Fast and accurate short read alignment with Burrows–Wheeler transform (Summary)

The Illumina sequencing technology led to new alignment programs, which map many short reads to a genome. They work either by hashing the sequencing reads and scanning through the genome, by hashing the genome, by merge-sorting the reads and the genome or by using Burrows-Wheeler Transform (BWT).

The aligner discussed, BWA, is developed on the string matching theory using BWT. Its implementation resembles that of a top-down traversal on the genome's prefix trie with exact and inexact matches having their distinct sampling methods. A prefix trie of a string can be used to test a query W against a string for an exact substring with the node representing W being found in O(|W|) time. A suffix array (SA) is constructed taking *n* bits of space; i.e. <1GB memory at peak time of the human genome, followed by the generation of BWT. The search for SA intervals of matches can be equalled to sequence alignment.

Both exact and inexact matching can be done by a procedure which tests whether a string W is a substring of another string X. Exact matching is done using backward search which can be compared with exact string matching on the prefix trie without directly putting the trie in memory. On the other hand, inexact matching is done using bounded traversal/backtracking which allows for no more than *z* differences while using backward search to sample different substrings from the genome, similar to a DFS on the prefix trie.

Illumina reads may be problematic due to their ambiguous bases, the support for paired-end mapping (Smith-Waterman alignment), the allowed maximum number of mismatches or gaps being calculated according to the length of the read in question and the generation of mapping quality scores being calculated using the Phredd-scaled probability of the alignment being inaccurate while assuming that the true hit can always be found (may lead to overestimation). SOLiD reads are mapped in a colour space, generated by converting the reference genome to dinucleotide 'colour' sequence and constructing the BWT index for said colour genome. Its paired-end mapping is correct for certain cases and the Smith-Waterman alignment is also applied in the colour space. Dynamic programming is implemented to convert the colour read sequences to their nucleotide counterparts with the help of the Phredd-scaled probabilities to approximate base qualities.

BWA features gapped alignment for single-end reads, paired-end mapping, mapping quality and multiple hits with SAM being the default format for output alignment, having standard tools being available for use by the users. It was evaluated against three alignment programs: MAQ, Bowtie and Soapv2 on both simulated and real data. The simulated reads' results showed BWA to have: similar alignment accuracy to MAQ, more confidence in the mapped reads and error rates than Bowtie and Soapv2, more speed than MAQ and less problems in terms of memory on modern servers due to the support of multithreading. The real reads' results showed BWA to be faster than MAQ with the same alignment accuracy and functionality but slower than Soapv2, and Bowtie to have a smaller alignment error rate with the exception of BWA if its mapping quality threshold is increased.

When the sequencing error rate is high, BWA's performance decreases as it always needs the alignment of the whole read. Longer reads are more prone to interruptions by variations or mis-assemblies in the genome making BWA crash. A possible solution would be to divide the long read into smaller reads, align them and then join to result in the full read alignment.