# Evaluation of *de novo* Transcriptome Assemblies from RNA-Seq Data

with Bo Li and Colin Dewey

CIBM Seminar, February 12, 2012

# Background - The Transcriptome

Definition: The collection of RNA molecules in a cell or sample.

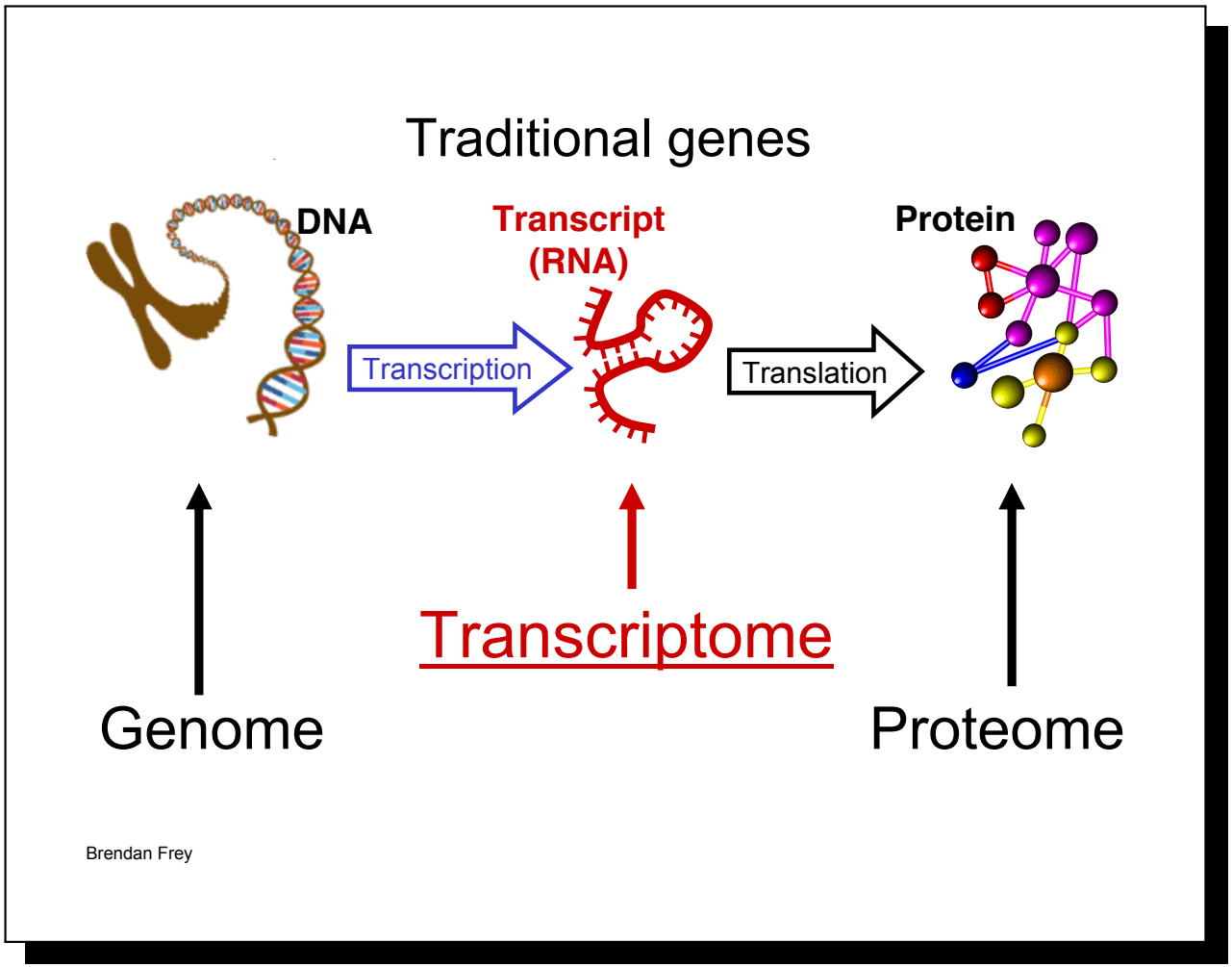



Figure from Brendan Frey, http://www.psi.toronto.edu/isit2006/.

# Background - RNA-Seq

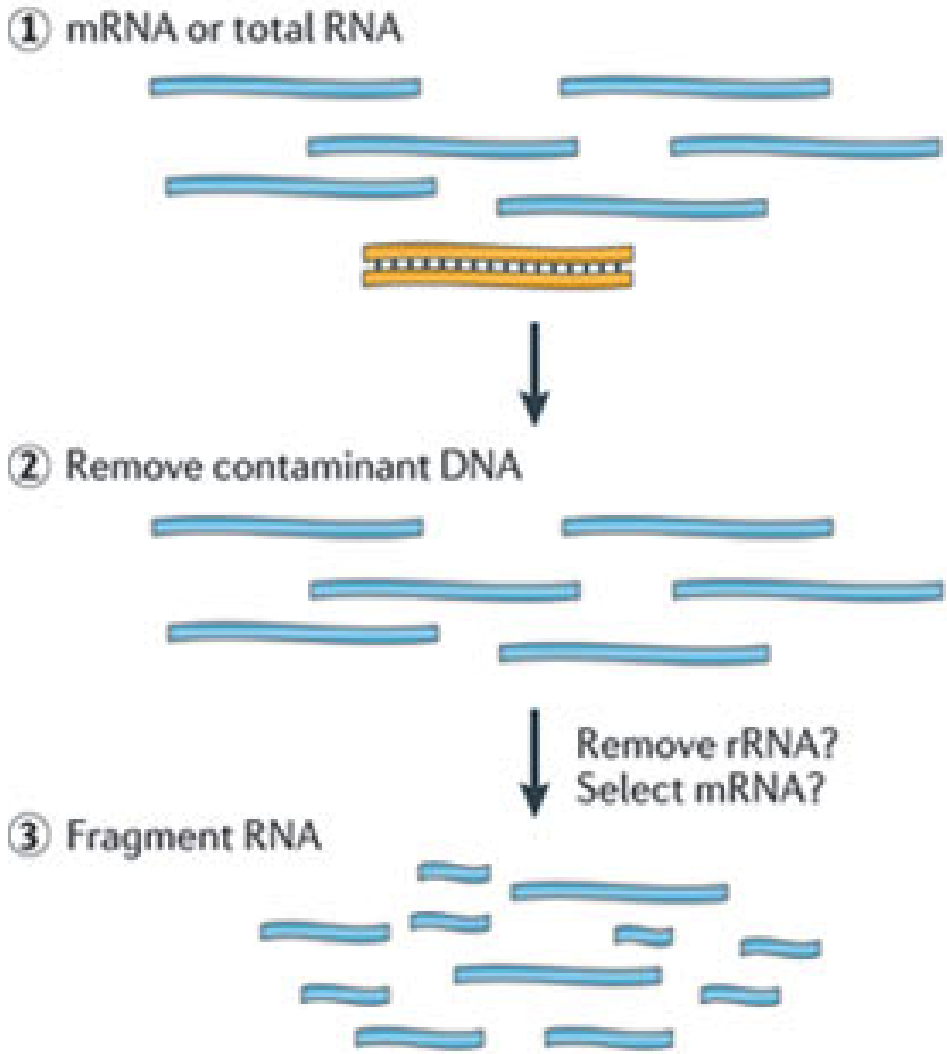



Figure from Martin and Wang, “Next-generation transcriptome assembly”, Nature Reviews Genetics, 2011.

# Background - RNA-Seq

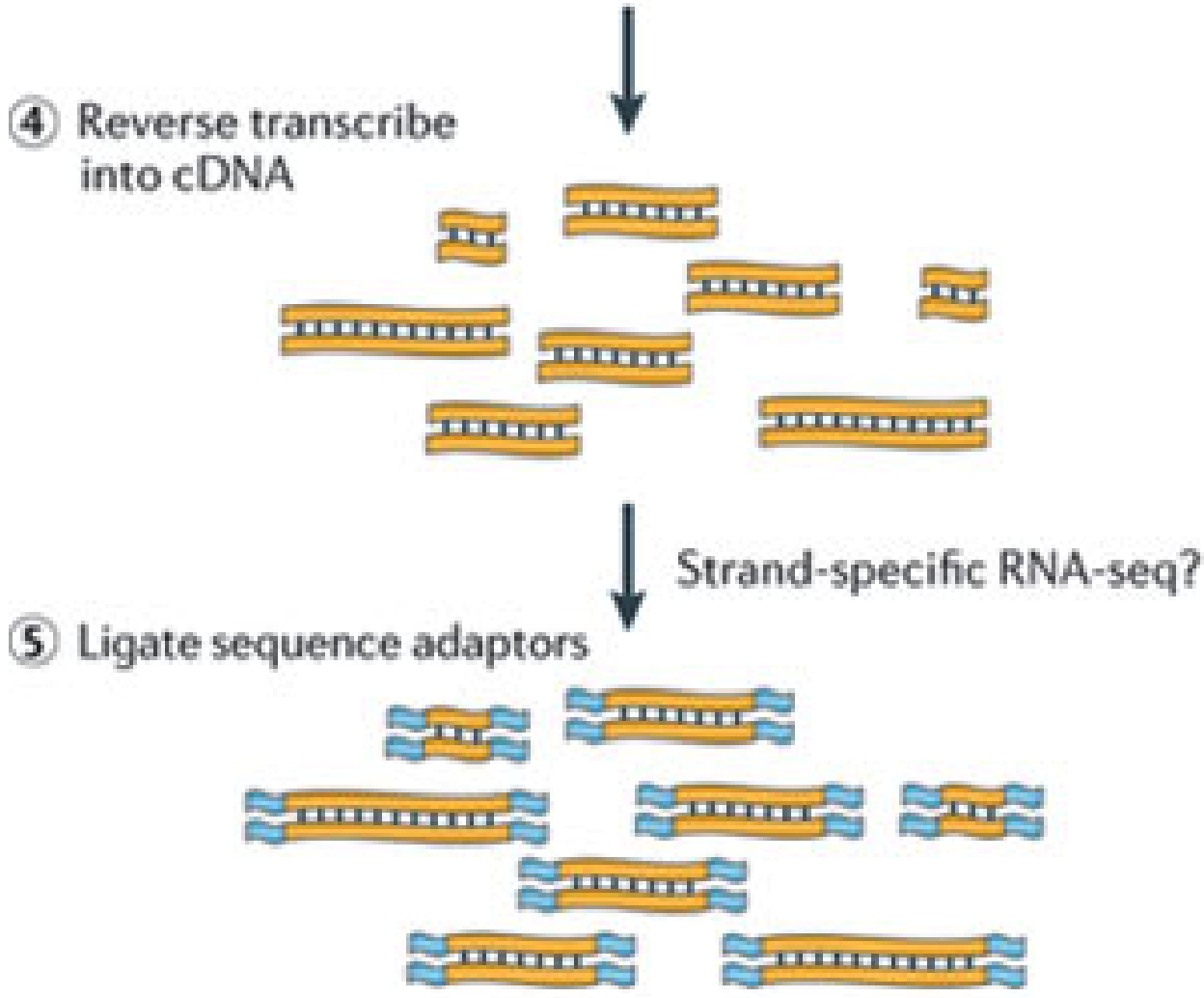



Figure from Martin and Wang, “Next-generation transcriptome assembly”, Nature Reviews Genetics, 2011.

# Background - RNA-Seq

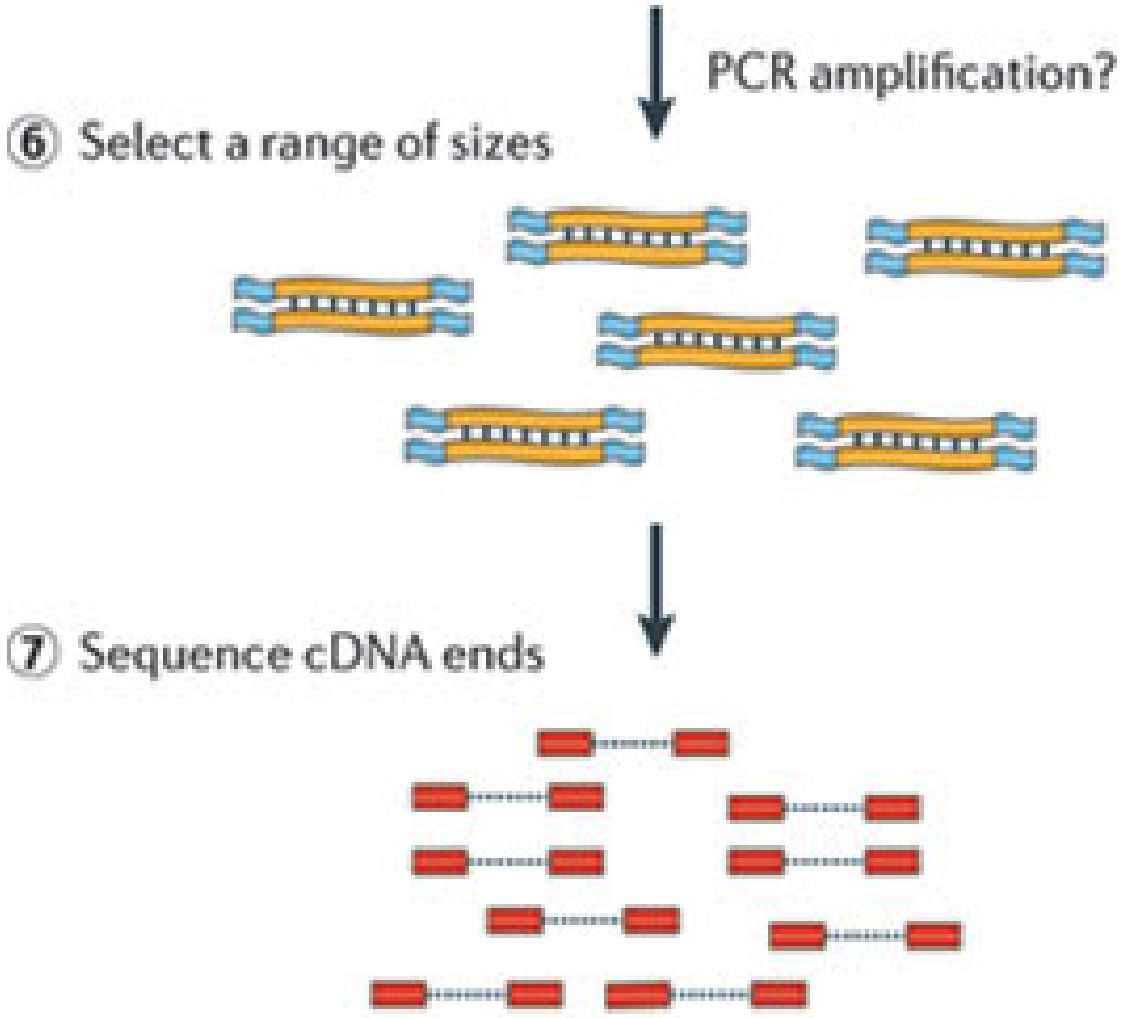
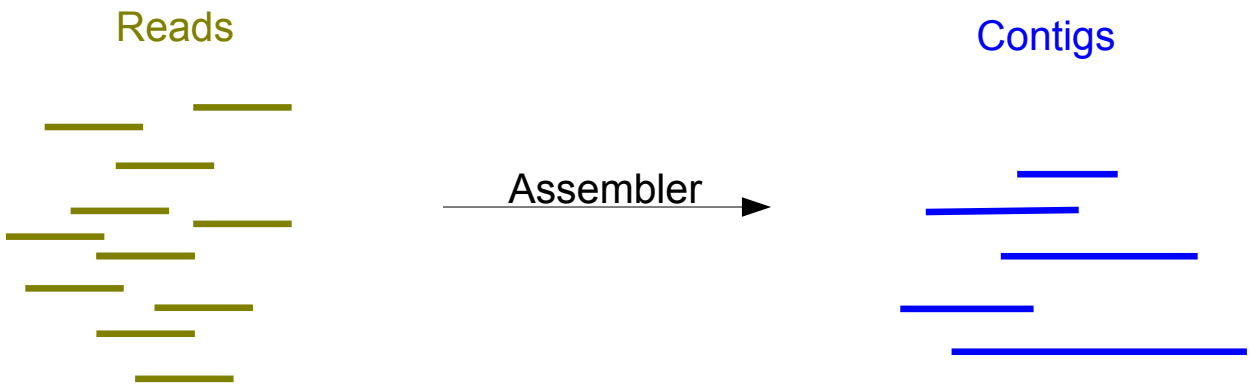


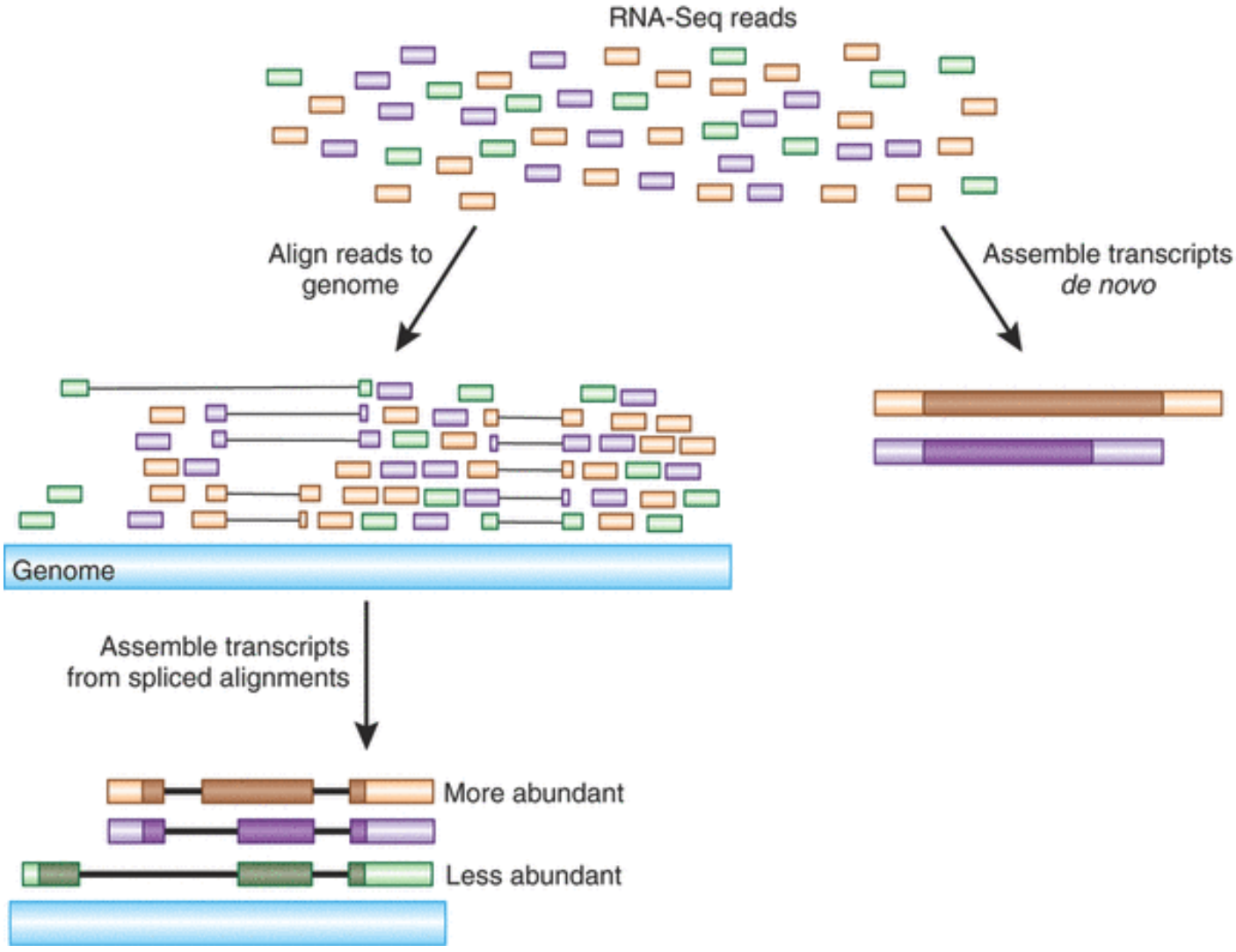

Figure from Martin and Wang, “Next-generation transcriptome assembly”, Nature Reviews Genetics, 2011.

# Background - Assembly




A contig is a continguous subsequence of a transcript sequence.

# Background - Assembly




Edited figure from Haas and Zody, “Advancing RNA-Seq analysis”, Nature Biotechnology, 2010.

# Background - Complications

- Non-uniform expression.
- Alternative splicing.

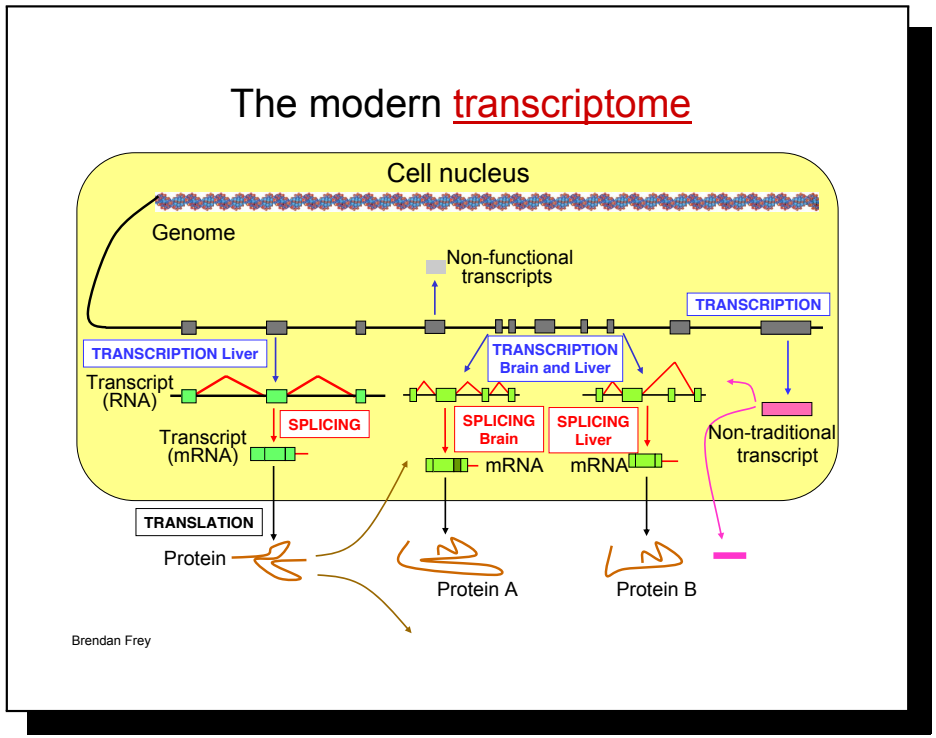



Figure from Brendan Frey, http://www.psi.toronto.edu/isit2006/.

# Background - *de novo* Assembly

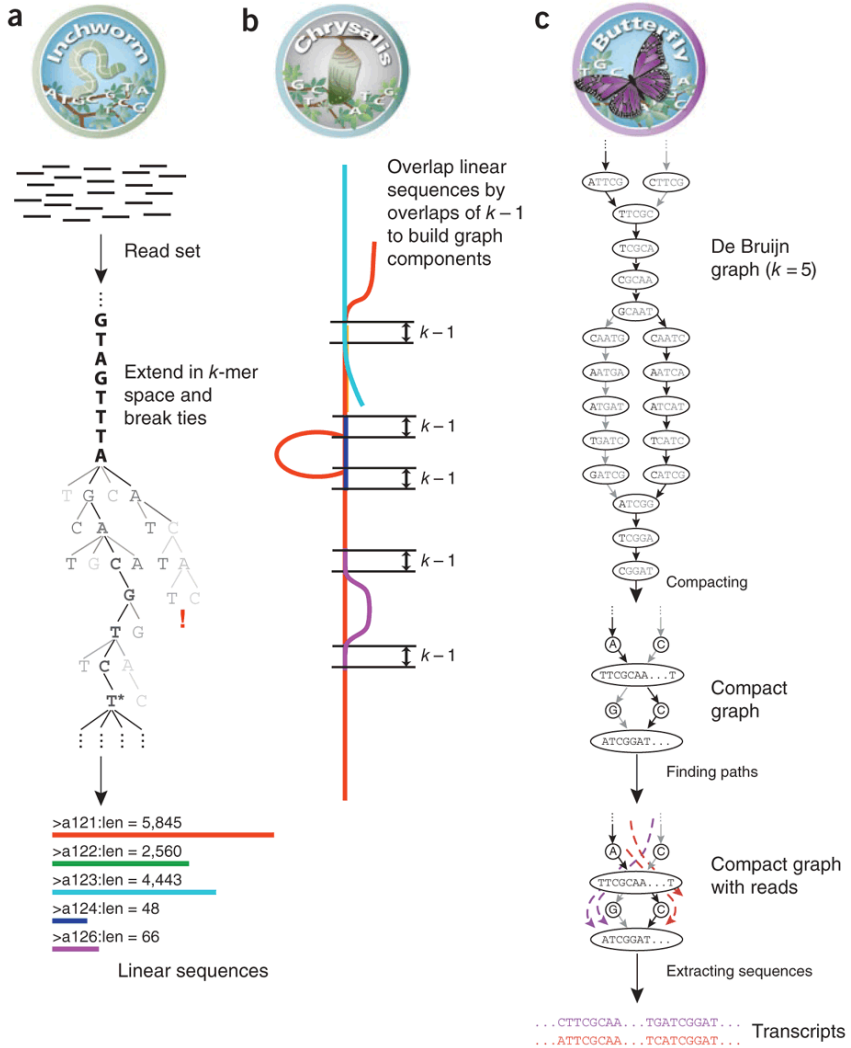



Figure from Grabherr et al., “Full-length transcriptome assembly from RNA-Seq data without a reference genome”, Nature Biotechnology,

# Evaluation Problem

Without the ground truth reference transcript set,

determine which assembly is best

based only on the RNA-Seq data

from which the assemblies were constructed.

# Evaluation Problem - Desiderata

- Start from first principles.
- Avoid trivialalities.
- Achieve the same ordering as a simple reference-based score.

Non-solution: N50, the largest $n$ such that the contigs with length $\geq n$ compose at least 50% of the total bases of the contigs set.

# Our Contributions

- A score that satisfies the given desiderata.
- A reference-based precision/recall framework for transcriptome assembly.
- A software package, DETONATE, that implements the above.
- A comprehensive meta-evaluation of the score.

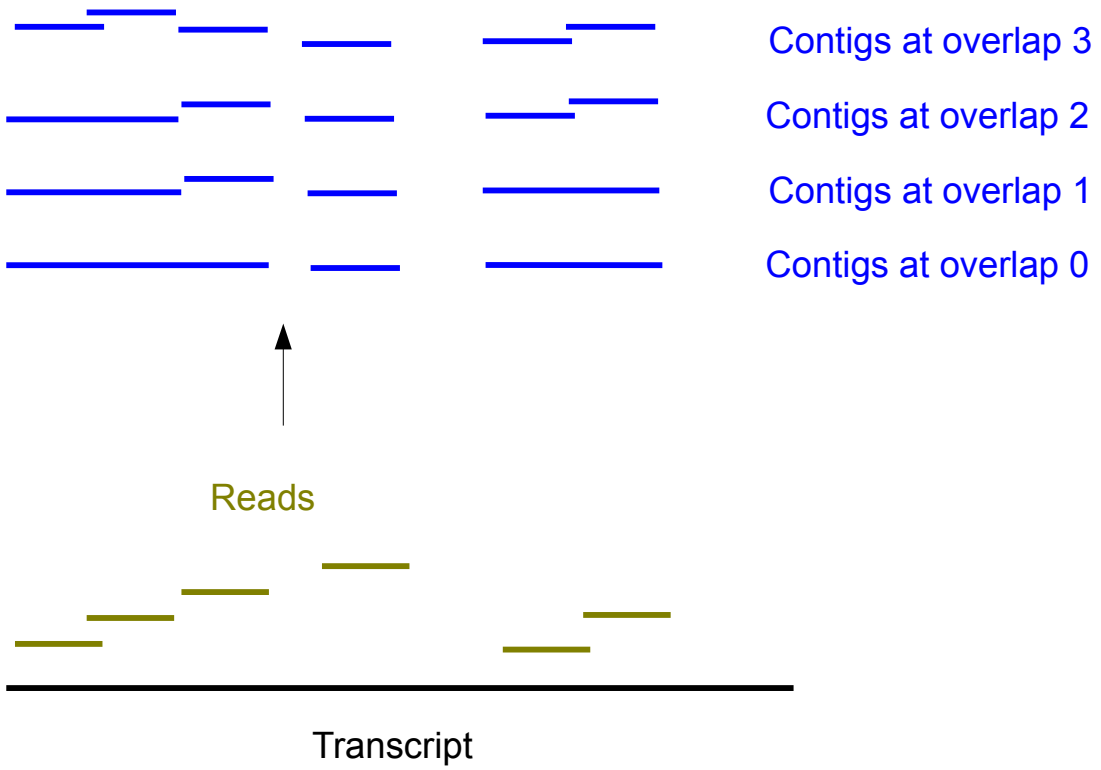
# The Score Is Based on a Probability Model

Our score: $P(\text{assembly}, \text{reads}) \propto P(\text{assembly}|\text{reads})$.

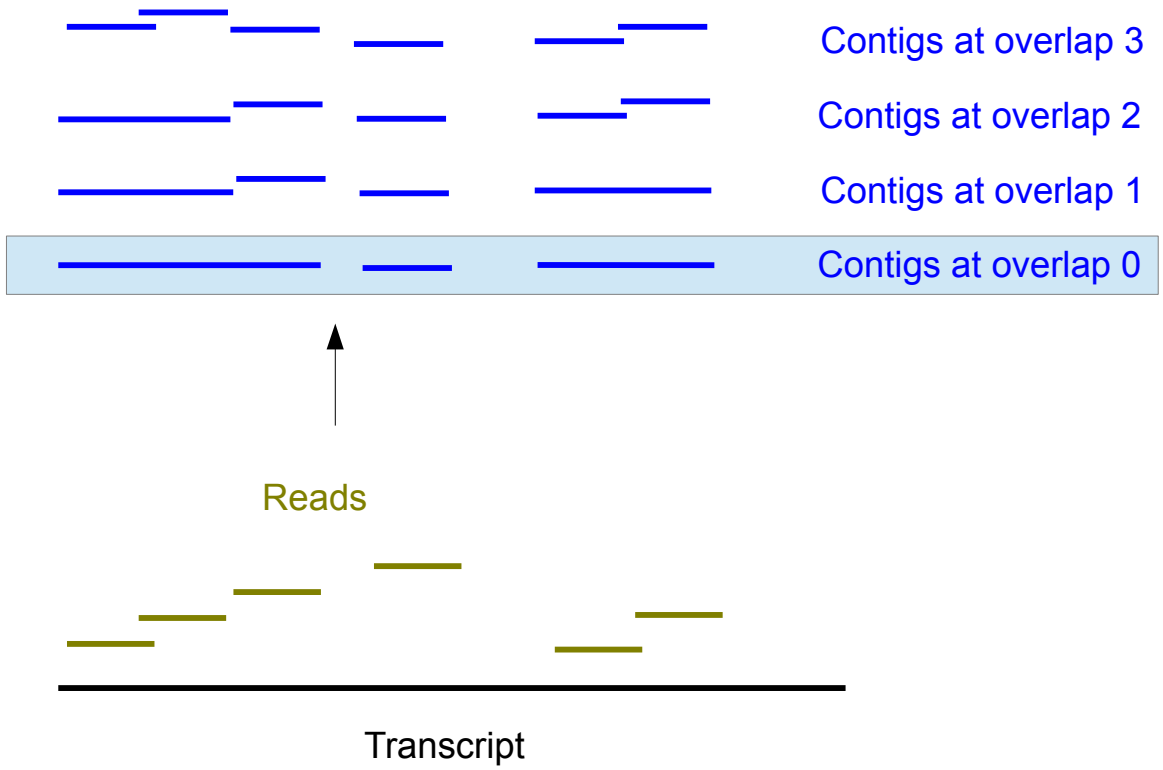
$$
\begin{aligned}
&P(\text{assembly}, \text{reads}) \\
&= \int P(\text{assembly}, \text{coverage}, \text{reads})\, d\text{coverage} \\
&= \int \underbrace{P(\text{assembly}, \text{coverage})}_{\text{prior}} \underbrace{P(\text{reads}|\text{assembly}, \text{coverage})}_{\text{likelihood}}\, d\text{coverage}
\end{aligned}
$$

A contig's "coverage" is the expected number of reads generated from each position of the contig's original transcript.

# The Probability Model Is Based on Ideal Assembly



# The Probability Model Is Based on Ideal Assembly



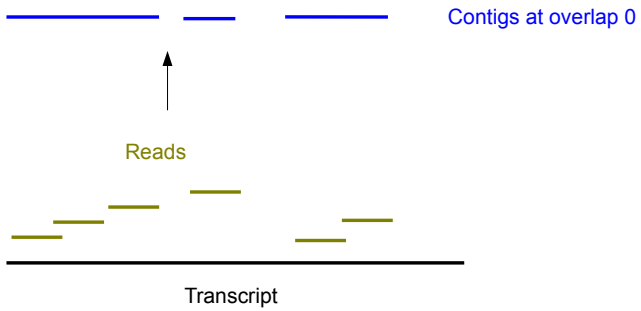
# The Probability Model - Prior

$$
\begin{aligned}
& P(\text{assembly}, \text{reads}) \\
&= \int P(\text{assembly}, \text{coverage}, \text{reads})\, d\text{coverage} \\
&= \int \underbrace{P(\text{assembly}, \text{coverage})}_{\text{prior}} \underbrace{P(\text{reads} | \text{assembly}, \text{coverage})}_{\text{likelihood}}\, d\text{coverage}
\end{aligned}
$$

# The Probability Model - Prior

Generative story:

- Transcript lengths $\sim_{iid}$ negative binomial.
- Given the transcript lengths:
  - Transcript sequences $\sim_{iid}$ uniform.
  - Number of reads starting at each position of a transcript $\sim_{iid}$ Poisson (mean = coverage).
- The assembly is formed from the reads at overlap 0.



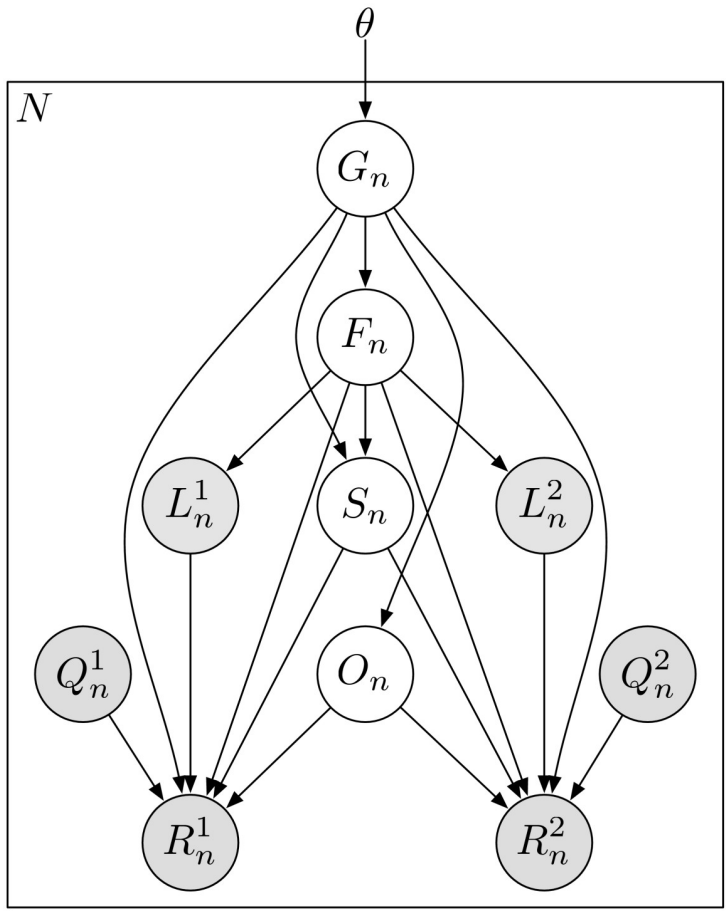

One can work out a recurrence for the prior probability of the assembly and coverage.

# The Probability Model - Likelihood

$$
\begin{aligned}
& P(\text{assembly}, \text{reads}) \\
&= \int P(\text{assembly}, \text{coverage}, \text{reads})\, d\text{coverage} \\
&= \int \underbrace{P(\text{assembly}, \text{coverage})}_{\text{prior}} \boxed{\underbrace{P(\text{reads}|\text{assembly}, \text{coverage})}_{\text{likelihood}}}\, d\text{coverage}
\end{aligned}
$$

# The Probability Model - Likelihood

Previous work, RSEM, introduced a generative model of reads, given transcripts and their expression.



# The Probability Model - Likelihood

Key observation:

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

Therefore, we stipulate:

$$P(\text{reads}|\text{assembly}, \text{coverage}) = \frac{P_{RSEM}(\text{reads}|\text{transcripts} = \text{assembly}, \text{expression} = f(\text{coverage}))}{P_{RSEM}(\text{reads cover assembly}|\text{transcripts} = \text{assembly}, \text{expression} = f(\text{coverage}))}$$

# The Probability Model - Marginalization

$$P(\text{assembly}, \text{reads})$$

$$= \int P(\text{assembly}, \text{coverage}, \text{reads}) \, d\text{coverage}$$

$$= \int \underbrace{P(\text{assembly}, \text{coverage})}_{\text{prior}} \underbrace{P(\text{reads}|\text{assembly}, \text{coverage})}_{\text{likelihood}} \, d\text{coverage}$$

# The Probability Model - Marginalization

Approximate the integral by BIC:

$$\begin{aligned}
&\log P(\text{assembly}, \text{reads}) \\
&= \log \int P(\text{assembly}, \text{coverage}, \text{reads})\, d\text{coverage} \\
&= \log P(\text{assembly}, \text{reads} | \text{coverage}^*) - \frac{1}{2} M \log N
\end{aligned}$$

where $M =$ number of contigs, $N =$ number of reads,
coverage$^*$ = maximum likelihood estimate.

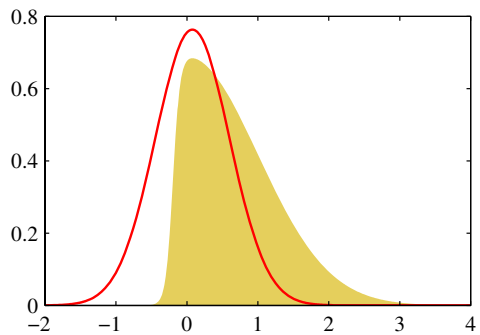
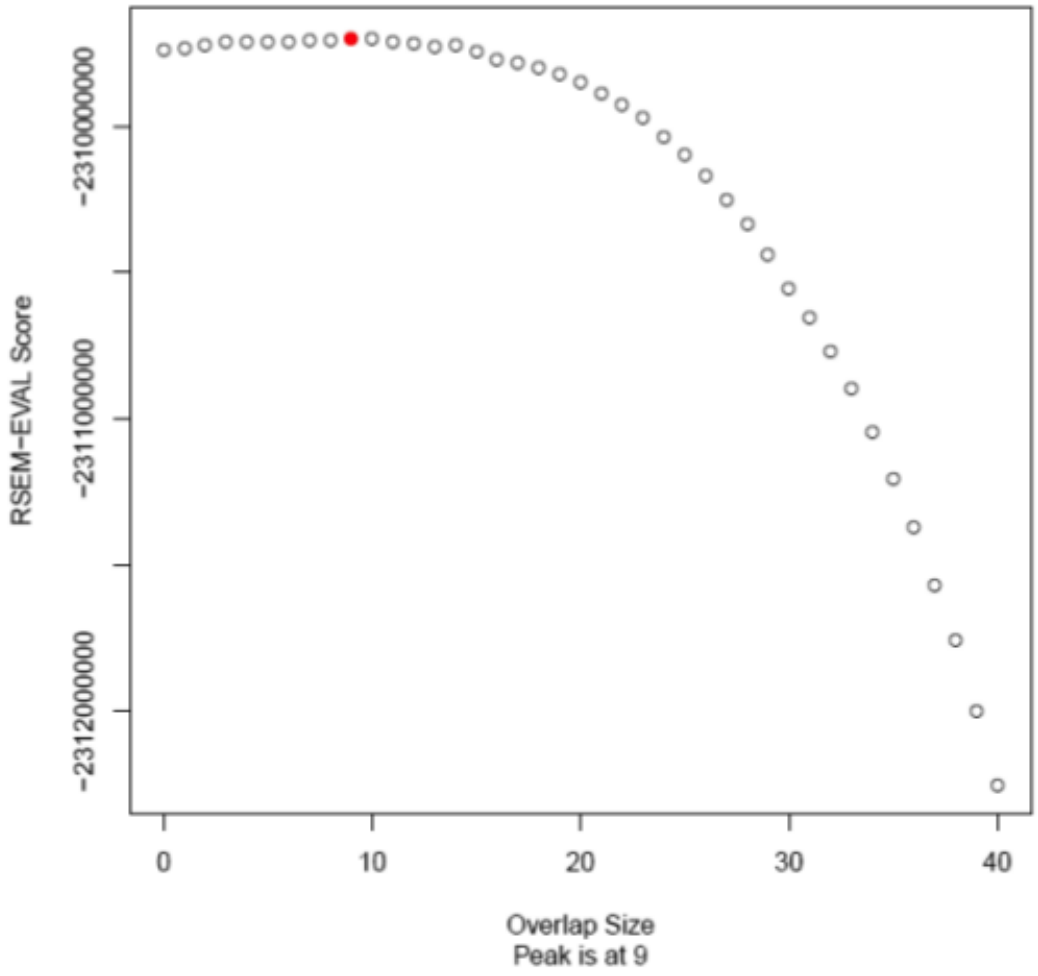



Figure from Bishop, *Pattern Recognition and Machine Learning*, Springer, 2009.

# Experiment 0 - Setup

- Goal: Make sure we have avoided trivialities.
- Procedure:
  - Construct ideal assembly at every possible overlap.
  - Compute score.
- Desired result: Best overlap is fairly close to 0.

# Experiment 0 - Results



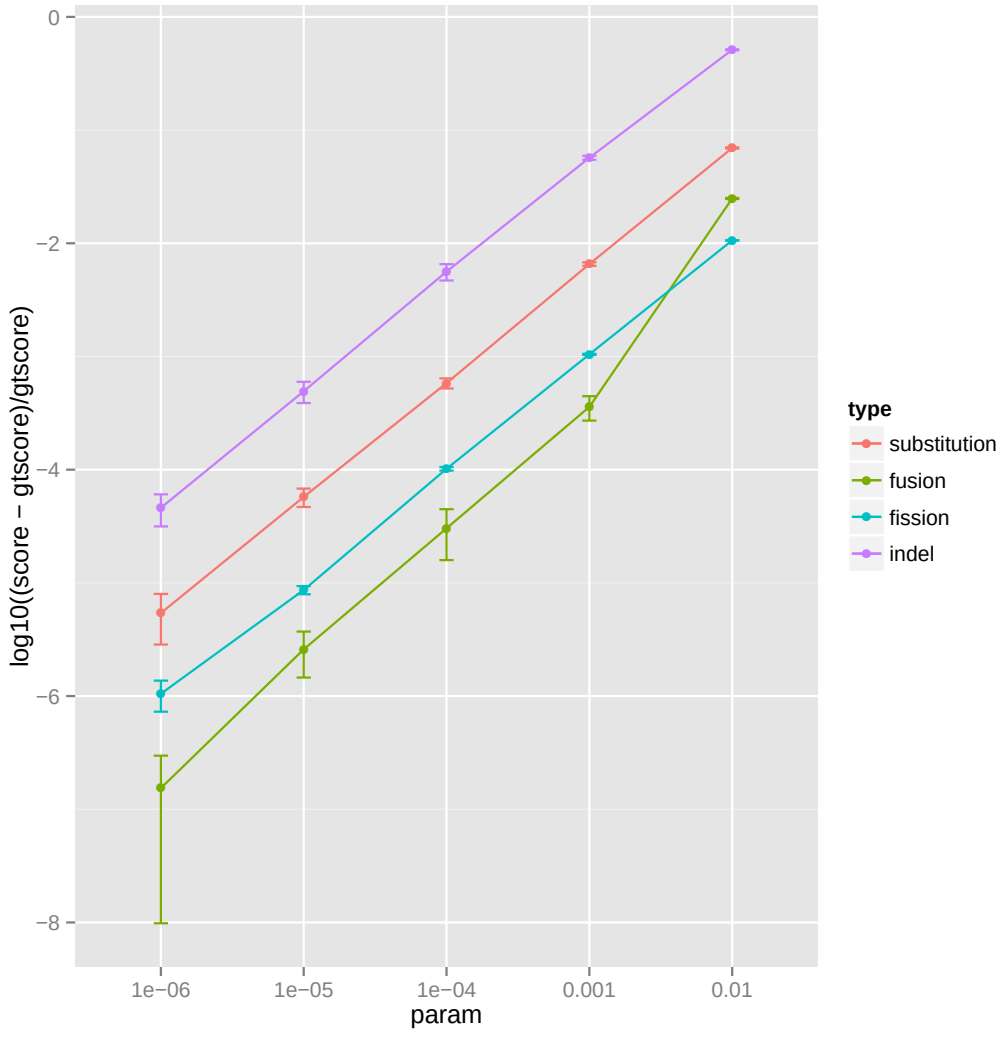
# Experiment 1 - Setup

- Goal: Make sure the true best assembly has the best score, on average.

- Procedure:
  - Construct ideal assembly at overlap 0.
  - 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 ideal and perturbed assemblies.

- Desired result: The ideal assembly has the best score.

# Experiment 1 - Results



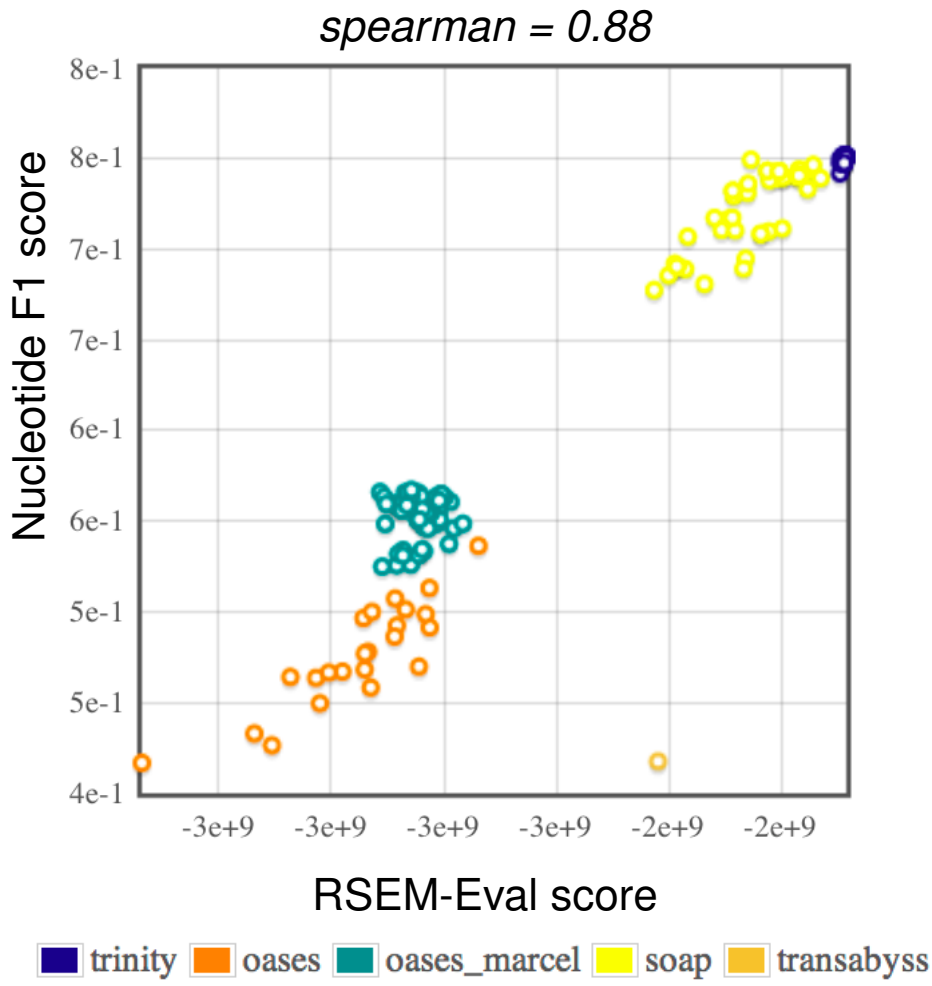
# Experiment 2 - Setup

- Goal: Study the correlation between our score and simple reference-based scores.
- Five datasets:
  - Mouse from Trinity paper.
  - Mouse from Oases paper.
  - Yeast from Trinity paper.
  - Axolotl from Thompson lab.
  - Simulated mouse.
- $\sim$100 assemblies per dataset, using:
  - Trinity.
  - Oases.
  - SOAPdenovo-trans.
  - Trans-ABySS.

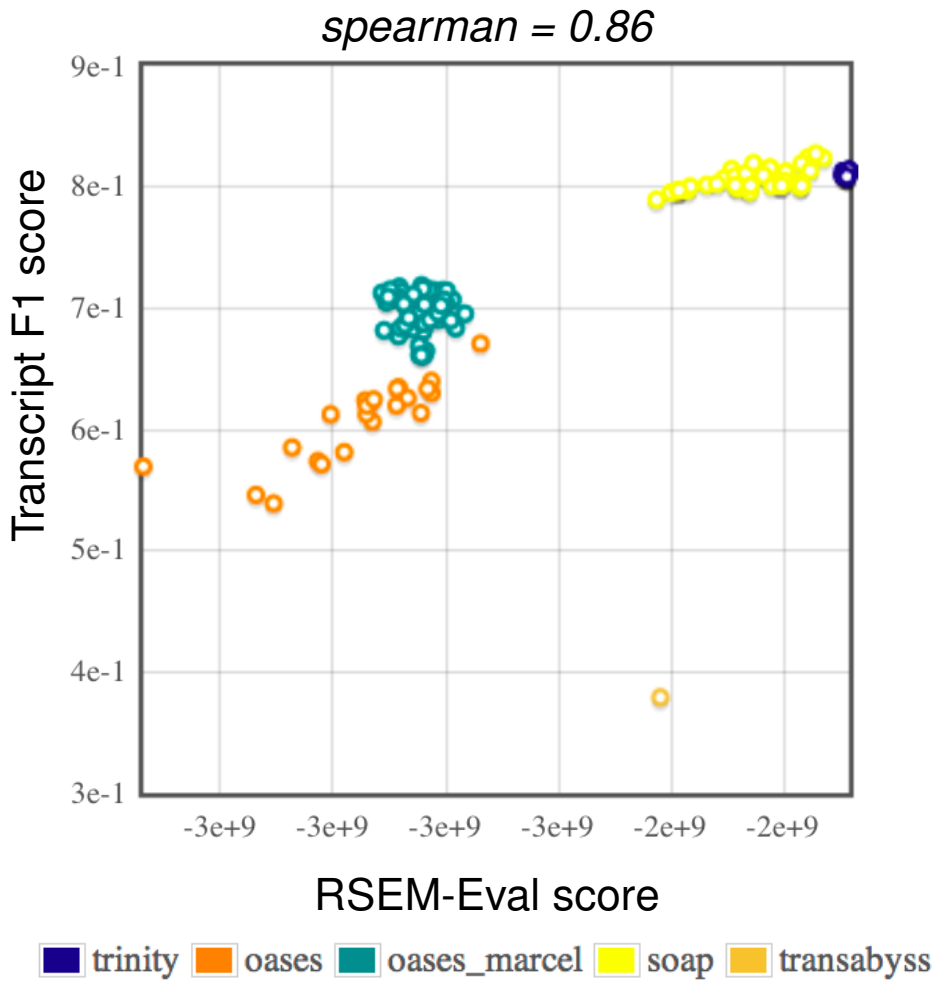
# Experiment 2 - Setup

- 3 reference-based F1 scores (harmonic mean of precision and recall):
  - Nucleotide F1.
  - Transcript F1.
  - Pair F1.
- 1 reference-based “k-mer” score:
  - Jensen-Shannon divergence between k-mer distributions.
- Procedure:
  - For each assembly: compute our *de novo* score and each reference-based score.
- Expected result:
  - Monotone relationship between the scores.

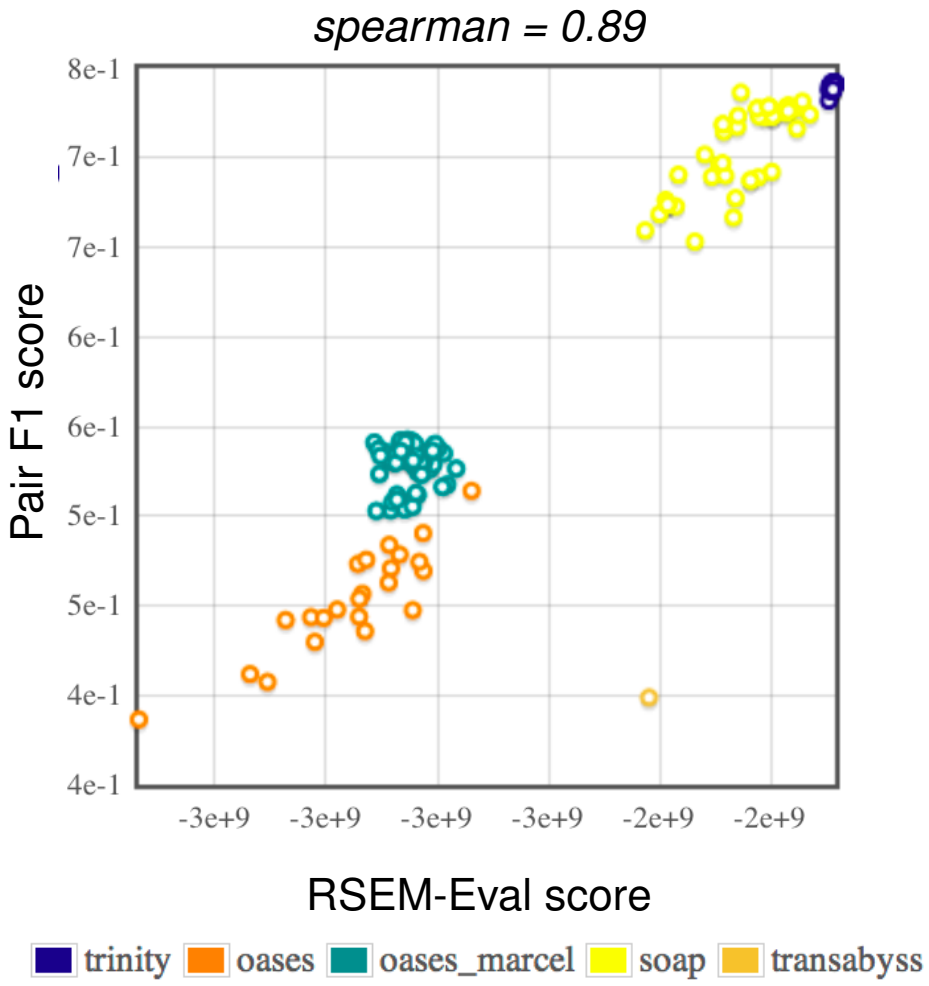
# Experiment 2 - Results



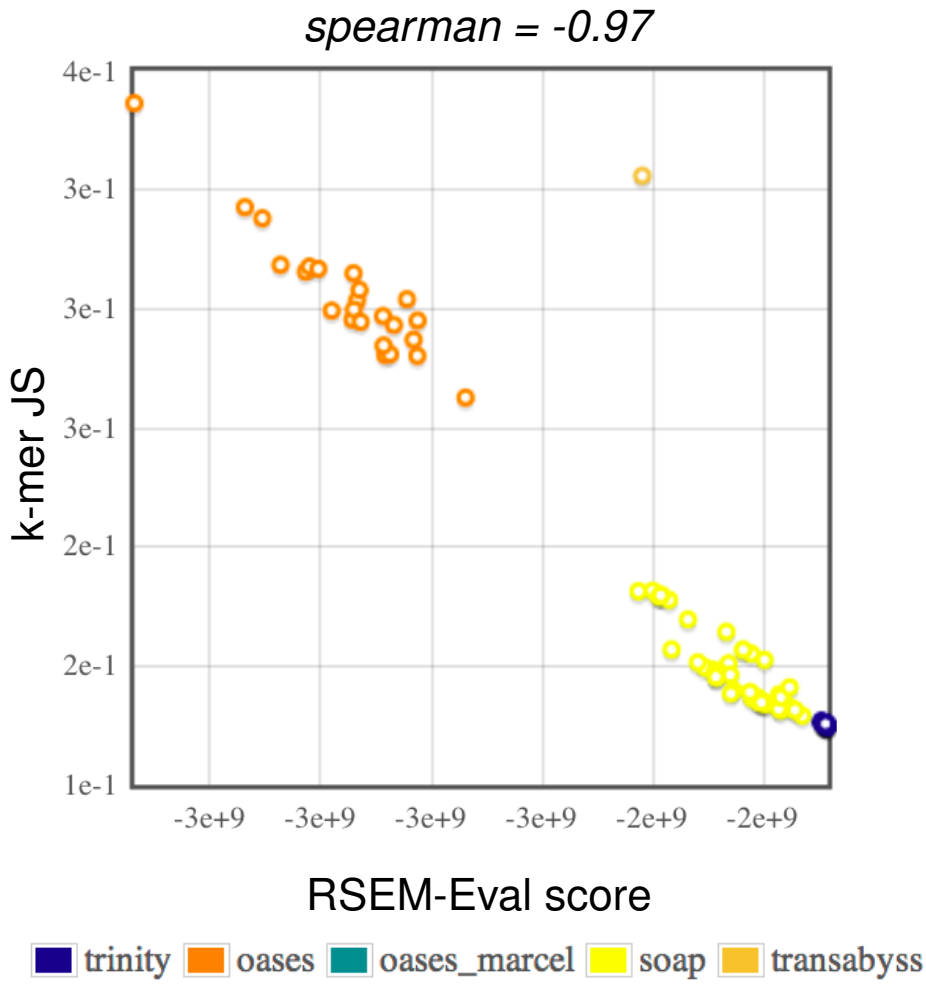
# Experiment 2 - Results



# Experiment 2 - Results



# Experiment 2 - Results



Thanks.

Nathanael Fillmore has been funded by NLM training grant 5T15LM007359. Bo Li has been funded by Morgridge Institute for Research support for Computation and Informatics in Biology and Medicine. Colin Dewey has been partially funded by NIH grant 1R01HG005232.