DOI 10.18699/WIBSB-2018-02

Statistical approaches for analysis of mapping quality for single-cell sequencing data

I.I. Abnizova[1, 2]*, R. te Boekhorst[3], N. Beka[3], F.M. Naumenko[4], A.V. Tsukanov[4, 5], Y.L. Orlov[4, 5, 6]*

[1] *Wellcome Trust Sanger Institute, Cambridge, UK*
[2] *Babraham Institute, Cambridge, UK*
[3] *University of Hertfordshire, Hertfordshire, UK*
[4] *Novosibirsk State University, Novosibirsk, Russia*
[5] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[6] *Institute of Marine Biology Researches of the RAS, Sevastopol, Russia*
* *e-mail: irina_abnizova@yahoo.co.uk;orlov@bionet.nsc.ru*

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

Motivation and Aim: Bioinformatics analysis is essential in providing biological insights for single-cell experiments, such as detecting variants, quantifying gene expression, and subpopulation detection. However, conventional tools developed for bulk-cell genomics cannot be directly applied to single-cell sequencing data [1].

Methods and Algorithms: This low coverage characteristic of single-cell sequencing data has posed difficulties in the variant calling procedure. Most bioinformatics tools employ sequence read density to call variants [2]. Single nucleotide polymorphisms and small insertions/deletions with low read support are excluded in conventional bioinformatics tools. In genome assemblies, the low coverage and heterogeneity of single-cell sequencing data also bring substantial disadvantages, leading to truncated sequences with high numbers of sequencing artefacts. Recently, single-cell assemblers such as SPAdes and IDBA-UD have been specifically developed to overcome the challenge of amplification artefacts in single-cell sequencing and generate more precise single-cell genomic assemblies. Common gene expression metrics such as Fragments Per Kilobase Million/Reads Per Kilobase Million (FPKM/RPKM) do not address these 3′-end biases and thus have a limited application for scRNA sequencing. Using own scripts we investigated chromosome mapping quality and possible artefacts [3].

Results and conclusion: We applied or approaches to study Differentially Chromatin accessed regions (DARs) and Diff Methylated Regions (DMR). The generation of artificial data by mapping of generated reads to a reference genome is justified from the point of view of reducing the benchmarking time. We will review current state of art of mapping programs in this research area.

Acknowledgements: The research has been supported by RFBR. Computing done at Siberian Supercomputer center SB RAS was supported by budget project No. 0324-2018-0017.

References

1. Yuan Y. et al. (2018) Single-Cell Genomic Analysis in Plants. Genes (Basel). 9(1):E50.
2. Valihrach L. et al. (2018) Platforms for Single-Cell Collection and Analysis. Int J Mol Sci. 19(3). pii: E807.
3. 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.