DOI 10.18699/BGRSSB-2018-023

PyMPFA: python pipeline for massively parallel functional assays used for characterization of DNA regulatory elements

A. Ivankin[1]*, A. Pindyurin[1, 2]
[1] *Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia*
[2] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
** e-mail: anton.ivankin@gmail.com*

Key words: massively parallel functional dissection assay, DNA regulatory elements, python

Motivation and Aim: The spatio-temporal regulation of gene expression is a complex process that determines the destiny and function of various cells and tissues. The regulation of this process consists of many stages and to a large extent is orchestrated by diverse DNA regulatory elements (promoters, terminators, enhancers, etc.) and epigenetic factors associated with them. Recently, a number of massively parallel functional assays (MPFAs) aimed to identify and dissect DNA regulatory elements were developed. Most of these assays make use of short DNA sequences, which are most frequently referred to as "tags" or "barcodes", to track the expression activity (and in some assays genomic localization) of individual gene reporter constructs. Typically, as assay readout, barcodes are PCR-amplified from cDNA and DNA samples obtained from the studied cells and subsequently identified and counted by using the high-throughput sequencing.
Methods and Algorithms: Python language (v. 2.7.6) was used to implement the pyMPFA algorithm designed to extract barcodes and associated with them variable DNA sequences (hereafter, "features", which are mutant variants or genomic locations) from the high-throughput sequencing reads and subsequently identify and report the most reliable barcode-feature combinations. The follow-up analysis of the data was carried out in R language (v. 3.4.4).
Results: We have developed a flexible pyMPFA pipeline, which is applicable to high-throughput sequencing reads of different structure. As a pilot test, we applied this pipeline to datasets generated by MPFAs aimed to characterize (1) the influence of minor DNA sequence variations in the 3′ downstream region of the reporter gene on its episomal expression in cultured human cells and (2) the influence of local chromatin environment of the reporter gene activity in cultured drosophila cells. As a result, we found that the running time of the pyMPFA pipeline strongly depends on the number of unique barcodes present in a dataset. There is an exponential dependence between the number of processed unique barcodes and the time of the pipeline run. For example, reads with about 200 thousand unique barcodes can be processed by the pyMPFA pipeline (on CPU Intel Core-i7 3770K, 32Gb DDR3) in single-threaded mode in 4–5 hours.
Conclusion: Despite the flexibility of the pipeline, it still operates slowly when processing large datasets. The processivity of the pyMPFA pipeline will be improved in its future releases. Source code of the pipeline is available on github: https://github.com/wiw/pyMPFA/tree/trip_0.3/pyMPFA.
Acknowledgements: Supported by the RSF (16-14-10288).