# Transcriptome Assembly Quality Evaluation

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

## A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

CSH Cold Spring Harbor Laboratory

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

New Results

## Establishing evidenced-based best practice for the de novo assembly and evaluation of transcriptomes from non-model organisms

Matthew D MacManes

**doi:** https://doi.org/10.1101/035642

# Transcriptome Assembly Quality Evaluation

- How many genes/transcripts/fragments do I have in my assembly?
  - *Trinity scripts or TransRate*
- How many full-length transcripts did I assemble?
  - *BLAST or DIAMOND*
- How well does my assembly represent the sequenced reads?
  - *Bowtie2 & Trinity scipts or TransRate*
- How complete is my transcriptome? i.e. how many of the highly conserved “benchmark” genes does it contain?
  - *BUSCO*

# Assembly QC

- How many transcripts do I have?

  $TRINITY_HOME/util/TrinityStats.pl  <assembly.fa>

```
################################
Counts of transcripts, etc.
################################
Total trinity 'genes':      333939
Total trinity transcripts: 480312
Percent GC: 48.76
```

**TransRate gives more information**

```
#################################
Stats based on ALL transcript contigs:
#################################

        Contig N10: 8379
        Contig N20: 6325
        Contig N30: 4969
        Contig N40: 3944
        Contig N50: 3062

        Median contig length: 439
        Average contig: 1213.23
        Total assembled bases: 582728498
```

# Assembly QC

- How well does my assembly represent the sequencing reads I put in?

```
$ bowtie2-build assembly.fa assembly.fa
$ bowtie2 -p 10 -q -x assembly.fa -1 left.fq -2 right.fq
        2>&1 1> /dev/null | tee align_stats.txt
```

**Or *TransRate***

```
374663449 reads; of these:
  374663449 (100.00%) were paired; of these:
    87397904 (23.33%) aligned concordantly 0 times
    71727817 (19.14%) aligned concordantly exactly 1 time
    215537728 (57.53%) aligned concordantly >1 times
    ----
    87397904 pairs aligned concordantly 0 times; of these:
      7264984 (8.31%) aligned discordantly 1 time
    ----
    80132920 pairs aligned 0 times concordantly or discordantly; of these:
      160265840 mates make up the pairs; of these:
        48961820 (30.55%) aligned 0 times
        23974234 (14.96%) aligned exactly 1 time
        87329786 (54.49%) aligned >1 times
93.47% overall alignment rate
```

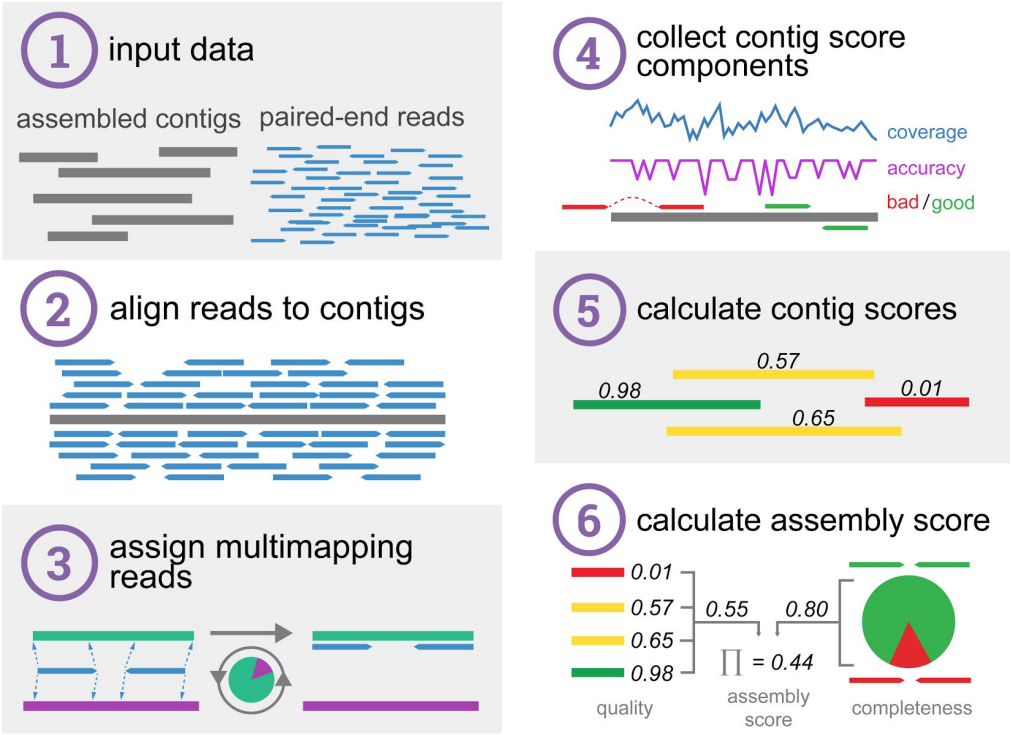
**Ideally >80%**

# *TransRate*: Types of assembly errors

| Error type | Transcripts | Assembly | Read evidence | |
|---|---|---|---|---|
| Family collapse | geneAA, geneAB, geneAC; n=3 | n=1 | bases in reads; agreement | Multiple members of a gene family assembled into a single hybrid contig. |
| Chimerism | geneC, geneB; n=2 | n=1 | coverage | Multiple transcripts concatenated into one contig |
| Unsupported insertion | n=1 | n=1 | no reads align to insertion | Bases are inserted into contig that are not supported by read evidence. |
| Incompleteness | n=1 | n=1 | read pairs align off end of contig | Reads align off ends of contigs. |
| Fragmentation | n=1 | n=4 | bridging read pairs | Reads bridge two contigs. Detect: read mapping |
| Local misassembly | n=1 | n=1 | read pairs in wrong orientation | Inversions and other de novo assembly problems |
| Redundancy | n=1 | n=3 | all reads assign to best contig | Transcript represented in multiple contigs |

Smith-Unna et al 2015, http://dx.doi.org/10.1101/021626

# *TransRate* assembly evaluation




Smith-Unna et al 2015, http://dx.doi.org/10.1101/021626

# *TransRate* example data

Contig metrics:

| | |
|---|---|
| n seqs | 510060 |
| largest | 36322 |
| n bases | 660425775 |
| mean len | 1294.8 |
| n under 200 | 0 |
| n over 1k | 156034 |
| n over 10k | 3652 |
| n with orf | 108295 |
| mean orf percent | 32.84 |
| n90 | 418 |
| n70 | 1513 |
| n50 | 3186 |
| n30 | 5234 |
| n10 | 8930 |
| gc | 0.49 |
| gc skew | 0.01 |
| at skew | 0.0 |
| cpg ratio | 1.42 |
| bases n | 0 |
| proportion n | 0.0 |
| linguistic complexity | 0.19 |

Read mapping metrics:

| | |
|---|---|
| fragments | 147213266 |
| fragments mapped | 142143230 |
| p fragments mapped | 0.97 |
| good mappings | 136385958 |
| p good mapping | 0.93 |
| bad mappings | 5757272 |
| potential bridges | 146664 |
| bases uncovered | 136521497 |
| p bases uncovered | 0.21 |
| contigs uncovbase | 233236 |
| p contigs uncovbase | 0.46 |
| contigs uncovered | 45698 |
| p contigs uncovered | 0.09 |
| contigs lowcovered | 431416 |
| p contigs lowcovered | 0.85 |
| contigs segmented | 28608 |
| p contigs segmented | 0.06 |

| | |
|---|---|
| **TRANSRATE ASSEMBLY SCORE** | **0.4221** |
| TRANSRATE OPTIMAL SCORE | 0.5469 |
| TRANSRATE OPTIMAL CUTOFF | 0.0426 |
| good contigs | 475768 |
| p good contigs | 0.93 |

**mean of all contig scores x p mapped reads**

**what assembly score would be if all “bad” contigs were removed**

**cutoff score for “bad” contigs**

# *TransRate* example data: good and less good

Read mapping metrics:

| | |
|---|---|
| p fragments mapped | 0.97 |
| p good mapping | 0.93 |
| bad mappings | 5757272 |
| potential bridges | 146664 |
| p bases uncovered | 0.21 |
| p contigs uncovbase | 0.46 |
| p contigs uncovered | 0.09 |
| p contigs lowcovered | 0.85 |
| p contigs segmented | 0.06 |

| | |
|---|---|
| TRANSRATE ASSEMBLY SCORE | 0.4221 |
| TRANSRATE OPTIMAL SCORE | 0.5469 |
| TRANSRATE OPTIMAL CUTOFF | 0.0426 |
| p good contigs | 0.93 |

Read mapping metrics:

| | |
|---|---|
| p fragments mapped | 0.25 |
| p good mapping | 0.21 |
| bad mappings | 13948087 |
| potential bridges | 0 |
| p bases uncovered | 0.67 |
| p contigs uncovbase | 0.68 |
| p contigs uncovered | 1.0 |
| p contigs lowcovered | 1.0 |
| p contigs segmented | 0.08 |
| TRANSRATE ASSEMBLY SCORE | 0.0312 |
| TRANSRATE OPTIMAL SCORE | 0.0838 |
| TRANSRATE OPTIMAL CUTOFF | 0.0119 |
| p good contigs | 0.73 |

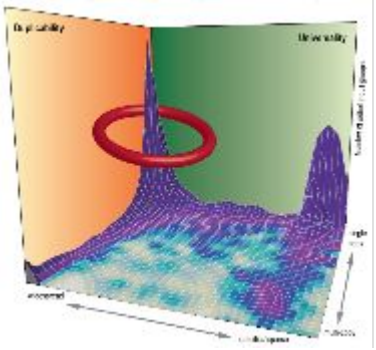
# Comparing assemblies using TransRate

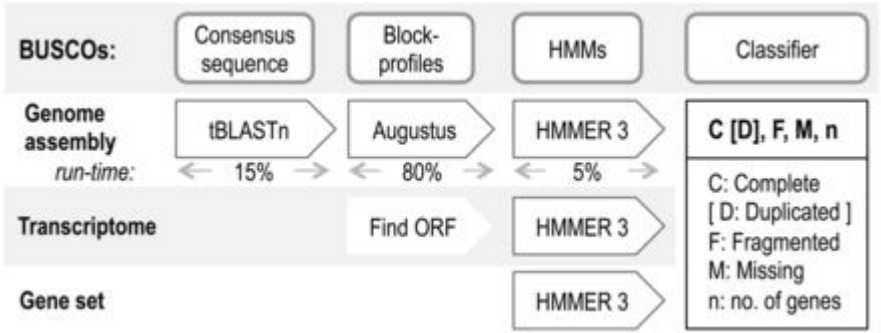
- Looks for similarities between 2 assemblies using CRBB (Conditional Reciprocal Best BLAST)
  - Conservative method for finding orthologs for annotation
  - Compare assembly1 to assembly2/reference using blastx
  - Compare assembly2/reference to assembly1 using tblastn
  - Conditional = e-value (similarity) cutoff is not user-defined
  - Learned by algorithm, accounting for sequence length and overall “relatedness” of the 2 datasets
- Tells you about relative completeness of assemblies
  - How much of assembly1 has hits to assembly2/reference & vice versa

```
Comparative metrics:
-----------------------------
CRBB hits                38678
p contigs with CRBB       0.34
n contigs with CRBB      38678
p refs with CRBB          0.33
n refs with CRBB         12874
reference coverage        0.65
rbh per reference         0.71
cov25                    11303
cov50                     9336
cov75                     6908
cov85                     5357
cov95                     2841
p cov25                   0.29
p cov50                   0.24
p cov75                   0.18
p cov85                   0.14
p cov95                   0.07
```

# BUSCO Evaluation of Transcriptome Completeness

- Benchmarking Universal Single-Copy Orthologs (BUSCO)
- Groups of genes with single-copy orthologs in >90% of species (OrthoDB)
- Expected to be present in any newly sequenced species
- 3023 genes for vertebrates, 843 for metazoans, 429 for eukaryotes






Simão et al 2015 https://doi.org/10.1093/bioinformatics/btv351

# BUSCO Evaluation of Transcriptome Completeness

| Species | Size | BUSCO notation assessment results |
|---|---|---|
| *D. mela* | 139 Mbp | C:98% [D:6.4%], F:0.6%, M:0.3%, n:2 675 |
| | 13 918 genes | C:99% [D:3.7%], F:0.2%, M:0.0%, n:2 675 |
| *C. eleg* | 100 Mbp | C:85% [D:6.9%], F:2.8%, M:11%, n:843 |
| | 20 447 genes | C:90% [D:11%], F:1.7%, M:7.5%, n:843 |
| *H. sapi* | 3 381 Mbp | C:89% [D:1.5%], F:6.0%, M:4.5%, n:3 023 |
| | 20 364 genes | C:99% [D:1.7%], F:0.0%, M:0.0%, n:3 023 |
| *L. giga* | 359 Mbp | C:89% [D:2.3%], F:4.3%, M:5.8%, n:843 |
| | 23 349 genes | C:90% [D:13%], F:7.8%, M:2.1%, n:843 |
| *A. nidu* | 30 Mbp | C:98% [D:1.8%], F:0.9%, M:0.2%, n:1 438 |
| | 10 534 genes | C:95% [D:7.3%], F:3.8%, M:0.9%, n:1 438 |

Simão et al 2015 https://doi.org/10.1093/bioinformatics/btv351

**C: complete**

- length of aligned sequence is within 2 SD of the BUSCO group's mean length (i.e. 95% expectation)

**D: duplicated**

- multiple copies of complete gene found in dataset (should be 0 or very low)

**F: fragmented**

- not complete

**M: missing**

- expected BUSCO missing from data set

# More example assembly stats

| Name | Num. Reads | Num. Contigs | Assembly Size | Score | BUSCO |
|---|---|---|---|---|---|
| **Single Ind.** | 38M | 205812 | 131.6Mb | 0.3064 | C:81%,D:41%,M:9% |
| **Subsampled** | 38M | 304162 | 183.8Mb | 0.2619 | C:84%,D:47%,M:8.4% |
| **10 Ind.** | 269M | 913295 | 440.2Mb | 0.22011 | C:88%,D:51%,M:5% |

MacManes 2016, doi: https://doi.org/10.1101/035642

Assemblies made from larger # of biological reps have lower TransRate scores due to higher polymorphism but recover more BUSCOs.