A shiny app for the integration and enrichment analysis of genomic region sets by NGS data

Davi Toshio[1], Henrique Cursino Vieira[2], Christiane Bezerra de Araujo[3], Maria C. Elias[3], Bruno Ferreira de Souza[4], Hugo A. Armelin[1], Milton Yutaka Nishiyama Junior[5],

1 Instituto Butantan
2 LECC-CeTICS, Instituto Butantan
3 LECC-CeTICS, Butantan Institute
4 ECC-CeTICS, Instituto Butantan
5 LETA-CeTICS, Instituto Butantan

Abstract

The new biotechnology advances has allowed studies of the systems involved in the DNA integrity, stability, replication, demethylation; recent discoveries have related them to genomic and transcription stability, besides relations between inflammation and DNA damage, which are essential aspects of molecular biology that underlies developmental processes and disease etiology. This work is part of The Center of Toxins, Immune-response and Cell Signaling (CeTICS), which aims to understand the behavior of biological systems based on analysis of -omics data and signaling networks. Several genomic techniques including MFA-seq, MNase-seq, ChIP-seq, DNase-seq and ATAC-seq have been developed to experimentally identify genome-wide profiles of regulatory regions and experiments have been profiled by the CeTICS project and thousands of samples can be found in the ENCODE and Roadmap Epigenomics consortia. The genomic datasets has grown rapidly and has been used as reference databases; they are essential to retrieve information on gene name, protein product, transposable elements, motifs, molecular markers and others and are important in the approaches to identify enriched regions. To allow the identification and characterization of enriched regions from high-throughput sequencing data, which can be increased with visual inspection of the data, we present a Shiny application for an interactive representation and analysis tools based on Fold Change and MACS2 software peak predictions. The app has been developed in R language, using the Shiny framework, integrating multiple Bioconductor packages. The first step is the alignment and coverage estimation, followed by the upload to the results and annotation file to the app for the downstream statistical and graphical analysis. The input files are composed by: i) the fold change and coverage estimation; ii) the MACS2 peak detection and coverage estimation; ii) Genome size and annotation. As a study case, was used the investigation of the DNA replication features of T. cruzi in order to verify possible links between genetic instability and DNA replication, using the high throughput analysis approach MFA-seq (Multiple Frequency Analysis). To do that, epimastigotes of T. cruzi were sorted in early S and G2/M phases and the DNA was extracted from each group and analyzed by MFAseq. The aim of this integrative approach is to allow the identification of enriched genomic regions distinct to each pair of conditions, integrating multiple annotations, coverage, predicted peaks, SNP's, genes, transcripts and others and the integration of Machine Learning methods to improve the data integration analysis.

Funding: #2013/07467-1, São Paulo Research Foundation (FAPESP)