Fast and accurate mapping of Complete Genomics reads

High throughput sequencing (HTS) sequences large groups of human genomes; the problem lying in the read mapping stage where, although tools for the most used HTS methods have been developed, open source aligners are still deficient for the Complete Genomics (CG) platform. It provides paired-end reads of 29-35 base-pairs (*bps*) and a fragment size of 400-1500 bps like other HTS technologies with the difference that a CG read does not have consecutive bases but instead it is composed of multiple sub-reads which may either overlap or have gaps between them

Burrows-Wheeler Transformation with FM-indexing (BWT-FM) is thus not practical for CG data due to its gapped nature and its inability to scale well with indels (**in**sertion or **del**etion of bases in the DNA of an organism). For this reason a sensitive read mapper sirFast was developed to align these reads to the reference assembly using a hash based seed-and-extend algorithm, supporting both SAM and DIVET file formats.

A seed-and-extend mapper first chooses the seeds and then checks their genome locations using a hash table. An extension step called verification, where the similarity score is measured between the read and reference, is taken where a positive score is dealt due to insertions, deletions or substitutions while a zero score is kept by each matching base; the smaller the score, the higher the similarity between the read and reference. The mentioned alignment algorithms (bp-granularity mapping algorithms) are from dynamic programming with the verification step complexity per location being quadratic. This complexity may be reduced to O(kn) if k is an upper bound for allowed indels.

The mappers' performance and accuracy is dependent on how the seeds are selected in the first stage and on the seed type; spaced seeds are not suitable for CG reads so consecutive seeds are used with length 10 and seed index via a hash table. The concept of combined seeds is implemented; its main benefits being the reduction of potential locations and of flexible-sized expected gaps (to decrease mapping time and overhead). The selected seeds then search for read potential locations which are verified using an alignment algorithm – Levenshtein, Smith-Waterman or their variants. The high number of expected gaps leads to the usage of Hamming distance.

Both simulated and real data sets were used to assess method performance. When mapping reads independently sirFast showed full map-ability and a potential use for duplication detection. When implementing paired-end read mapping and precision/recall tests sirFast showed high precision and recall in paired-end mode. When mapping a real dataset and comparison against the CG mapper the reference human genome assembly was used to show that the CG mapper is more error-prone when compared to sirFast.

Theoretically, full dynamic programming requires O(n^2) run time and the combined seed method takes O(kl) where k is the total gap size and l is the seed size. Also, hash tables are used for seed queries as they perform in O(1) time while suffix index binary search takes O(log n) time.

Therefore, the cost of generating CG data is relatively low. However the lack of analysis tools limit research in proprietary algorithm discovery among others. Thus, sirFast should improve the variation detection mechanism for the CG platform.