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

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 and library implemented in the Ruby and C languages 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 score**: a novel summary statistic 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 (e.g. [*list a few nice transcriptomics papers]).

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.

Compared to transcriptome assemblies, 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)).

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. As well as reporting many existing statistics about an

assembly, transrate implements two novel reference-free statistics: the transrate contig score and the transrate assembly score. These allow for optimisation within and between assemblers 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.

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 required chimera splitting, 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 show that transrate improves upon the state of the art in several key ways: Firstly, we show that transrate is more accurate than existing reference-free measures when tested using real and simulated data, as well as being significantly faster. Secondly, we demonstrate that transrate accurately identified several types of recoverable misassembly. Finally, we demonstrate that using the transrate contig score to select the optimal subset of contigs from an assembly improves the biological utility of assemblies.

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 score

We developed a reference-free statistical measure of assembly quality, the transrate score.

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:

a. each base in the contig must be derived from only one transcript
b. all bases in the contig must be derived from the same transcript

2. unambiguously and accurately representing the identity of each base in the transcript
3. 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 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 ϱk. The maximum possible edit distance for alignment is fixed by the aligner, 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 letters 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 composition 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 assigned to a contig that are valid.

The assembly score

We model a perfect assembly descriptively such that:

1. a perfect assembly is made up of a perfect contig representing each transcript that it contains (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 algorithm, we opted to use data from previously published assembly papers. Two different strands of analysis were performed: a detailed evaluation of the algorithm using ten assemblies from four species, and a broader survey of the range of assembly scores achievable using the entire NCBI Transcriptome Shotgun Archive.

Detailed algorithm evaluation

Using real data 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 transcripts 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 has at least one hit was given a reference score by selecting the alignment with the lowest bitscore 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, 8 million 100bp paired-end reads were simulated using a learned error profile from real Illumina reads.

From each set of simulated reads, an assembly was generated using 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 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.

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 *method* section, but we summarise it briefly here: A contig is considered perfect if it represents all the bases in a single 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.

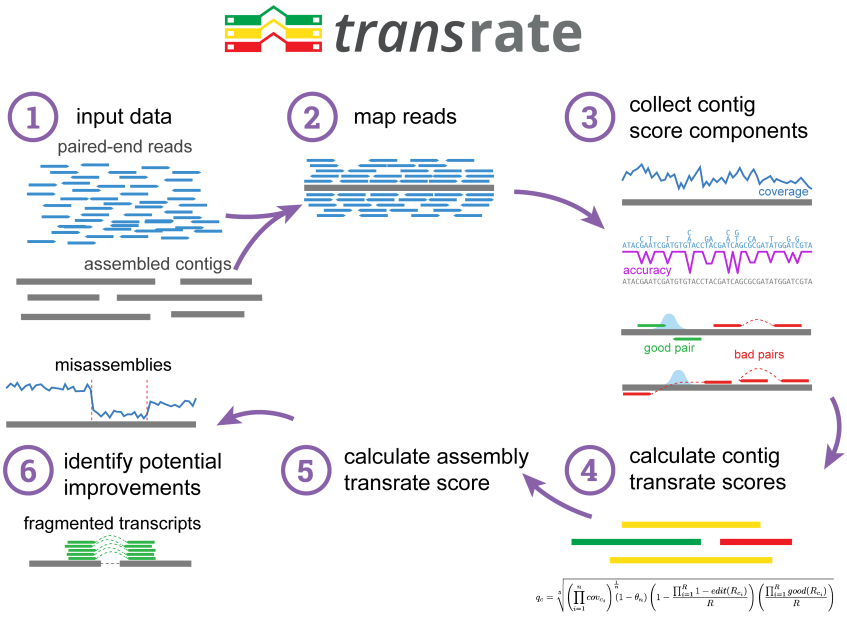

Figure 1: The Transrate workflow. (1.) Transrate takes as input one or more *de-novo* transcriptome assemblies and the paired-end reads used to generate them. (2.) The reads are aligned to the contigs with SNAP, and multi-mapping reads are assigned to their most likely contig of origin with Salmon. (3.) 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 entire assembly. (6.) Contigs are classified according to whether they are well-assembled, poorly assebled and unfixable, or poorly assembled and potentially fixable by either reassembly, chimera splitting, or targeted scaffolding.

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 haplotypes, 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 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 geniune 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 can 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 can 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. 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*.

We sought to ensure whether each of the contig score components was contributing useful information. Figure 2a 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 the the components capture real variability present in assemblies, and that the components are useful for classification.

In order to determine whether the all the components were necessary, we examined correlation between the components. To avoid giving preference to larger assemblies, we sampled 5000 contigs from each assembly. Pearson correlation was calculated between each pair of score components (Figure 2b) 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 3a, 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.

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.

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 novely, including unannotated transcripts from

known genes, transcripts from unannotated genes, and lncRNAs, we considered only the set of

The assembly score allows comparison between assemblies

Contig classification leads to assembly improvement

The contig score accurately predicts true assembly quality.

In order to quantify the accuracy of the transrate score, for each contig in each assembly we sought to establish whether it was an accurate reconstruction of a genome-encoded RNA product.

Because RNAseq reads capture not only protein-coding sequences, but also lncRNAs and other RNAs, we first sought to discover the true set of genome-encoded RNA products represented in each RNAseq dataset. For each species, we aligned the RNAseq reads to the genome and performed a genome-guided assembly of the aligned reads to extend the existing annotation. We then filtered the annotations to remove those with no reads mapping.

For each assembly, we performed gapped alignment of contigs directly to the genome reference using exonerate. Contigs were classified as 'complete' if they aligned full-length to an annotated feature, with gaps corresponding to annotated introns. This allowed complete specification of accuracy as depicted in table 3:

–
Transrate score
High
Low

Using this scheme we evaluated precision, recall and accuracy for each assembly (figure 6). Transrate score was strongly correlated with genome-based completeness (figure 6a).

To evaluate the predictive power of transrate scores, we varied the cutoff for 'high' transrate scores (figure 6). In all cases, the maximum accuracy was > [SOME NUMBER], with precision and recall in the ranges of [V-X] and [Y-Z].

Transrate enables optimisation across the assembler parameter space

Massive-scale analysis of assemblies provides guidance for using the transrate score.

To provide examples of real datasets and scores we downloaded all publicly available assembled transcriptomes that are available on the NCBI Transcriptome Shotgun Assembly database (ftp://ftp.ncbi.nlm.nih.gov/genbank/tsa/). Assemblies from this database were selected only if the following criteria were met:

1. The assembly program was listed.
2. Paired-end reads were available for download.
3. The final assembly contained at least 5000 contigs.

TODO:

- describe the overall distribution of transrate scores - what makes a good score?
- does the score differ by assembly method?
- does it differ by phylogeny?

Future work

Tables

Table 1

Basic metrics as reported in the original publications for assemblies from the 4 species in analysis 1.

Table 2

Summary of the transrate analysis for analysis 1.

Table 3

Summary of the method used to measure accuracy of the transrate score.

Figures

Figure 1

Flow diagram explaining that reads are mapped to reference etc.

Figure 2

Existing commonly used metrics used on the tests datasets used in this paper to describe the datasets.

Figure 3

A) Plot of distribution of values per contig of all selected scores.
B) Correlation between all scores.
C) still to be decided.

Figure 4

Distribution of all contig scores for each of the species for each of the assemblies. Small plots two column figure.

Figure 5

A) Plot of score v contig length
B) score v FPKM, (demonstrate effect of sequence length and expression level bias on score).

Figure 6

A) Exonerate validation percentage of transcripts in each decile that align full length to the genome using exonerate.
B) MCC plot varying the transrate score cut off for inclusion of transcripts.

Figure 7

A) Distribution of all scores for the NCBI TSA database.
B) Assembly score V number of reads.
C) Partitioned by assembly method.
D) Average time taken to run analysis using transrate vs number of reads, transrate score.

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.

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 Dzurisin, Rory D. Carmichael, Neil F. Lobo, Scott J. Emrich, and Jessica J. Hellmann. 2010. "Population-Level Tran-

scriptome Sequencing of Nonmodel Organisms Erynnis Propertius and Papilio Zelicaon." *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.