Comparative analysis of algorithms for next-generation sequencing read alignment (Summary)

Next-generation sequencing techniques support many applications which produce a large amount of data, leading to many difficulties. Re-sequencing is a process whereby short reads from an individual's genome is compared/mapped against an already sequenced reference genome. Multiple factors must be taken into account for this process to occur, namely: sequencing error, short read length and volume of reads. A lot of short read alignment algorithms are available and so, a simulation and evaluation suite SEAL which simulates short read sequencing and analyses the performance of the algorithm was constructed.

Some of the read alignment programs already available are discussed. Bowtie uses a BWT index having small memory consumption but with no assurance of a high quality read mapping if no exact match exists and if it is configured for maximum speed. BWA also uses a BWT index (unlike its predecessor MAQ which uses a hash-based index) having fast searching and reliable quality scores. The mr- and mrs-Fast tools report all mappings of a read rather than only the best one, thus aiding in structural variant detection. They use a seed-and-extend method along with a hash-based index, i.e. a kmer index. Novoalign uses a hash-based index similar to MAQ with a high accuracy rate. SHRiMP uses q-gram filters, spaced seeds and a faster Smith-Waterman algorithm. SOAPv2 uses a BWT hash-based index having a fast alignment speed and a larger memory than its predecessor SOAPv1.

SEAL evaluates the programs' performance by first simulating some reads from a reference genome, with the user having the option of altering certain parameters such as: read length, sequencing error rate (current platforms reporting it at 1%), indel rate, indel length and coverage. It should be kept in mind that coverage does not directly affect accuracy but it should still be realistic due to performance. As most tools report a mapping quality score using Phred scores, the evaluation considers only those reads whose score is greater than a certain value implying a high quality – and thus high accuracy and performance rates. Also, two evaluation methods are taken as some tools report all matching positions while others report only the best matches. In fact, the accuracy of mappings is defined to be either correctly mapped, strict incorrectly mapped, relaxed incorrectly mapped or unmapped. Thus, the strict and relaxed reads provide an accuracy interval if all positions are reported.

Accuracy results were computed according to varying error rate, varying indel sizes and varying indel frequencies. The former shows that Bowtie, BWA and Novoalign are the most sensitive to mapping quality threshold at high error rates and SOAP has a high accuracy even at the lowest possible threshold. The second shows SOAP failing to align reads as it is not quite suited for indel calling, Bowtie, BWA and Novoalign have low accuracy when the threshold is low but have many incorrect mappings with low scores and mr/s-Fast are better with longer indels. The latter shows that all programs' accuracy depends on indel rate – they are inversely proportional. Runtime results were computed according to indexing time and alignment time. Most of the programs exhibit a linear relationship between genome length and index construction time. It can also be observed that most of them have a trade-off between the runtimes (speed versus accuracy) in order to optimise variation detection.

These results should prove beneficial for genomic researchers, keeping in mind that not all experimental scenarios and hardware characteristics could be simulated.