Transcriptome Assembly and Quantification from Ion Torrent RNA-Seq Data

Serghei Mangul
Department of Computer Science
University of California
Los Angeles, CA 90095
Email: serghei@cs.ucla.edu

Sahar Al Seesi and Ion Mandoiu
Deptment of Computer Science &
Engineering, University of Connecticut
Storrs, CT 06269
Email: {sahar.alseesi,ion}@engr.uconn.edu

Adrian Caciula, Alex Zelikovsky
Department of Computer Science
Georgia State University
Atlanta, Georgia 30303
Email: {acaciula, alexz}@cs.gsu.edu

Dumitru Brinza
Ion Bioinformatics,
Life Technologies Corporation,
Foster City, CA 94404
Email: Dumitru.Brinza@lifetech.com

***Abstract*—We propose novel method for transcriptome reconstruction and quantitation of both known and novel transcripts from Ion Torrent RNA-Seq reads.**

Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, has become the technology of choice for performing gene expression profiling. However, reconstruction of full-length novel transcripts from RNA-Seq data remains challenging due to the short read length delivered by most existing sequencing technologies. Ion Torrent has a competitive advantage for this application due to the longer reads it generates, however, tools for novel transcript reconstruction from Ion Torrent reads are yet to be developed.

In this work, we propose novel method for transcriptome reconstruction and quantitation of both known and novel transcripts from Ion Torrent RNA-Seq reads. Our method, referred as **Ma**ximum **L**ikelihood **T**ranscriptome **A**ssembly (MALTA), incorporates maximum likelihood model for candidate transcript expression estimation. To reconstruct novel transcipts, we use a splice graph to represent alternatively spliced transcripts for every gene in a sample. A splice graph is a directed acyclic graph where each vertex in the graph represents a segment of a gene, called pseudo-exon. A pseudo-exon is a region of a gene between consecutive transcriptional or splicing events. Transcriptional events include: alternative first exon, alternative last exon. Splicing events include: exon skipping, intron retention, alternative 5 splice site(A5SS), and alternative 3 splice site (A3SS). Two pseudo-exons are connected by an edge if they are adjacent in at least one transcript (which is witnessed by at least one (spliced) read)). First we infer exon-exon junction from mapped reads, this information is used to build the splice graph. Then we enumerate all maximal paths in the splice graph using a depth-first-search algorithm. These paths correspond to putative transcripts and are the input for the MALTA algorithm.

To solve the transcriptome reconstruction problem we must select a set of putative transcripts with the highest support from the RNA-Seq reads. We formulate this problem as a transcriptional-splicing cover problem. The objective is to select the smallest set of putative transcripts that will cover all transcriptional and splicing events presented in a sample. To solve this problem we use greedy algorithm that greedily selects candidate transcripts ordered by inferred IsoEM expression levels until all transcriptional and splicing events in a sample are explained. To infer transcript expression levels we use modified version of IsoEM capable to handle read alignments with insertion and deletions which makes suitable for use with Ion Torrent RNA-Seq reads. IsoEM is an expectation maximization algorithm for expression levels estimation of alternative spliced transcripts.

We use GOG-350 RNA-Seq dataset from Ion Community to evaluate the accuracy of transcriptome assembly tools. For this study, we compare MALTA and Cufflinks. Comparison with IsoLasso was omitted due to technical problems, results were consistently incomparable to other methods. GOG-350 dataset includes 4.5 millions reads with average read length 121 bp. Reads are mapped to reference genome using TopHat2 which is able to produce splice alignment used by transcriptome assembly tools. Transcripts assembled by both methods are compared to annotated transcripts from human reference genome. MALTA and Cufflinks assemble 17378 and 15385 alternative spliced transcripts. Among those 4555(26%) and 2031(13%) match annotated transcripts.

To evaluate transcriptome quantification accuracy of the methods we tested IsoEM and Cufflinks on Ion Torrent RNA-Seq data generated from two commercially available reference RNA samples that have been well-characterized by quantitative real time PCR (qRT-PCR) as part of the MicroArray Quality Control Consortium (MAQC); namely an Ambion Human Brain Reference RNA, Catalog #6050), henceforth referred to as HBRR and a Stratagene Universal Human Reference RNA (Catalog #740000), henceforth referred to as UHRR. used in the comparison consisted of five HBRR datasets and five UHRR datasets. Accuracy was assessed by calculating the correlation (R^2) between the (qRT-PCR) and Fragment Per Kilobase of exon length per Million reads (FPKM) estimates, calculated by IsoEM. The results were compared with the FPKM values generated by Cufflinks for the same datasets. IsoEM estimates correlate better with qPCR measurements compared to Cufflinks. Additionally, IsoEM has a consistent accuracy, across datasets.

978-1-4673-1321-6/12/$31.00