Comparison of RNA-Seq Aligners for Low-Abundance Transcripts

Theo Nelson (tmn2126@columbia.edu)

Mentor: Thomas Postler, PhD – Columbia Department of Microbiology and Immunology

Abstract: The characterization of low-expression genes by RNA sequencing (RNA-seq) is becoming increasingly important as the cost of next-generation sequencing continues to decrease. The most complex processing step within standard analysis of RNA-seq data involves the accurate alignment of reads to a reference genome. Numerous approaches and programs have been developed to accomplish this task. Ten are evaluated here. The aligners were examined on the basis of speed, read mapping ability, variability and impact on downstream analysis. A clear hierarchy emerges in terms of both speed and read mapping ability. There is significant variability between all of the different alignment programs especially among the set of low-expression genes. The two pseudo-aligners examined within the comparison, Kallisto and Salmon, display the greatest divergences from the other aligners among low-expression genes. The variability in alignment results in significant effects on downstream analysis, especially Differential Gene Expression analysis. Caution should be exercised when interpreting the results of individual low-expression genes from an RNA-seq experiment.

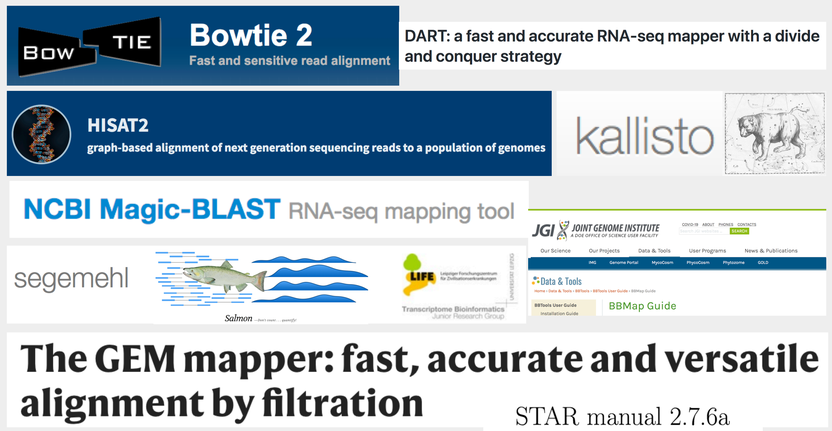

Figure 1: Compilation of the Alignment Programs Evaluated: BBMap, BowTie2, Dart, GemMapper3, HiSat2, Kallisto, MagicBlast, Salmon, Segemehl and Star. The aligners were selected based on factors such as algorithmic distinctness, hardware compatibility, novelty and popularity. They sample large classes of aligners including splice-aware aligners, gapped aligners, pseudo-aligners, short-read aligners and long-read aligners.

Methods: Bulk RNA-seq samples were downloaded from the NCBI GEO database accession number GSE158983 utilizing sra-tools version 2.10.1 (Edgar, 2002; Howard et al., 2021; Leinonen et al., 2011). All timed processes were run on a single Late 2013 Mac Pro with a 2.7 GHz 12-Core Intel Xeon E5 processor and 128 GB 1066 MHz DDR3 memory. Where possible multithreading was enabled, giving the programs access to the 24 available threads. Time estimates were made with the UNIX "time" command which is placed directly in front of the relevant command. Fastqc version 0.11.9 was utilized to perform comprehensive pre-processing QC analysis on the samples (Simon Andrews, 2010). fastp version 0.20.1 was run to trim adapters from the samples (Chen et al., 2018). Alfred version 0.2.1 was used to measure post-alignment quality metrics (Rausch et al., 2019). Lastly, multiqc version 1.10 was run in order to compile the generated quality reports and output from the aligners regarding performance (Ewels et al., 2016). BBMap version 38.87, BowTie2 version 2.4.2, Dart version 1.4.6, GemMapper3 version 3.6.1, HiSat2 version 2.2.1, Kallisto version 0.44.0, MagicBlast version 1.5.0, Salmon version 1.4.0, Segemehl version 0.2.0 and Star version 2.7.5c were run consecutively on all six samples (Boratyn et al., 2019; Bray et al., 2016; Brian Bushnell, 2014; Dobin et al., 2013; Kim et al., 2015; Langmead and Salzberg, 2012; Lin and Hsu, 2018; Marco-Sola et al., 2012; Otto et al., 2014; Patro et al., 2017). The featureCounts command from the subread package version 2.0.1 was utilized in order to generate count matrices from the traditional aligners (Liao et al., 2014). We summarized the transcript-level abundances from the pseudoaligners Kallisto and Salmon as gene-level counts utilizing the tximport package available in R's BiocManager (Soneson et al., 2016). This operation was performed in RStudio with R version 4.0.3 (RStudio Team, 2020). Differential Gene Expression Analysis was performed with SarTools version 1.7.3, using DESeq2 version 1.2.8 and the SERE statistic (Love et al., 2014; Schulze et al., 2012). These operations were also performed in RStudio with R version 4.0.3 (RStudio Team, 2020).

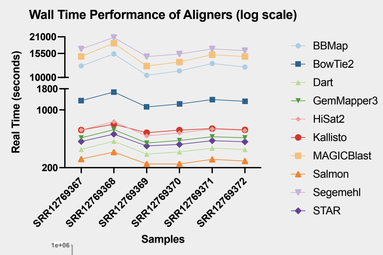

Figure 2: Wall Time Performance for the Aligners with Default Settings (split across six RNAseq samples)

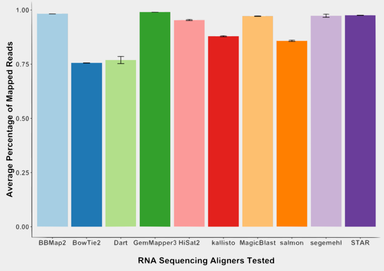

Figure 3: Percentage of Mapped Reads for the Aligners with Default Settings (averaged across RNAseq samples)

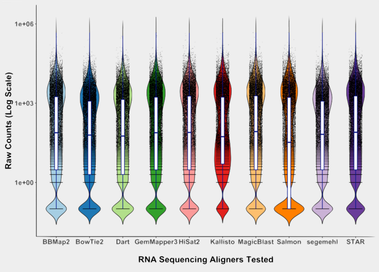

Figure 4: Distribution of Counts for Detected Genes (raw value > 0 for at least one aligner)

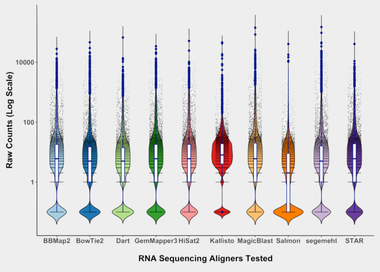

Figure 5: Distribution of Counts for Low-Expression Genes (raw 22 >= value >= 3 for at least one aligner)

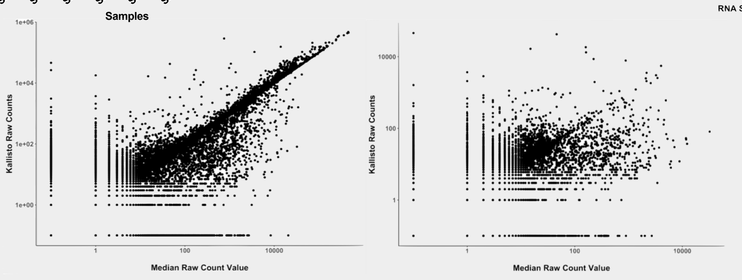

Figure 7: Alignment profiles comparing Kallisto pseudoalignment (y-axis) and the median across ten aligners (x-axis) for Detected (left) and Low-Expression (right) genes

Results: Figures 2-7 provide representative visualizations of the different statistics except for sensitivity. Figure 8 (flowchart, right) provides a summary of the results. Speed refers to wall time performance. Mapping % is defined as the number of single-end reads mapped over the total number of reads in a sample. Detection refers to the average overlap of the set of low-expression genes for a particular aligner with that of all other aligners tested. Median Deviation refers to the change in the correlation value between given aligners and the median count across aligners. Sensitivity refers to the overlap of the set of low-expression genes which are also differentially expressed for a particular aligner with that of all other aligners tested. Across all metrics, the traditional aligners STAR and HiSat2 achieved reasonable speed and mapping percentages within our hardware environment without significant deviations in terms of their detection and differential gene expression analysis of low-expression genes. Furthermore, these results reinforce the distinctness of pseudo-aligners compared to conventional aligners especially in the context of low-expression genes. These results provide a cautionary warning in the interpretation of the results of individual genes in the context of RNA sequencing experiments. Given adequate computational resources calculating expression values with multiple alignment programs can provide further confidence that results are not simply technical artifacts.