RNA-Bloom: *de novo* RNA-seq assembly with Bloom filters

Ka Ming Nip[1,2], Justin Chu[1,2], Inanç Birol[1,3]
[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada
[2]Bioinformatics Graduate Program, the University of British Columbia, Vancouver, BC, Canada
[3]Department of Medical Genetics, the University of British Columbia, Vancouver, BC, Canada

RNA-seq is primarily used in measuring gene expression, quantification of transcript abundance, and building reference transcriptomes. Without bias from a reference sequence, *de novo* RNA-seq assembly is particularly useful for building new reference transcriptomes, detecting fusion genes, and discovering novel transcripts. A number of approaches for de novo RNA-seq assembly were developed over the past six years, including Trans-ABySS, Trinity, Oases, IDBA-tran, and SOAPdenovo-Trans. Using 12 CPUs, it takes approximately a day to assemble a human RNA-seq sample and require up to 100GB of memory. While the high memory usage may be alleviated by distributed computing, access to a high-performance computing environment is a strict requirement for RNA-seq assembly.

Here, we present a novel de novo RNA-seq assembler, "RNA-Bloom," that utilizes Bloom filter-based data structures for compact storage of k-mer counts and the de Bruijn graph of two k-mer sizes in memory. Compared to existing approaches, RNA-Bloom can assemble a human RNA-seq sample with comparable accuracy using merely 10GB of memory, which is readily available on modern desktop computers. The de Bruijn graph of two k-mer sizes allows RNA-Bloom to effectively assemble both lowly and highly expressed transcripts. In addition, RNA-Bloom can assemble and quantify transcript isoforms without alignment of sequence reads, thus resulting in a quicker run-time than existing alignment-based protocols.