Read mapping tool for AB SOLiD data

Marta Gîrdea, Laurent Noé, and Gregory Kucherov*

INRIA Lille - Nord Europe, LIFL/CNRS, Université Lille 1, 59655 Villeneuve d'Ascq, France

AB SOLiD sequencers [1] produce reads encoded in a 4 color space, whose error-correcting properties allow to better distinguish SNPs from reading errors. We developed a tool for mapping AB SOLiD reads, that aims at obtaining better results by making full use of the encoding properties and the read qualities.

Seeds for finding candidate mapping positions We use *Iedera* [2] to design efficient seeds, based on a model that reflects the reading error distribution observed on SOLiD reads: high chance of errors appearing at distances multiple of 5, and increasing error probability towards the end of the read. *Iedera* can associate to each seed the list of relevant positions on the read to which the search should be restricted, optimized according to the seed patterns. If configured not to provide this list, the positions are established on each read according to read qualities, basically by avoiding to apply a seed if it requires matches on low quality colors. The false positive hits are detected and eliminated by a second filter that performs a fast SIMD bandwith alignment of the read and the reference.

Alignment algorithm We designed a "base intelligent" alignment algorithm which explicitly manages reading errors, base substitutions and up to 7 indels, being able to handle consecutive substitutions even in the presence of reading errors. This is achieved by processing the aligned color pairs in a sliding window rather than individually, and keeping a limited color change history that allows, based on the color code properties, to determine the nature of a color mismatch.

The match/mismatch scores depend on the read qualities: a lower quality implies a lower match reward or mismatch penalty when aligning that color. The scores currently used vary from 0 to 3 for match and from 0 to -3 for mismatch.

Mapping algorithm The reads are mapped in decreasing order of their score (most "trusted" reads first). When mapping a new read, its score and traceback are adjusted according to a score-weighted multiple alignment of the already mapped reads it overlaps, improving the choice of candidate mapping positions.

References

1. Biosystems, A.: A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction. Methods for Annotating 2 Base Color Encoded Reads in the SOLiD^{TM}System (2008)
2. Kucherov, G., Noé, L., Roytberg, M.: A unifying framework for seed sensitivity and its application to subset seeds. Journal of Bioinformatics and Computational Biology **4**(2) (2006) 553–570

* On leave in J.-V.Poncelet Lab, Moscow, Russia