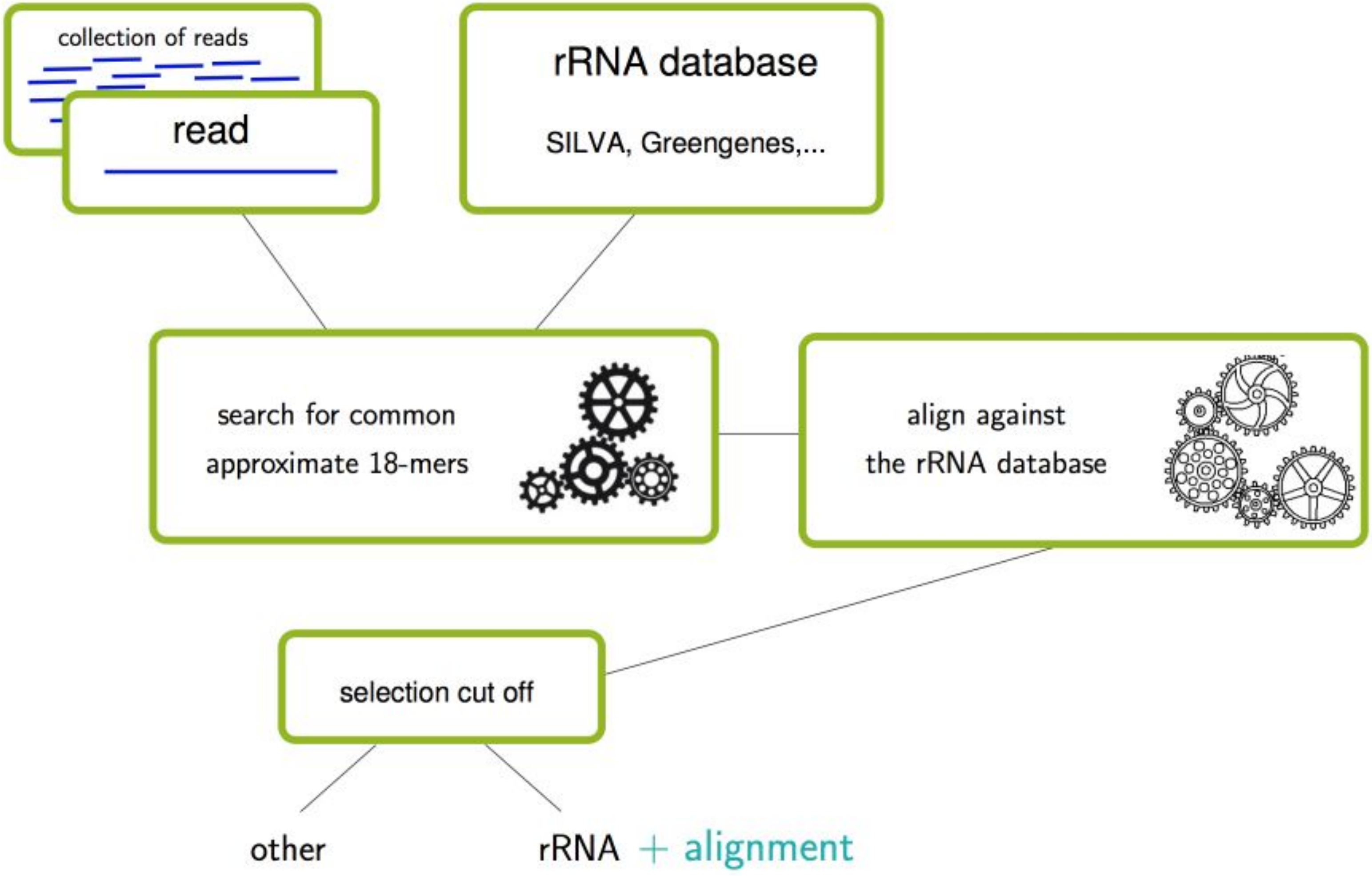
Prior to sequencing :

- Ribodepletion kits
- Selection polyA

After sequencing :

- Remove rRNA reads from raw reads
- Detect rRNA transcripts

SortMeRNA


collection of reads
read
rRNA database
SILVA, Greengenes,...
search for common approximate 18-mers
align against the rRNA database
selection cut off
other
rRNA + alignment

SortMeRNA commands v4.x

```
>sortmerna -fastx -a 4 -paired_out
```

Reference DB

```
\-ref silva-bac-16s-id90
\-ref silva-arc-16s-id95
\-ref silva-euk-18s-id95
\-ref silva-bac-23s-id98
\-ref silva-arc-23s-id98
\-ref silva-euk-28s-id98
\-ref rfam-5s-id98
\-ref rfam-5.8s-id98
```

reads

```
-reads reads1.fq.gz -reads reads2.fq.gz
```

output

```
-other output_mRNA.fastq fastq
-aligned output_aligned.fastq
```

```
>unmerge-paired-reads.sh output_mRNA.fastq
read-sortmerna_1.fq read-sortmerna_2.fq
```

SortMeRNA results

```
Results:
Total reads = 34 196 864
Total reads for de novo clustering = 4 084 914
Total reads passing E-value threshold = 30 122 173 (88.08%)
Total reads failing E-value threshold = 4 074 691 (11.92%)
Minimum read length = 150
Maximum read length = 150
Mean read length = 150
By database:
silva-bac-16s-id90.fasta      6.95%
silva-bac-23s-id98.fasta      18.75%
silva-euk-18s-id95.fasta      9.97%
silva-euk-28s-id98.fasta      52.42%
rfam-5s-database-id98.fasta   0.00%
rfam-5.8s-database-id98.fasta 0.00%

Total reads passing %id and %coverage thresholds = 26 037 259
```

Detect rRNA transcripts : RNAMMER

RNAmmer

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

```
# ---------------------------------------------------------------------------------------
##gff-version2##source-version RNAmmer-1.2##date 2009-11-16
##Type DNA
# seqname    source     feature    start          end      score         +/-      frame  attribute
# ---------------------------------------------------------------------------------------
AE000511     RNAmmer-1.2rRNA   448462         44857749.2   +        .          5s_rRNA
AE000511     RNAmmer-1.2rRNA   1473564        1473679        49.2    -        .          5s_rRNA
AE000511     RNAmmer-1.2rRNA   1045067        1045183        40.3    +        .          5s_rRNA
AE000511     RNAmmer-1.2rRNA   445339         448223         3056.5 +        .          23s_rRNA
AE000511     RNAmmer-1.2rRNA   1473918        1476803        3032.8 -        .          23s_rRNA
AE000511     RNAmmer-1.2rRNA   1207586        1209074        1801.4 -        .          16s_rRNA
AE000511     RNAmmer-1.2rRNA   1511140        1512627        1803.6 -        .          16s_rRNA
```

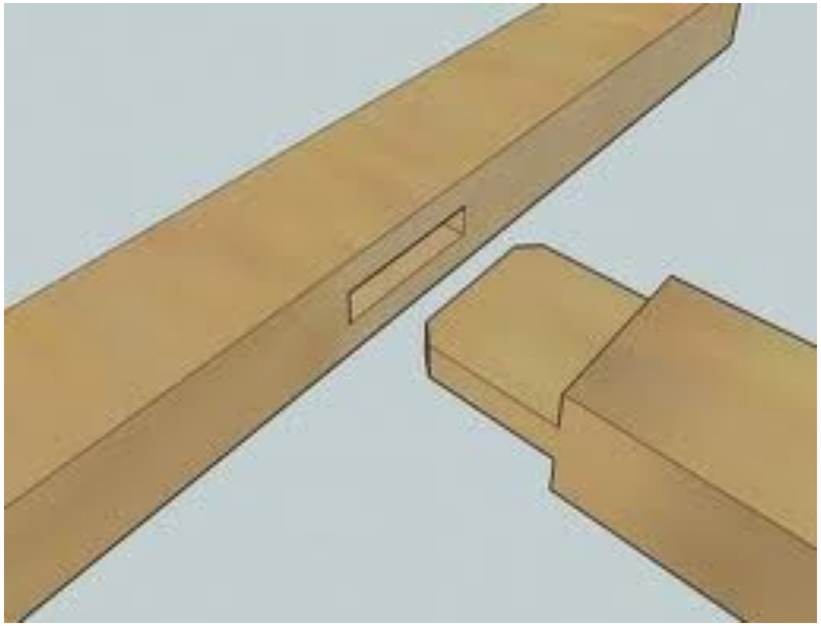
Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW **RNammer: consistent annotation of rRNA genes in genomic sequences**
Nucleic Acids Res. 2007 Apr 22.

Alternative Barrnap : *BAsic Rapid Ribosomal RNA Predictor*
https://github.com/tseemann/barrnap

RNAmmer - Barrnap

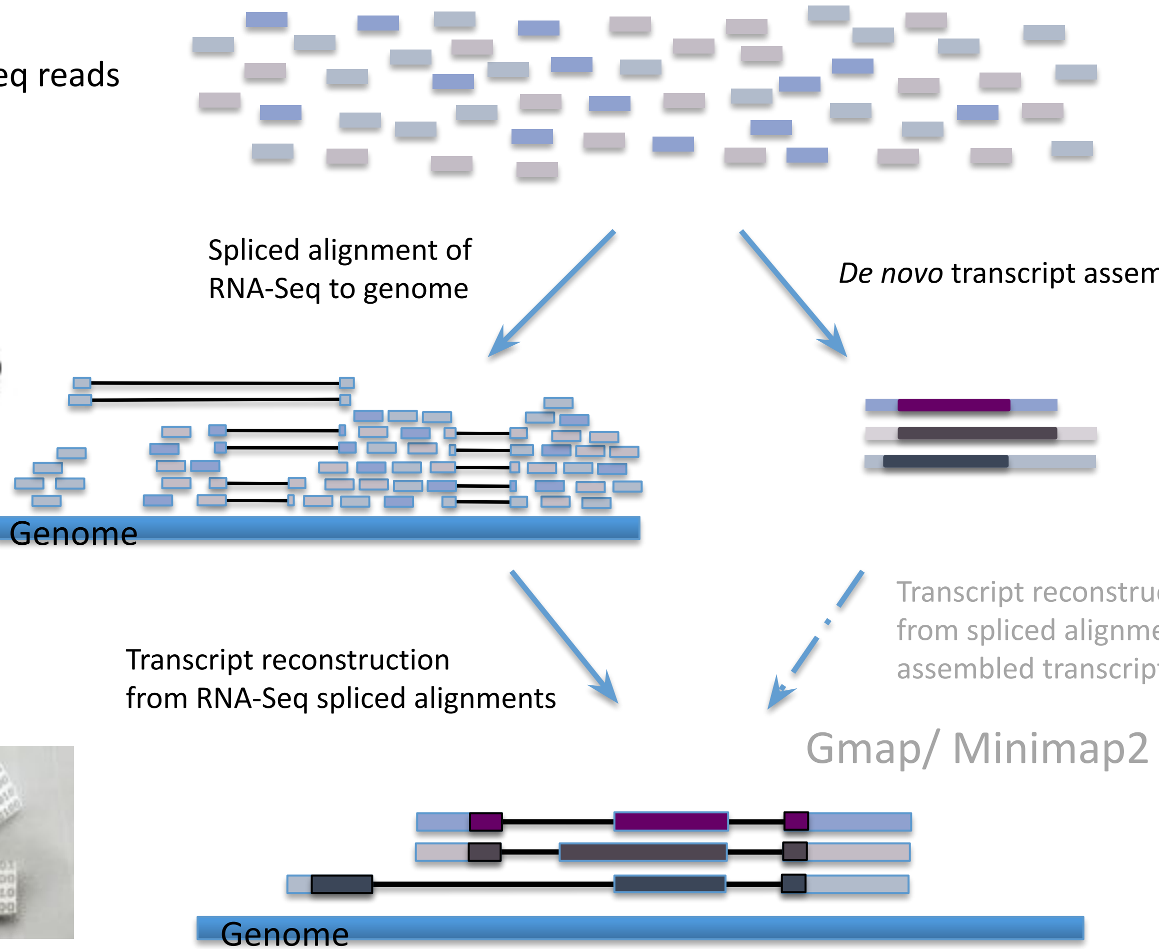
```
> Trinotate-3.0.1/util/rnammer_support/RnammerTranscriptome.pl
--transcriptome Assembly.fasta  —org_type
(arc|bac|euk) --path_to_rnammer
/usr/local/genome2/rnammer/rnammer

>bedtools getfasta -fi Assembly.fasta -bed
rnammer_predictions.gff > trancripts_rrna.fasta
```

```
> barnap --kingdom bac --threads 10 --outfasta rrna_bact.fasta
Assembly.fasta
```



TRANSCRIPTOME ASSEMBLY STRATEGIES

Contemporary strategies for transcript reconstruction from RNA-Seq



RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Transcript reconstruction from RNA-Seq spliced alignments

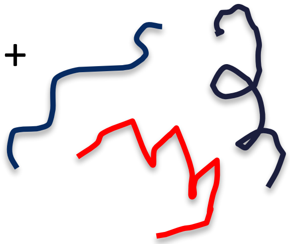
Transcript reconstruction from spliced alignment of assembled transcripts to genome

Gmap/ Minimap2

STAR
HISAT2
Tophat

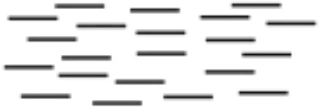
rnaSPAdes
Oases
SoapDenovoTrans
AbyssTrans
IDBA-Tran
Shannon
BinPacker
Bridger

Stringtie
Cufflinks
IsoLasso
Bayesembler
Trip
Traph
CEM
TransComb

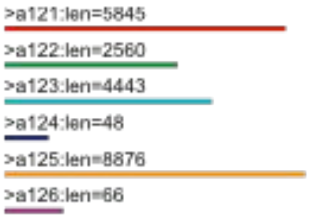


Trinity – How it works:

RNA-Seq reads

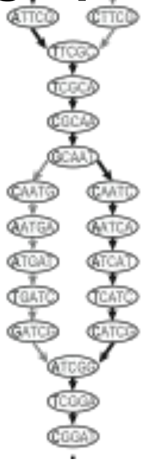


Linear contigs



>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

de-Bruijn graphs



Thousands of disjoint graphs

Transcripts + Isoforms

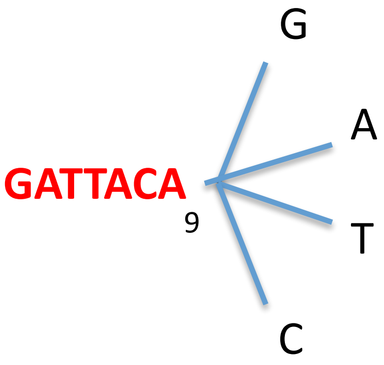
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

Inchworm Algorithm

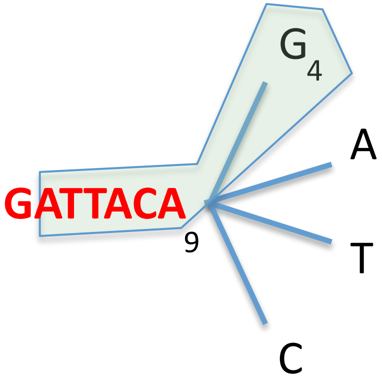
Decompose all reads into overlapping Kmers (25-mers) and count them : Jellyfish

Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

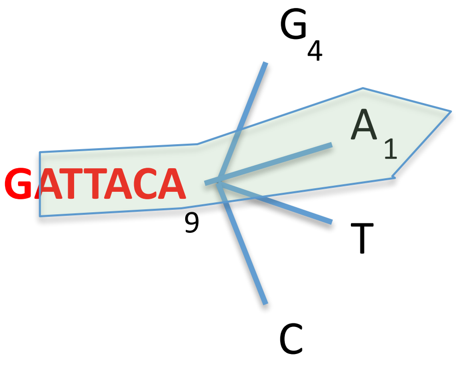
Extend kmer at 3’ end, guided by coverage.



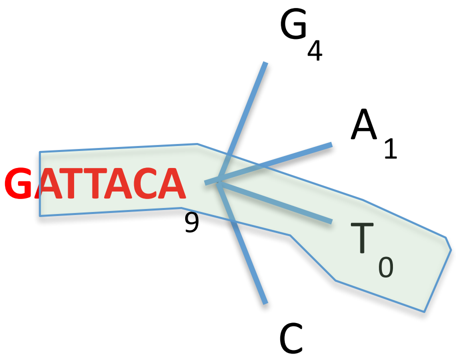
Inchworm Algorithm



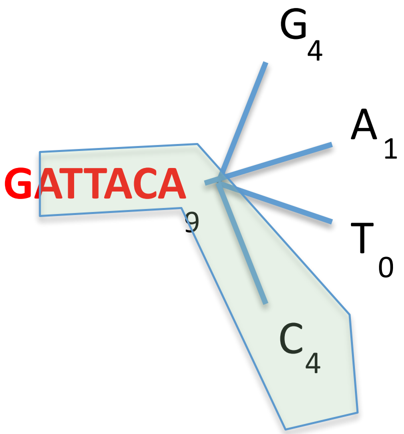
Inchworm Algorithm



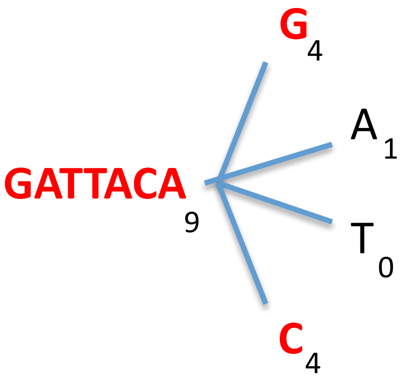
Inchworm Algorithm



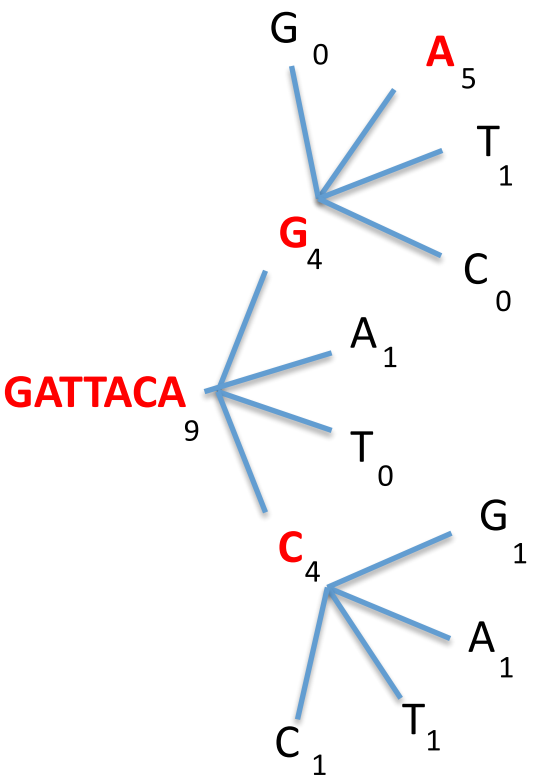
Inchworm Algorithm



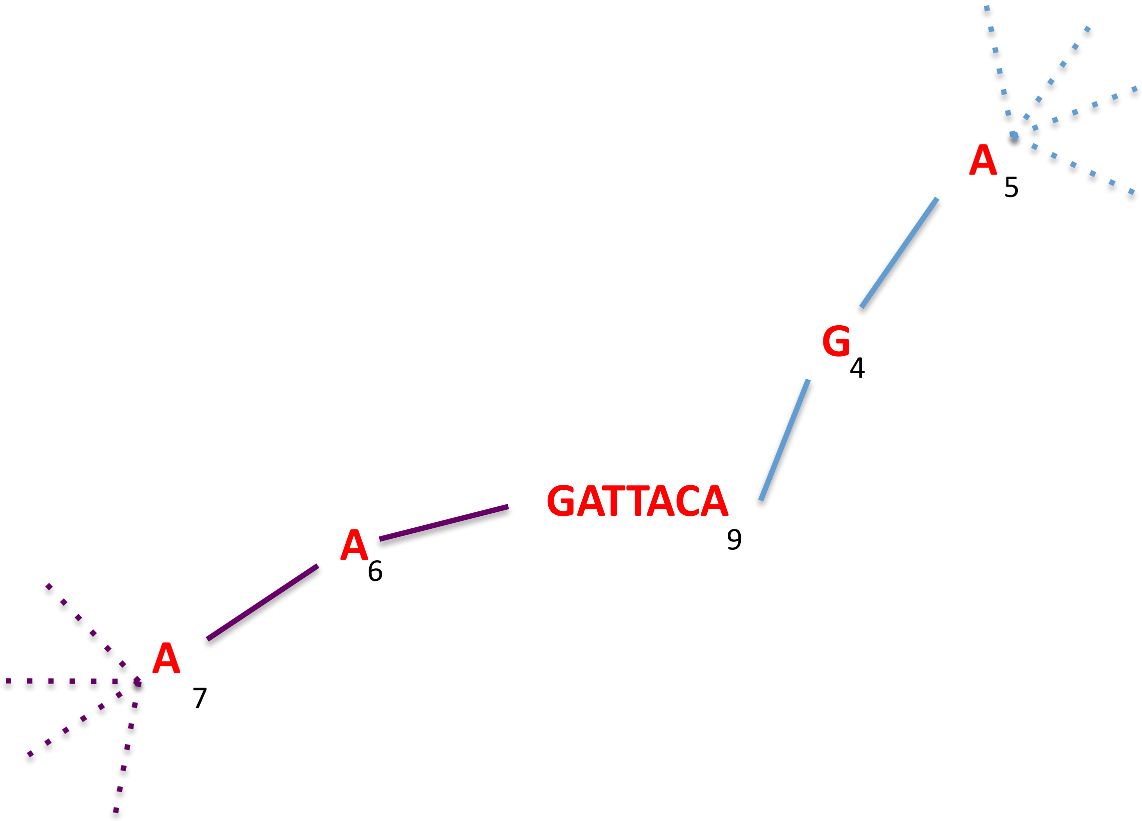
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm

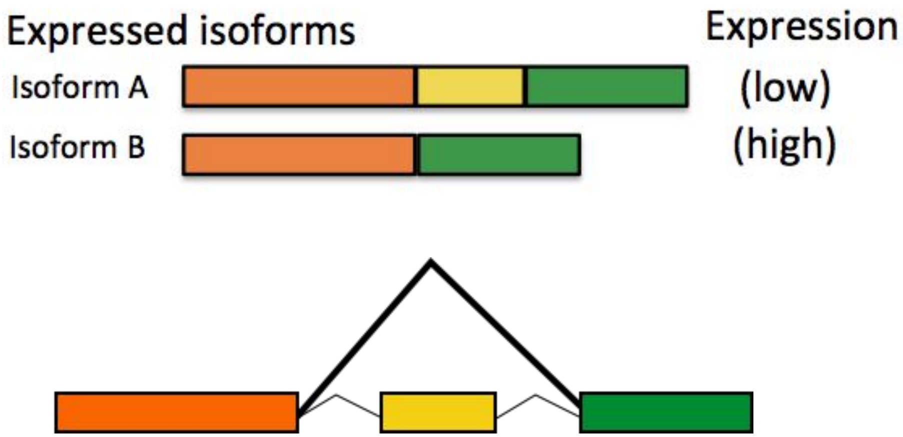


Report contig:

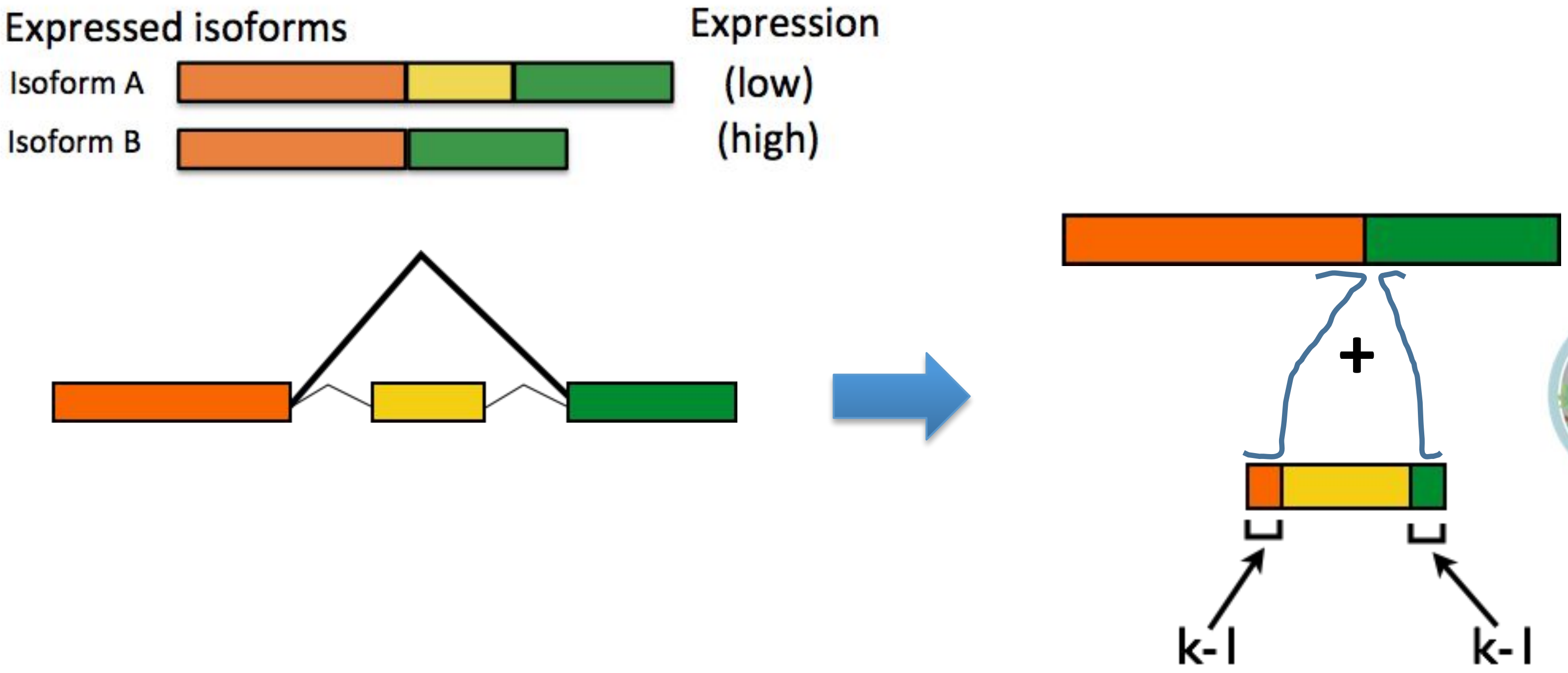
....AAGATTACAGA....

Remove assembled kmers from catalog, then repeat the entire process.

Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts

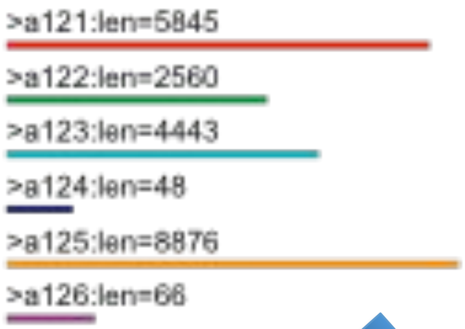


Inchworm can only report contigs derived from unique kmers.

Alternatively spliced transcripts :

- the more highly expressed transcript may be reported as a single contig,
- the parts that are different in the alternative isoform are reported separately.

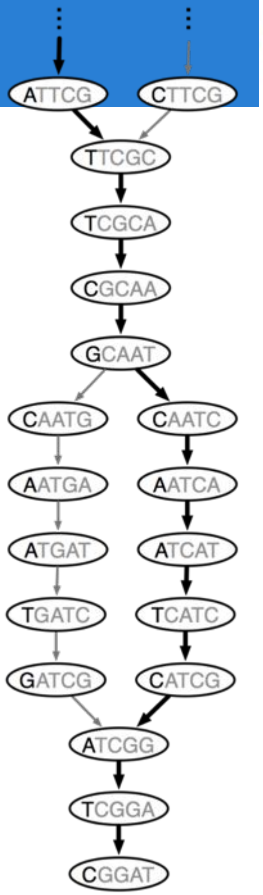
Chrysalis

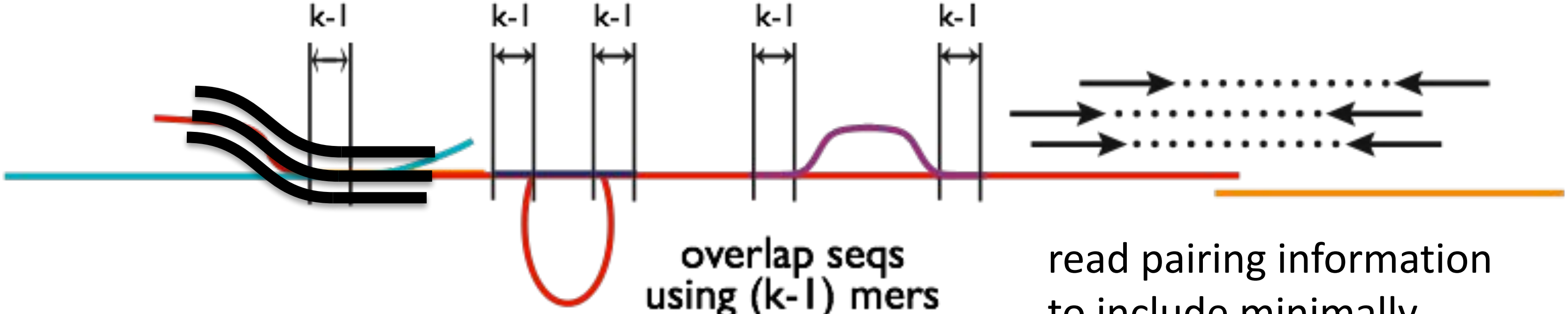


Integrate (clustering)
Isoforms via k-1 overlaps

Verify via "welds"

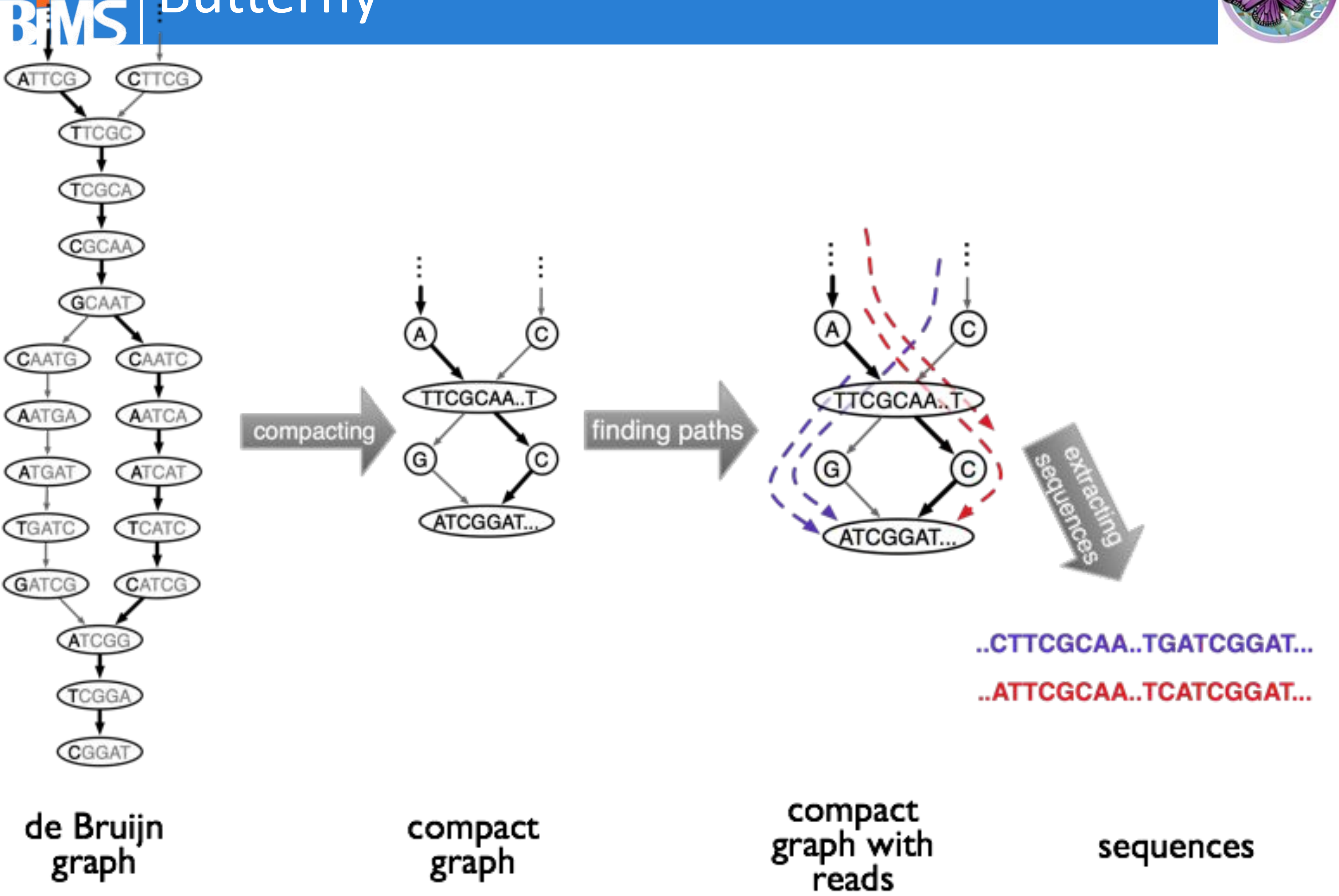
Build de Bruijn Graphs
(ideally, one per gene)





read pairing information
to include minimally
overlapping contigs

Butterfly


ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
compacting
A
C
TTCGCAA..T
G
C
ATCGGAT...
finding paths
A
C
TTCGCAA..T
G
C
ATCGGAT...
extracting sequences
..CTTCGCAA..TGATCGGAT...
..ATTCGCAA..TCATCGGAT...
de Bruijn graph
compact graph
compact graph with reads
sequences

Trinity usage and options

Typical Trinity command

```
Trinity --seqType fq --max_memory 50G
\--left A_rep1_left.fq  --right A_rep1_right.fq --CPU 4
```

```
Trinity --seqType fq --max_memory 50G --single single.fq
--CPU 4
```

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

The assembled transcripts will be found at 'trinity_out_dir/Trinity.fasta'.

Results

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

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

>TRINITY_DN810_c0_g1_i2 len=226 path=[407:0-225] [-1, 407, -2]
GATGATATCAACAATGAGACTTGTGAACCAGGTGAAGAAAACTCTTTCTTTGTATGCGAC
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

Typical rnaSPAdes command

```
spades.py -- rna --pe1-1 A_rep1_left.fq  --pe1-2
A_rep1_right.fq  -o spades_directory -t 4 -m 50
```

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 transcipts, 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 statistics

```
TRINITY_HOME/util/TrinityStats.pl Trinity.fasta
```

```
################################
## Counts of transcripts, etc.
################################
Total trinity 'genes': 7648
Total trinity transcripts: 7719
Percent GC: 38.88
########################################
Stats based on ALL transcript contigs:
########################################
Contig N10: 4318
Contig N20: 3395
Contig N30: 2863
Contig N40: 2466
Contig N50: 2065
Median contig length: 1038
Average contig: 1354.26
Total assembled bases: 10453524
```

```
#####################################################
## Stats based on ONLY LONGEST ISOFORM per 'GENE':
#####################################################
Contig N10: 4317
Contig N20: 3375
Contig N30: 2850
Contig N40: 2458
Contig N50: 2060
Median contig length: 1044
Average contig: 1354.49
Total assembled bases: 10359175
```

Trinity usage and options

Typical Trinity command with multiple samples

```
Trinity --seqType fq --max_memory 50G --CPU 4
\--left A_rep1_left.fq,A_rep2_left.fq
\--right A_rep1_right.fq,A_rep2_right.fq
```

```
sample.txt
cond_A    cond_A_rep1    A_rep1_left.fq    A_rep1_right.fq
cond_A    cond_A_rep2    A_rep2_left.fq    A_rep2_right.fq
cond_A    cond_A_rep3    A_rep3_left.fq    A_rep3_right.fq
cond_B    cond_B_rep1    B_rep1_left.fq    B_rep1_right.fq
cond_B    cond_B_rep2    B_rep2_left.fq    B_rep2_right.fq
cond_B    cond_B_rep3    B_rep3_left.fq    B_rep3_right.fq
```

```
Trinity --seqType fq --max_memory 50G --CPU 4
\--samples_file sample.txt
```

Trinity « genome guided »

If your RNA-Seq **sample differs sufficiently** from your reference genome and you'd like **to capture variations** within your assembled transcripts

De novo assembly is restricted to only those reads that map to the genome.

The advantage is that **reads that share sequence in common but map to distinct parts of the genome** will be targeted separately for assembly.

The disadvantage is that reads that do not map to the genome will not be incorporated into the assembly.
-> Unmapped reads can, however, be targeted for a separate genome-free de novo assembly.

Genome guided Trinity command

```
Trinity --genome_guided_bam rnaseq_alignments.csorted.bam
--max_memory 50G --genome_guided_max_intron 10000 --CPU 6
```

The assembled transcripts will be found at 'trinity_out_dir/Trinity-GG.fasta'.

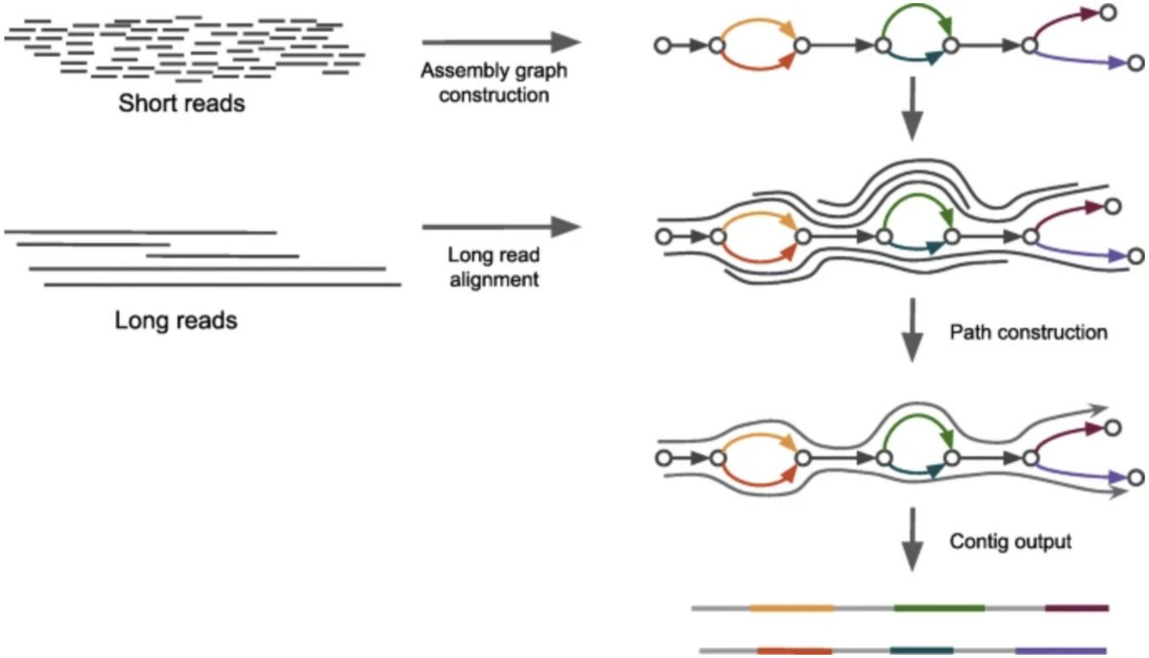
Trinity « longreads »

```
Trinity --seqType fq --max_memory 50G --CPU 4
\--samples_file sample.txt --long_reads contigs.fasta
```

Still Under development ☹

`contigs.fasta:`
fasta file containing error-corrected or circular consensus (CCS) PacBio reads

In short, the Trinity v2.4.0 version uses the pacbio reads mostly for path tracing in a graph that's built based on the illumina reads (not build using illumina AND pacbio) .

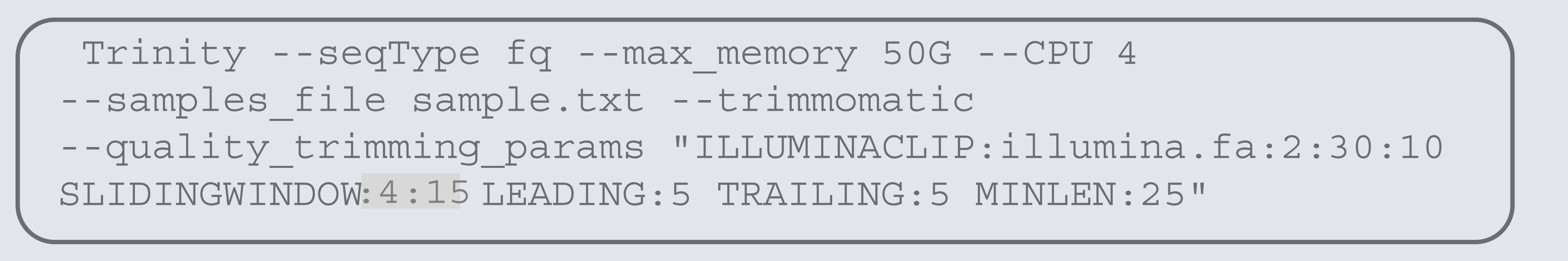


rnaSPAdes mode hybrid assembly you can use PacBio or Oxford Nanopore reads ☺ !

Prjibelski, A.D., Puglia, G.D., Antipov, D. *et al.* Extending rnaSPAdes functionality for hybrid transcriptome assembly. *BMC Bioinformatics* **21,** 302 (2020). https://doi.org/10.1186/s12859-020-03614-2

Trinity including trimming and normalisation

- Trimming



```
 Trinity --seqType fq --max_memory 50G --CPU 4
--samples_file sample.txt --trimmomatic
--quality_trimming_params "ILLUMINACLIP:illumina.fa:2:30:10
SLIDINGWINDOW:4:15 LEADING:5 TRAILING:5 MINLEN:25"
```

Trinity including trimming and normalisation

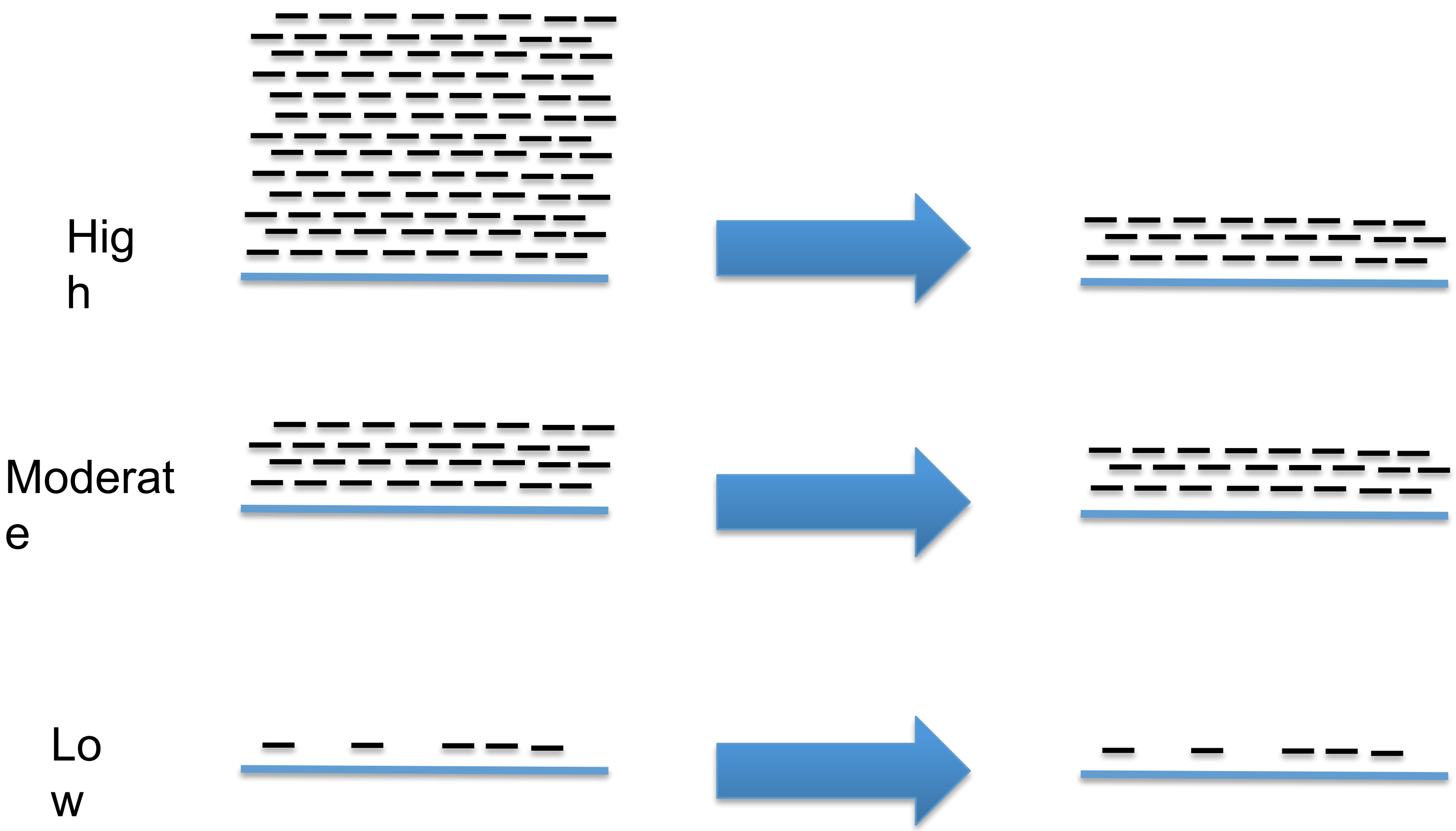
- Trimming

```
 Trinity --seqType fq --max_memory 50G --CPU 4
--samples_file sample.txt --trimmomatic
--quality_trimming_params "ILLUMINACLIP:illumina.fa:2:30:10
SLIDINGWINDOW:4:15 LEADING:5 TRAILING:5 MINLEN:25"
```

- Normalisation:
  - 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



NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):
   - with k = 25

2. For each read, compute the `median`, `average` and `stdev` kmers coverage

3. Accept a read with a probability of:

```
max coverage/median
```

NGS reads normalization (by Trinity)

3. Accept a read with a probability of:
   $max\_coverage/median$

e.g. with $max\ coverage = 30$

Read_A: $median\ coverage = 60 \rightarrow \frac{max\_coverage}{median} = 0.5$

➔ Read_A has a 50% chance of being kept

Read_B: $median\ coverage = 10 \rightarrow \frac{max\_coverage}{median} = 3$

➔ Read_B has a 300% chance of being kept ;-)

➔ Read_B will be kept

NGS reads normalization (by Trinity)

3. Accept a read with a probability of:

$$\mathit{max\_coverage}/\mathit{median}$$

Reads coming from a highly expressed transcript and are several times more covered than the threshold.

□ Its information is also contained by other reads.

□ So it has less chance to be kept.

Reads coming from a low expressed transcript, way below the threshold.

□ Its information is not very redondant, need it for the assembly.

□ So it will absolutly be kept

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):
   - with k = 25

2. For each read, compute the `median`, `average` and `stdev` kmers coverage

3. Accept a read with a probability of: *maxcov/median*

4. Remove a read if: *standartdev/average* (CV)> 1 (100%)

A high variability in a read kmer coverage means there is probably a lot of sequencing errors in this read

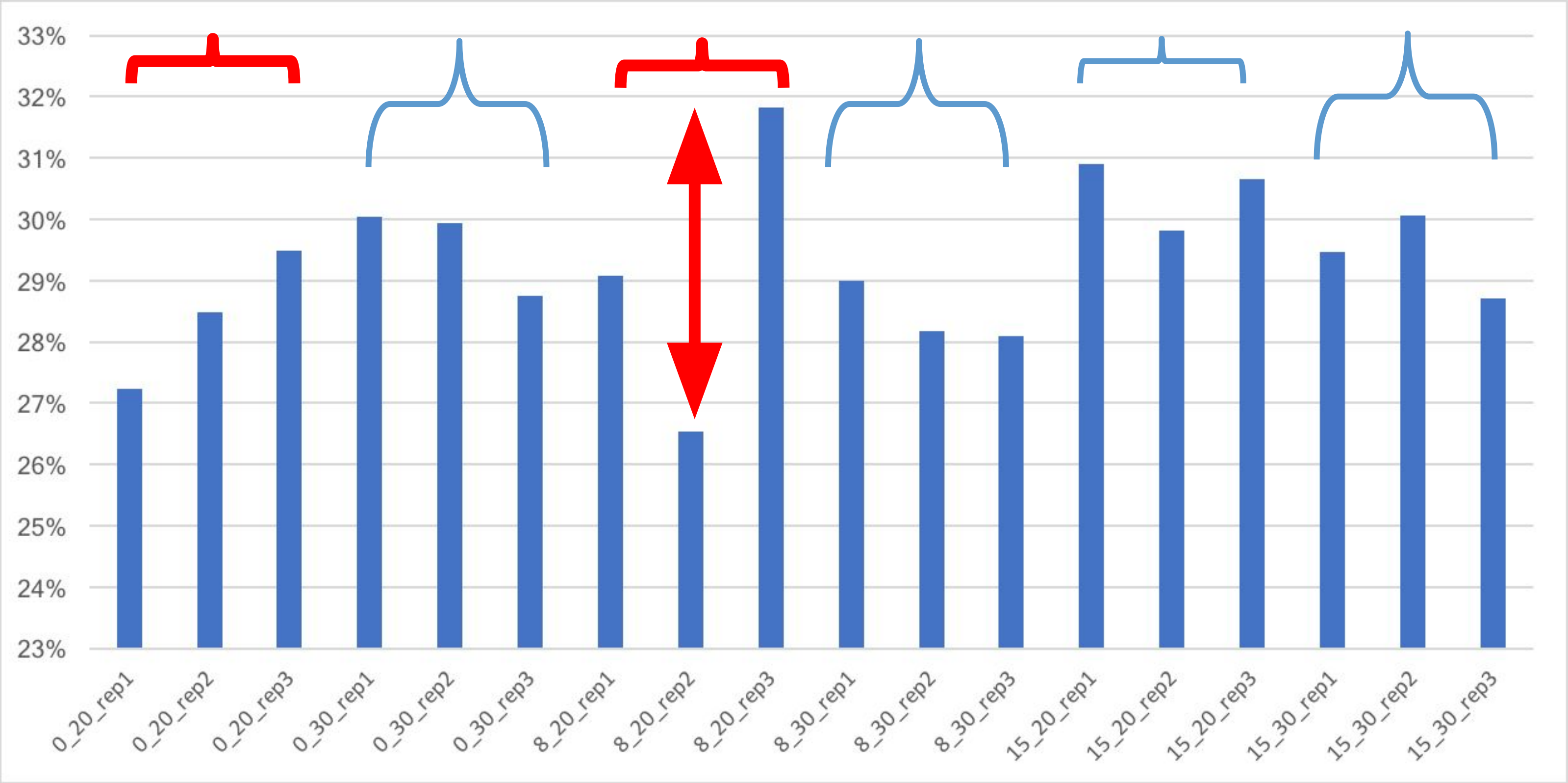
Stand alone normalisation

```
$TRINITY_HOME/util/insilico_read_normalization.pl
\ --seqType fq --JM 1G --max_cov 50
\ --left lib1_1.P.qtrim --right lib2_2.P.qtrim
\ --pairs_together --output insil_norm_ex
```

```
1189570 / 1879312 = 63.30% reads selected during normalization.
1094 / 1879312 = 0.06% reads discarded as likely aberrant based on
coverage profiles.

Normalization complete. See outputs:
insil_norm_ex/lib1_1.P.qtrim.normalized_K25_C50_pctSD200.fq
insil_norm_ex/lib1_2.P.qtrim.normalized_K25_C50_pctSD200.fq
```

Trinity normalisation

```
 Trinity --seqType fq --max_memory 50G --CPU 4
--samples_file sample.txt --trimmomatic
--quality_trimming_params "ILLUMINACLIP:illumina.fa:2:30:10
SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25
--normalize_by_read_set
```

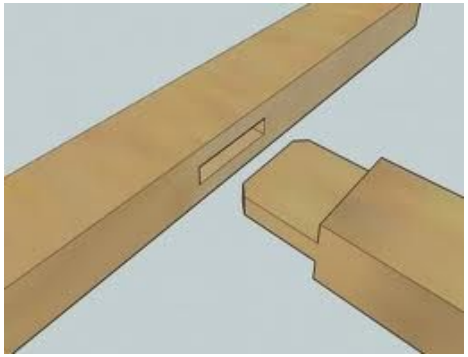


Reads *in silico* normalisation

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

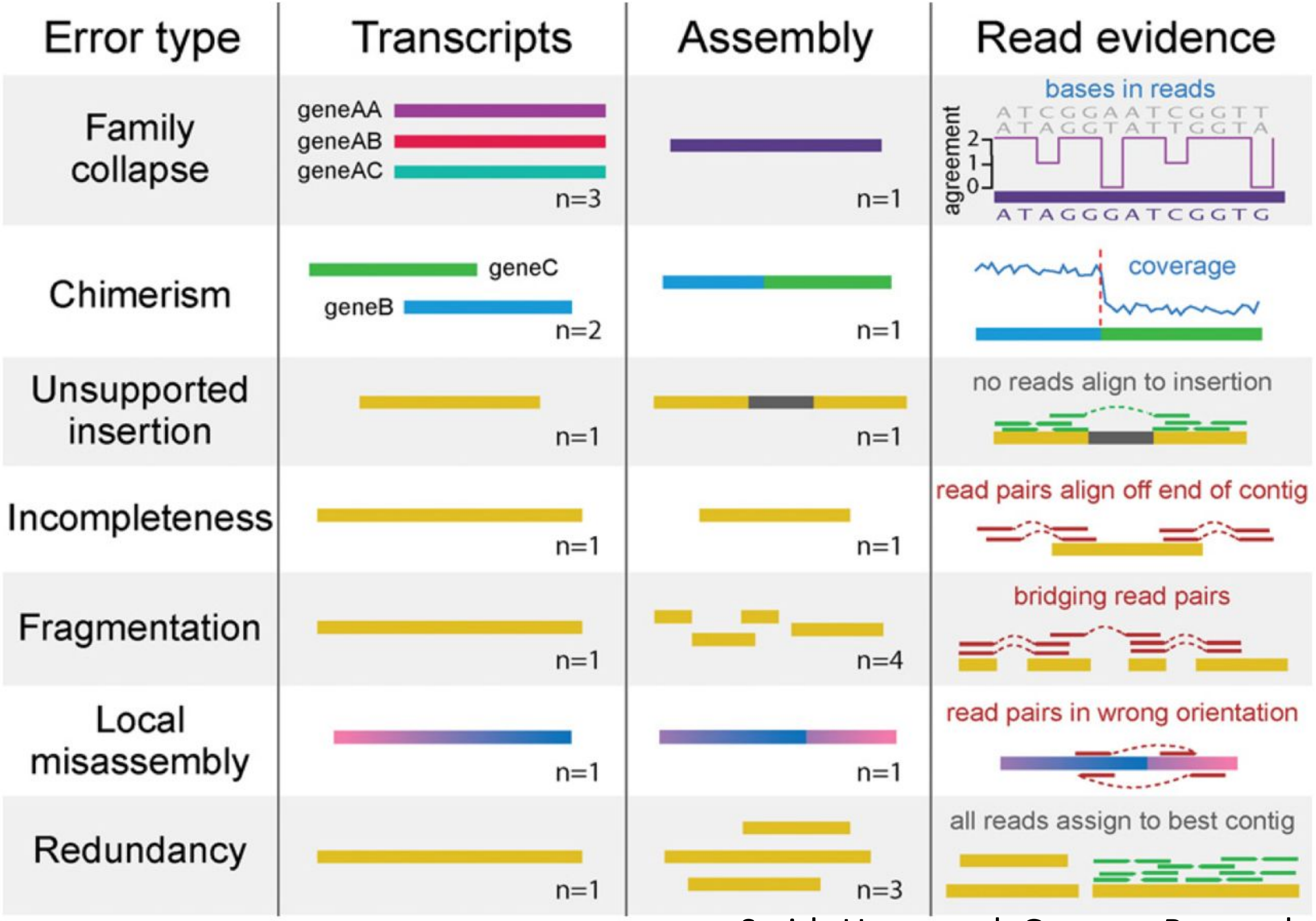
RNA Seq analysis



Transcriptome assembly

ASSEMBLY QUALITY ASSESSMENT AND CLEANNING

De novo Transcriptome Assembly is Prone to Certain Types of Errors



Smith-Unna et al. Genome Research, 2016

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.

Metrics

- 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 Assembly is a sum-up.

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

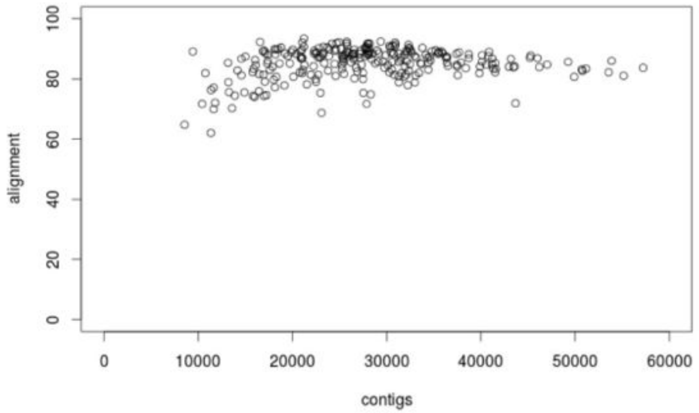
-> compute percentage of reads mapped

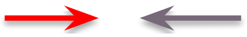
Factors affecting realignment rate:

- Presence of highly expressed genes
- Contamination by building blocks (adaptors)
- Reads quality

Realignment metrics

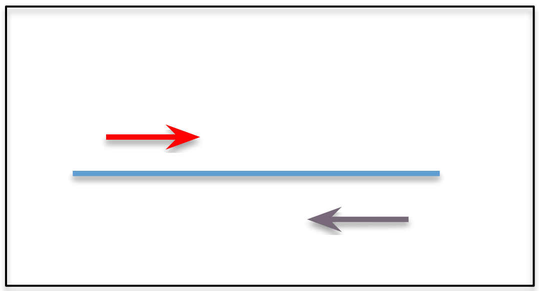
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.



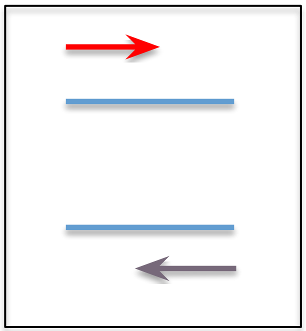
Given read pair: 

Possible mapping contexts in the Trinity assembly are reported:

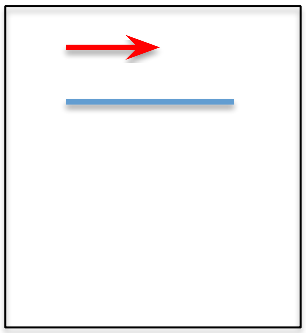
Proper pairs



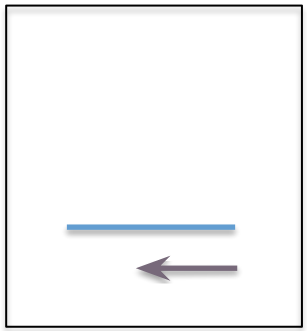
Improper pairs



Left only



Right only



Assembly evaluation : read remapping

Alignment methods : bowtie2 -RSEM

```
$TRINITY_HOME/util/align_and_estimate_abundance.pl --seqType fq
--transcripts Trinity.fasta --est_method RSEM --aln_method bowtie2
--prep_reference --trinity_mode --samples_file samples.txt
--seqType fq
```

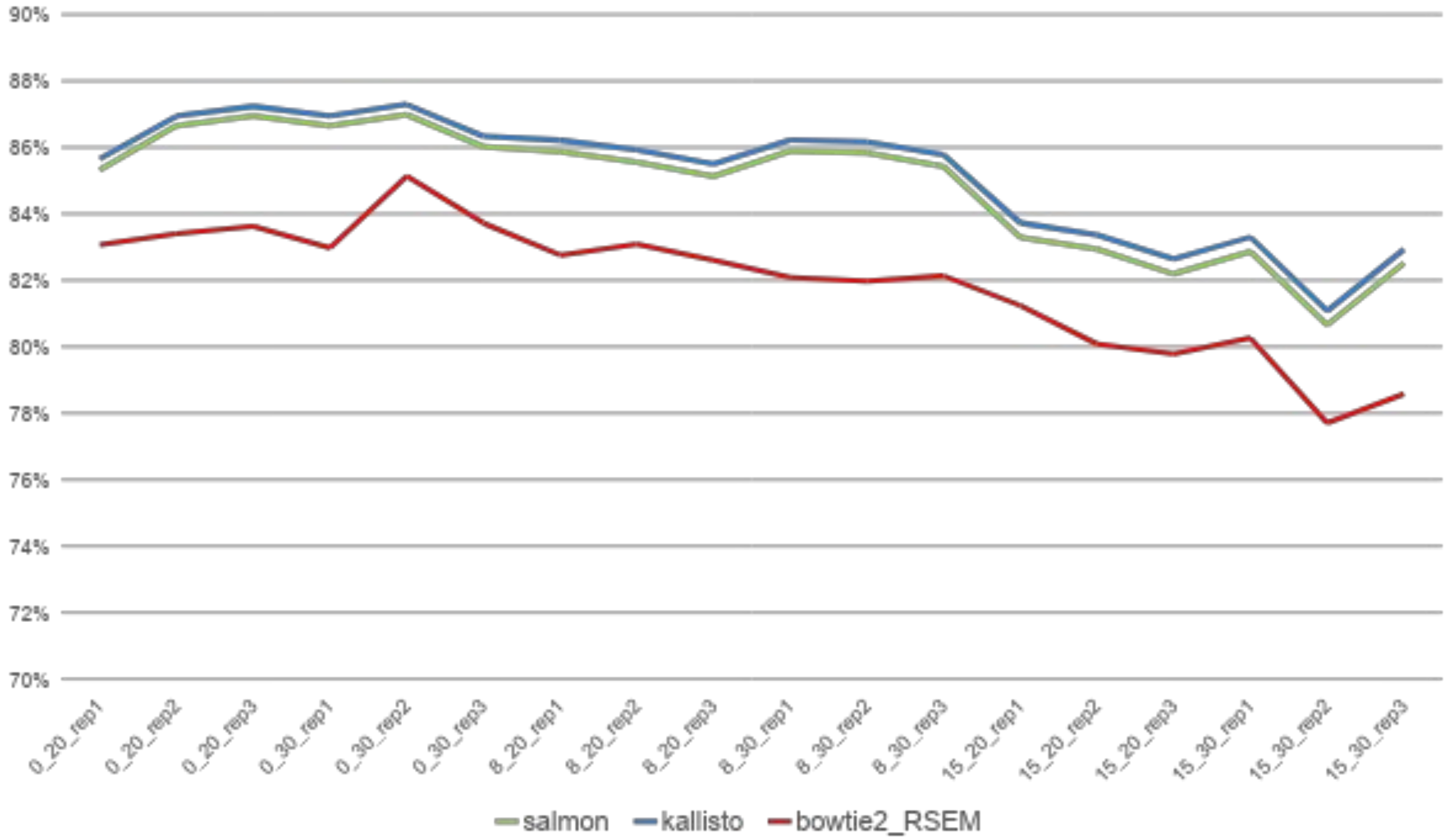
Pseudo-Alignment methods : kallisto

```
$TRINITY_HOME/util/align_and_estimate_abundance.pl --seqType fq
--transcripts Trinity.fasta --est_method kallisto --prep_reference
--trinity_mode --samples_file samples.txt --seqType fq
```

Pseudo-Alignment methods : salmon

```
$TRINITY_HOME/util/align_and_estimate_abundance.pl --seqType fq
--transcripts Trinity.fasta --est_method salmon --prep_reference
--trinity_mode --samples_file samples.txt --seqType fq
```

Realignment metrics



Assembly evaluation : read remapping

Pseudo-Alignment methods : kallisto (salmon : quant.sf ; quant.sf.genes)

```
head cond_A_rep1/abundance.tsv | column -t
Or
head cond_A_rep1/abundance.tsv.genes | column -t
```

```
target_id             length     eff_length      est_counts       tpm
TRINITY_DN144_c0_g1_i1     4833  4703.42         138        16.266
TRINITY_DN144_c0_g2_i1     2228  2098.42         0.000103136      2.72479e-05
TRINITY_DN179_c0_g1_i1     1524  1394.42         227        90.2502
TRINITY_DN159_c0_g1_i1     659   529.534         7.75713          8.12123
TRINITY_DN159_c0_g2_i1     247   119.949         0.24287          1.12251
TRINITY_DN153_c0_g1_i1     2378  2248.42         16         3.9451
TRINITY_DN130_c0_g1_i1     215   89.2898         776        4818.09
TRINITY_DN130_c1_g1_i1     295   166.986         216        717.115
TRINITY_DN106_c0_g1_i1     4442  4312.42         390        50.137
```

```
target_id             length     eff_length      est_counts       tpm
TRINITY_DN2774_c0_g1       2926.00    2796.42          31.00           6.15
TRINITY_DN5482_c0_g1       3064.00    2934.42          344.00          64.99
TRINITY_DN6803_c0_g1       1439.00    1309.42          1379.00         583.85
TRINITY_DN386_c0_g2        4279.00    4149.42          3.23       0.43
TRINITY_DN23_c0_g2         632.00     502.53           9.99       11.02
TRINITY_DN5348_c0_g1       2091.00    1961.42          264.00          74.62
TRINITY_DN5222_c0_g1       2416.00    2286.42          148.00          35.89
TRINITY_DN4680_c0_g1       1420.00    1290.42          167.00          71.75
TRINITY_DN2900_c0_g1       283.00     155.12           1.00       3.57
```

Expression matrix construction

```
$TRINITY_HOME/util/abundance_estimates_to_matrix.pl
\ --est_method kallisto --out_prefix Trinity_trans
\ --name_sample_by_basedir
\ cond_A_rep1/abundance.tsv
\ cond_A_rep2/abundance.tsv
\ cond_B_rep1/abundance.tsv
\ cond_B_rep2/abundance.tsv
```

Two matrices,

- one containing the estimated counts,
- one containing the TPM expression values that are cross-sample normalized using the TMM method.

TMM normalization assumes that most transcripts are not differentially expressed, and linearly scales the expression values of samples to better enforce this property.

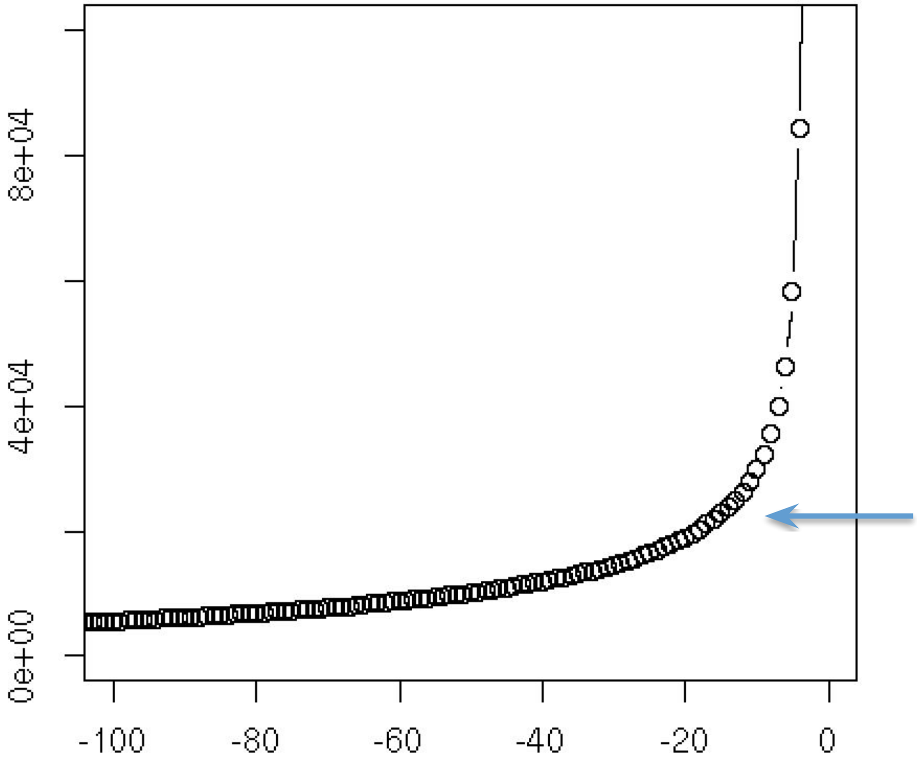
A scaling normalization method for differential expression analysis of RNA-Seq data, Robinson and Oshlack, Genome Biology 2010.

Alternative to N50 ?

Often, most assembled transcripts are *very* lowly expressed

(How many ‘transcripts & genes’ are there really?)

Cumulative # of Transcripts



-1 * minimum TPM

1.4 million Trinity transcript contigs

N50 ~ 500 bases

20k transcripts

Expression

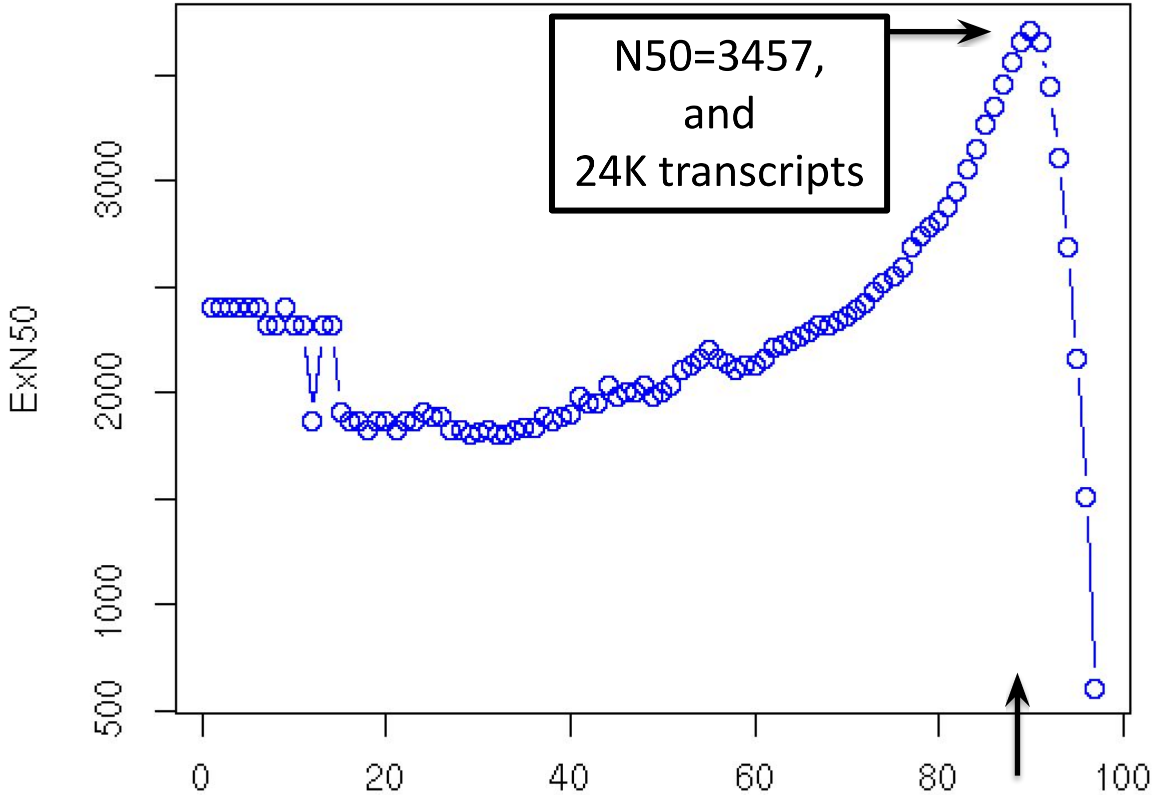
* Salamander transcriptome

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

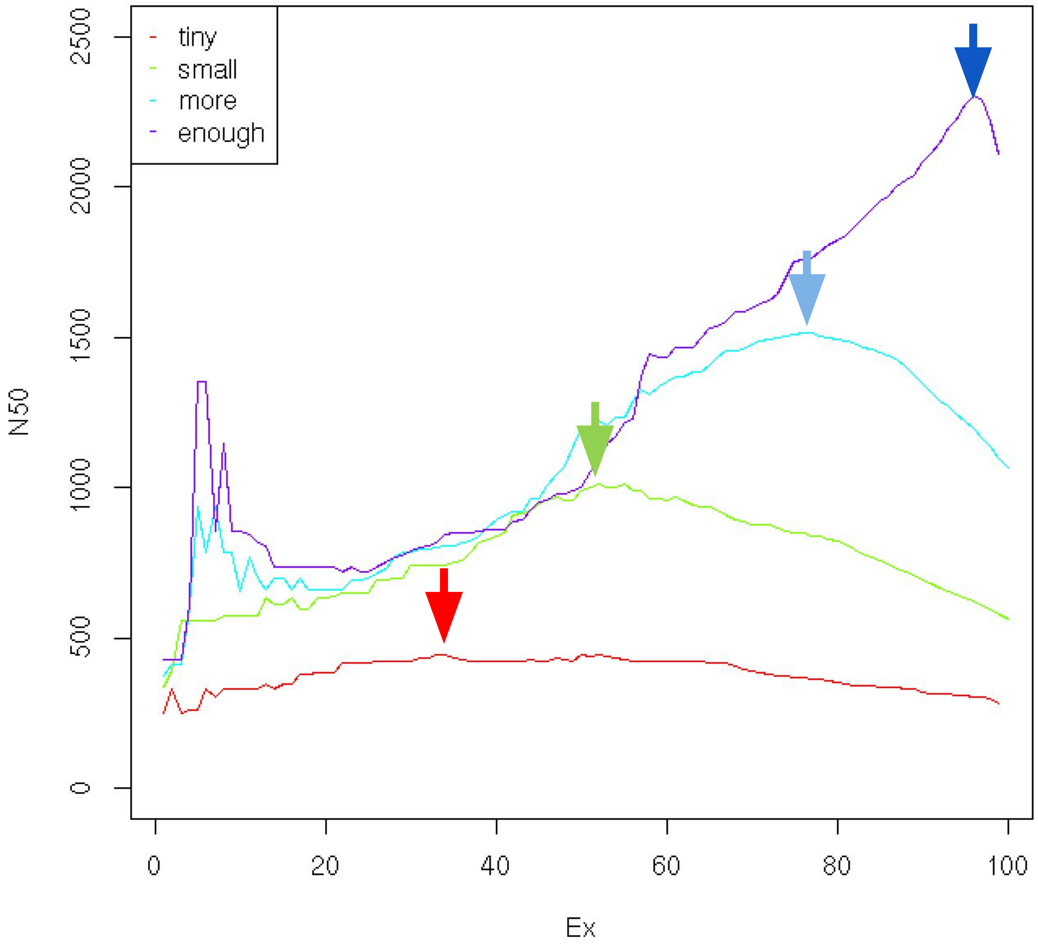


N50=3457, and 24K transcripts

90% of expression data

```
$TRINITY_HOME/util/misc/contig_ExN50_statistic.pl
\Trinity_trans.TMM.EXPR.matrix Trinity.fasta > ExN50.stats
```

ExN50 Profiles for Different Trinity Assemblies Using Different Read Depths



Millions of Reads

Thousands of Reads

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

* Candida transcriptome

Tools to evaluate transcriptomes

Transrate: understand your transcriptome assembly. http://hibberdlab.com/transrate

Transrate analyses a transcriptome assembly in three key ways:

- by inspecting the contig sequences
- by mapping reads to the contigs and inspecting the alignments
- by aligning the contigs against proteins or transcripts from a related species and inspecting the alignments
  - Assemblies score
  - Contigs score
  - Optimised assemblies score (filter out bad contigs from an assembly, leaving you with only the well-assembled ones)

Aternatives :
- Detonate
- RNAquast (https://github.com/ablab/rnaquast)

BUSCO analysis

CEGMA (http://korflab.ucdavis.edu/datasets/cegma/)

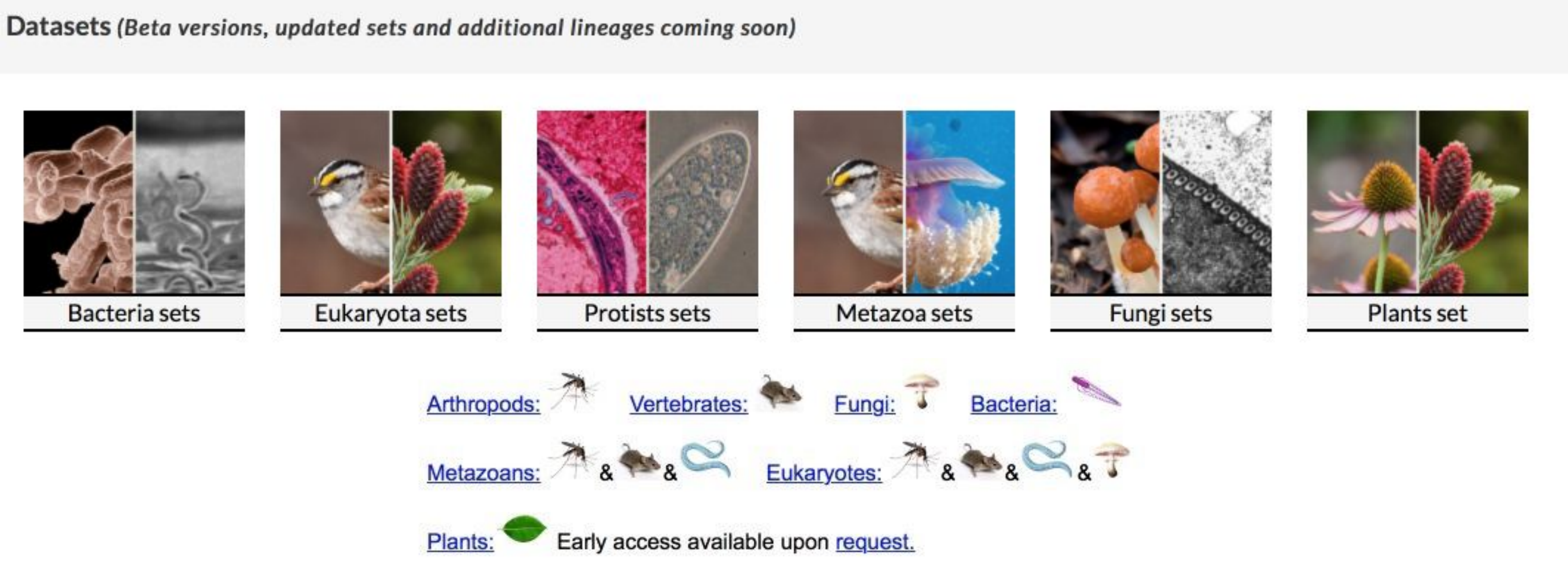
HMM:s for 248 core eukaryotic genes aligned to your assembly to assess completeness of gene space

“complete”: 70% aligned
“partial”: 30% aligned

BUSCO

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

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



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:

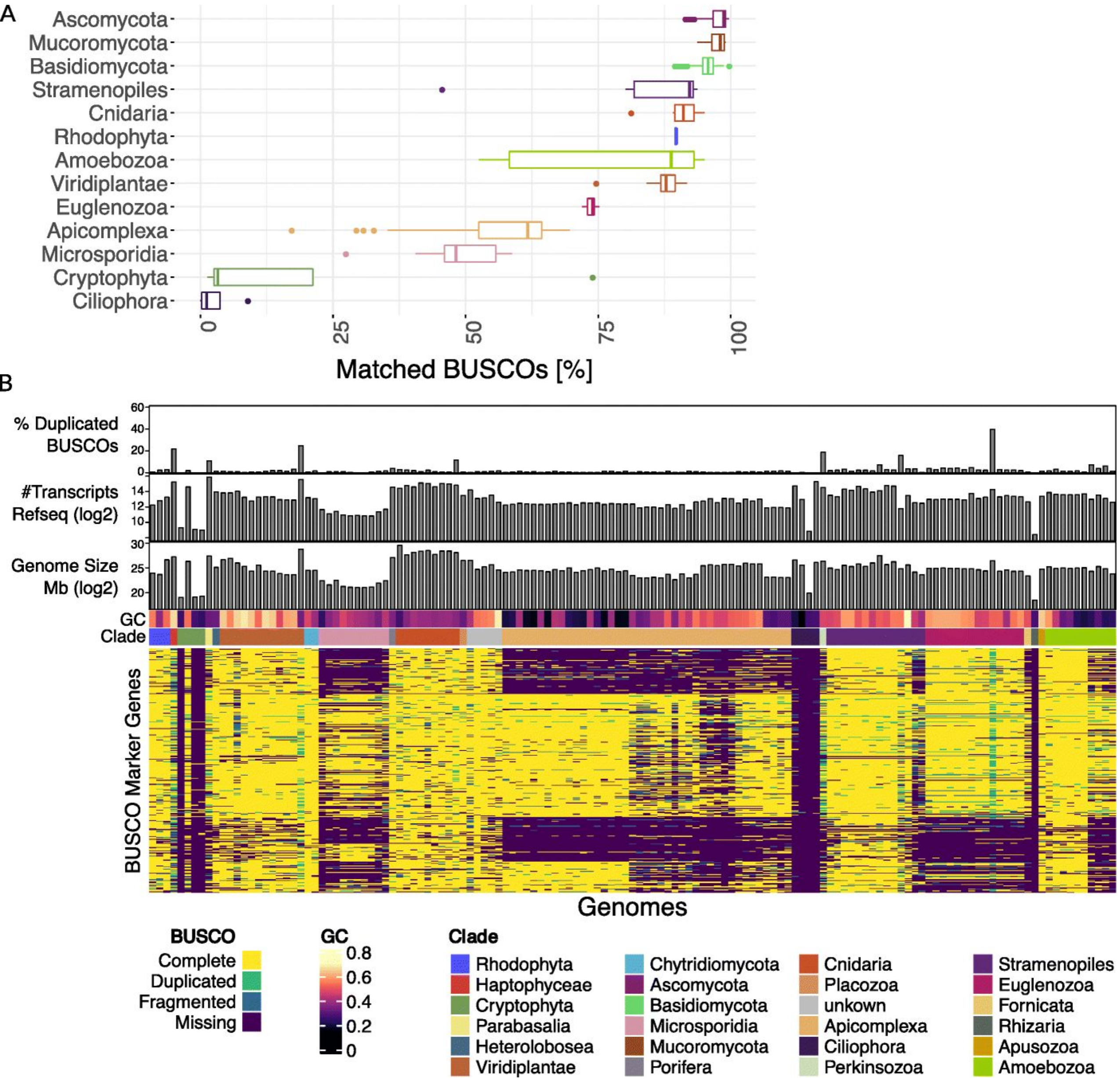
Metazoans: & & Eukaryotes: & & &

Plants: Early access available upon request.

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 limitation

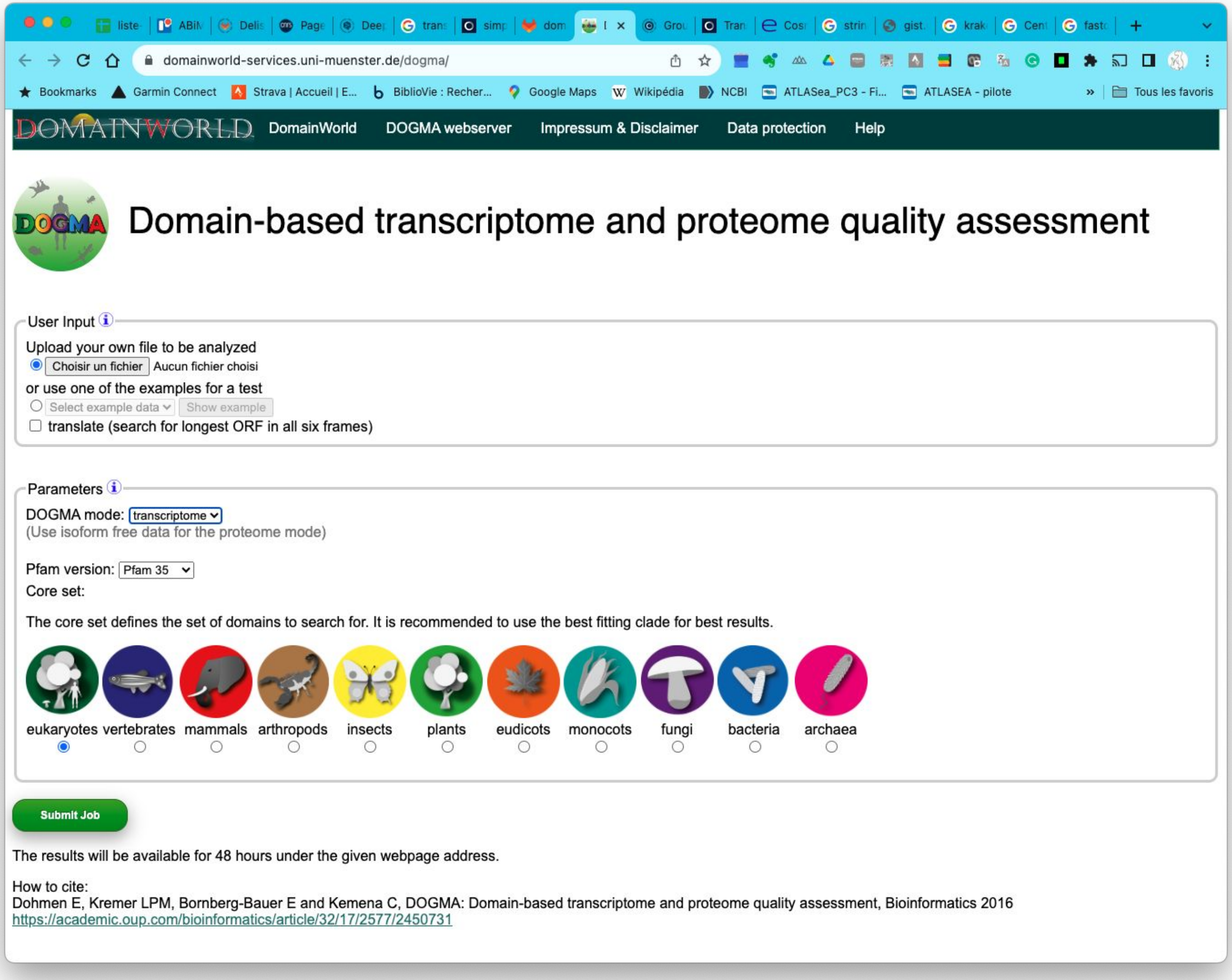


https://github.com/Finn-Lab/EukCC/

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/s13059-020-02155-4

BUSCO Alternative

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



Domain-based transcriptome and proteome quality assessment

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)

Pfam version: Pfam 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 ○

Submit Job

The results will be available for 48 hours under the given webpage address.

How to cite:
Dohmen E, Kremer LPM, Bornberg-Bauer E and Kemena C, DOGMA: Domain-based transcriptome and proteome quality assessment, Bioinformatics 2016
https://academic.oup.com/bioinformatics/article/32/17/2577/2450731

Transcriptome thinning / redondancy reduction

de novo assembled contigs include transcriptional artifacts,

- pre-mRNA and ncRNA in addition to the protein-coding transcripts [61].
- alternative splicing which manifests as transcript isoforms.

It may not always be necessary to retain all such sequences

Most of the de novo assembler identify isoformes

SPAdes:

```
NODE_217_length_5488_cov_28.472321_g133_i0
NODE_260_length_5302_cov_30.767137_g133_i1
```

Trinity :

```
TRINITY_DN98_c5_g1_i0
TRINITY_DN98_c5_g1_i1
```

A representative isoform can be chosen in several different ways:

- the isoform with the highest read support,
- the longest isoform,
- the isoform that produces the longest translated amino acid sequence
- the isoform whose coding sequence (CDS) has the highest read support.

Transcriptome thinning / redondancy reduction

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

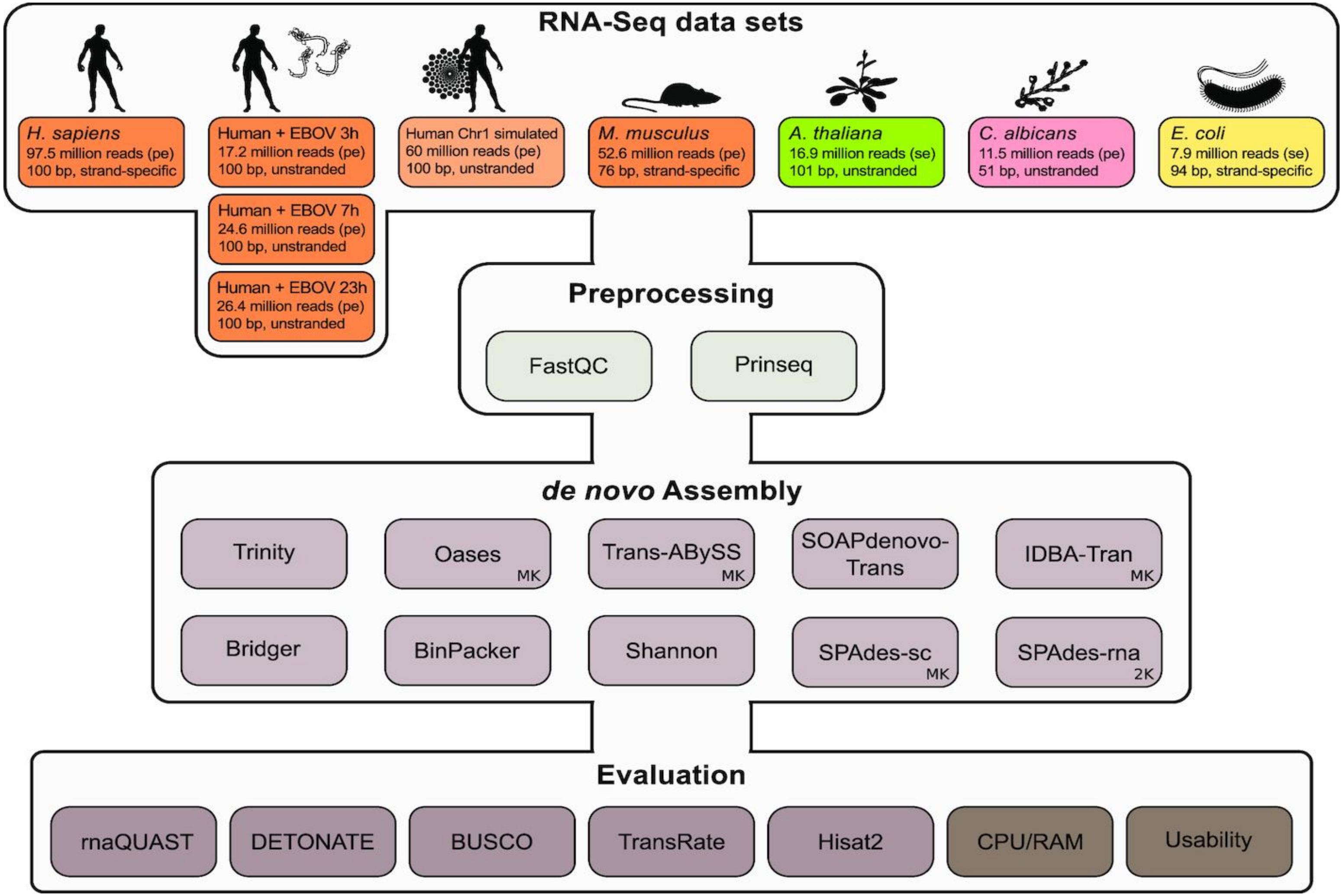
TransLig - https://sourceforge.net/projects/transcriptomeassembly/

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

Assemblers comparison

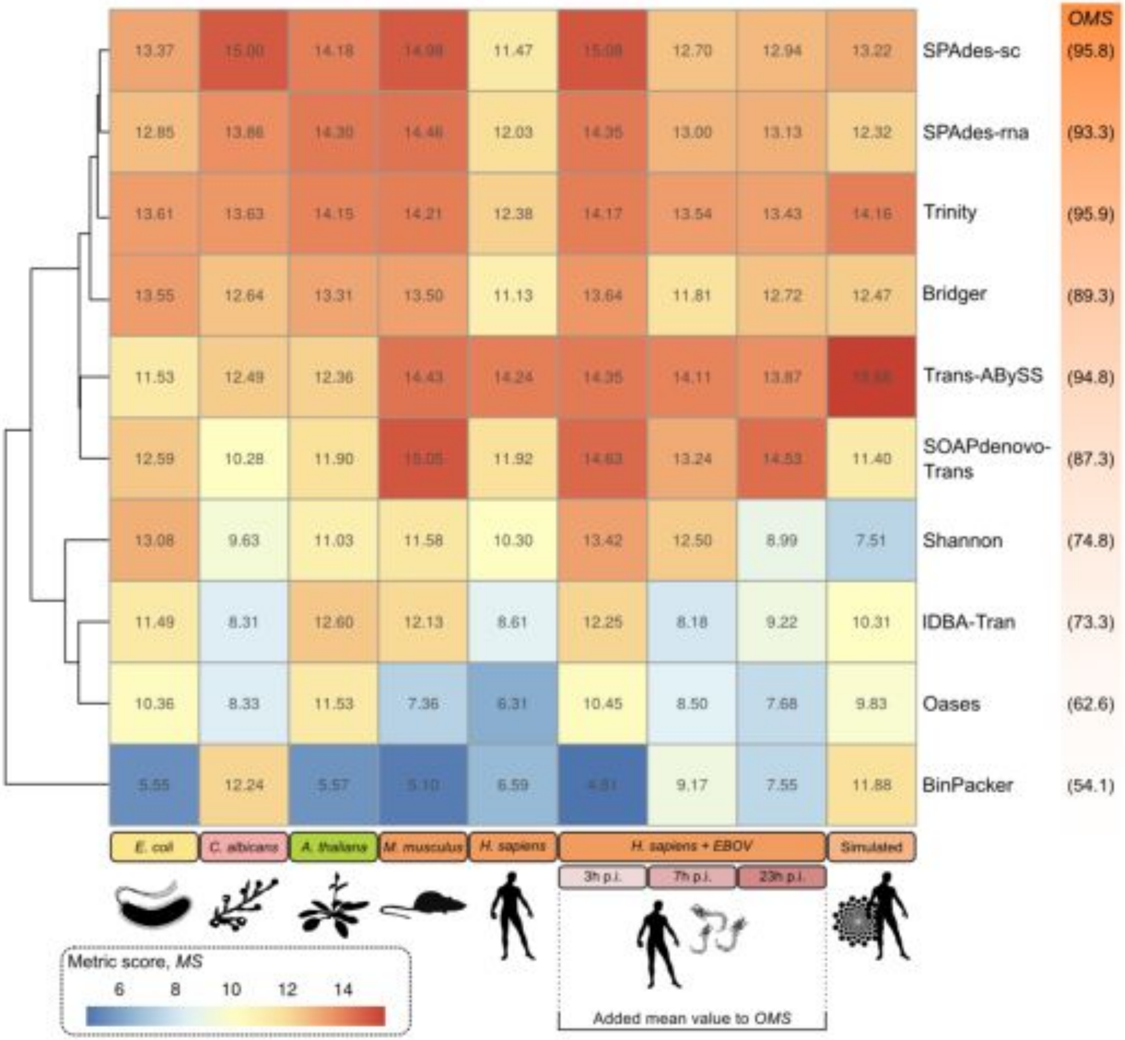
- Qiong-Yi Zhao et al., Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 2011, 12(Suppl 14):S2
- Clarke, K., Yang, Y., Marsh, R., Xie, L., & Zhang, K. K. (2013). Comparative analysis of de novo transcriptome assembly. Science China Life Sciences, 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

Assemblers comparison

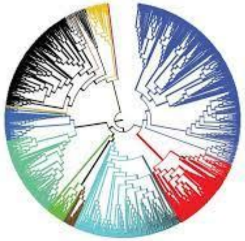

RNA-Seq data sets
H. sapiens 97.5 million reads (pe) 100 bp, strand-specific
Human + EBOV 3h 17.2 million reads (pe) 100 bp, unstranded
Human Chr1 simulated 60 million reads (pe) 100 bp, unstranded
M. musculus 52.6 million reads (pe) 76 bp, strand-specific
A. thaliana 16.9 million reads (se) 101 bp, unstranded
C. albicans 11.5 million reads (pe) 51 bp, unstranded
E. coli 7.9 million reads (se) 94 bp, strand-specific
Human + EBOV 7h 24.6 million reads (pe) 100 bp, unstranded
Human + EBOV 23h 26.4 million reads (pe) 100 bp, unstranded
Preprocessing
FastQC
Prinseq
de novo Assembly
Trinity
Oases MK
Trans-ABySS MK
SOAPdenovo-Trans
IDBA-Tran MK
Bridger
BinPacker
Shannon
SPAdes-sc MK
SPAdes-rna 2K
Evaluation
rnaQUAST
DETONATE
BUSCO
TransRate
Hisat2
CPU/RAM
Usability

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

Assemblers comparison



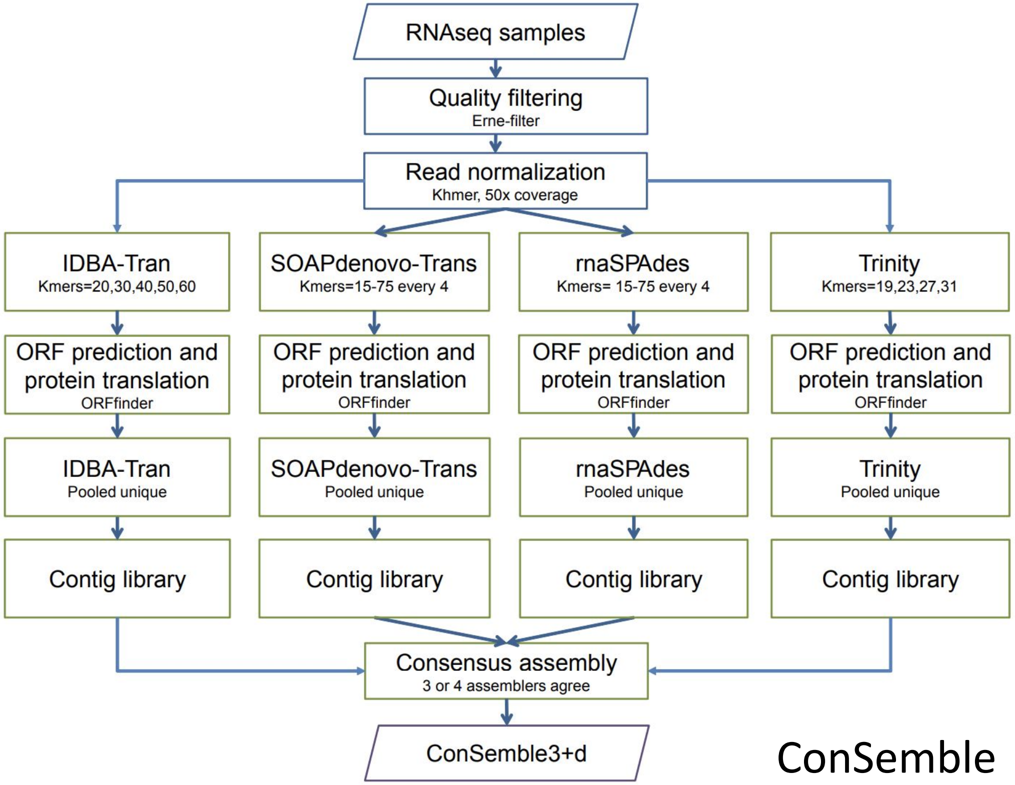
	E. coli	*C. albicans*	*A. thaliana*	*M. musculus*	*H. sapiens*	*H. sapiens + EBOV* 3h p.i.	*H. sapiens + EBOV* 7h p.i.	*H. sapiens + EBOV* 23h p.i.	Simulated		*OMS*
	13.37	15.00	14.18	14.98	11.47	15.08	12.70	12.94	13.22	SPAdes-sc	(95.8)
	12.85	13.86	14.30	14.46	12.03	14.35	13.00	13.13	12.32	SPAdes-rna	(93.3)
	13.61	13.63	14.15	14.21	12.38	14.17	13.54	13.43	14.16	Trinity	(95.9)
	13.55	12.64	13.31	13.50	11.13	13.64	11.81	12.72	12.47	Bridger	(89.3)
	11.53	12.49	12.36	14.43	14.24	14.35	14.11	13.87	[illegible]	Trans-ABySS	(94.8)
	12.59	10.28	11.90	15.05	11.92	14.63	13.24	14.53	11.40	SOAPdenovo-Trans	(87.3)
	13.08	9.63	11.03	11.58	10.30	13.42	12.50	8.99	7.51	Shannon	(74.8)
	11.49	8.31	12.60	12.13	8.61	12.25	8.18	9.22	10.31	IDBA-Tran	(73.3)
	10.36	8.33	11.53	7.36	6.31	10.45	8.50	7.68	9.83	Oases	(62.6)
	5.55	12.24	5.57	5.10	6.59	4.81	9.17	7.55	11.88	BinPacker	(54.1)



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

CNRS • SORBONNE UNIVERSITE Station Biologique de Roscoff ABiMS

New strategies



ConSemble

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

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

Mathieu c'est à toi !!

- 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