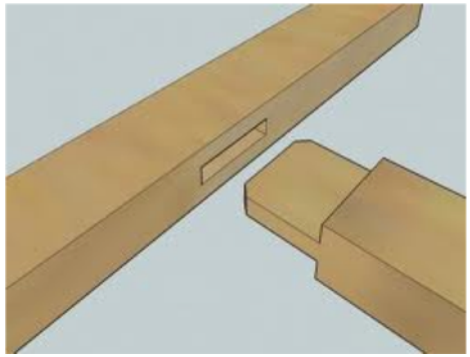
25/09/2019

RNA Seq analysis

Assembly quality assesment

ABiMS – Station Biologique Roscoff

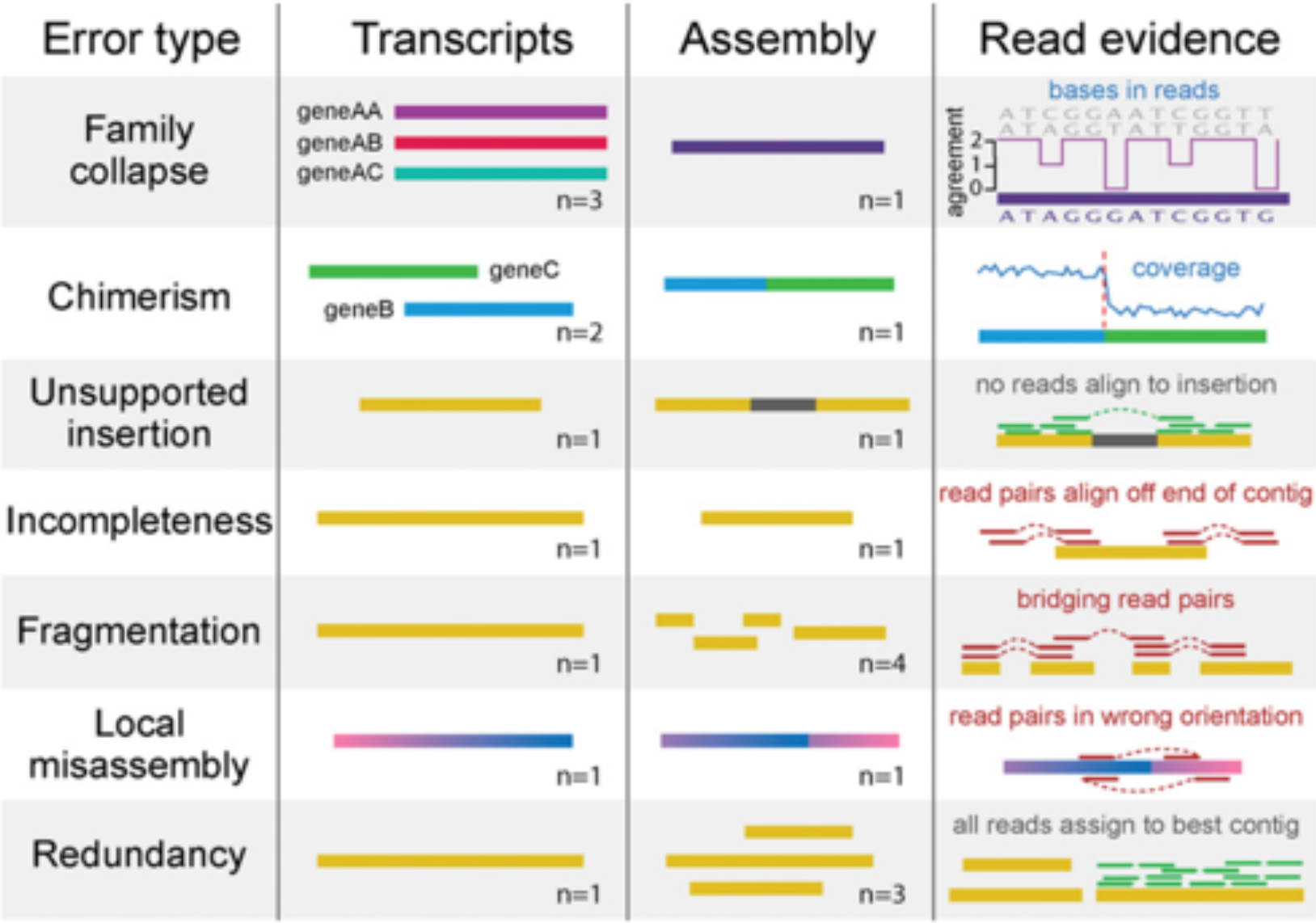
RNA Seq analysis



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

- Assembly metrics
- Contigs length histogram and proteome comparison
- Reads mapping back rate

Metrics

The possible metrics derived from genome assembly:

- Idea of global size (# bases)
- Idea of number of elements (#contigs/scaffolds)
- Idea of compactness (N50):

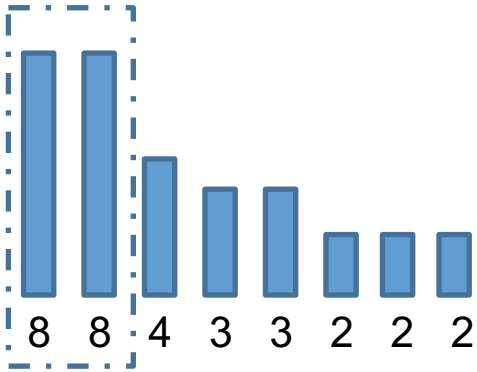
Metrics

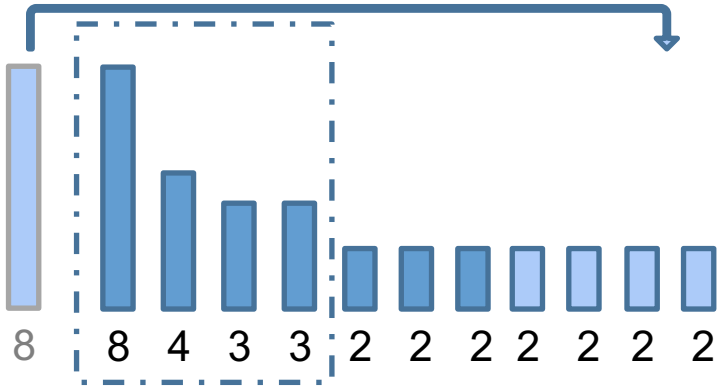
- The number of contigs in the assembly
- The size of the smallest contig
- The size of the largest contig
- The number of bases included in the assembly
- The mean length of the contigs
- The number of contigs <200 bases
- The number of contigs >1,000 bases
- The number of contigs >10,000 bases
- The 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
- At skew
- The number of bases that are N
- The proportion of bases that are N
- The total linguistic complexity of the assembly

N50

- **N50**: given a set of contigs of varying lengths, the N50 length is defined as the length N for which 50% of all bases in the contigs are in contigs of length L < N

contig size list L = (8,8,4, 3, 3, 2, 2, 2) = 32
we have 50% of total length (16/32) above 4 -> **N50** is equal to 8





N50 = 8

Average : 32/8 = 4

Mediane = 3

N50 = 3

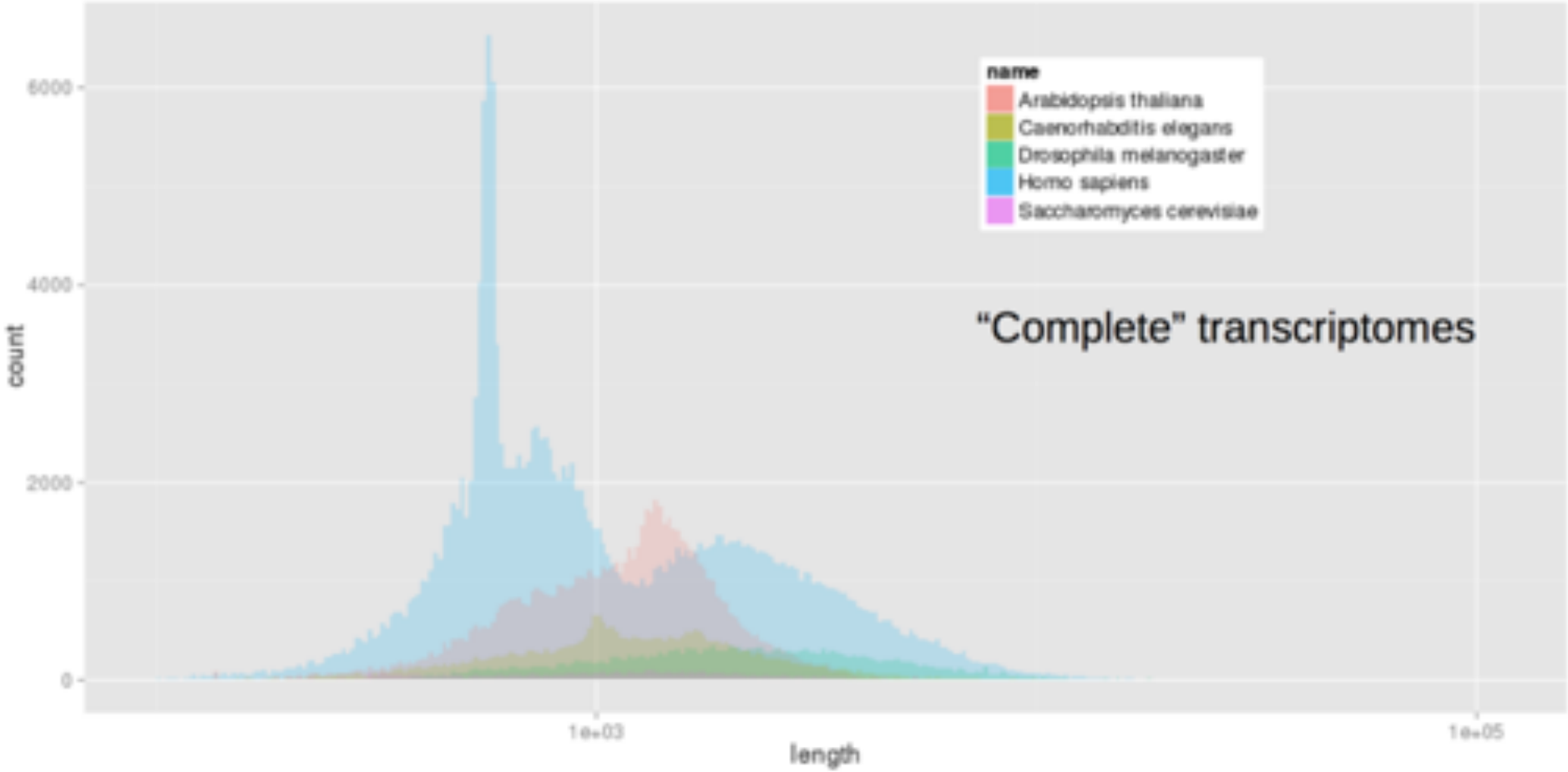
Average : 32/11 = 2.9

Mediane = 2

much more difficult to predict with transcriptome data

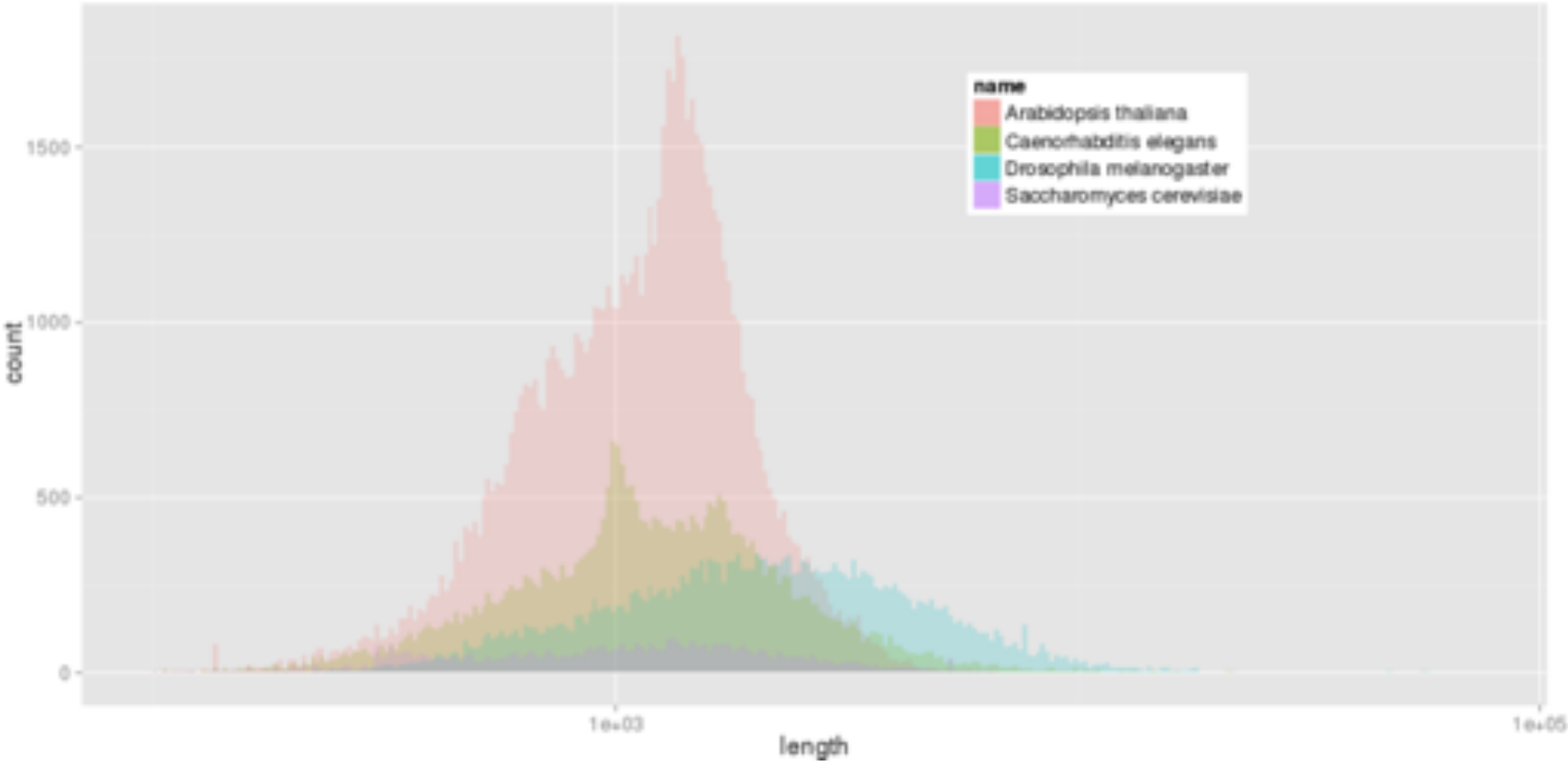
Transcripts length histogram

Transcript lengths are not randomly distribute :
-> We should get a known distribution shape



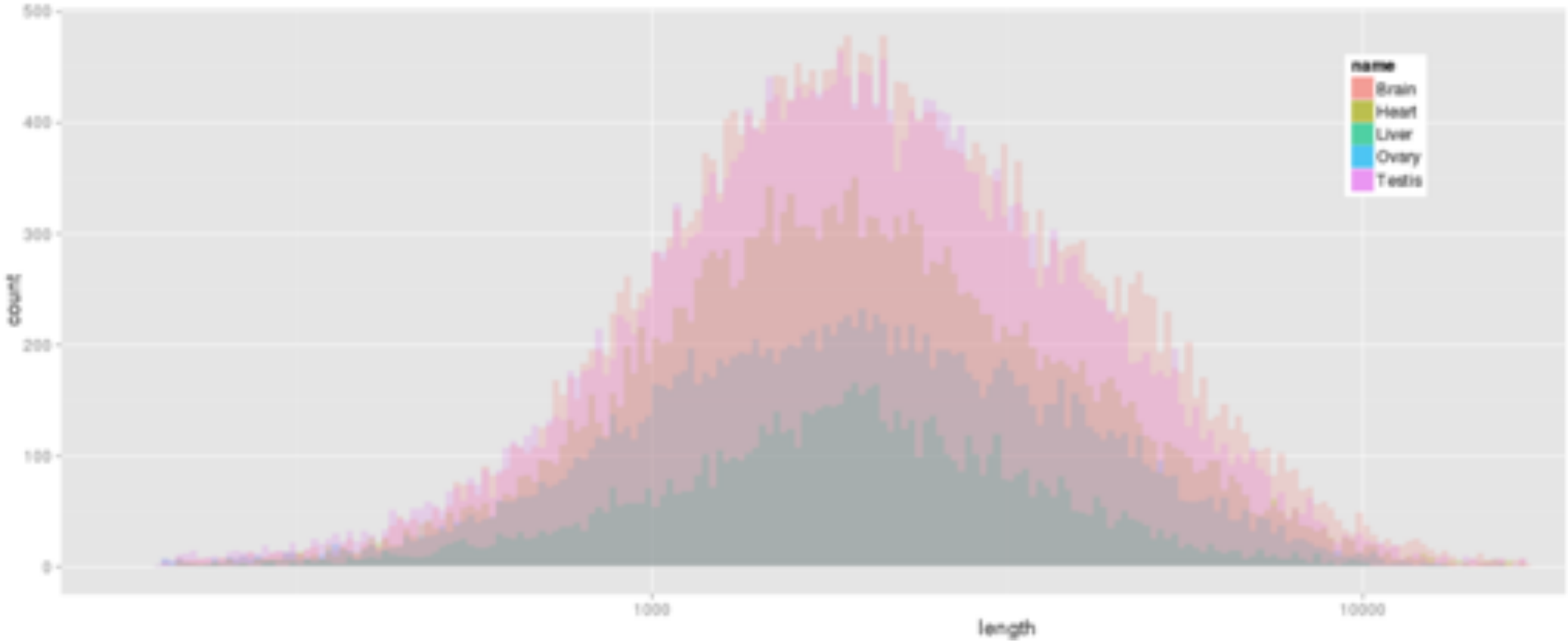
Transcripts length histogram

RNAseq data



Transcripts length histogram

Zebrafish tissue specific assembled transcriptomes : not so different



Practice

3

Aller sur la practice 3 Assessing transcriptome assembly quality *du github*

3.1 Getting basic Assembly metrics with the trinity script TrinityStats.pl

3.2 Reads mapping back rate and abundance estimation using the trinity script align_and_estimate_abundance.pl

Tools to evaluate transcriptomes

Since a reference genome is not available, the quality of computer-assembled contigs may be verified :

- by comparing the assembled sequences to the reads used to generate them (reference-free)

- by aligning the sequences of conserved gene domains found in mRNA transcripts to transcriptomes or genomes of closely related species (reference-based).

Realignment metrics

The assembly is a sum-up. The realignment rate gives how much of the initial information is inside the contigs.

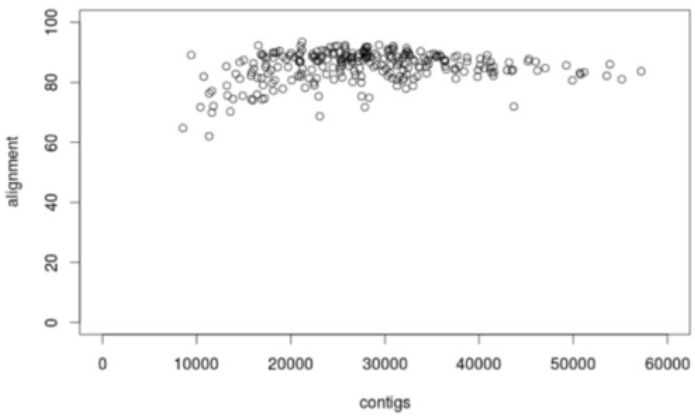
Reads mapped back to transcripts (RMBT)

- align reads against assembly generated transcripts
- compute percentage of reads mapped

Realignment metrics

Factors affecting realignment rate:

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

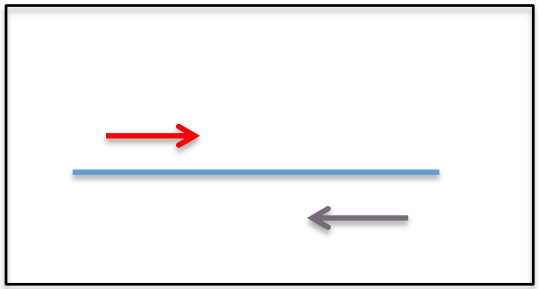


A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

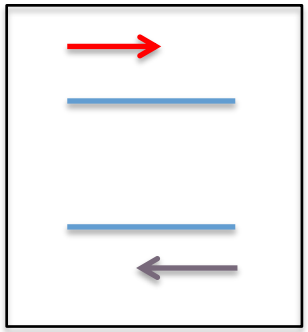
Given read pair: 

Possible mapping contexts in the Trinity assembly are reported:

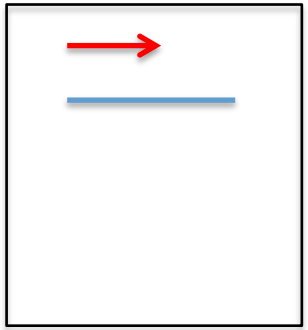
Proper pairs



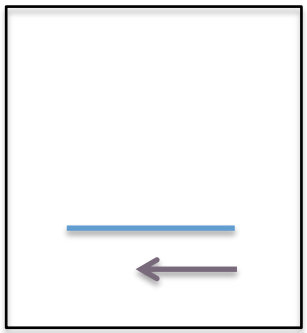
Improper pairs



Left only



Right only



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)

Assembly evaluation : read remapping

Alignment methods : bowtie -RSEM

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

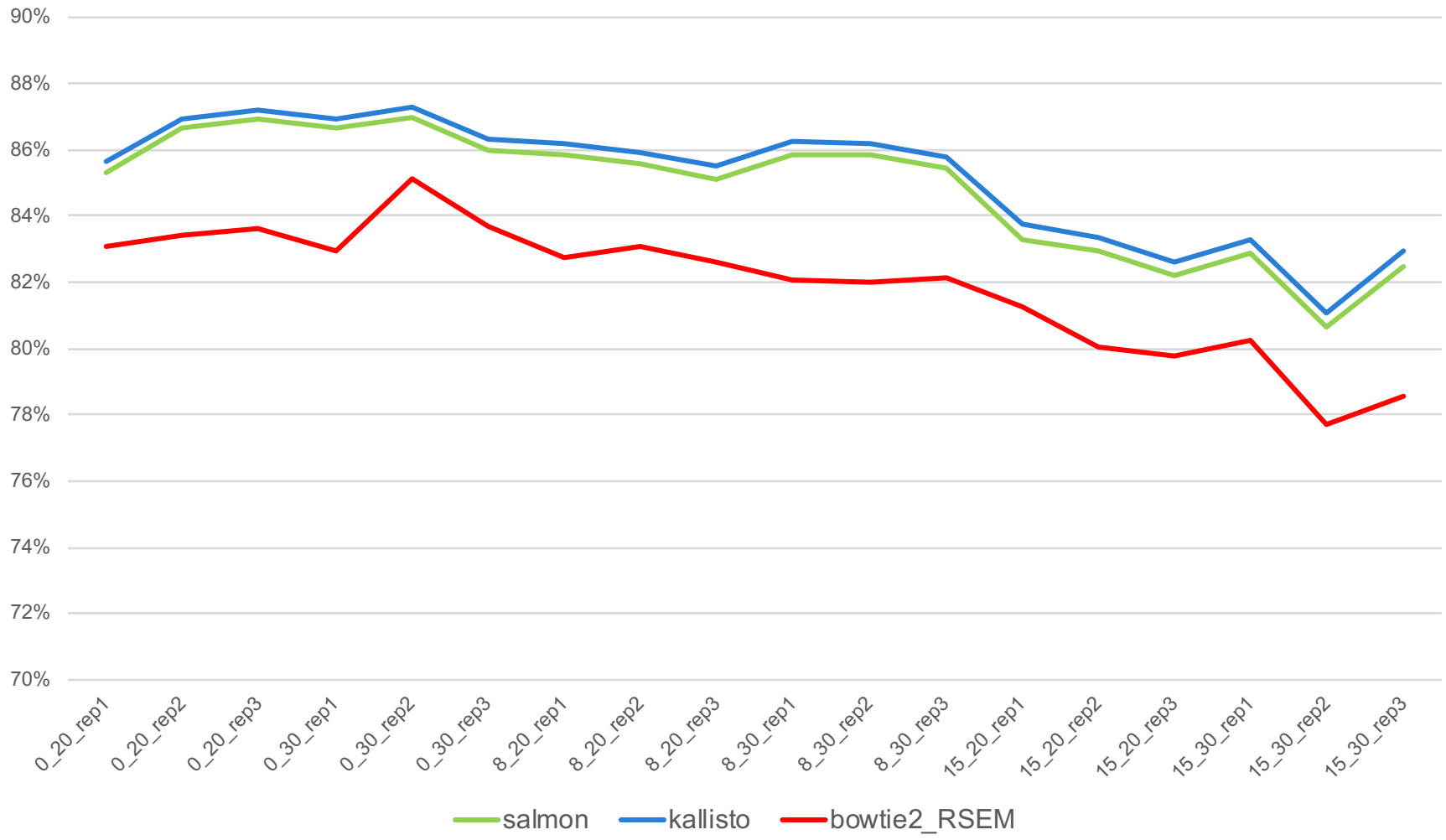
Pseudo-Alignment methods : kallisto

```
$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
```

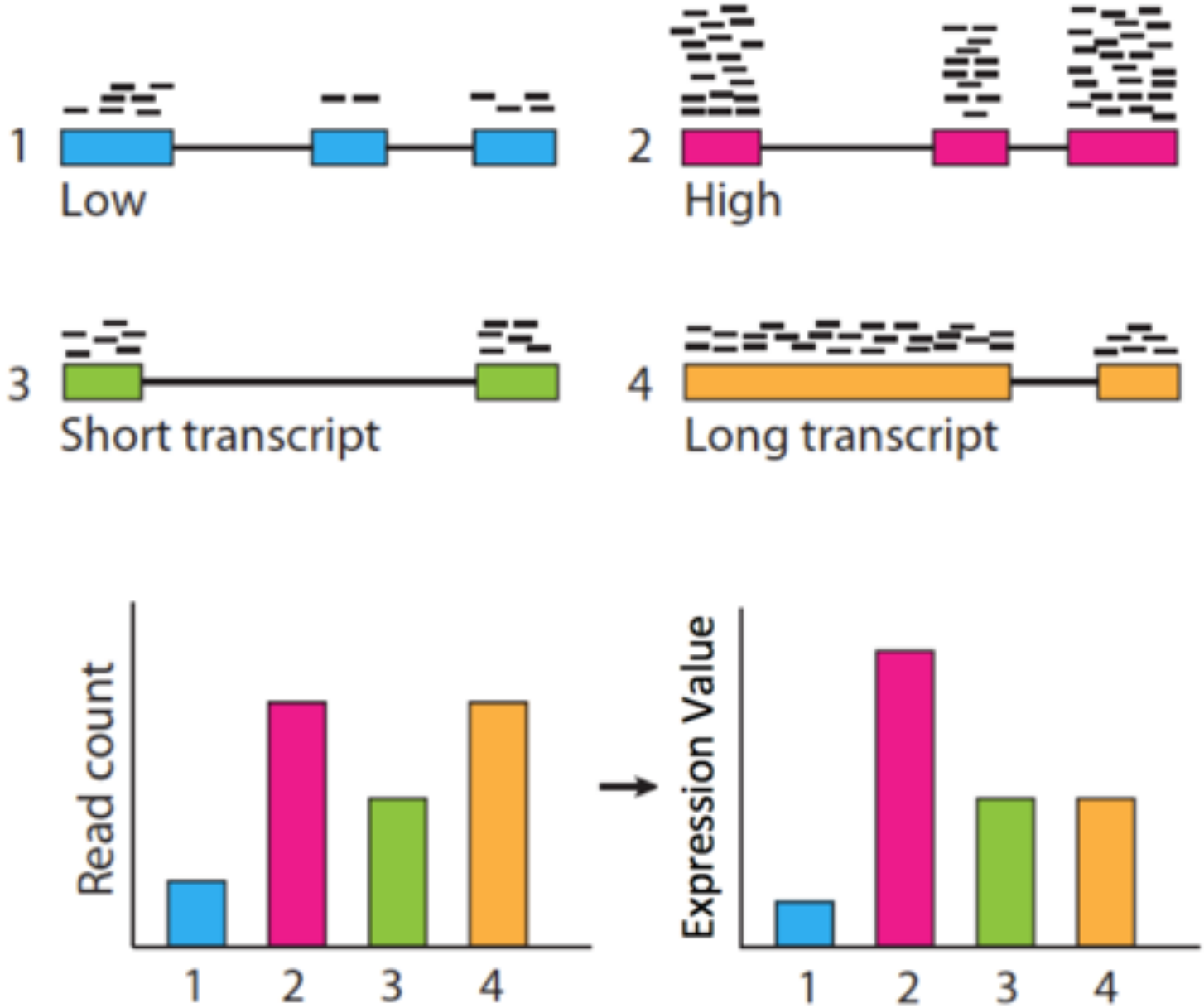
Two matrices,

- 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.

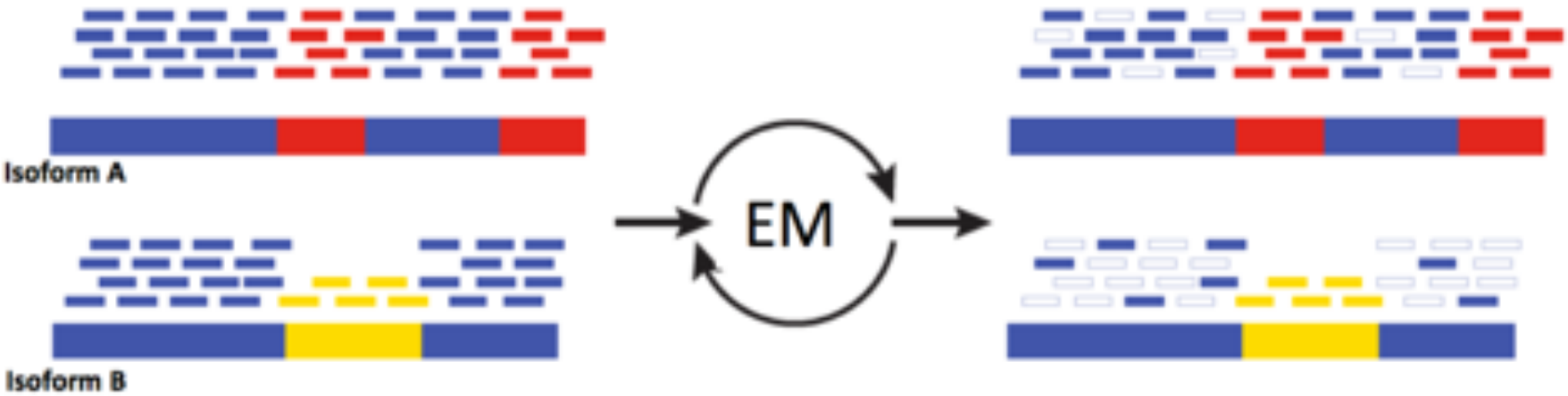
A scaling normalization method for differential expression analysis of RNA-Seq data, Robinson and Oshlack, Genome Biology 2010.

Calulating Expression of genes and transcripts



Calulating Expression of genes and transcripts

Multiply-mapped Reads Confound Abundance Estimation : RSEM Count



Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

ML abundance estimates using the Expectation-Maximization (EM) algorithm to find the most likely assignment of reads to transcripts

RSEM.isoforms.results

Transcript name

Gene name

Real length

Length minus fragment size

Comptage attendu (*)

Transcripts Per Million

Fragments Per Kilobase of transcript per Million mapped reads

isoform percentage

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
comp100000_c0_seq1	comp100000_c0	340	239	13.09	2.79	3.29	100
comp10000_c0_seq1	comp10000_c0	353	252	43.44	8.84	10.43	100
comp10001_c0_seq1	comp10001_c0	569	468	48.01	5.61	6.62	100
comp10002_c0_seq1	comp10002_c0	1563	1462	197.27	7.78	9.19	93.26
comp10002_c0_seq2	comp10002_c0	1563	1462	0	0	0	0
comp10002_c0_seq3	comp10002_c0	1087	986	9.73	0.56	0.66	6.74
comp10002_c0_seq4	comp10002_c0	1087	986	0	0	0	0
comp10004_c0_seq1	comp10004_c0	661	560	105.99	10.48	12.37	100
comp100058_c0_seq1	comp100058_c0	879	778	45	3.26	3.85	100
comp10005_c0_seq1	comp10005_c0	274	173	28	7.82	9.23	100
comp10006_c0_seq1	comp10006_c0	309	208	42	10.07	11.88	100
comp10007_c0_seq1	comp10007_c0	477	376	66	9.42	11.11	100
comp100094_c0_seq1	comp100094_c0	279	178	14	3.82	4.51	100
comp10009_c0_seq1	comp10009_c0	256	155	13.77	4.2	4.96	100
comp1000_c0_seq1	comp1000_c0	292	191	20	5.15	6.08	100

(*)Because 1) each read aligning to this transcript has a probability of being generated from background noise; 2) RSEM may filter some alignable low quality reads, the sum of expected counts for all transcript are generally less than the total number of reads aligned.

Normalized Expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.

- Reported as: Number of RNA-Seq **F**ragments
  Per **K**ilobase of transcript
  per total **M**illion fragments mapped

FPKM

RPKM (reads per kb per M) used with Single-end RNA-Seq reads
FPKM used with Paired-end RNA-Seq reads.

RPKM vs TPM

RPKM

Running SUM of reads in a sample

↓

Total Reads (sample) / 1 million
= Scaling Factor (SF)

↓

Normalize for read depth

Each Gene Reads / SF = RPM

↓

Normalize for gene length

RPM / Gene Length (kb) = RPKM

www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Transcripts per Million (TPM)

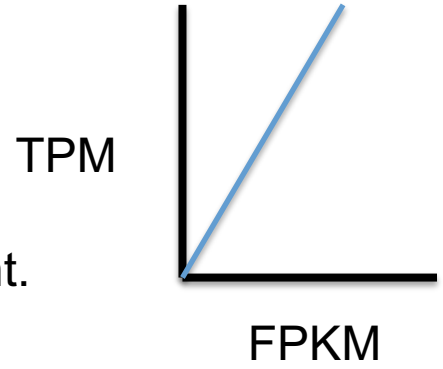
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

Preferred metric for measuring expression

- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.



RPKM vs TPM

RPKM

Running SUM of reads in a sample

↓

Total Reads (sample) / 1 million
= Scaling Factor (SF)

↓

Normalize for read depth

Each Gene Reads / SF = RPM

↓

Normalize for gene length

RPM / Gene Length (kb) = RPKM

TPM

Scaled by gene length

Each Gene Reads / Gene Length (kb)
= RPK

↓

Running SUM of RPK (sample)

↓

RPK / 1 million = RPK-SF

↓

Normalize for sequencing depth

Each Gene RPK / RPK-SF = TPM

www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

RPKM vs TPM

Gene	Gene length (KB)	Rep 1 counts	Rep 2 counts	Rep 3 counts
A	2	10	12	30
B	4	20	25	60
C	1	5	8	15
D	10	0	0	1

RPKM

Gene	Gene length (KB)	Rep 1 RRPM	Rep 2 RRPM	Rep 3 RRPM
A	2	1.43	1.33	1.42
B	4	1.43	1.39	1.42
C	1	1.43	1.78	1.42
D	10	0.00	0.00	0.01
	SUM of RPKM	**4.29**	**4.50**	**4.25**

TPM

Gene	Gene length (KB)	Rep 1 TPM	Rep 2 TPM	Rep 3 TPM
A	2	3.33	2.96	3.33
B	4	3.33	3.09	3.33
C	1	3.33	3.95	3.33
D	10	0.00	0.00	0.02
	SUM of TPM	**10.00**	**10.00**	**10.00**

www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

RSEM.isoforms.results and RSEM.genes.results

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
c128_g0_i1	c128_g0	209	1.73	0.00	0.00	0.00	0.00
c13_g0_i1	c13_g0	235	7.16	1.00	12561.51	5282.75	100.00
c22_g0_i1	c22_g0	215	2.62	0.00	0.00	0.00	0.00
c28_g0_i1	c28_g0	329	54.60	4.00	6591.85	2772.21	100.00
c33_g0_i1	c33_g0	307	40.30	3.00	6697.56	2816.66	100.00
c35_g0_i1	c35_g0	219	3.33	0.00	0.00	0.00	0.00
c35_g1_i1	c35_g1	204	1.19	1.00	75295.99	31665.75	100.00
c39_g0_i1	c39_g0	348	68.20	1.00	1319.32	554.84	100.00
c39_g0_i2	c39_g0	255	13.97	0.00	0.00	0.00	0.00
c41_g0_i1	c41_g0	592	295.77	12.00	3650.37	1535.16	100.00
c44_g0_i1	c44_g0	361	78.10	1.00	1151.96	484.46	100.00
c44_g1_i1	c44_g1	280	25.22	1.00	3568.05	1500.54	100.00

Transcripts

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
c128_g0	c128_g0_i1	0.00	0.00	0.00	0.00	0.00
c13_g0	c13_g0_i1	235.00	7.16	1.00	12561.51	5282.75
c22_g0	c22_g0_i1	0.00	0.00	0.00	0.00	0.00
c28_g0	c28_g0_i1	329.00	54.60	4.00	6591.85	2772.21
c33_g0	c33_g0_i1	307.00	40.30	3.00	6697.56	2816.66
c35_g0	c35_g0_i1	0.00	0.00	0.00	0.00	0.00
c35_g1	c35_g1_i1	204.00	1.19	1.00	75295.99	31665.75
c39_g0	c39_g0_i1,c39_g0_i2	348.00	68.20	1.00	1319.32	554.84
c41_g0	c41_g0_i1	592.00	295.77	12.00	3650.37	1535.16
c44_g0	c44_g0_i1	361.00	78.10	1.00	1151.96	484.46
c44_g1	c44_g1_i1	280.00	25.22	1.00	3568.05	1500.54

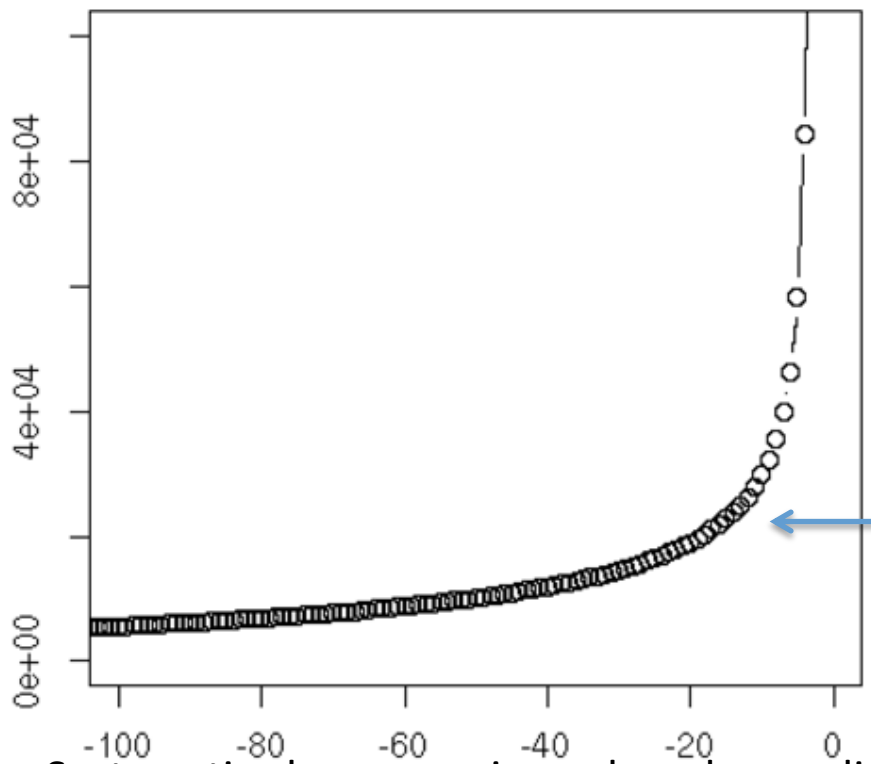
Genes

Alternative to N50 ?

Often, most assembled transcripts are *very* lowly expressed

(How many 'transcripts & genes' are there really?)

Cumulative # of Transcripts



1.4 million Trinity transcript contigs

N50 ~ 500 bases

20k transcripts

Sort contigs by expression value, descendingly.
-1 * minimum TPM

Expression

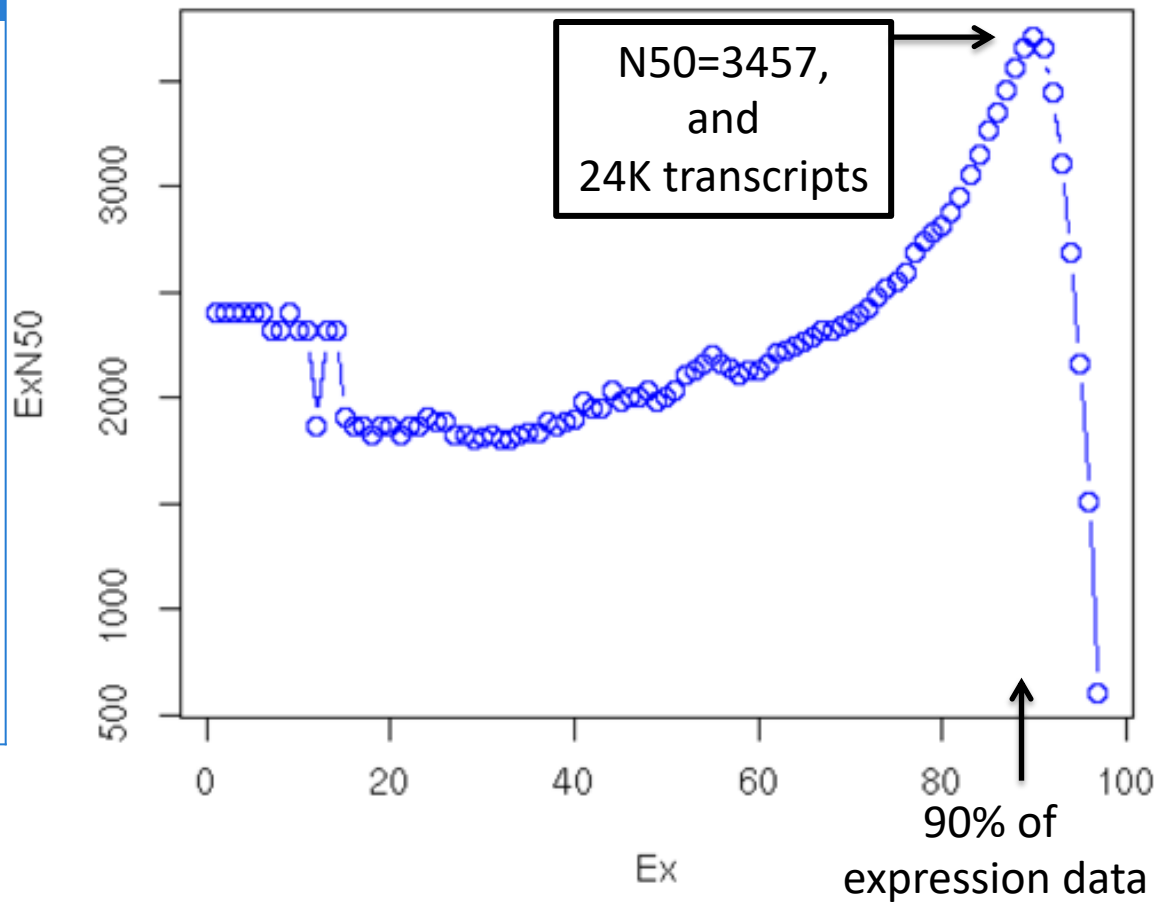
* Salamander transcriptome

Alternative to N50 : ExN50 – E90N50

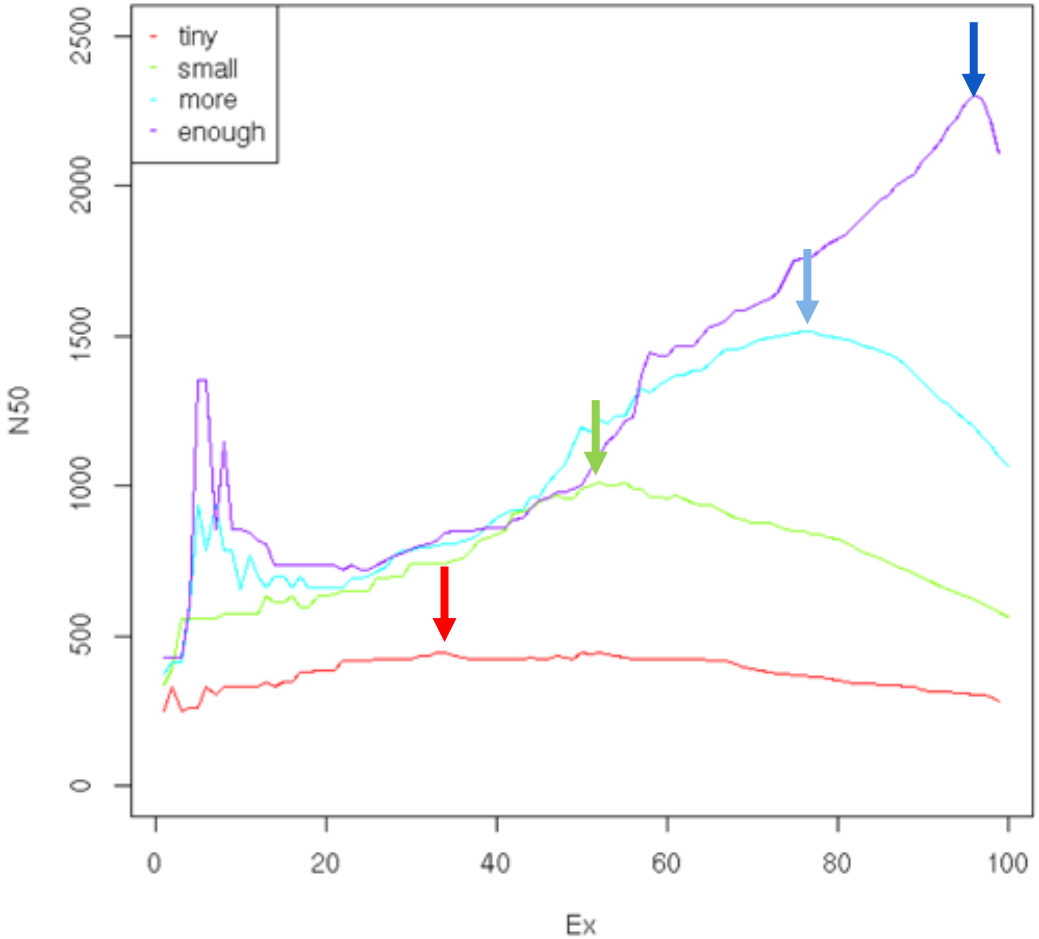
Compute N50 Based on the Top-most Highly Expressed Transcripts (ExN50)

- 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



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

A Trinity alternative

BlastX of Trinity.fasta against uniprot
Script Trinity : *analyze_blastPlus_topHit_coverage.pl*

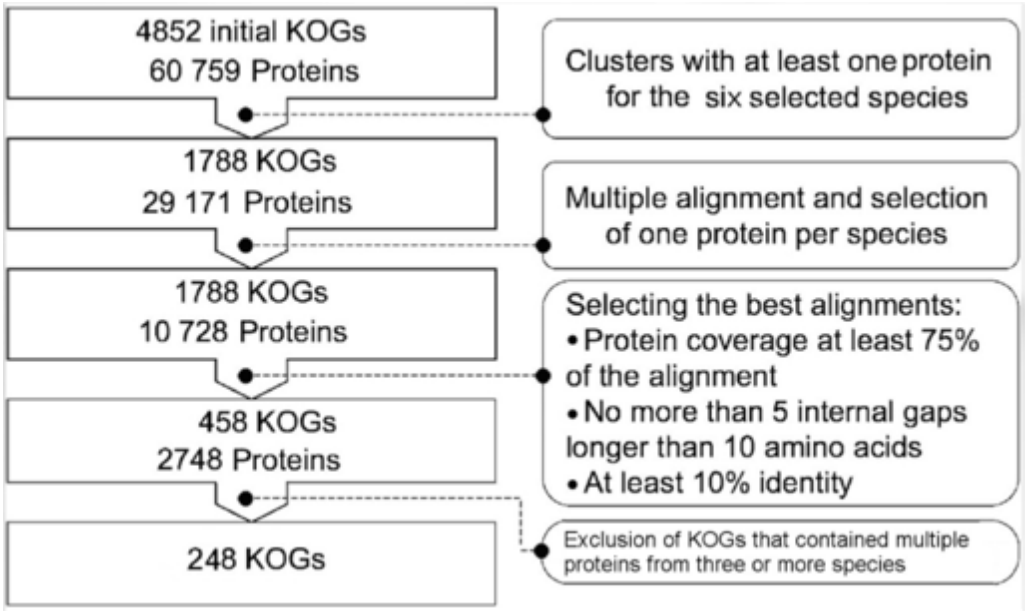
hit_pct_cov_bin	count_in_bin	>bin_below
100	3242	3242
90	268	3510
80	186	3696
70	202	3898
60	216	4114
50	204	4318
40	164	4482
30	135	4617
20	76	4693
10	0	4693
0	0	4693

- There are 268 proteins that each match a Trinity transcript by >80% and ⇐ 90% of their protein lengths.
- There are 3510 proteins that are represented by nearly full-length transcripts, having >80% alignment coverage.
- There are 3242 proteins that are covered by more than 90% of their protein lengths.

CEGMA analysis

Core **E**ukaryotic **G**enes **M**apping **A**pproach : http://www.iplantcollaborative.org

Mapping a set of conserved protein families that occur in a wide range of eukaryotes onto assembly to assess completeness .
A set of eukaryotic core proteins (KOG = euKaryotic Orthologous Groups) from 6 species: H. sapiens, D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, S.pombe



First set of 458 core genes

First set of 248 core genes with less paralogs

Genis Parra, Keith Bradnam and Ian Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. » . Bioinformatics, 23: 1061-1067 (2007)
Genis Parra, Keith Bradnam, Zemin Ning, Thomas Keane, and Ian Korf. Assessing the gene space in draft genomes » . Nucleic Acids Research, 37(1): 298-297 (2009)

CEGMA analysis : Output example (output.completeness_report)

- Complete (70% of the protein length)
- Partial (not matching “complete” criteria but exceed a pre-computed alignment score)

```
#      Statistics of the completeness of the genome based on 248 CEGs      #

              #Prots  %Completeness  -  #Total  Average  %Ortho

  Complete      245       98.79      -   593     2.42     64.90

   Group 1       66      100.00      -   146     2.21     60.61
   Group 2       56      100.00      -   129     2.30     60.71
   Group 3       58       95.08      -   140     2.41     67.24
   Group 4       65      100.00      -   178     2.74     70.77

   Partial      245       98.79      -   631     2.58     67.76

   Group 1       66      100.00      -   152     2.30     62.12
   Group 2       56      100.00      -   142     2.54     64.29
   Group 3       58       95.08      -   148     2.55     68.97
   Group 4       65      100.00      -   189     2.91     75.38

#    These results are based on the set of genes selected by Genis Parra   #

#    Key:                                                                  #
#    Prots = number of 248 ultra-conserved CEGs present in genome          #
#    %Completeness = percentage of 248 ultra-conserved CEGs present        #
#    Total = total number of CEGs present including putative orthologs     #
#    Average = average number of orthologs per CEG                         #
#    %Ortho = percentage of detected CEGS that have more than 1 ortholog   #
```

BUSCO analysis

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

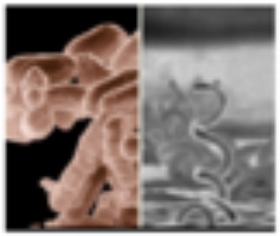
HMM:s for 248 core eukaryotic genes aligned to your assembly to assess completeness of gene space

"complete": 70% aligned
"partial": 30% aligned

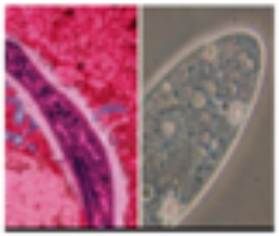
BUSCO(http://busco.ezlab.org/)

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

BUSCO analysis

Bacteria

- bacteria
- proteobacteria
- rhizobiales
- betaproteobacteria
- gammaproteobacteria
- enterobacteriales
- deltaepsilonsub
- actinobacteria
- cyanobacteria
- firmicutes
- clostridia
- lactobacillales
- bacillales
- bacteroidetes
- spirochaetes
- tenericutes

Eukaryota

- eukaryota **(303)**
- fungi **(290)**
- microsporidia
- dikarya
- ascomycota
- pezizomycotina
- eurotiomycetes
- sordariomyceta
- saccharomyceta **(1759)**
- saccharomycetales
- basidiomycota
- metazoa
- nematoda
- arthropoda
- insecta
- endopterygota
- hymenoptera
- diptera
- vertebrata
- actinopterygii
- tetrapoda
- aves
- mammalia
- euarchontoglires
- laurasiatheria
- embryophyta
- protists_ensembl
- alveolata_stramenophil es_ensembl

Practice

3

Aller sur la practice 3 Assessing transcriptome assembly quality *du github*

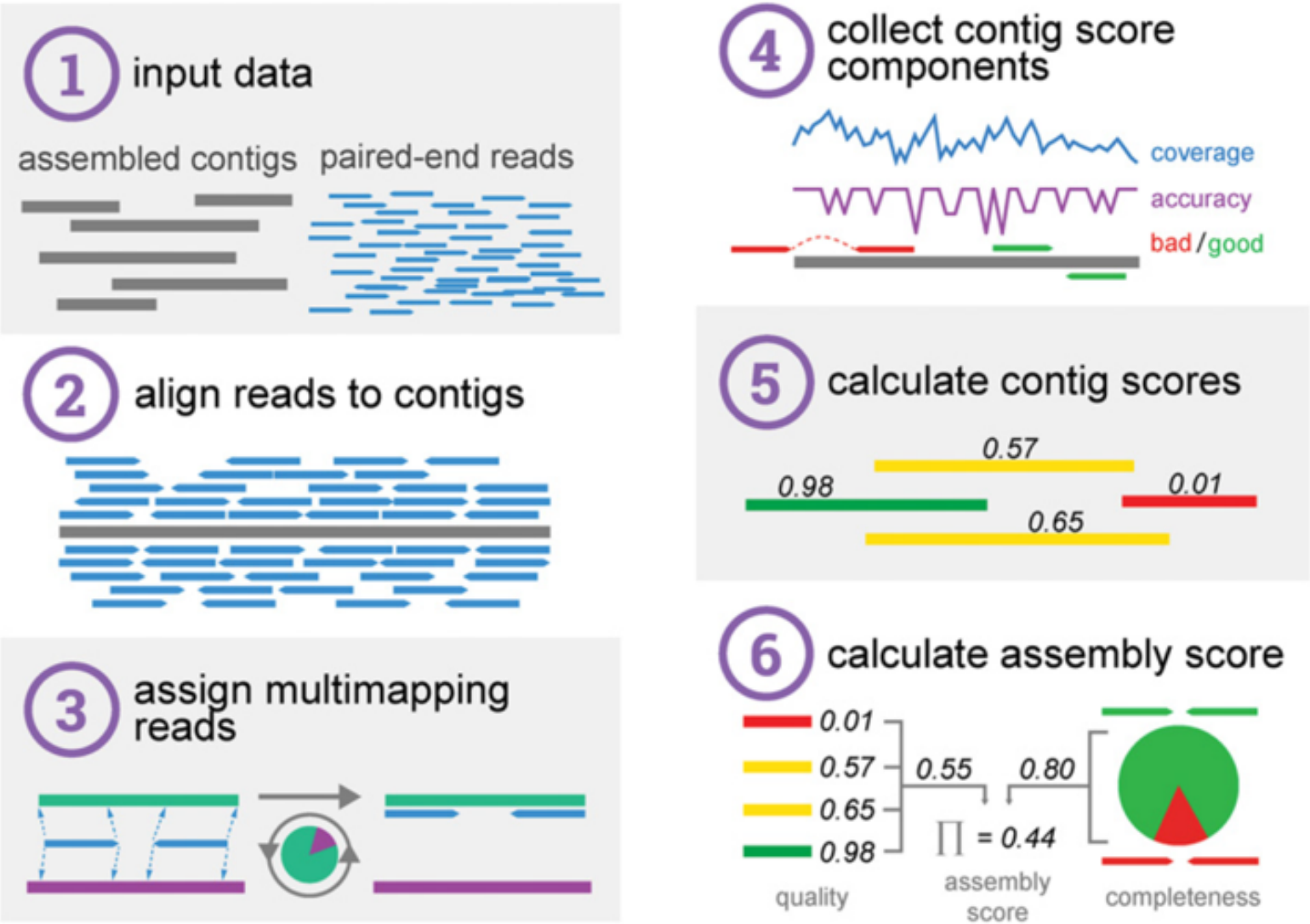
3.2 Analysis of remapping results

3.3 Quantifying completness using BUSCO

3.4 BLASTX comparison to known protein sequences database

Station Biologique Roscoff

TransRate



Smith-Unna et al. Genome Research, 2016

Tools to evaluate transcriptomes

Detonate: Li, B et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology 2014, 15:553

A methodology and corresponding software package for evaluating de novo transcriptome assemblies, which can compute both reference-free and reference-based measures. DETONATE consists of two component packages, RSEM-EVAL and REF-EVAL

	CLC	SOAP de novo trans	Trinity
Score	-13777089814	-13270583330	-10037861970
BIC_penalty	-941678.17	-2443248.59	-2106368.55
Prior_score_on_contig_lengths	-746170.82	-926991.89	-7415766.35
Prior_score_on_contig_sequences	-126215414.1	-201779663.6	-408041405.4
Data_likelihood_in_log_space_without_correction	-13649697269	-13066158028	-9627819309
Correction_term	-510717.95	-724602.05	-7520878.54
Number_of_contigs	98684	256044	220740
Expected_number_of_aligned_reads_given_the_data	121502964.5	127676508.9	157057277.9
Number_of_contigs_smaller_than_expected_read/fragment_length	0	147623	0
Number_of_contigs_with_no_read_aligned_to	74	530	31212
Maximum_data_likelihood_in_log_space	-13644505579	-12932075677	-9620152715
Number_of_alignable_reads	122079646	129886064	157696259
Number_of_alignments_in_total	123076291	179395943	448982192
Transcript_length_distribution_related_factors	-479292.41	-506592.48	-881127.96

RNA Quast

Publications

Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A**.** rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*, 2016

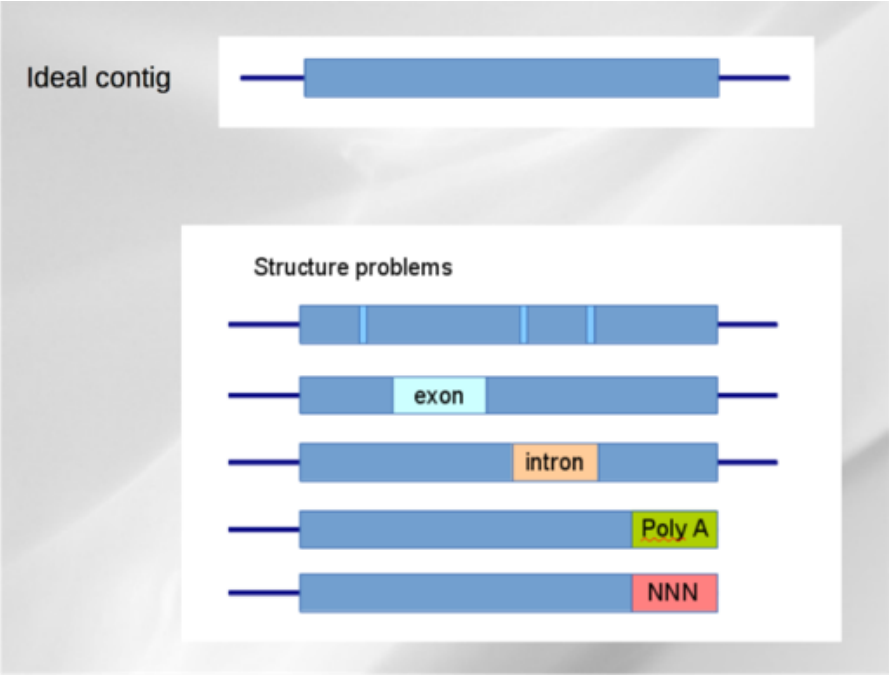
tblastn, HMMER and transeq.
GeneMarkS-T
STAR aligner (or alternatively TopHat)
BUSCO v1.1b1

Transcriptome assembly

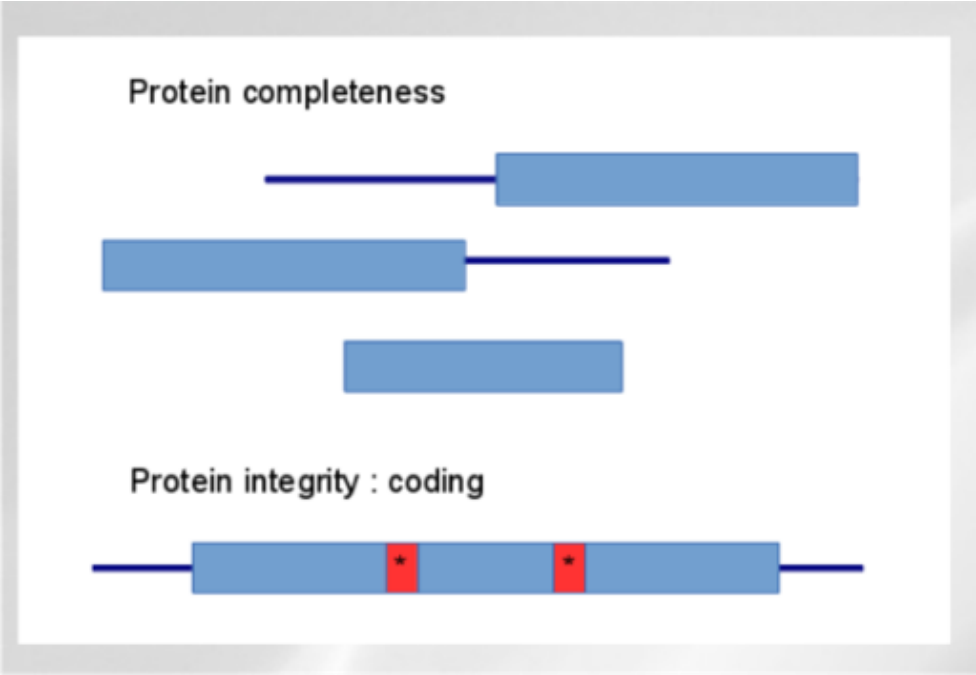
CLEANING THE ASSEMBLY

Cleaning the assembly

Transcripts



Proteins



- cleaning polyA tails, terminal N blocks, low complexity areas
- insertion/deletion correction using the alignment
- cis or trans-chimera detection
- low fold coverage filtering (graph data)
- low expression filtering
- possible filtering of contigs which do not have a long enough ORF (phylogenomy)

Transcriptome cleaning

- Remove remaining polyA tails
- Remove blocks of Ns located at the extremities
- Remove low complexity areas

Seqclean: a script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences.

- Finding frame-shifts :
- Insertion/deletion correction

- Going back to alignment reads vs transcripts to find INDEL
- Using a proteic reference to find frame-shifts

- Detect splice form

- Going back to alignment reads vs transcripts to find splice
- Isoforms alignments + reads
- Alignment against « close » reference genome

Transcriptome cleaning : Chimera

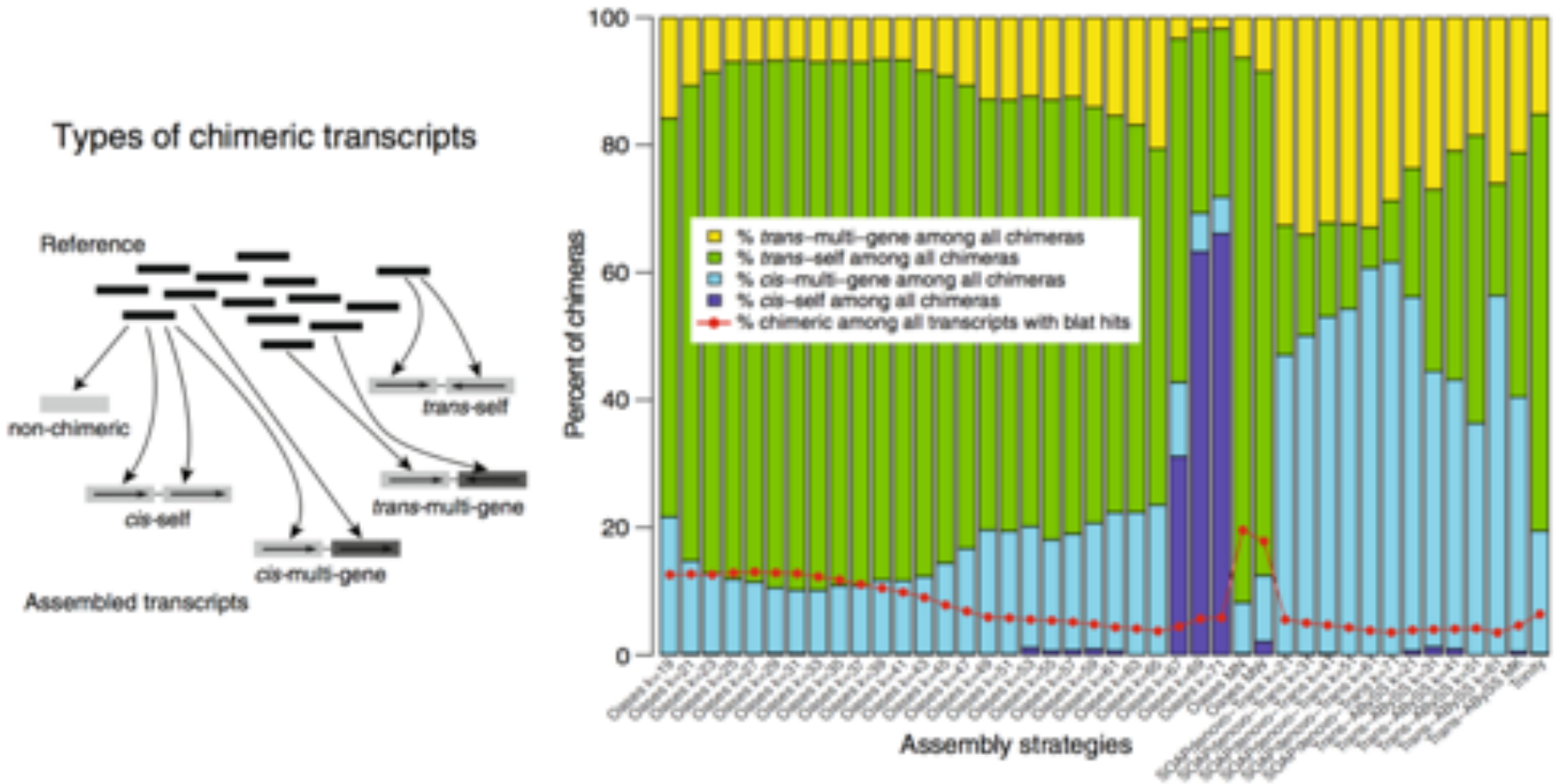


Figure 1 Chimera compositions among assembled transcripts before post-processing. Oases MN: Oases-M merging single *k*-mer assemblies of 21, 31, 41, 51 and 61; MW: Oases-M merging single *k*-mer assemblies of 19–71, with increment of 2; Trans-ABySS MK: Trans-ABySS merging single *k*-mer assemblies of 21, 31, 41, 51 and 61.

Majority of trans-self chimeras for small-middle k-mers
Majority of cis-self chimeras for large k-mers and oases merge
Chimeras increase with merging and small kmer

Without reference, cannot tackle multi-gene chimeras
Blast against itself EBARD de novo

Yang, Y., & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics, 14(1), 328.

ChimPipe

Cancer Gene Profiling pp 239-253 | Cite as

Transcriptome Sequencing for the Detection of Chimeric Transcripts

Authors | Authors and affiliations

Hsueh-Ting Chu

SCIENTIFIC REPORTS

Article | Open Access | Published: 10 February 2016

Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data

Shailesh Kumar, Angie Duy Vo, Fujun Qin & Hui Li

Scientific Reports **6**, Article number: 21597 (2016) | Download Citation

BMC Genomics

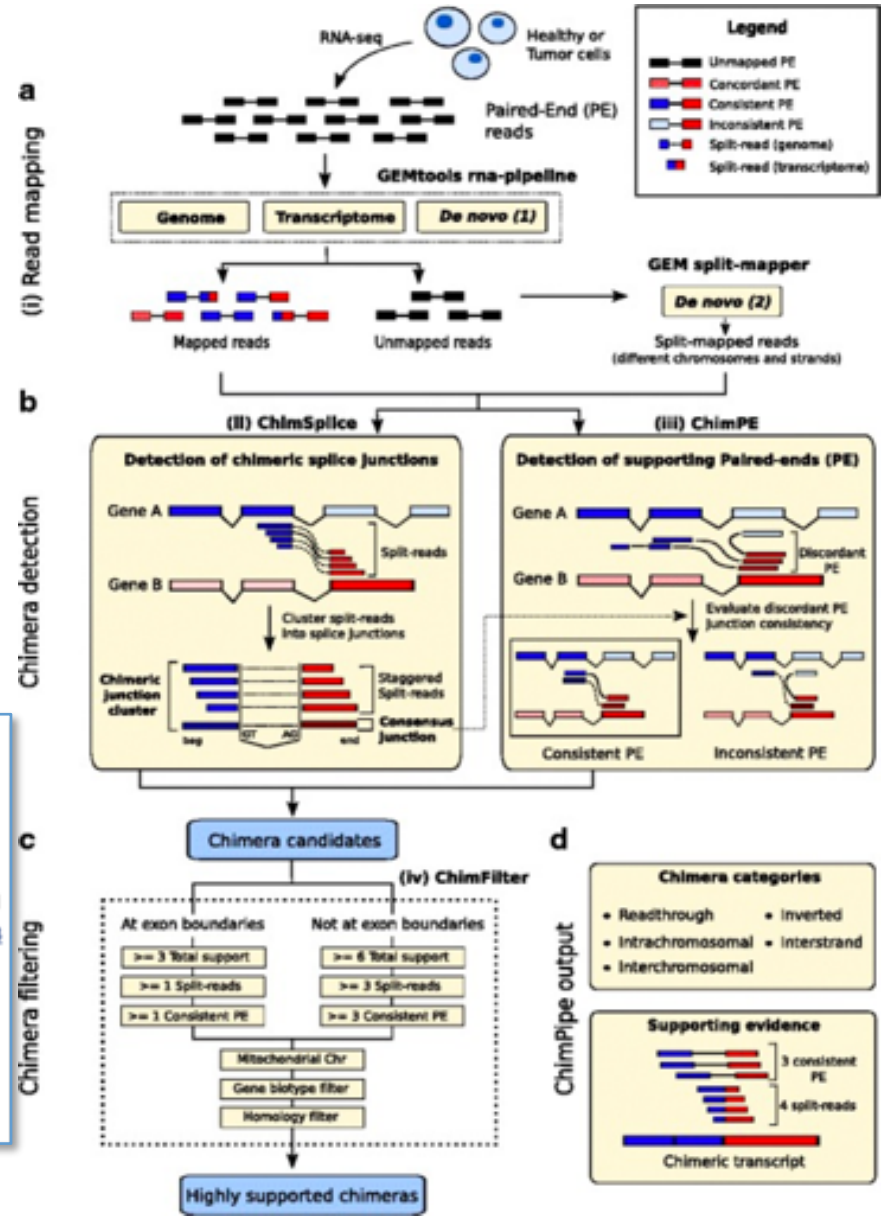
BMC

BMC Genomics. 2017; 18: 7.
Published online 2017 Jan 3. doi: 10.1186/s12864-016-3404-9

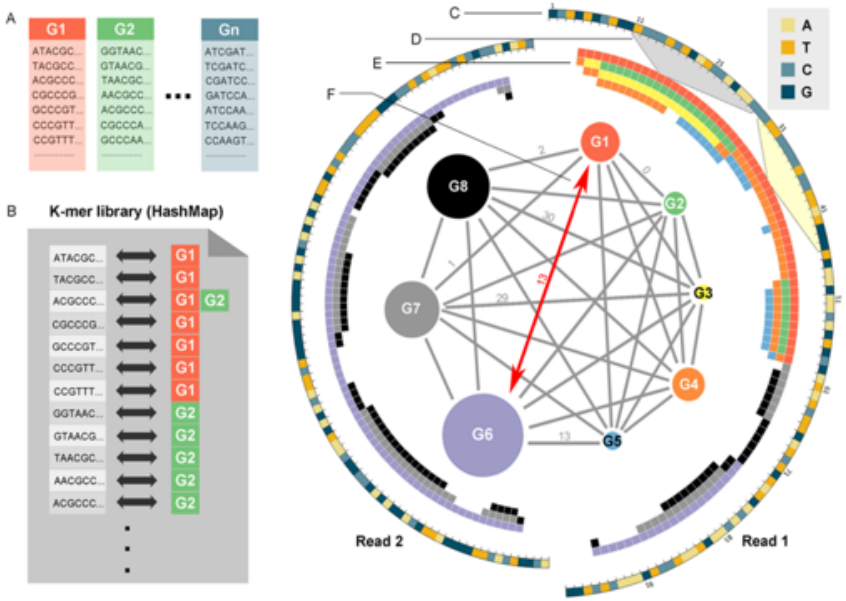
PMCID: PMC5209911
PMID: 28049418

ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data

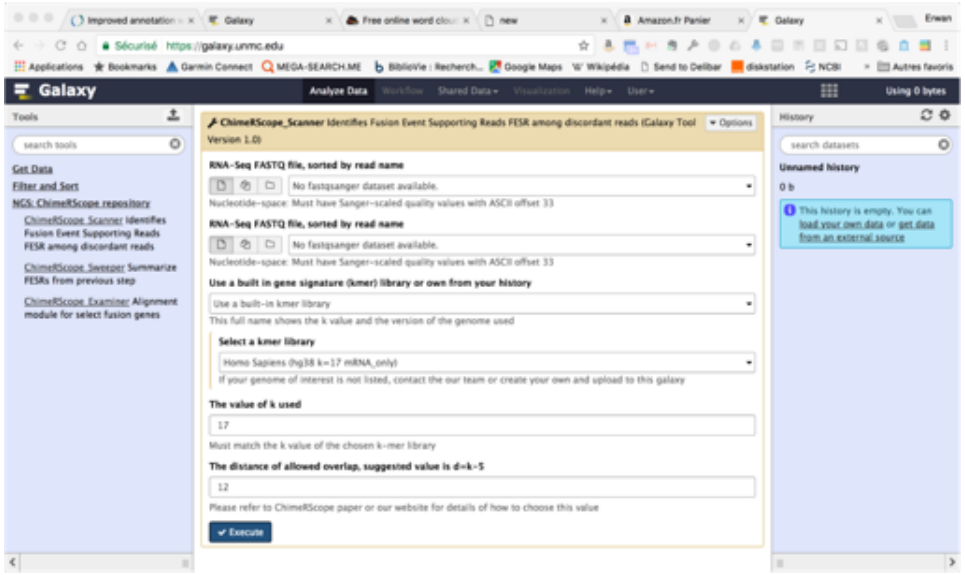
Bernardo Rodríguez-Martín,[1,2,3] Emilio Palumbo,[1,2] Santiago Marco-Sola,[4] Thasso Griebel,[4] Paolo Ribeca,[4,5] Graciela Alonso,[6] Alberto Rastrojo,[6] Begoña Aguado,[6] Roderic Guigó,[1,2,7] and Sarah Djebali[1,2,8]



ChimeRScope



A novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data



https://galaxy.unmc.edu/

Li Y, Heavican TB, Vellichirammal NN, Iqbal J, Guda C. (2017) **ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data.** *Nucleic Acids Res*.

Transcriptome redundancy

Trinity is often criticized for his verbosity

- *Lots* of transcripts is the rule rather than the exception.
- Most of the transcripts are very lowly expressed.
- The **deeper you sequence** and **the more complex your genome**, the **larger the number of lowly expressed transcripts** you will be able to assemble.

- *Trinity transcripts are not scaffolded across sequencing gaps : **smaller transcript fragments may lack enough properly-paired read** support to show up as expressed, **but are still otherwise supported by the read data**.*

- Biological relevance of the lowly expressed transcripts could be questionable - some are bound to be very relevant.

Transcriptome cleaning : Redondancy

- Consider results at genes level

- Filtering base upon expression and % isoforms

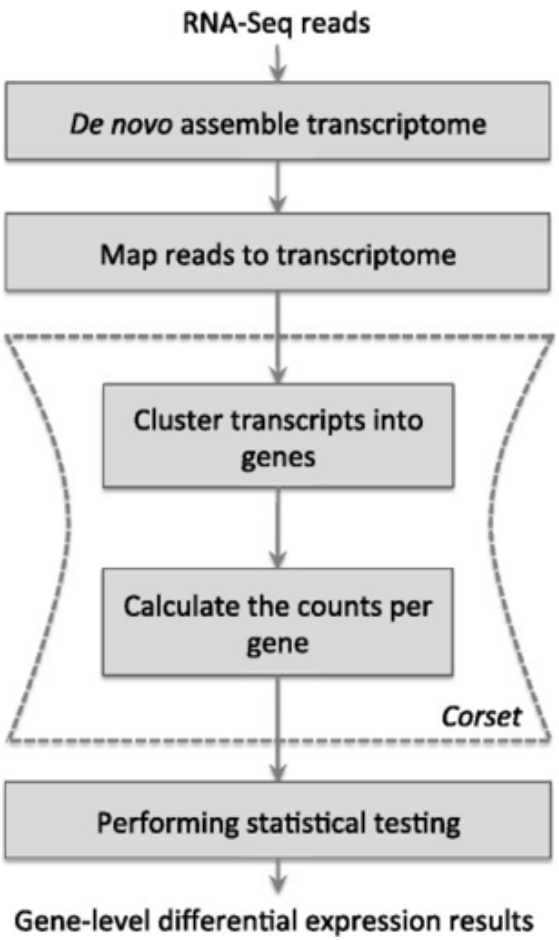
*«- **retaining only those that represent at least 1% of the per-component** (IsoPct) expression level. : filter artifacts and lowly expressed transcripts*

*- Therefore**, filter cautiously** and we don't recommend discarding such lowly expressed (or seemingly unexpressed) transcripts, but rather putting them aside for further study »*

- CDHIT-EST + TGICL :

cd-hit-est -o cdhit -c 0.98 -i Trinity.fasta -p 1 -d 0 -b 3 -T 10

Transcriptome cleaning : Redondancy



- Corset :

Davidson and Oshlack *Genome Biology* 2014 **15**:410
doi:10.1186/s13059-014-0410-6

- DRAP : **D**e novo **R**NA-seq **A**ssembly **P**ipeline :

Cabau C, et al. PeerJ 5:e2988 (2017). Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies.

- See example :

Context

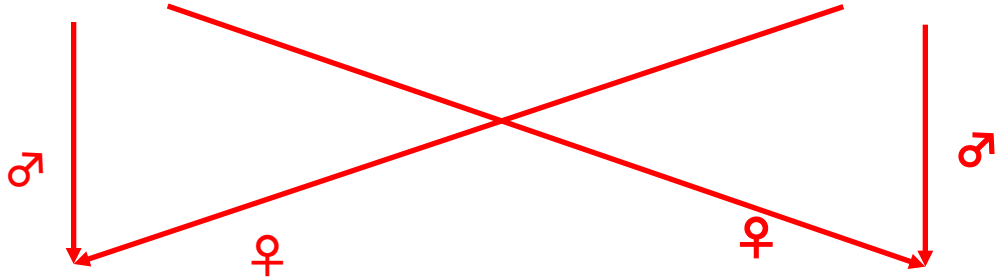
Anas platyrhynchos

Pekin : Canard de Pékin

Cairina moschata

Muscovy : Canard musqué

Genome sequence of the duck (*Anas platyrhynchos*).
An et al. . GigaScience Database. 2014
http://dx.doi.org/10.5524/101001



Hinny overfeed :

Production of foie gras -

TG secretion, peripheral fattening +++

Mulard overfeed:

Production of foie gras +++

TG secretion, peripheral fattening +

"Foie gras" production
Mulard > Hinny
Muscovy > Pekin

2 very close species and 2 sort of mating species : but only one describe genome

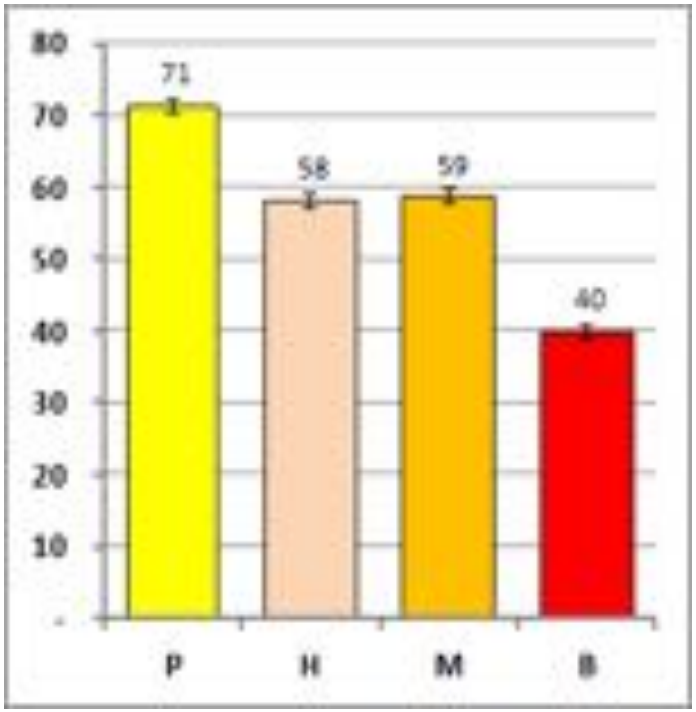
Objectives

Compare gene expressions in duck livers

- Of these four genotypes,
- Fed *ad libitum* or force-fed

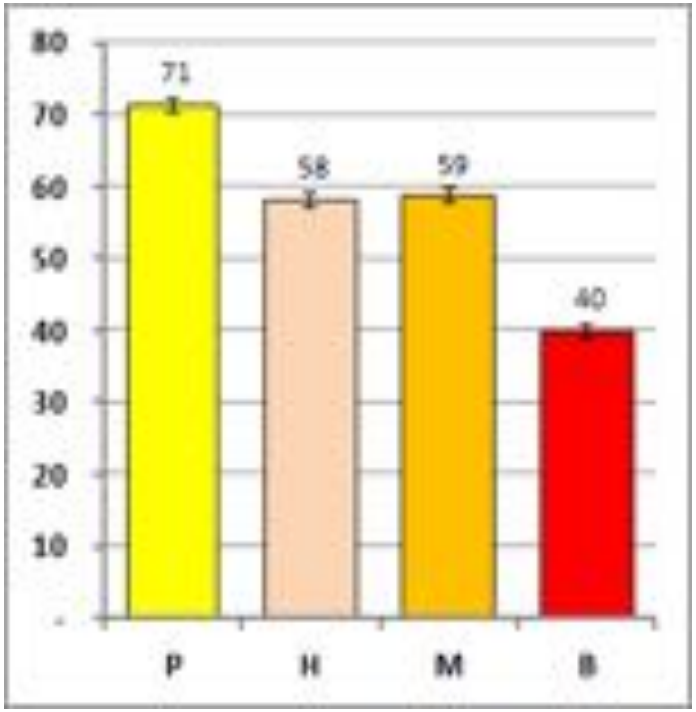
In order to understand the phenotypic differences

A first analyse was perfom using a reference approach
Lot of reads excluded from the initial analysis



% remapping on ref. genome

Objectives



% remapping on ref. genome

Compare gene expressions in duck livers

- Of these four genotypes,
- Fed *ad libitum* or force-fed

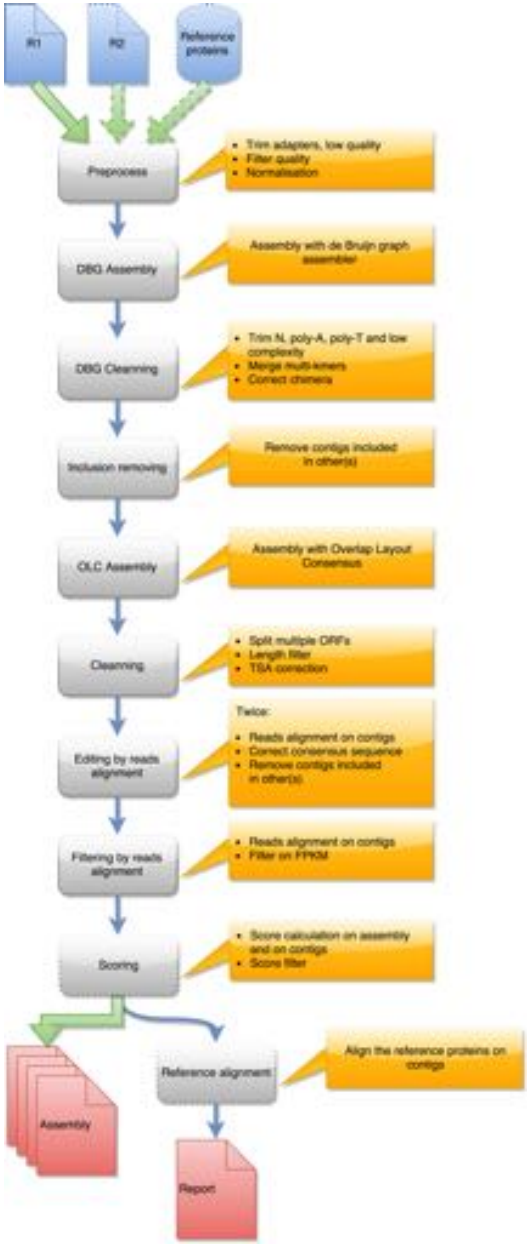
In order to understand the phenotypic differences

A first analyse was perfom using a reference approach
Lot of reads excluded from the initial analysis

A second analysis performed using a full *de novo* approach.

How to create an hydrid transcriptome from 4 differents genotypes ?

DRAP : **D**e novo **R**NA-Seq **A**ssembly Pipeline



Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. Cabau et al. 2017 DOI - 10.7717/peerj.2988

Step1 in runDRAP workflow.

This workflow is used to produce an assembly from one sample/tissue/development stage. It take as input R1 from single-end sequencing or R1 and R2 from paired-end sequencing and eventually a reference proteins set from closest species with known proteins.

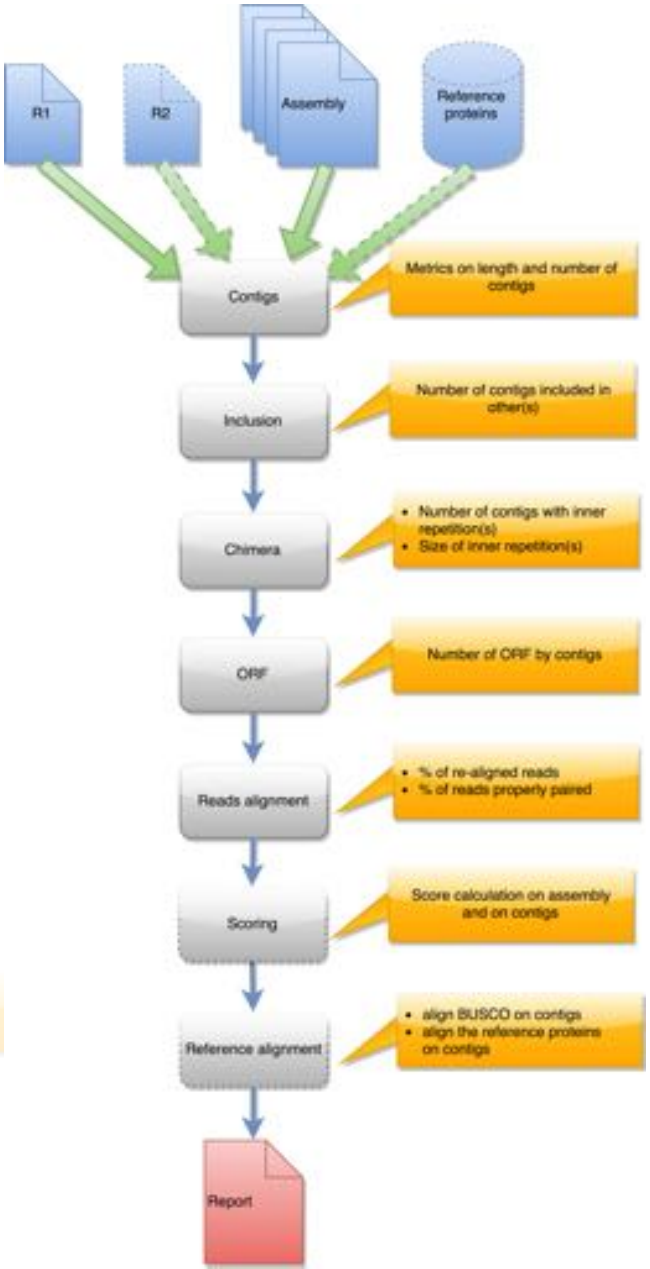
DRAP : **D**e novo **R**NA-Seq **A**ssembly Pipeline

Step 2 in runMeta workflow.

This workflow is used to produce a merged assembly from several samples/tissues/development stage outputted by runDRAP. Inputs are runDRAP output folders and eventually a reference protein set.

DRAP : **D**e novo **R**NA-Seq **A**ssembly Pipeline

Steps 3 in runAssessment workflow.
This workflow is used to evaluate quality for one assembly or for compare several assemblies produced from the same dataset. Inputs are the assembly/ies, R1 and eventually R2, and a reference protein set



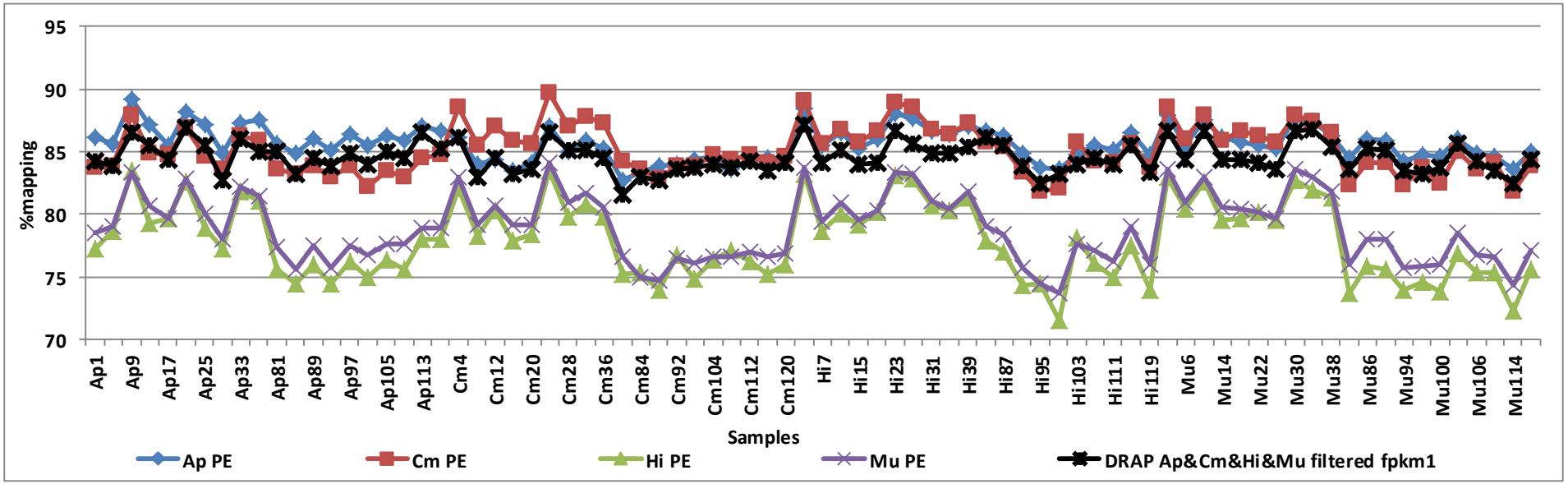
Anas_platyrhynchos. BGI_duck_1.0.cdna.all.fa

Anas_platyrhynchos. cufflink.merge.fasta

DRAP_ApCmHiMu transcripts_fpkm_1.fa

```
Results:
C:81.5%[S:42.9%,D:38.6%],F:11.2%,M:7.3%,n:3
247 Complete BUSCOs (C)
130 Complete and single-copy BUSCOs (S)
117 Complete and duplicated BUSCOs (D)
34 Fragmented BUSCOs (F)
22 Missing BUSCOs (M)
303 Total BUSCO groups searched
```

Completeness: transcriptome de novo is better than reference



Higher remmaping rate on the hybrid *de novo* transcriptome

DEG analysis comparison

Transcriptome de novo	DEG	Pekin	Muscovy	Mule	Hinny	common
edgeR	up-regulated	2281	3450	4907	3901	539
	down-regulated	1468	2717	4013	3795	364
	all	3749	6167	8920	7696	906
Mapping ref genome Ap	DEG	Pekin	Muscovy	Mule	Hinny	common
edgeR	up-regulated	1553	1371	1592	1314	520
	down-regulated	680	773	953	924	235
	all	2233	2144	2545	2238	758

There is a slight increase of DEG in the reference specie (+68%) and especially large increases in the others (+188 %, +250%, +244%).

Mapping agaisnt genome is quite relevant in homologue to identify the DEG, but definitly not heterologous species