DOI 10.18699/WIBSB-2018-36

Analysis of possible sequence aligner artefacts using novel read density distribution

F.M. Naumenko[1]*, I.I. Abnizova[2, 3], N. Beka[4], Y.L. Orlov[1, 5, 6]**
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Wellcome Trust Sanger Institute, Cambridge, UK*
[3] *Babraham Institute, Cambridge, UK*
[4] *University of Hertfordshire, Hertfordshire, UK*
[5] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[6] *Institute of Marine Biology Researches of the RAS, Sevastopol, Russia*
** e-mail: fedor.naumenko@gmail.com; ** orlov@bionet.nsc.ru*

Key words: next-generation sequencing, DNA alignment, read density distribution

Motivation and Aim: The use of artificial data to evaluate the performance of aligners and peak callers not only improves its accuracy and reliability, but also makes it possible to reduce the computational time. One of the natural ways to achieve such time reduction is by mapping a single chromosome.
Methods and Algorithms: Using own program scripts we investigated whether a single chromosome mapping causes any artefacts in the alignments' performances [1]. We applied our benchmarking tests on 7 open source DNA sequencing mapping tools, namely Bowtie (1.1.1), Bowtie2 (2.2.4), BWA (0.7.5 and 0.7.12 applying two algorithms), MAQ (0.7.1), MOSAIK (2.2.3), SMALT (0.7.6).
Results: In this paper, we compared the accuracy of the performance of seven aligners on well-controlled simulated benchmark data which was sampled from a single chromosome and also from a whole genome. The generation of artificial data by mapping of reads generated from a single chromosome to a reference chromosome is justified from the point of view of reducing the benchmarking time. The proposed quality assessment method allows to identify the inherent shortcoming of aligners that are not detected by conventional statistical methods, and can affect the quality of alignment of real data.
Conclusion: We found that commonly used statistical methods are insufficient to evaluate an aligner performance, and applied a novel measure of a read density distribution similarity, which allowed to reveal artefacts in aligners' performances. We also calculated some interesting mismatch statistics, and constructed mismatch frequency distributions along the read. The approach could be used for analysis of transcriptomics data in plants [2].
Acknowledgements: The research has been supported by RFBR 18-04-00483. Computing done at Siberian Supercomputer center SB RAS was supported by budget project 0324-2018-0017.

References

1. Naumenko F.M. et al. (2018) Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. BMC Genomics. 19(Suppl 3):92.
2. Wang J. et al. (2017) Non-coding RNAs and their roles in stress response in plants. Genomics Proteomics Bioinformatics. 15(5):301-312.