Caleb Lareau

BST 290 Advanced Computational Biology
Prepared 13 September 2016

Primary Summary–Bowtie

For the first presentation of the BST 290 class, I discussed Bowtie, one of the early aligners popularized by the explosion of next-generation sequencing data over the past decade. The tool was developed by Ben Langmead and Steven Salzberg as a fixture in a suite of tools to, in part, quantify gene expression values from RNA-Sequencing data. Other tools developed by the Salzberg lab included Cufflinks and Tophat. The development of this aligner was well-received by the community and was widely-used as next-generation sequencing data became more affordable.

Algorithmically, Bowtie uses the Burrows-Wheeler (BW) Transformation to facilitate rapid string identification and matching. Simply speaking, aligning DNA to a reference genome is simply a string matching problem. The BW Transformation was first proposed as a memoryless compression algorithm that maintains the original order of a string with only a minor computational burden. As Ben and Steven were both computer scientists, they were able to take the application of this algorithm to string compression and rededicate it to the issue of rapidly aligning next-gen sequencing reads to the genome.

One of the noteworthy discussions in class was whether RNA-Seq or microarrays was more accurate in calculating the true abundance of RNA in a cell. While RNA-Seq correlated better with qt-PCR, RNA-Seq is inaccurate in quantifying expression at low transcript abundances. This bias is a consequence of RNA-Seq pulling out transcripts that have greater expression, including transcripts that exceed lowly expressed genes by several orders of magnitude. In contrast, microarrays are specific to a particular transcript and thus can more robustly identify lowly expressed genes though the dynamic range for moderately and highly expressed transcripts is worse.

Another useful discussion was centered around the flexible alignment implementation that facilitates base mismatches. While exact string matching is relatively straightfoward and computationally friendly, inexact matching requires a much larger search space of possible alignments, necessarily leading to errors and increased computational depend. Bowtie permutes mismatched bases starting at the lower quality base calls that are indicated in the .fastq file. This provides an efficient solution but does not guarantee alignment as the permutations are capped.

Overall, bowtie and related aligners remain in the backbone of of most analyses of genomics, transcriptomics, and epigenomics data. While other aligners such as BWA, Bowtie2, etc. have been optimized for longer reads, they fundamentally work very similar to bowtie. Since I introduced so many auxiliary topics, the didn't get a chance to discuss the concept of pseudoalignment from tools like kallisto like I had wanted. These tools show great promise in changing the way alignment algorithms are conducted due to their speed and flexibility. However, the modern implementations bowtie and similar algorithms remain the gold standard (though they had notorious hiccups along the way) for sequencing alignment.

Secondary Summary–Bowtie 2 and Applications

Rather than emphasizing a particular secondary paper, I spent time discussing Bowtie2 as well as some applications. As was discussed in class, there were some notorious issues with Bowtie that were the consequence of low confidence sequencing estimates leading to indel errors. In fact, the reputation of Bowtie in early development phases was that it was somewhat inaccurate. Overall, as this algorithm became very popular, Ben and Steven began writing a second implementation of the tool that both facilitated longer reads and made the alignment more resilient to errors. Another key feature was the utilization of ambiguous bases in the reference genome, which is not supported in the first implementation of Bowtie.

The ambiguous base in the reference genome was a critical addition as it facilitates the unbiased determination of allele-specific effects. To demonstrate this, I discussed a couple of cool papers that I found that attempt to identify variants that affect transcription factor binding or that modulate the three-dimensional topology of DNA. Each of the techniques involved benefit from the unbiased reference genome that was supported in Bowtie2 but not Bowtie1. Otherwise, the discrepancies (at least those laid out in the Bowtie2 paper in Nature Methods) were somewhat underwhelming. The support for longer reads became a necessity due to the evolution of the sequencing technologies. Overall, if I had known that I would for sure be my own secondary presenter, I would have linked to spend more time on pseudoalignment and contrasting the styles for what pseudoalignment and regular alignment mean in the context of RNA-Seq data. However, the additional discussions of implementations of aligners in class (e.g. STAR) also provided some nice secondary insight into the utilization of Bowtie. However, as Bowtie and Bowtie2 have been sited a combined 12,000 times, it's extremely easy to find a study that directly used the alignment innovations that were produced by this sequencing tool.