Accurate and efficient transcript identification and quantification using RNA-seq data

Mihaela Pertea[1,2,3,5], Geo M. Pertea[1,2], and Steven L. Salzberg[1,2,4,5,6]

[1]Center for Computational Biology, [2]McKusick-Nathans Institute of Genetic Medicine, [3]Department of Medicine, [4]Department of Biomedical Engineering, [5]Department of Computer Science, and [6]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205

Contact: mpertea@jhu.edu

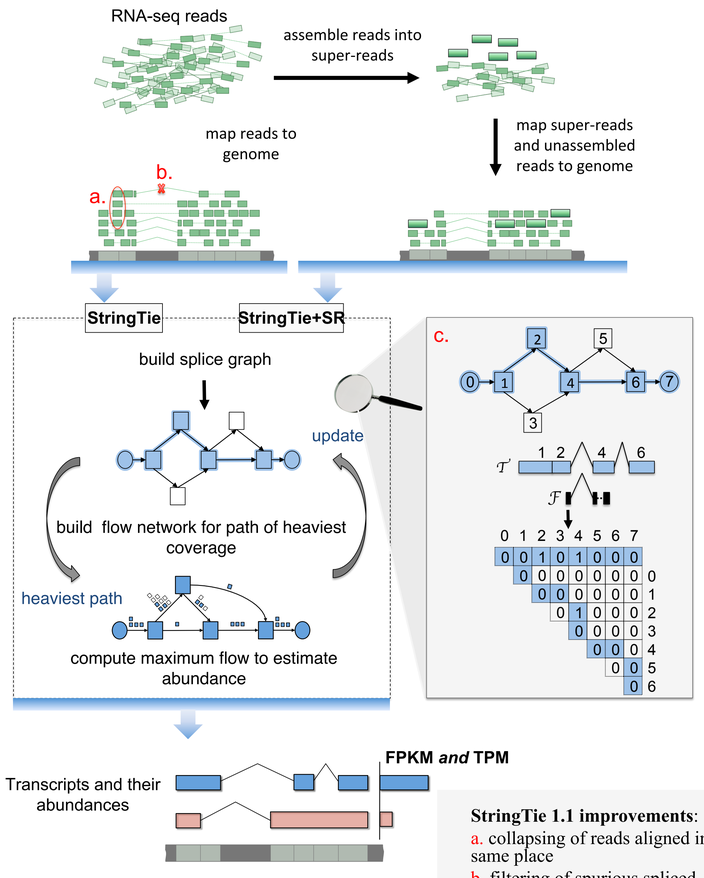
Background

Transcriptome assembly and gene expression profiling are key components in a vast range of biological experiments today, playing a central role in unraveling the complexity of cell type, cell differentiation, responses to stress, and myriad other conditions. Although transcript assemblers have been developed previously, most of them perform poorly on real, large-scale RNA-seq data sets, severely limiting their impact.

Over the last decade, multiple studies have revealed an astonishing degree of complexity in the transcriptomes of eukaryotes. First, we now know that most plant and animal protein coding genes occur in multiple splice variants, most of which are not yet annotated. Second, a significant number of transcribed elements are never translated into proteins, but instead function as non-coding RNA genes that show complex patterns of expression and regulation. These genes also are still largely unannotated. Because we still have an incomplete picture of the exon-intron structure of most transcripts, transcriptome assembly is a critical necessity for analysis of gene transcription.

The StringTie Algorithm

Our method – StringTie - is the first transcript assembler that uses an optimization technique known as maximum flow in a specially-constructed flow network to determine gene expression levels, and it does this at the same time as it is assembling each splice variant of a gene. It is also the first genome-guided transcript assembler to incorporate techniques from whole-genome assembly, which has the potential to dramatically improve our ability to resolve alternative splice variants.

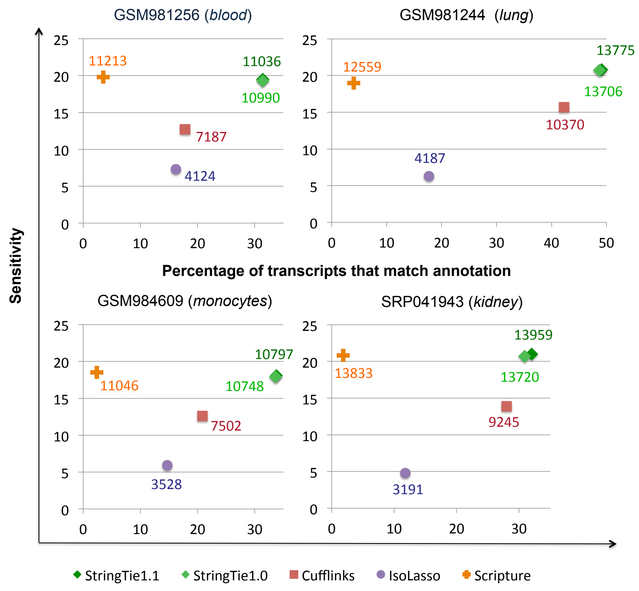


StringTie 1.1 improvements:
a. collapsing of reads aligned in the same place
b. filtering of spurious spliced alignments within large bundles
c. better representation of internal data structures

Transcriptome Assembly Accuracy

Our results on both simulated and real data demonstrate that, as compared with other leading transcript assembly programs, StringTie produces more complete and more accurate reconstructions of genes and better estimates of expression levels.

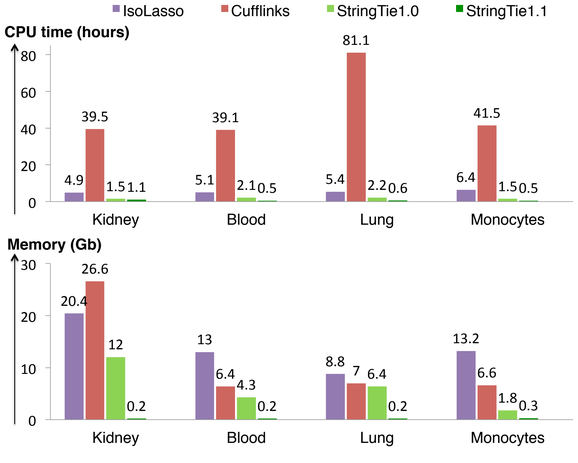
Our real data includes three human RNA-seq data sets from the ENCODE project, all of them strand-specific; and one unstranded RNA-seq data set that we generated for this study from a human kidney cell line.



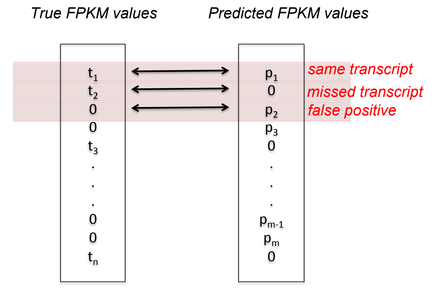
Accuracy of various transcript assemblers at assembling known transcripts, measured on real data sets from four different tissues. Transcript sensitivity (y-axis) measures the percentage of known transcripts that were correctly assembled. Note that many isoforms are not expressed in a given tissue; thus maximum sensitivity may be very low. The x-axis shows the percentage of all predicted transcripts that match an annotated transcript. Labels next to data points represent the number of correctly predicted transcripts.

Speed and Memory Efficiency

StringTie is much faster and more memory efficient than other programs, and StringTie 1.1 provides another 10-100 fold improvement in memory usage.



Transcript Quantification

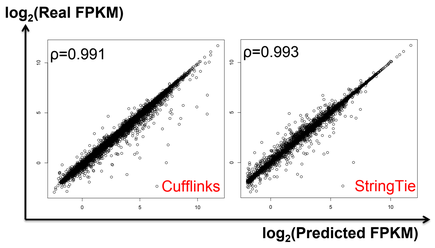


If a predicted transcript P fails to match any true expressed transcript, we match P with a transcript that has an expression level of zero. If a true transcript T is not covered (even in part) by any prediction, we match T with a prediction that has an expression level of zero. If multiple predicted transcripts (transfrags) are contained within a single true transcript, we sum all the reads assigned to the predicted transfrags and correlate this sum with the expression level of the true transcript.

We compute the Spearman correlation coefficient between the true and estimated expression levels for each set of transcripts. Specifically, we compare the expression level of each predicted transcript with the true transcript that it matches.

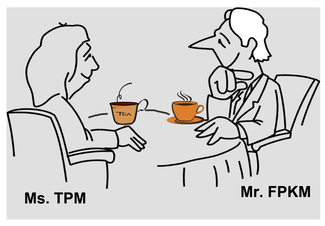
Measure	StringTie 1.1	Cufflinks 2.2.1
ρ_{all}	0.789	0.720
$\rho_{predicted}$	0.905	0.883

StringTie and Cufflinks quantification performances on a 150 million 75-bp paired-end reads simulated data set. Results show the Spearman correlation coefficient between the real and predicted number of reads (measured by FPKM values). Rows labeled "ρ_{all}" include all true and predicted transcripts. Rows labeled "$\rho_{predicted}$" include all predictions but exclude true transcripts that were not predicted by a given program.



Correlation between real (y-axis) and predicted (x-axis) expression levels on simulated data using only transcripts that were assembled correctly end-to-end by both Cufflinks 2.2.1 and StringTie 1.1. ρ represents the Spearman correlation coefficient between real and predicted FPKM values.

StringTie 1.1: now serving both FPKM and TPM!



Availability

StringTie is a free, open source software available from:

http://ccb.jhu.edu/software/stringtie

Get this poster from here:

Reference

M Pertea, GM Pertea, CM Antonescu, TC Chang, JT Mendell & SL Salzberg. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads", *Nature Biotechnology* 2015, 33 (3), 290-295.

Acknowledgements

This work was supported in part by NIH grant R01-HG006677 and NSF grant DBI-1458178.