TITLE: DEVELOPMENT OF A WEB PLATFORM FOR ENHANCEMENT OF PAN-GENOME ANALYSIS FROM IMPROVED GENOME ASSEMBLY

AUTHORS: PANTOJA, Y. P. M.; DANTAS, C. W. D.; PINHEIRO, K. C.; SILVA, A.; RAMOS, R. T. J.

INSTITUTION: UNIVERSIDADE FEDERAL DO PARÁ, BELÉM, PA (R. AUGUSTO CORRÊA, 1, CEP 66075-110, GUAMÁ, BELÉM – PA, BRAZIL)

ABSTRACT:

New generation sequencers are characterized by high sequencing and high cost sequencing, contributing to a significant increase in the amount of genomic data deposited in public databases, which made it possible to perform comparative analysis, such as pan-genome analysis. Despite the high coverage of sequencing, coverage biases, such as those related to GC content, can hinder the assembly process, and consequently subsequent analysis, such as pan-genomics. In addition, current approaches to pan-genome analysis focus exclusively on regions with potential for protein coding, which excludes non-coding regions (Intergenic Regions), which normally occupy about 15% of the bacterial genome, and are regulatory elements, such as ncRNAs. In this way, PanWeb2 was developed, a Web platform for the reduction of the effects of GC bias on the assembly of genomes, to be submitted to pan-genome analysis, and for the pan-genome analysis of intergenic regions. For the validation of the platform, 6 genomes of the organism *Escherichia coli* were used**.** A comparison was made with the results obtained through a standard/manual assembly and pan-genome analysis and a automatically assembly and pan-genome analysis performed by PanWeb2. As a result of the assembly step of the standard approach, for the strain *E. coli* P12b, for example, 209 contigs were obtained, whereas in the analysis done by PanWeb2, 153 contigs were obtained. And in relation to the numbers of N50, number of bases and number of genes, all increased in the analysis done by PanWeb2. The same behavior was repeated for the other 5 strains of *E. coli*, number of contigs decreased and number of bases, number of genes and N50 increased. It was observed that assembly of genomes with the reduction of the effects of GC bias becomes more representative, which directly impacts the results of the posterior pan-genome analysis.

Keywords: genome assembly, GC bias, pan-genome analysis, intergenic regions

Development Agency: CAPES, CNPq