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 number of significant transcriptomic 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.

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.

Transcript Quantification
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.

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.

Availability
StringTie is a free, open source software available from: http://ccb.jhu.edu/software/stringtie

Reference

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