Evaluation of *de novo* transcriptome assemblies from RNA-Seq data

Nathanael Fillmore,
based on work with Bo Li, Yongsheng Bai, Mike Collins,
James A. Thomson, Ron Stewart, and Colin N. Dewey

University of Wisconsin, Madison

Apr 11, 2015

De novo transcriptome assembly

```
read 1:  AGCATCGCGT
read 2:  CGTTGCGTCC
read 3:  CGTCCCGCGC
read 4:  GCGCGCTTAG
read 5:  GCTACTCTCA
read 6:  TACTCTCACA
```

Reads

⇓

```
AGCATCGCGT
       CGTTGCGTCC
            CGTCCCGCGC
                  GCGCGCTTAG
GCTACTCTCA
  TACTCTCACA
```

⇓

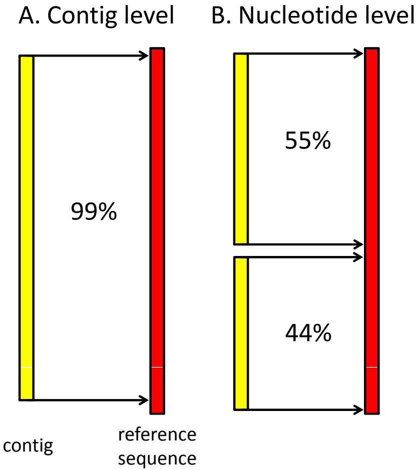
```
contig 1:  AGCATCGCGTTGCGTCCCGCGCTTAG
contig 2:  GCTACTCTCACA
```

De novo assembly

Two classes of assembly evaluation measures

- Reference-based: compare assembly to ground truth reference.
- Reference-free: evaluate assembly without reference.

Reference-based evaluation



Recall values

	Contig	Nucleotide
A	100%	99%
B	0%	99%

Reference-free evaluation

- N50: length of the longest contig such that all contigs of at least that length compose at least 50% of the bases of the assembly.

Reference-free evaluation

- N50: length of the longest contig such that all contigs of at least that length compose at least 50% of the bases of the assembly.
- Statistical model-based scores for evaluating genome (CGAL: Rahman and Pachter, 2013) and metagenome (Genovo: Laserson et al., 2011; ALE: Clark et al., 2013) assemblies.

Our contribution

Our contribution is a reference-free transcriptome assembly scoring function, which can be used to choose the best assembly from a collection of candidate *de novo* assemblies when no ground-truth reference is available. The score is based on a statistical model of the process of RNA-Seq read generation and of “true” transcriptome assembly.

Our reference-free score

$$\begin{aligned}
&\text{score}(\text{assembly}, \text{reads}) = \log P(\text{assembly}, \text{reads}) \\
&= \log \int P(\text{assembly}|\lambda) P(\text{reads}|\text{assembly}, \lambda)\, dP(\lambda) \\
&\approx \log \underbrace{P(\text{assembly}|\lambda^*)}_{\text{prior}} + \log \underbrace{P(\text{reads}|\text{assembly}, \lambda^*)}_{\text{likelihood}} \\
&\qquad - \underbrace{\frac{1}{2} N_{\text{contigs}} \log N_{\text{reads}}}_{\text{BIC penalty}}
\end{aligned}$$

A contig's "coverage" λ_i is the expected number of reads generated from each position of the contig's parent transcript, and λ^* is the maximum likelihood estimate.

The assembly prior $P(\text{assembly}|\lambda)$

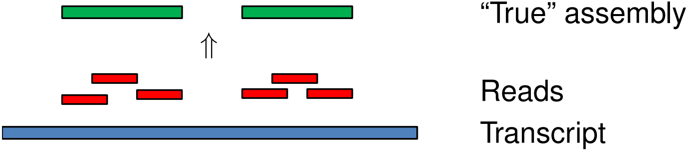
The prior distribution over assemblies is specified indirectly:

- We specify a simple parametric distribution over transcriptomes and reads from them.

The assembly prior $P(\text{assembly}|\lambda)$

The prior distribution over assemblies is specified indirectly:

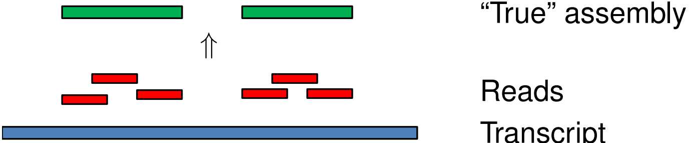
- We specify a simple parametric distribution over transcriptomes and reads from them.
- We define the “true” assembly, formed by joining reads whose true positions (within the transcript set) overlap or are contiguous.



The assembly prior $P(\text{assembly}|\lambda)$

The prior distribution over assemblies is specified indirectly:

- We specify a simple parametric distribution over transcriptomes and reads from them.
- We define the “true” assembly, formed by joining reads whose true positions (within the transcript set) overlap or are contiguous.



- The above induces a distribution over assemblies.

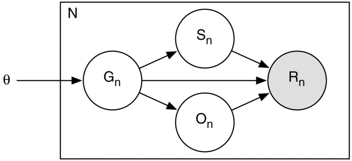
The assembly prior $P(\text{assembly}|\lambda)$

Practical contribution of the prior:

- Penalizes assemblies whose contigs have aberrant lengths relative to the coverage.
- Penalizes assemblies with too many nucleotides.

The data likelihood $P(\text{reads}|\text{assembly}, \lambda)$

RSEM (Li et al., 2010), introduced a generative model of reads, given transcripts and their expression:



where

- θ_j is the expression of transcript j.
- N is the number of reads.
- G_n is the transcript read n comes from.
- S_n is the start position of read n within its transcript.
- O_n is the orientation of read n within its transcript.
- R_n is read n.

The data likelihood $P(\text{reads}|\text{assembly}, \lambda)$

Key observation:

- Generating from contigs $\equiv$ generating from transcripts, except that contigs are guaranteed to be covered by reads.

The data likelihood $P(\text{reads}|\text{assembly}, \lambda)$

Key observation:

- Generating from contigs $\equiv$ generating from transcripts, except that contigs are guaranteed to be covered by reads.

Therefore, we define the likelihood to be the probability of the reads given the contigs, according to RSEM's model, divided by the probability that the contigs are covered by reads.

The data likelihood $P(\text{reads}|\text{assembly}, \lambda)$

Key observation:

- Generating from contigs $\equiv$ generating from transcripts, except that contigs are guaranteed to be covered by reads.

Therefore, we define the likelihood to be the probability of the reads given the contigs, according to RSEM's model, divided by the probability that the contigs are covered by reads.

Practical contribution of the likelihood:

- On one hand, the likelihood penalizes contigs that are not well-supported by reads.
- On the other hand, the likelihood penalizes assemblies that do not make use of all the reads.

Experiment 1 - Random perturbation - Setup

The “true” assembly an approximate local maximum of the score.

Experiment 1 - Random perturbation - Setup

The “true” assembly an approximate local maximum of the score.

Procedure:

- Simulate RNA-Seq data.

Experiment 1 - Random perturbation - Setup

The “true” assembly an approximate local maximum of the score.

Procedure:

- Simulate RNA-Seq data.
- Construct the “true” assembly.

Experiment 1 - Random perturbation - Setup

The “true” assembly an approximate local maximum of the score.

Procedure:

- Simulate RNA-Seq data.
- Construct the “true” assembly.
- Perturb this assembly:
    - Substitution - substitute a base.
    - Fusion - join two contigs into one contig.
    - Fission - split one contig into two contigs.
    - Indel - insert or delete a fragment from a contig.

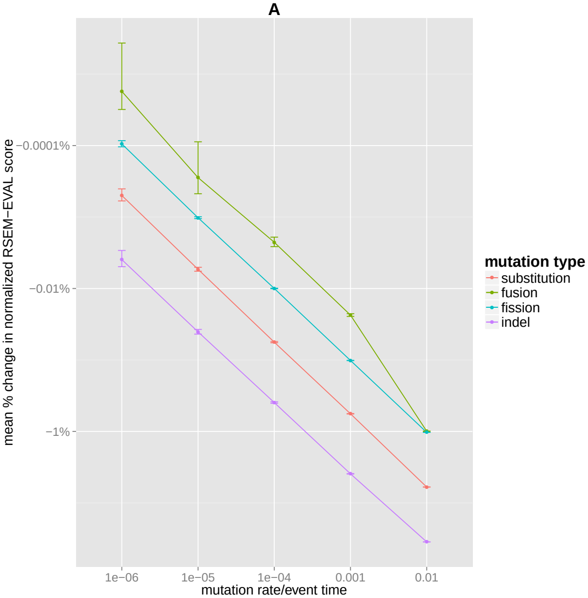
Experiment 1 - Random perturbation - Setup

The “true” assembly an approximate local maximum of the score.

Procedure:

- Simulate RNA-Seq data.
- Construct the “true” assembly.
- Perturb this assembly:
  - Substitution - substitute a base.
  - Fusion - join two contigs into one contig.
  - Fission - split one contig into two contigs.
  - Indel - insert or delete a fragment from a contig.
- Compute score for “true” and perturbed assemblies.

Experiment 1 - Random perturbation - Results



Experiment 2 - Correlation - Setup

Our reference-free score correlates well with simple reference-based scores.

Experiment 2 - Correlation - Setup

Our reference-free score correlates well with simple reference-based scores.

Procedure:

- For each dataset (real mouse and simulated mouse):

Experiment 2 - Correlation - Setup

Our reference-free score correlates well with simple reference-based scores.

Procedure:

- For each dataset (real mouse and simulated mouse):
  - Create $\sim$200 assemblies, by running several assemblers with different parameter settings.

Experiment 2 - Correlation - Setup

Our reference-free score correlates well with simple reference-based scores.

Procedure:

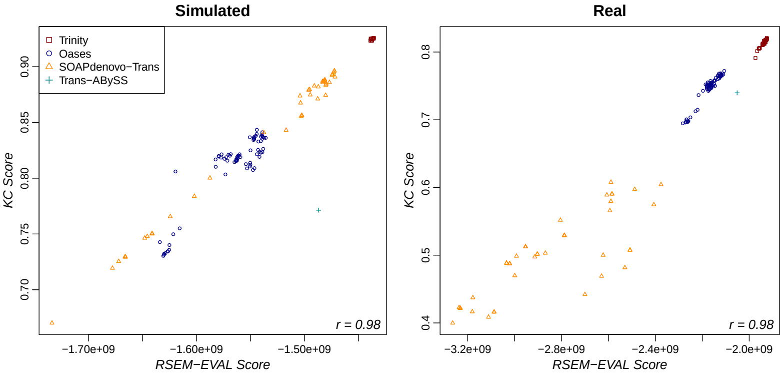
- For each dataset (real mouse and simulated mouse):
    - Create $\sim$200 assemblies, by running several assemblers with different parameter settings.
    - For each assembly, compute:
        - Our model-based score.
        - Contig and nucleotide F1.
        - Our reference-based k-mer compression score (next slide).

Definition of *k*-mer compression score

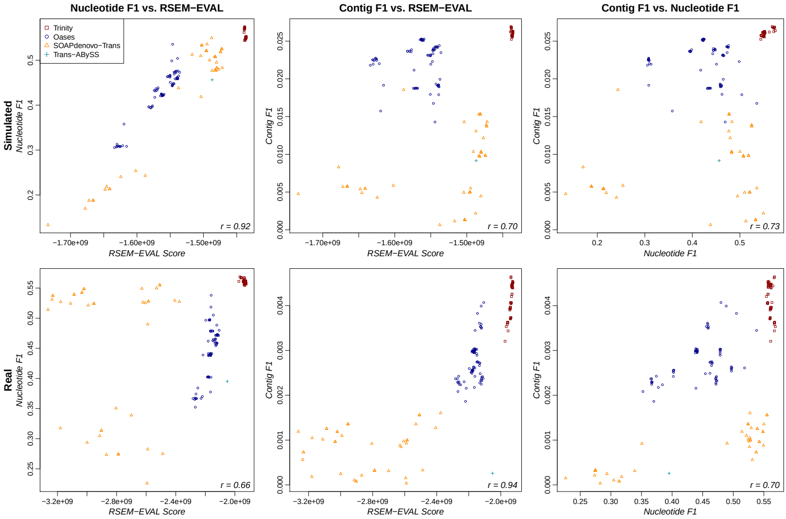
$$
\begin{aligned}
k\text{-mer compression (KC) score} = {} & \text{weighted } k\text{-mer recall (WKR)} \\
& - \text{inverse compression ratio (ICR).}
\end{aligned}
$$

- WKR = assembly's recall of the *k*-mers present in the reference sequences, with each *k*-mer weighted by its relative frequency within the reference transcriptome.
- ICR = $\dfrac{\text{number of bases in the assembly}}{\text{number of bases in the set of reads}}$.

Experiment 2 - Correlation - Results



Experiment 2 - Correlation - Results

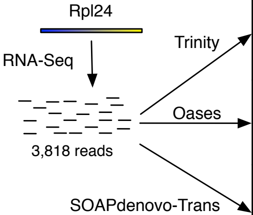


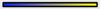
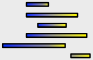
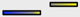
Thanks

Software:

- DETONATE: http://deweylab.biostat.wisc.edu/detonate/

Extra slide - RSEM-EVAL Example



Score Assembly	Likelihood score	Prior score	BIC score	RSEM-EVAL score
	−86542	−876	−8	−87426
	−198666	−2771	−29	−201466
	−254310	−613	−12	−254935

Extra slide - Runtime comparison

	Assembly T		Assembly O		Assembly S	
Program	Runtime	Memory	Runtime	Memory	Runtime	Memory
RSEM-EVAL*	1h 4m 57s	2.02 GB	4h 40m 36s	8.18 GB	34m 57s	1.23 GB
Genovo	6d 11h 54m 3s	192.23 GB	> 1 week	–	4d 15h 3m 3s	188.79 GB
ALE*	12h 39m 36s	0.67 GB	6d 23h 23m 13s	2.31 GB	7h 33m 1s	0.59 GB
REF-EVAL, contig**	3s	0.19 GB	8s	0.33 GB	2s	0.2 GB
REF-EVAL, nucleotide**	8s	0.39 GB	33s	1.27 GB	6s	0.33 GB
REF-EVAL, KC score	1m 18s	2.09 GB	1m 30s	2.37 GB	1m 13s	2.03 GB
Bowtie	15m 42s	0.11 GB	1h 1m 38s	0.31 GB	11m 16s	0.1 GB
Blat	35m 14s	0.0 GB	1h 51m 1s	0.01 GB	28m 19s	0.0 GB

* Plus time to run Bowtie. We calculate Bowtie statistics separately because ALE takes Bowtie alignments as input.

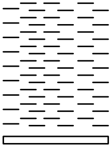
** Plus time to run Blat.

Extra slide - Correlation comparison

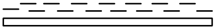
	KC Score	Contig F1	Nucleotide F1
RSEM-EVAL Score	0.99	0.83	0.46
Genovo Score	0.96	0.80	0.53
ALE Score	0.64	0.45	0.62
N50	0.22	0.33	-0.31
Number of Nucleotides in Assembly	0.13	0.29	-0.21
Number of Unique Proteins Matched	0.68	0.81	0.73
Average Ortholog Hit Ratio	0.31	0.31	-0.19

Table 1 The Spearman rank correlation coefficient of the scores assigned by several alternative transcriptome assembly evaluation measures, described in the main text, to the reference-based scores from REF-EVAL. The evaluated assemblies were produced by Trinity, Oases, SOAPdenovo-Trans, and Trans-ABySS, based on the subset of reads in the real (strand non-specific) mouse data that align to genes on chromosome 1. This subset was used in the interest of computational efficiency of the alternative measures.

Extra slide - Comparison on an example



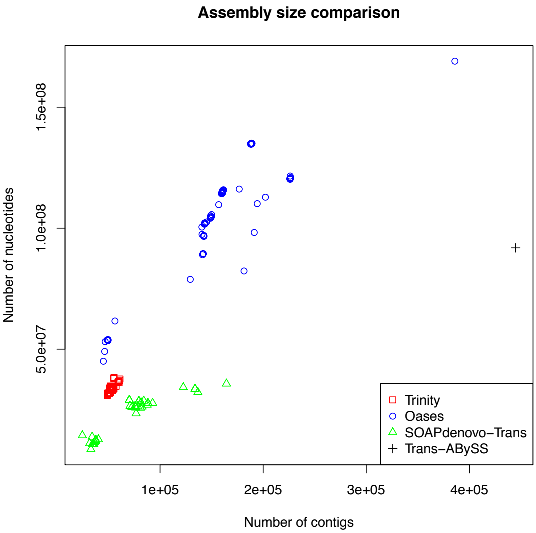
Isoform 1 (90% relative abundance)



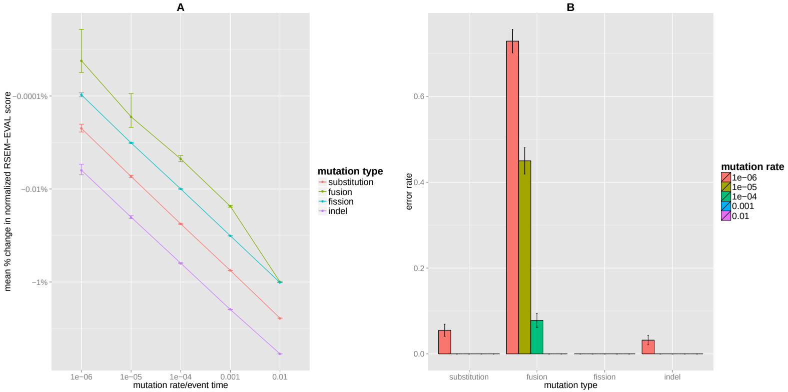
Isoform 2 (10% relative abundance)

Score / Assembly	RSEM-EVAL	GENOVO	ALE
Truth	−43720	−19557	−116316
Long only	−44403	−18199	−88905
Short only	−104963	−68997	−52090

Extra slide - Different parameters give very different assemblies



Extra slide - Experiment 1 - Random perturbation - Results



Extra slide - The assembly prior $P(\text{assembly}|\lambda)$

The prior distribution is specified as follows:

- Transcript lengths follow a negative binomial distribution, iid.

Extra slide - The assembly prior $P(\text{assembly}|\lambda)$

The prior distribution is specified as follows:

- Transcript lengths follow a negative binomial distribution, iid.
- Given the transcript lengths:
  - Transcript sequences follow a uniform distribution, iid.

Extra slide - The assembly prior $P(\text{assembly}|\lambda)$

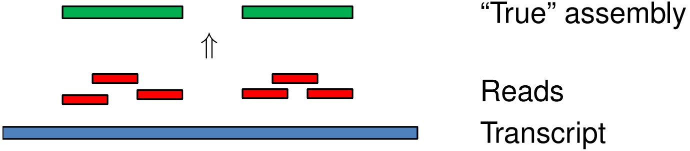
The prior distribution is specified as follows:

- Transcript lengths follow a negative binomial distribution, iid.
- Given the transcript lengths:
  - Transcript sequences follow a uniform distribution, iid.
  - The number of reads starting at each position of a transcript follows a Poisson distribution (mean = coverage), iid.

Extra slide - The assembly prior $P(\text{assembly}|\lambda)$

The prior distribution is specified as follows:

- Transcript lengths follow a negative binomial distribution, iid.
- Given the transcript lengths:
  - Transcript sequences follow a uniform distribution, iid.
  - The number of reads starting at each position of a transcript follows a Poisson distribution (mean = coverage), iid.
- The "true" assembly is formed by joining reads whose true positions (within the transcript set) overlap or are contiguous.



- Based on the above, one can work out a recurrence for the prior probability of the assembly.