Single-cell DNA Sequencing Data: a Pipeline for Multi-Sample Analysis

Marilisa Montemurro
marilisa.montemurro@polito.it

Gianvito Urgese
gianvito.urgese@polito.it

Elena Grassi
elena.grassi@ircc.it

Andrea Bertotti
andrea.bertotti@ircc.it

Elisa Ficarra
elisa.ficarra@polito.it

ABSTRACT

In order to help cancer researchers in understanding tumor heterogeneity and its evolutionary dynamics, we propose a software pipeline to explore intra-tumor heterogeneity by means of scDNA sequencing data.

KEYWORDS

single-cell sequencing, pipeline, evolution, phylogenetic tree, tumor heterogeneity

1 INTRODUCTION

One of the main challenges for cancer researchers is understanding the evolutionary dynamics of the disease. In fact, it is largely known that cancer is not a static disease, but behaves like a living organism, which grows, evolves and shapes itself, to adapt to the external environment and survive to all possible threats (e.g. the patient immune system). Single-cell sequencing offers the possibility to better investigate the complexity and the heterogeneity of the internal structure of tissues, thanks to the high-resolution data it produces.

2 THE PIPELINE

We propose a software pipeline capable of producing multi-sample copy-number variation (CNV) analysis on large-scale single-cell DNA sequencing data (in the order of some thousands of cells) and investigate spatial and temporal tumor heterogeneity.

Single-sample analysis. The pipeline has been designed to run on the scDNA sequencing data produced by *10X Genomics*[1] technology, which is, at the time of writing, the only existing solution capable of performing large-scale sc-sequencing. The company provides a proprietary pipeline, *Cell Ranger DNA*, that aligns reads, identifies CNV events and infers phylogenetic trees. This tool allows to execute only single-sample analysis and, being a closed platform, it is not possible to modify it with new features. For this reason, we embedded it into a pipeline, which exploits the functionalities of an open-source single-cell CNV analysis tool, named *Ginkgo*[2], to re-analyze data and use them to perform additional processing. The two applications are integrated by means of a demultiplexer, which, starting from the aligned reads produced by Cell Ranger DNA, filters out the multi-mapped ones and splits them by cell identifier. The demultiplexed data are provided to Ginkgo, which computes its own copy-number profiles. Ginkgo results are validated by computing Jaccard similarity scores between the CNV calls produced for each cell by the two applications. Ginkgo CNV profiles are, then, used to build a new phylogenetic tree and a new heatmap, for each sample: in this way, it is possible to compare the results of the two tools, also, in a qualitative way. Moreover, in order to provide a functional information, CNV events are annotated with the corresponding gene symbols. Finally, density plots highlight, for each sample, the clusters of cells with a similar mean ploidy, allowing to set a threshold and filter out cells which ploidy is considered irrilevant for the following analysis.

Multi-sample analysis. Multi-sample data are, finally, aggregated to evaluate evolutionary distance among them. This task is accomplished by another module which performs the phylogenetic reconstruction algorithm on all cells, from all different samples, and produces one single tree and one single heatmap. In order to evaluate intra and inter-tumor heterogeneity, two distance metrics have been defined: (i) an intra-tumor (intra-TH) heterogeneity measure, which tells how distant cells within the same sample are, and (ii) an inter-tumor heterogeneity (inter-TH) measure, which tells how distant cells from different sample are. Both measures are based on the distances computed to construct the phylogenetic tree and associated to its branches.

3 CONCLUSION AND FUTURE WORK

We have developed a platform for large-scale single-cell DNA CNV analysis, capable of enriching the results of a proprietary scDNA pipeline, Cell Ranger DNA, and exploit them to evaluate intra and inter-tumor heterogeneity, with the help of a flexible and open source tool, Ginkgo. Additionally, we plan to implement a new module capable of performing cell clustering starting from their CNV profiles: this should allow to better investigate spatial segregation of clones. In conclusion, this pipeline is a ready-to-use instrument for researchers who are interested in exploring temporal and spatial tumor evolution and need automatic tools to help them managing data and performing analyses.

ACKNOWLEDGMENTS

This work has been supported by the SmartData@PoliTO center on Big Data and Data Science. Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it).

REFERENCES

[1] 10x Genomics. [n. d.]. 10x Genomics: Biology at True Resolution. Retrieved April 15,2 2019 from https://www.10xgenomics.com/

[2] Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. 2015. Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods* 12 (Sept. 2015), 1058–1060. https://doi.org/10.1038/nmeth.3578