General scheme of existing mappers

To be able to handle the requirements of nanopore sequencing mapping, new properties of read-to-reference alignment had to be considered to result in a stable sensitivity and recall for very long reads with a low identity. Typical for these alignments are programs such as BLAST, which provide high sensitivity, but become infeasible for larger genomes for example the human genome. Differently to that BWA-MEM and BLASR provide a different tradeoff, scaling well to large genomes but have lower sensitivity and a low precision for high-error rate reads. This can be explained by the specific parameter settings of BLASR, which are designed for PacBio reads and can be chosen when using BWA-MEM for different sequencing technologies. DALIGNER, a highly sensitive overlapper which additionally supports read mapping, also provided precision and recall that degraded quickly with read error rate and genome size. Other mappers, such as LAST, are originally designed for aligning genomes, traded better with these setting, but still exhibits lower recall for large genomes.

Having a closer look on mapper that had a good performance in mapping read-to-reference on a larger genome, approaches of LAST, BWA-MEM and GraphMap for seed finding, seed extension and determining regions on the reference genome seem to be interesting and will be compared in the following section. As a first step of the paradigms of these mappers, region selection relies on finding seeds between the query sequence and the reference. Those can be distinguished by the scoring of the seeds, BWA-MEM uses maximal exact matches (MEMs) or supermaximal exact matches (SMEMs), LAST Hamming distance based spaced seeds and GraphMap a form of gapped spaced seeds, similar to gapped q-gram filters for Levenshtein distance. Fixed length seeds were found to be either not sensitive enough or not specific enough in the presence of high error rates (as used in LAST). BWA-MEM is reducing mismappings caused by missing seeds by re-seeding. The next step is filtering those seeds into candidate regions, with different criteria, like BWA-MEM is greedily chaining seeds that are co-linear and close to each other, and filtering short chains that are largely contained in a long chain and are much worse than the long chain to reduce unsuccessful seed extension in later steps. GraphMap is using for binning seed hits instead the concept of a Hough transform (HT), which corresponds to the main diagonal in the dynamic programming alignment matrix and is clustering the seeds into anchors under graph-based vertex-centric construction. For extending these anchors the Longest Common Substring of kmers (LCSk) had been built with variable lengths corresponding to the anchors, and are filtered with L1 regression to determine the region on the reference genome. A ranking of those anchors is made by the number of exact kmers covered by the anchor, the length of the query sequence which matched the target, the number of bases covered by anchors and the read length. BWA-MEM got a similar approach of ranking those seeds, criteria are length of the chain it belongs to and then by the seed length, after dropping seeds that are contained in an alignment found before, those seeds are extended with banded affine-gap-penalty dynamic programming matrices. Later heuristics avoids extension through a poorly aligned region with good flanking alignment.

Quellen:

<https://www.nature.com/articles/ncomms11307>

<https://arxiv.org/pdf/1303.3997.pdf>