Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis (Summary)

Next-generation sequencing (NGS) technology is used for areas pertaining to genome evolution and genomic variation among others. The generated data is aligned and mapped on the reference genome using different alignment algorithms, all of which are evaluated according to their feature, performance and accuracy hereunder.

An algorithm-based classification of multiple aligners was designed for three NGS platforms – Roche454, Illumina and ABI SOLiD – to adapt them to high-throughput data. The two main classifiers were based on the hash table-based algorithm and the Burrows-Wheeler Transform (BWT) based backtracking algorithm. The former employs the seed-and-extend strategy with k-mer indexing to improve high-throughput short reads as well as a dynamic programming algorithm (Smith-Waterman or Needleman-Wunsch) to extend the alignment; also being able to align multiple-error reads. The latter employs the prefix/suffix tree and suffix array data structures with FM-indexing for fast read searching and to solve alignment to genome copies as well as a reversible data compression algorithm (BWT) to decrease the memory usage of the previously mentioned data structures.

With regards to application-specific features, most of the aligners were found to support paired-end alignment for repetitive areas and a few of the aligners did not have the function for SNPs and structural variation discovery. The most important support needed was deemed to be gapped alignment, paired-end alignment, trimming alignment and bisulfite alignment. With regards to computational performance, computation time, maximum memory usage and mapped read counts were analysed as described below.

Most of the aligners exhibited a linear relationship between the computation time and the reference genome size. Also, most read counts effected computation time in that most aligners seemed to depend more of these counts than on the genome size. Overall, BWT-based backtracking algorithms (Illumina) had a faster computation time when compared with the others regardless of size and reads. It was also found that if multiple threads are utilised, computation time either increases or decreases depending on the aligners being implemented.

Variation of maximum memory usage was analysed by comparing the aligners against the server's memory usage percentage. It was showed that said usage was quite low and independent of the genome size so even a low RAM could run them. In the case of the human genome, memory usage increased dramatically. Also, utilisation of multiple threads again showed either an increase or a decrease of usage due to the differing algorithms.

Mapped read counts can generally lead to evaluation of read density. Most aligners showed similar results for short-read datasets but only a handful showed this for long-read datasets. That being said, an accurate deduction for aligner capability and sensitivity could not be made as real-life data's true alignment locations are unknown. To compensate, the accuracy was instead evaluated with *in silico* data. Most aligners showed a high sensitivity as well as precision increase with multi-mapped reads. Datasets having differing error rates, indel sizes and read lengths showed high percentages of total and corrected multi-mapped reads.

In conclusion, this study should aid the user (mainly biologists) in choosing which aligner is most suitable for their research area with regards to NGS data; i.e. an appropriate selection of an aligner for the application in question can be made.