

# Transrate: Quality assessment of *de-novo* transcriptome assemblies

Richard Smith-Unna      Chris Boursnell      Julian M Hibberd  
Steven Kelly

January 27, 2015

## Abstract

Improvements in short-read sequencing technology combined with rapidly decreasing prices have enabled the use of RNA-seq to assay the transcriptome of species whose genome has not been sequenced.

*De-novo* transcriptome assembly attempts to reconstruct the original transcript sequences from short reads.

Such transcriptome assemblies are relied upon for gene expression studies, phylogenetic analyses, and molecular tooling.

It is therefore important to ensure that assemblies are as accurate as possible, but to date there are few published tools for deep quality assessment of *de-novo* transcriptome assemblies, and none that allow the identification of useful parts of an assembly.

We present **transrate**, an open source command-line program that automates deep analysis of transcriptome assembly quality.

Transrate evaluates assemblies based on inspecting contigs, paired-read mapping, and optionally comparison to reference sequences with an extensive suite of established and novel metrics.

We introduce the **transrate scores**: novel summary statistics based on an explicit, intuitive statistical model of transcriptome assembly that captures many aspects of assembly quality.

Individual contigs and entire assemblies can be scored, enabling quality filtering of contigs and comparison and optimisation of assemblies.

Uniquely, the components of the transrate score quantify specific common problems with individual contigs, allowing the identification of subsets of contigs

that can be improved by post-processing, and those that are already suitable for downstream analysis.

We demonstrate using real and simulated data that the transrate score accurately assesses contig and assembly quality, identifies the strengths and weaknesses of different assembly strategies, and classifies contigs.

## Background

The use of RNA-seq for de-novo transcriptome assembly is a complex procedure, but if done well can yield valuable, high throughput biological insights at relatively low cost.

A transcriptome assembly pipeline might include trimming adapters and low quality bases, read error correction, digital normalisation, assembly and post-assembly improvements such as scaffolding and clustering.

Because the computational problems involved in these steps are hard to solve [\*cite?], there are many competing approaches. For example, popular tools for the assembly step include Trinity (Grabherr et al. 2011), Oases (Schulz et al. 2012), Trans-AbySS (Robertson et al. 2010), IDBA-tran (Peng et al. 2013) and SOAPdenovo-Trans (Xie et al. 2014), among many others. Each of these tools implements a complex algorithm with many heuristics and parameters that can often be user-controlled. Furthermore, because each organism has unique genomic properties, the algorithms need to be selected and tuned carefully for each experiment.

These conditions mean that any de-novo transcriptome experiment should ideally involve comparison of assemblies from across a potentially vast parameter space. For this to be tractable, a method is required to accurately judge the quality of transcriptome assemblies.

In addition to quality varying between assemblies, contigs within an assembly can have varying levels of usefulness. Any transcriptome assembly is likely to contain well-assembled contigs that represent full-length transcripts, as well as poorly assembled contigs that are incomplete or erroneous reconstructions of transcripts, and nonsense contigs that are artefacts of the assembly algorithm. It is therefore desirable to be able to select out the well-assembled contigs, likely to be of use in downstream biological interpretation, from those that are not suitable for downstream use.

Evaluation of the quality of genome and metagenome assemblies is a relatively mature field. Approaches include providing a range of basic metrics about assemblies (Gurevich et al. 2013), or explicitly modelling the sequencing and assembly process to provide a likelihood-based measure of quality (Clark et al. 2013, Rahman and Pachter (2013)). By comparison, there has been little work on detailed quality analysis of transcriptome assemblies. Some authors have used

reference-based measures for evaluation of de-novo transcriptome assemblies (Elijah K Lowe, Billie J Swalla, and C. Titus Brown 2014; O’Neil and Emrich 2013; O’Neil et al. 2010). However, in most cases, a high-quality, closely related reference transcriptome is not available, limiting the usefulness of these metrics in practice (B. Li et al. 2014). To date, only a single published reference-free transcriptome assembly evaluation tool, RSEM-EVAL (B. Li et al. 2014), takes a statistically principled approach to transcriptome assembly quality evaluation.

In this work we describe Transrate, our software for deep quality analysis of transcriptome assemblies. Transrate implements two novel reference-free statistics: the Transrate contig score and the Transrate assembly score. These allow for optimisation within and between assemblies respectively, using only the assembly and the paired-end reads used to generate it.

Unlike existing reference-free statistical approaches to assembly evaluation, the Transrate scores are made up of components that are independently useful in identifying specific problems with contigs, namely gene family collapse, fragmentation or chimerism.

Uniquely, Transrate uses these components to classify the contigs in an assembly, outputting separate files containing the well-assembled contigs, those that could be improved by scaffolding, those that require chimera splitting, those that might benefit from targetted reassembly, and those that are poor quality in multiple or other ways. Furthermore, the Transrate model is descriptive rather than generative, making it considerably easier for users to understand and interpret than existing methods.

We demonstrate the value of Transrate in several ways. Firstly, we explain the use of Transrate for analysing the quality of individual contigs in an assembly. Next, we show that Transrate is more accurate at evaluating relative contig quality than existing reference-free measures when tested using real and simulated data, and is uniquely able to measure absolute quality of contigs to a high degree of accuracy. We demonstrate that Transrate accurately classifies contigs to identify several types of recoverable misassembly, and show that using the Transrate contig score to select the optimal subset of contigs from an assembly improves the biological utility of assemblies. By conducting a large-scale analysis of assemblies from simulated data, we demonstrate the utility of Transrate for comparing and optimising pipelines. Finally, we conduct a survey of over 150 assemblies from the literature to provide a benchmark distribution against which users can compare their Transrate assembly scores.

# Methods

## Transrate

### Overview

Transrate takes as input one or more transcriptome assemblies generated from the same set of paired-end reads, and the reads used to generate the assemblies.

Analysis proceeds by aligning the reads to the assemblies. For reads with multiple alignments within an assembly, only the most likely alignment is chosen. For each contig, the reads aligning to it are inspected to accumulate the components of the contig score. The assembly score is calculated using the contig scores and the full set of reads and alignments, including reads that did not align. Finally, contigs are classified according to whether they are (a) well-assembled, poorly assembled but could be improved by either (b) scaffolding, (c) chimera splitting, (d) reassembly or (d) poorly assembled and unable to be improved.

### Implementation

Transrate is written in Ruby and C++. It is open source, released under the MIT license. Code is available at <http://github.com/Blahah/Transrate>, while help and full documentation are available at <http://hibberdlab.com/Transrate>. The code is fully covered by automated tests. The software is operated via a user-friendly command line interface and can be used on OSX and linux. Transrate can also be used programmatically as a Ruby gem.

### Read alignment and assignment

Reads are aligned to each assembly using SNAP v1.0.0.dev67 (Zaharia et al. 2011). Alignments are reported up to a maximum edit distance of 30. Up to 10 multiple alignments are reported per read where available (`-omax 10`), up to a maximum edit distance of 5 from the best-scoring alignment (`-om 5`). Exploration within an edit distance of 5 from each alignment is allowed for the calculation of MAPQ scores (`-D 5`).

BAM-format alignments produced by SNAP are passed to Salmon, part of the Sailfish suite, (Patro, Mount, and Kingsford 2014), to assign multi-mapping reads to their most likely contig of origin.

### The Transrate scores

The Transrate scores, the contig score and the assembly score, estimate confidence in the quality of contigs and assemblies respectively.

An assembly consists of a set of contigs  $C$  derived from a set of reads  $\hat{R}$ . Reads are aligned and assigned to contigs such that  $\forall c_i \in C, \exists R_i \in \hat{R} : R_i$  is the set of reads assigned to  $c_i$ .

### For contigs

We model a perfect contig as:

1. being a representation of a single transcript such that:
2. each base in the contig must be derived from only one transcript
3. all bases in the contig must be derived from the same transcript
4. unambiguously and accurately representing the identity of each base in the transcript
5. being structurally accurate and complete, such that the ordering of bases in the contig faithfully recreates the ordering of bases in the transcript

The Transrate contig score is an estimate of the probability that a contig is perfect, i.e. meets all these criteria, using the aligned, assigned reads as evidence. We estimate the contig score  $p(c_i)$  by taking the product of the probability of the components  $S_1..S_4$ , mapping approximately to the criteria above.

To estimate our confidence  $p(S_1)$  that each base in the contig is derived from a single transcript, we use the alignment edit distance, i.e. the number of changes that must be made to a read in order for it to perfectly match the contig sequence. We denote the edit distance of an assigned read  $r_{ij} \in R_i$  as  $e_{r_{ij}}$  and the set of reads that cover base  $k$  ( $k \in [1, n]$ ) as  $\varrho k$ . The maximum possible edit distance for alignment is fixed during alignment, denoted as  $\hat{e}$ . Then the probability  $p(b)$  that a base is derived from a single transcript is estimated as the arithmetic mean of  $1 - \frac{e_{r_{ij}}}{\hat{e}}$  for each  $r_{ij} \in \varrho k$ , and the probability  $p(S_1)$  that each base in a contig is derived from a single transcript is then the root mean square of  $p(b)$ .

We adapt the Bayesian segmentation algorithm of J. S. Liu and Lawrence (1999) to estimate  $p(S_2)$ , our confidence that all bases in a contig derive from the same transcript. We assume that a contig that represents a single transcript will have a read coverage related to the expression level of that transcript in the sequenced sample. A contig that is a chimera derived from concatenation two or more transcripts will have multiple levels of read coverage representing the expression levels of its component transcripts. We therefore approximate  $p(S_2)$  by the probability that the read coverage over a contig has a single level. To make the computation tractable, we further simplify the problem by treating the read coverage along the contig as a sequence of symbols in an unordered alphabet. We achieve this representation by discretising the coverage at each base by taking its base-2 logarithm, rounded to the nearest integer.  $p(S_2)$  can then be stated as the probability that the sequence of coverage values does not change composition at any point along its length, i.e. that it is composed of a

single segment. The Liu and Lawrence (1999) algorithm is applied to find this probability.

Whether the contig accurately represents base identity of the transcript of origin is partially captured in  $p(S_1)$  for bases that have reads assigned to them. We thus capture the missing information required to include this confidence in the score as  $p(S_3)$ , which is estimated as the proportion of bases that are supported by assigned reads.

Confidence in the structural accuracy and completeness of a contig,  $p(S_3)$ , is estimated using the pairing information of reads. We classify alignments of read pairs according to whether they are biologically plausible if we assume that the contig is structurally accurate and complete. Thus a read pair must meet all the following criteria to be ‘valid’: (a) both reads in the pair align to the same contig, (b) in an orientation that matches the sequencing protocol, (c) within a plausible distance given the fragmentation and size selection applied in the sequencing protocol.  $p(S_3)$  is then approximated by the proportion of reads  $R_i$  that are valid.

### The assembly score

We model a perfect assembly descriptively such that:

1. a perfect assembly is made up of a single perfect contig representing each transcript (and thus has high per-contig scores)
2. has all transcripts represented (and thus incorporates a high proportion of the experimental evidence)

We take the geometric mean of the contig scores to represent (1), and use the proportion of read pairs that had at least one structurally valid alignment to represent the completeness of the assembly.

Our confidence  $p(C)$  in the quality of an assembly can therefore be expressed as:

$$q_A = \sqrt{\left(\prod_{c=1}^n q_c\right)^{\frac{1}{n}} R_{valid}}$$

### Evaluation

To evaluate the Transrate scores, several different strands of analysis were performed. We used assemblies and read data from previously published assembly papers, as well as simulated reads from reference data, to conduct a detailed evaluation of the contig and assembly scores. We then conducted a broader survey of the range of assembly scores achievable using the entire NCBI Transcriptome Shotgun Archive.

## Detailed algorithm evaluation

**Using real reads** To evaluate the contig and assembly scores using real data, we used Transrate to analyse assemblies from two previous publications: Xie et al. (2014), and Davidson and Oshlack (2014).

From Xie et al. (2014), assemblies were available for rice (*Oryza sativa*) and mouse (*Mus musculus*) that had been assembled using Oases, Trinity, and SOAPdenovo-Trans. From Davidson and Oshlack (2014), assemblies were available for human (*Homo sapiens*) and yeast (*Saccharomyces cerevisiae*) that had been assembled with Oases and Trinity.

These assemblies were chosen because they represent a phylogenetically diverse range of species assembled with several assemblers, with the read data and the transcriptome assemblies available to download, and with a relatively well-annotated reference genome available for each species.

Transrate was run separately for each species, with the full set of reads and all assemblies for that species as input.

For the pre-made assemblies, we generated a reference-based score for each contig in each of the ten assemblies. A reference dataset was compiled by including all transcripts plus any non-coding RNAs described in the reference annotation for each species.

Contigs were compared to the reference dataset by nucleotide-nucleotide local alignment with BLAST+ blastn version 2.2.29 (Camacho et al. 2009). Because no genome annotation is complete, de-novo transcriptome assemblies are likely to contain contigs that are well-assembled representations of real RNA species not present in the reference. We therefore only considered contigs for score comparison if they aligned successfully to at least one reference transcript.

Each contig that had at least one hit was given a reference score by selecting the alignment with the lowest e-value for each contig, then taking the product of the proportion of the reference covered, the proportion of the query covered, and the identity of the alignment.

**Using simulated data** We generated reads by simulated sequencing for each of the four species (rice, mouse, human and yeast) using flux-simulator v1.2.1 (Griebel et al. 2012). For each species, a total of 10 million mRNA molecules were simulated from across the full set of annotated mRNAs from the Ensembl annotation with a random (exponentially distributed) expression distribution. mRNA molecules were uniform-randomly fragmented and then size-selected to a mean of 400 and standard distribution of 50. From the resulting fragments, 4 million pairs of 100bp reads were simulated using the error profile included in flux-simulator, which is learned from real Illumina 76bp reads.

From each set of simulated reads, an assembly was generated using Velvet-Oases, with a kmer size of 23, and defaults for all other parameters.

Accuracy was evaluated as for real data, except that all contigs (including those that did not align) were incorporated into the accuracy calculation.

### **Assembly score survey**

A survey of the range of achievable assembly scores was conducted by analysing transcriptome assemblies from the Transcriptome Shotgun Archive (<http://www.ncbi.nlm.nih.gov/genbank/tsa>). Entries in the archive were filtered to retain only those where paired-end reads were provided, the file-size of the compressed reads was lower than 8GB, the assembler and species were named in the metadata, and the number of contigs was at least 5,000. For the retained entries, the assembly and reads were downloaded, and Transrate run to produce the assembly score for each entry.

## **Results and discussion**

### **Transrate is software for deep quality analysis of transcriptome assemblies**

We have developed Transrate, a method for detailed quality analysis of *de-novo* transcriptome assemblies and their constituent contigs without relying on a reference dataset of any kind. Transrate uses only the contigs themselves and the paired-end reads used to generate them as evidence. In the following sections we present the Transrate method. First we describe the Transrate contig and assembly scores, with a focus on how they can be used to identify misassemblies, select the most useful information from the assembly, and to improve and compare assemblies. Next, we perform experiments using real and simulated data across a range of species to evaluate the accuracy and usefulness of the method, and demonstrate its improvement over existing methods.

In transcriptome assembly experiments, the aim is to reconstruct as accurate a representation as possible of the true set of mRNAs present in biological sample. However, due to errors and noise in the sequencing process; incomplete coverage of all transcripts due to low expression or insufficient sequencing depth; and the computational complexity of assembly, an assembly is an imperfect reconstruction. The aim of Transrate is to enable iterative improvements towards a perfect assembly, regardless of the assembly pipeline used, and to quantify confidence in any given assembly or contig. Because the vast majority of transcriptomics experiments currently use Illumina paired-end sequencing, Transrate is focused on data of this type, although the method could be expanded to other types of sequencing.



# **transrate**

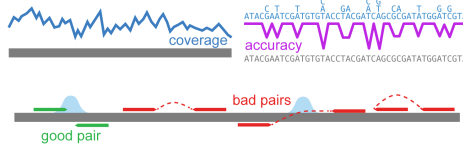
## 1 input data



## 2 map and assign reads



## 3 collect contig score components



## 4 calculate contig transrate scores

$$q_c = \sqrt{\left(\prod_{i=1}^n cov_{c_i}\right)^{\frac{1}{n}} (1 - \theta_n)} \left(1 - \frac{\prod_{i=1}^R 1 - edit(R_{c_i})}{R}\right) \left(\frac{\prod_{i=1}^R good(R_{c_i})}{R}\right)$$



## 5 calculate assembly transrate score

$$q_A = \sqrt{\left(\prod_{c=1}^n q_c\right)^{\frac{1}{n}} good(R)}$$



## 6 classify contigs

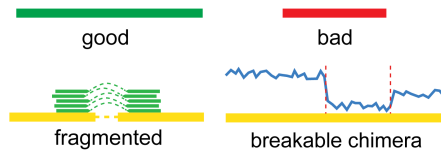


Figure 1: The Transrate workflow. (1.) Transrate takes as input one or more *de-novo* transcriptome assemblies and the paired-end reads used to generate them. (2.) The reads are aligned to the contigs with SNAP, and multi-mapping reads are assigned to their most likely contig of origin with Salmon. (3.) The assigned alignments are examined to measure per-base coverage, per-base edit distance, and the proportion of reads mapping to each contig that agree with the contig structure for each contig. Per-base coverage is analysed to determine segmentation. (4.) Score components are combined to score each contig. (5.) Contig scores are combined with the full set of reads and alignments to score the

## The Transrate contig score evaluates confidence in contigs

Transcriptome assemblies tend to contain characteristic errors that result from methodological constraints. Transrate evaluates each contig in an assembly to determine whether it shows any evidence of these errors when compared to the evidence of the aligned reads. A score between 0 and 1 is produced for each contig, estimating confidence that the contig is a perfect assembly of a transcript that was sequenced. The contig score is derived from a descriptive model that captures our definition of a “perfect” contig. This model is fully described in the *methods* section, but we summarise it briefly here: A contig is considered perfect if it represents all the bases in a single source transcript, with the identity and ordering of bases exactly matching the transcript of origin.

One aim of Transrate is to enable researchers to maximise the biological utility of their transcriptome assemblies by selecting out the high-confidence contigs. To this end, Transrate outputs a FASTA file containing the contigs whose score was  $> 0.5$ , that is, those contigs that are more likely than not to be well-assembled.







	Truth	Assembly
Family collapse	 n=3	 n=1
Chimerism	 n=2	 n=1
Fragmentation	 n=1	 n=4

Figure 2: Three types of common transcriptome assembly errors captured by Transrate. (1) Gene family collapse: multiple similar transcripts are collapsed into a single contig. (2) Chimeras: multiple transcripts are concatenated together into a single contig. (3) Fragmentation: a single transcript is represented by multiple contigs each representing different parts of the transcript.

In addition to providing users with the well-assembled contigs for downstream use, the way the contig score is constructed allows identification of specific kinds of misassembly that are potentially recoverable. Transrate outputs a FASTA file for each possible type of error containing contigs that exhibit only that error, as depicted in figure 2:

**Gene family collapse.** Transcripts from different genes in a family, from

homeologs, or from gene copies share a high level of sequence identity. The heuristics used by assemblers to avoid incorporating read errors can lead to this true biological information being collapsed, by outputting a single contig from reads that in reality originated from multiple similar transcripts. If groups of such contigs can be separated from the rest of the assembly, they could potentially be reassembled using more relaxed heuristics to achieve a better representation of the source transcripts.

**Chimeras.** Regions of repetitive sequence that are shared between multiple transcripts, especially in the polyA tails or UTRs, can be difficult for assemblers to distinguish from genuine connectivity. It is therefore common to find that a contig contains two or more otherwise well-assembled transcripts that have been concatenated together. If these contigs can be identified, they could potentially be examined and split at the point of concatenation to recover the useful biological information.

**Fragmentation.** Low coverage regions within a sequenced transcript can result from various phenomena including low sequencing depth, low abundance transcripts, and high or low complexity in the original sequence. Whatever the cause, low coverage can lead to incomplete assembly of a transcript, so that the transcript is present in several separate, non-overlapping contigs. Using the pairing of reads, it is common practise to scaffold these fragments. However, in our experience many scaffolded assemblies still contain them. By identifying all the contigs that show evidence of fragmentation, iterative targetted scaffolding could potentially be applied to improve contiguity.

### **The Transrate assembly score quantifies assembly quality**

Having evaluated all contigs in an assembly, Transrate produces an assembly score, capturing the overall quality of the assembly. This score allows comparison of assemblies from the same set of reads, enabling optimisation of assembly protocols.

When comparing two assemblies from the same reads, there are some situations in which we have a clear intuition about which assembly is better. If we consider two assemblies that each represent the same proportion of the sequenced transcripts, but where the contigs in one assembly tend to be less accurate reconstructions of their source transcripts, the assembly with the more accurate contigs should be preferred. Conversely, of two assemblies that have equally good quality contigs, but where one assembly captures more transcripts, it is the more complete assembly that should be preferred. In summary, a perfect assembly must be accurate and complete, with our confidence in the assembly being proportional to these two features.

The assembly score captures this intuition. The score is the product of two components: (1) the geometric mean of all the contig scores, representing the

quality of the contigs that were present, and (2) the proportion of input read pairs that supported the assembly, representing the completeness of the assembly.

### The Transrate score components are independent and classifiable

Key to the contig score, and the classification of contigs, is to capture different types of misassembly. To ensure that the score and its components captured phenomena present in real assemblies, we used Transrate to analyse 10 previously published assemblies from four species as described in *Methods*.

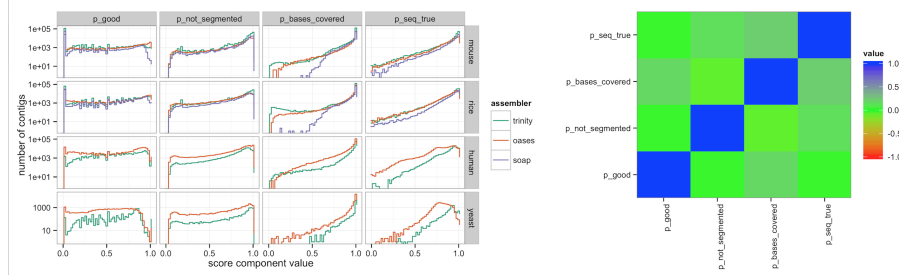


Figure 3: Detailed examination of the contig score components. (a) Distribution of each contig score component in ten assemblies across four species and three assemblers. (b) All-vs-all pairwise Spearman correlation of contig score components using 5,000 contigs samples from each of ten assemblies.

We sought to ensure whether each of the contig score components was contributing useful information. Figure 3a shows that all the components were distributed across the full range (0 to 1) in all assemblies, and that each exhibits enrichment towards the extremes. This suggests again that the components capture real variability present in assemblies, and that the components are useful for classification.

In order to determine whether all the components were necessary, we examined correlation between the components. To avoid giving undue weight to larger assemblies, we sampled 5000 contigs from each assembly. Pearson correlation was calculated between each pair of score components (Figure 3b) demonstrate that the score components are independent, i.e. that each captures unique information compared to the other components.

Next, we examined the distribution of contig scores to establish whether it allowed natural classification of contigs. As shown in figure 4, the contig score showed a bimodal distribution in most assemblies, with enrichment of scores close to the extremes. Some assemblies (rice-soap, human-trinity and human-yeast) did not show bimodal distributions, but were distributed across the range. This indicates that the contig score is capturing real variability between contigs in all the assemblies, and suggests a natural dichotomy between high and low quality contigs in many assemblies.

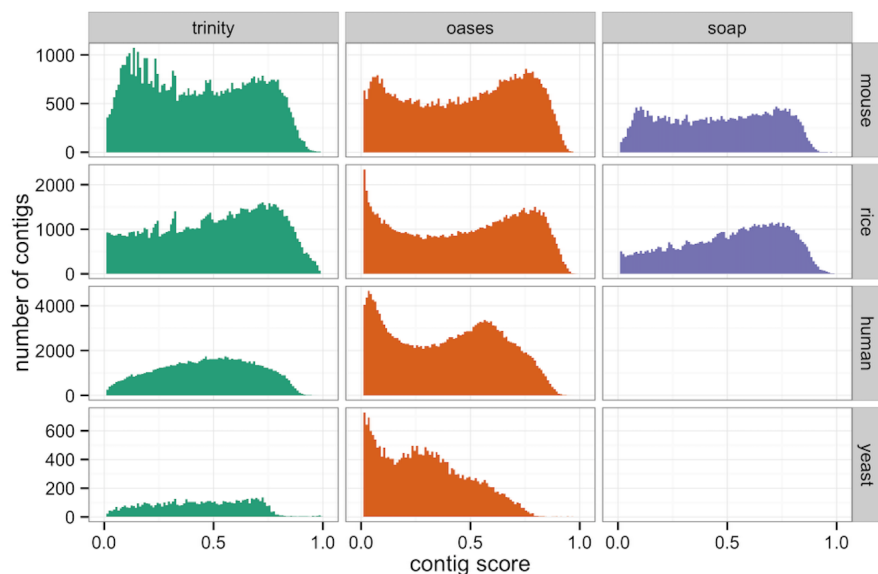


Figure 4: Distribution of contig scores across ten assemblies using real data from four species and three assemblers.

### The contig score is a highly accurate measure of contig quality

Having established that the contig score meets basic requirements for use as a classifier, we sought to quantify its accuracy using both real and simulated data.

**Using real data** For each of the four species for which we sourced assemblies from the literature, we downloaded the full set of annotated cDNAs for that species from the Ensembl Genomes v25 release to use as a reference. We aligned the contigs from each assembly to the reference using `blastn`. Because the *de-novo* assemblies are likely to contain genuine biological novelty, including unannotated transcripts from known genes, transcripts from unannotated genes, and lncRNAs, we considered only the set of contigs that aligned to a reference transcript.

We constructed a reference score that reflects our intuition about a perfect contig. In the reference context, a perfect contig aligns to a transcript so that the entire transcript is covered by the contig, the entire contig is covered by the transcript, and the aligned sequences have perfect sequence identity. We formalised this as a reference score as follows:

$$refscore = qcov * tcov * id$$

We calculated this score for every contig that mapped to a reference transcript

in each assembly. We then classified contigs as ‘true’ or ‘false’ according to the reference, with refscore  $\geq 0.9$  corresponding to ‘true’, and refscore  $< 0.9$  corresponding to ‘false’.

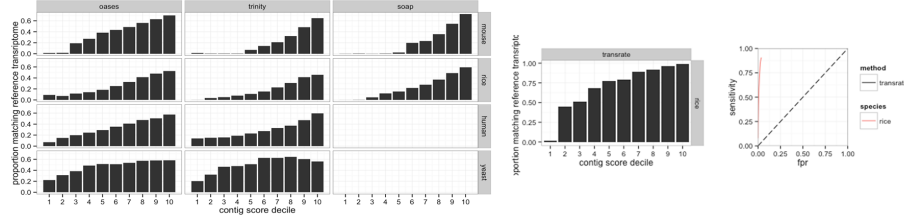


Figure 5: Accuracy of the reference-free Transrate contig score as compared to a reference-based evaluation. (a) Proportion of contigs in each Transrate score bin that are ‘true’ according to the reference, for contigs in assemblies generated from real sequencing data. Note that differences between the reference and the truth limit the maximum proportion of ‘true’ contigs in any bin. (b) As a, but using simulated reads where the ground truth is known. (c) Receiver Operator Characteristic (ROC) curves for assemblies of each species using simulated reads, where the ground truth is known.

As expected, we found that the higher Transrate score deciles tend to contain higher proportions of ‘true’ contigs when compared to the reference (figure 5a). Across rice, mouse and human there was a very clear positive trend with increasing proportions of true contigs in increasing score deciles. The trend was present but less clear in yeast.

Across all assemblies the maximum proportion of contigs that were ‘true’ in any decile was 0.7. We hypothesise that this discrepancy reflects the incompleteness of the reference annotations.

*note to coauthors: we should compare to RSEM-eval here:*

- in the RSEM-eval paper they focus almost exclusively on the assembly score, not on scoring contigs
- however there is a comment that they calculate a `contig_impact_score` which measures the contribution each contig made to the assembly score
- I took a look at those scores for some assemblies - they are range between -5000 and +10000000
- We can do a binning analysis of their `contig_impact_score` and compare it to ours (see example plot in the directory)
- If we do this, we should be sure to include assemblies where there are lots of different kinds of errors, as it would make them look artificially better to select assemblies with only collapse errors

**Using simulated data** To examine the hypothesis that the ~30% of ‘false’ contigs in the highest score bins in the analysis based on real data were caused by incomplete reference annotation, we conducted an experiment using simulated reads where we knew the exact set of possible true transcripts that the assembly might reconstruct.

We simulated 4 million read pairs from each species (rice, mouse, human, and yeast) as described in *Methods*, and assembled them using Velvet-Oases with default settings. We then analysed the assemblies using Transrate, and generated reference-based scores for each contig as in the real data analysis.

Score binning analysis on these data (figure 5b) showed that the highest Transrate contig score bins contained close to 100% ‘true’ contigs when compared to the reference, supporting our hypothesis that incomplete reference caused imperfect correlation in the analysis based on real data.

Because the exact set of possible true assembled sequences was known for the simulated datasets, we extended our analysis to perform a full binary classification accuracy test. Using the ‘true’/‘false’ classification as in the previous analysis, but this time using a naive reference score cutoff as 0.5, we varied the cutoff for classifying contigs using the Transrate contig score from 0.1 to 0.9 and generated receiver operator characteristic curves for each assembly (figure 5c). We also calculated the Matthews correlation coefficient (MCC) for each contig score cutoff. The Transrate score allowed extremely accurate classification of contigs, with an optimal MCC of 0.84, corresponding to a sensitivity of 0.89721695 and specificity of 0.9665546.

*note to coauthors: (this should be extended for the other species - currently just rice for demonstration purposes).*

*note to coauthors: after much thought, I realised it’s not valid to compare to RSEM-eval in the ROC analysis, because their contig\_impact\_score is a relative ranking or impact, not an objective quality score, and has no maximum (that I can discern from their paper or outputs), and so cannot be used for classification.*

## **The assembly score allows comparison between assemblies**

*note to coauthors: I think we should have this section. here’s what I think it should contain:*

- from each test species, we take one chromosome and simulate a small set of rnaseq reads from it, say 250,000 pairs

then:

1. we take those reads and do a parameter sweep with 3 different assemblers, say 40 assemblies from each

- Transrate them all, and also generate reference-based score (geometric mean of the refscores for all contigs as above)
  - show the correlation between change in Transrate assembly score and change in reference-based score
  - at this stage we could include RSEM-eval for comparison
2. take the reference contigs and introduce three kinds of errors at known rates: gene family collapse, chimerism, and fragmentation
    - then Transrate them all and show correlation between Transrate assembly score and the error rate for the different kinds of errors
    - could

### **Contig classification leads to assembly improvement**

*note to coauthors: I think we should do this analysis*

very simple procedure:

- take the assemblies and throw out the contigs with bad scores, show that the assembly score gets better
- show that there's an optimal cutoff? We've talked about this idea before. We could save this for the transfix paper, but the RSEM-eval paper does show a small amount of data where they throw away contigs that don't contribute much - if we want to do better it might be worth including this idea (simple for us to do)

### **Broad analysis of assemblies provides guidance for using the Transrate assembly score**

As we have shown, the Transrate contig score is highly accurate at identifying poor quality contigs, and the assembly score incorporates contig quality and assembly completeness. Assemblies with a very low Transrate score must therefore either contain mostly very poor quality contigs (increasing the difficulty and likelihood of error for downstream analysis), or incorporate very little of the read data (reducing the power of downstream analysis), or both.

Noting that, to date, no study has surveyed assembly quality across the literature, we downloaded and analysed transcriptomes available on the NCBI Transcriptome Shotgun Assembly database (TSA). Assemblies from this database were selected for further analysis only if the following criteria were met:

1. The assembly program was listed
2. The compressed read files were  $\leq 8$ GB in size



3. Paired-end reads were available for download
4. The final assembly contained at least 5000 contigs

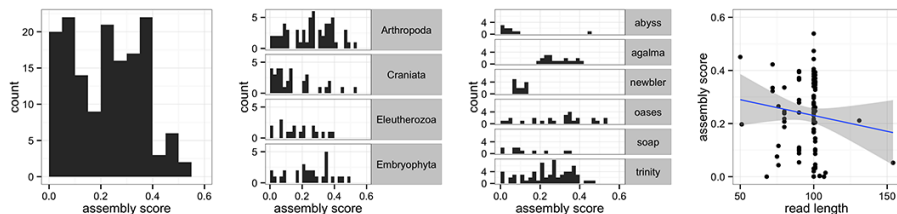


Figure 6: Distribution of assembly scores across 177 assemblies from the NCBI Transcriptome Shotgun Archive. (a) Transrate assembly score distribution across the whole dataset. (b) Assembly score distribution segmented by clade. (c) Segmented by assembler and (d) assembly score plotted against read length.

We ran Transrate on the 177 assemblies that met these criteria, and found that the resulting assembly scores ranged from 0.001 to 0.56 (figure 6a). This suggests that there are many assemblies of very poor quality in the wild. The data also suggest that a score of  $> 0.5$  is excellent compared to the quality of assemblies currently being used in the literature. This survey demonstrates the need for accurate quality assessment, and informed assembly improvement strategies.

Some authors (e.g. Martin and Wang (2011)) have suggested that particular clades of organism may present more of a challenge in transcriptome assembly than others. We examined the TSA analysis data segmented by clade at a depth that separated arthropods, vertebrates, and vascular plants, to see if there was evidence of worse assembly quality in particular clades. Taking only the clades with  $>10$  assemblies (figure 6b), we found no enrichment of poor-quality assemblies in any clade. Rather, we found that in every clade, a range of assembly qualities was present spanning from extremely poor ( $\sim 0.001$ ) to relatively good ( $>0.5$ ).

*note to coauthors: do we need some statistics in here? currently just subjective assessment of the plots*

Novel assembly tools tend to claim superiority over other assemblers when tested using a limited dataset. We used the TSA dataset to see if any assembler consistently produced higher-quality assemblies. Comparing only assemblers with  $>10$  assemblies in the dataset (figure 6c), we found that three of the most popular assemblers (Trinity, Oases, and SOAPdenovo-Trans) produced assemblies of variable quality across the full range of scores present, with no clear advantage of any one assembler. We note that Newbler, Agalma and Trans-Abyss assemblies tended to produce lower scores, but that these assemblies were each from a single study per assembler, limiting the range of scores that might be expected. We therefore refrain from concluding that these assemblers are necessarily worse.

Because contiguity in assembly is related to overlap of the information content in reads, it might be expected that longer reads would reduce the difficulty of the assembly process and lead to higher quality assemblies. We segmented the TSA dataset by read length (figure 6d), and found no evidence that increased read length was correlated with higher assembly scores. However, the power of the analysis was limited by the overrepresentation of reads of length 100 or 101 in the dataset.

### Future work

- automate fixing of the problems identified by transrate
- automate optimisation of assemblies
- automate writing of papers?

### References

- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: architecture and Applications.” *BMC Bioinformatics* 10 (December): 421. doi:10.1186/1471-2105-10-421. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2803857/>.
- Clark, Scott C., Rob Egan, Peter I. Frazier, and Zhong Wang. 2013. “ALE: a Generic Assembly Likelihood Evaluation Framework for Assessing the Accuracy of Genome and Metagenome Assemblies.” *Bioinformatics* 29 (4): 435–43. doi:10.1093/bioinformatics/bts723. <http://bioinformatics.oxfordjournals.org/content/29/4/435>.
- Davidson, Nadia M., and Alicia Oshlack. 2014. “Corset: enabling Differential Gene Expression Analysis for de Novo Assembled Transcriptomes.” *Genome Biology* 15 (7): 410. doi:10.1186/s13059-014-0410-6. <http://genomebiology.com/2014/15/7/410/abstract>.
- Elijah K Lowe, Billie J Swalla, and C. Titus Brown. 2014. “Evaluating a Lightweight Transcriptome Assembly Pipeline on Two Closely Related Ascidian Species.” doi:10.7287/peerj.preprints.505v1. <http://dx.doi.org/10.7287/peerj.preprints.505v1>.
- Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data Without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52. doi:10.1038/nbt.1883. <http://www.nature.com/nbt/journal/v29/n7/full/nbt.1883.html>.
- Griebel, Thasso, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. 2012. “Modelling and Simulating Generic RNA-Seq Experiments with the Flux Simulator.” *Nucleic Acids Research*

- 40 (20): 10073–83. doi:10.1093/nar/gks666. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488205/>.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75. doi:10.1093/bioinformatics/btt086. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624806/>.
- Li, Bo, Nathanael Fillmore, Yongsheng Bai, Mike Collins, James A. Thomson, Ron Stewart, and Colin N. Dewey. 2014. “Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data.” *Genome Biology* 15 (12): 553. doi:10.1186/s13059-014-0553-5. <http://genomebiology.com/2014/15/12/553/abstract>.
- Liu, J. S., and C. E. Lawrence. 1999. “Bayesian Inference on Biopolymer Models.” *Bioinformatics* 15 (1): 38–52. doi:10.1093/bioinformatics/15.1.38. <http://bioinformatics.oxfordjournals.org/content/15/1/38>.
- Martin, Jeffrey A., and Zhong Wang. 2011. “Next-Generation Transcriptome Assembly.” *Nature Reviews Genetics* 12 (10): 671–82. doi:10.1038/nrg3068. <http://www.nature.com/nrg/journal/v12/n10/full/nrg3068.html>.
- O’Neil, Shawn T., and Scott J. Emrich. 2013. “Assessing de Novo Transcriptome Assembly Metrics for Consistency and Utility.” *BMC Genomics* 14 (1): 465. doi:10.1186/1471-2164-14-465. <http://www.biomedcentral.com/1471-2164/14/465/abstract>.
- O’Neil, Shawn T., Jason DK Dzuris, Rory D. Carmichael, Neil F. Lobo, Scott J. Emrich, and Jessica J. Hellmann. 2010. “Population-Level Transcriptome Sequencing of Nonmodel Organisms Erynnis Propertius and Papilio Zelicson.” *BMC Genomics* 11 (1): 310. doi:10.1186/1471-2164-11-310. <http://www.biomedcentral.com/1471-2164/11/310/abstract>.
- Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. “Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.” *Nature Biotechnology* 32 (5): 462–64. doi:10.1038/nbt.2862. <http://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html>.
- Peng, Yu, Henry C. M. Leung, Siu-Ming Yiu, Ming-Ju Lv, Xin-Guang Zhu, and Francis Y. L. Chin. 2013. “IDBA-Tran: a More Robust de Novo de Bruijn Graph Assembler for Transcriptomes with Uneven Expression Levels.” *Bioinformatics* 29 (13): i326–34. doi:10.1093/bioinformatics/btt219. <http://bioinformatics.oxfordjournals.org/content/29/13/i326>.
- Rahman, Atif, and Lior Pachter. 2013. “CGAL: computing Genome Assembly Likelihoods.” *Genome Biology* 14 (1): R8. doi:10.1186/gb-2013-14-1-r8. <http://genomebiology.com/2013/14/1/R8/abstract>.
- Robertson, Gordon, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D. Jackman, Karen Mungall, et al. 2010. “De Novo Assembly and Analysis of RNA-Seq Data.” *Nature Methods* 7 (11): 909–12.

doi:10.1038/nmeth.1517. <http://www.nature.com/nmeth/journal/v7/n11/abs/nmeth.1517.html>.

Schulz, Marcel H., Daniel R. Zerbino, Martin Vingron, and Ewan Birney. 2012. "Oases: robust de Novo RNA-Seq Assembly Across the Dynamic Range of Expression Levels." *Bioinformatics* 28 (8): 1086–92. doi:10.1093/bioinformatics/bts094. <http://bioinformatics.oxfordjournals.org/content/28/8/1086>.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. "SOAPdenovo-Trans: de Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics*, February, btu077. doi:10.1093/bioinformatics/btu077. <http://bioinformatics.oxfordjournals.org/content/early/2014/03/07/bioinformatics.btu077>.

Zaharia, Matei, William J. Bolosky, Kristal Curtis, Armando Fox, David Patterson, Scott Shenker, Ion Stoica, Richard M. Karp, and Taylor Sittler. 2011. "Faster and More Accurate Sequence Alignment with SNAP." *arXiv:1111.5572 [Cs, Q-Bio]*, November. <http://arxiv.org/abs/1111.5572>.