TITLE: CLUSTGCLOSED: A CLUSTERING-BASED TOOL FOR CLOSING GAPS IN GENOME ASSEMBLY

AUTHORS: [1]NEGRÃO, D. M.; [1]FRANCO, E.F.; [1]PINHEIRO ,K.; [1]MIRANDA, F.; [1]ARAGÃO, G.; [1]SILVA, A.; [2]GHOSH, P.; [1]RAMOS, R.T.

INSTITUTION: [1]UNIVERSIDADE FEDERAL DO PARÁ - (R. Augusto Corrêa, 1 - Guamá, Belém - PA, 66075-110) - BRASIL.

[2]VIRGINIA COMMONWEALTH UNIVERSITY (907 FLOYD AVE, RICHMOND, VA 23284, EUA)

ABSTRACT:

The process of closing the gaps in genome assembly is both laborious and intensive step. Hence completion of a genome gets impeded specifically considering the large amount of data generated by Next Generation Sequencing (NGS) and different errors that are generated during the assembly process. Several tools developed for gaps closure have been reported to facilitate and improve the closing process of the gaps. However, the gap closure process continues to be a challenge to assess assembly completeness, because it demands a lot of time and sometimes it is not possible to close all the gaps present in the genome. Prevailing reasons for the generation of gaps in genome assembly relate to the limitation of the algorithms to recover repetitive and low coverage regions. Moreover, the GC bias can cause low coverage sequencing in certain regions of the genome , which prevents the representation of these regions during a traditional assembly due to the lack of a definition of a unique parameter of coverage and/or k-mer to represent the genome. This bias leads to the formation of gaps due to low coverage in regions where the GC-content may be high and/or low depending on the sample analyzed. We present CLustGClosed, a pipeline and web-tool to improve the process for closure of gaps in genome assembly. CLustGClosed uses the genome's GC content to devise clustering techniques, for generating contigs to close the gaps present in the draft genome. We have successfully validated the proposed pipeline using different genomes from SRA database and compared the results with other gap closure tools. In different validation tests, CLustGClosed reduced the number of gaps upto 75%, improved the N50 and decreased the number of Ns present in the genome by as much as 503%. The pipeline provides versatility in its use because it can be used through different forms such as API, web or CLI, and can take different input files which provides greater accessibility for the scientific community.

Keywords: Machine Learning, Gaps Closure, Clustering, Gap filling, Genome finishing.

Development Agency: CAPES, CNPq