TITLE: Case study of teaching in bioinformatics and evaluation of a low cost computational server for processing short reads

AUTHORS: S, R. S.[1]; V, M. C.[1]; OLIVEIRA, M. S. [2]; VERAS, A. A. O.[2]; ALVES, J. T. C.[3]; SÁ, P. H. C. G.[1]

INSTITUTION: [1]FEDERAL RURAL UNIVERSITY OF AMAZONIA (RODOVIA PA 140, CEP 68680-000, TOMÉ-AÇU, PARÁ, BRAZIL); [2]FEDERAL UNIVERSITY OF PARÁ (RUA ITAIPU, 36, CEP 68464-000, TUCURUÍ, PARÁ, BRAZIL); [3]PARÁ STATE UNIVERSITY (AV. HILÉIA, S/N, CEP 68502-100, MARABÁ, PARÁ, BRAZIL)

ABSTRACT:

The development of the Next Generation Sequencers (NGS) enabled a revolution in DNA and RNA sequencing techniques. This scenario had a direct impact on the development of bioinformatics techniques in the most diverse areas, such as health, environmental studies, forensic biology, among others. However, knowing that the importance of bioinformatics is still growing, it is necessary to focus on training professionals in this area. Thus the goal of this work is to evaluate the difficulty of performing the main bioinformatics analyzes for undergraduate students and at the same time evaluate the processing of these analyzes in a low cost computational server. Undergraduate students in biological sciences performed the main analyzes of bioinformatics and report the difficulties faced in the analysis execution and interpretation of the results. Four sets of simulated reads generated by the Art program were used, with different coverages (10x, 25x, 50x and 100x). The programs used by the students were: for quality analysis, FastQC; for filtering and trimming the reads, Fastxtoolkit, Prinseq, Trimommatic and Ngsshort; for alignment, Bowtie2; for genome assembly SPADES, Velvet, Soapdenovo and Megahit; and for evaluation of the genome assembly, the Quast. Also, the execution time of each analysis was evaluated, with each set of reads on the low cost computational server, a desktop with Core i7 processor and 20gb of RAM. The four sets of reads were analyzed in all programs cited, by the students separately. As a result of the difficulty assessment, the only program considered easy in execution and interpretation was FastQC. All genome assembly programs were considered difficult and the remaining programs range from moderate to difficult. And in the result of the server evaluation the only data that could not be processed was the 100x coverage data for the genome assembly analyzes. For the other analyzes all the datasets (10x, 25x, 50x and 100x) could be processed with different execution times. Thus, with the result of this work, it was possible to evaluate the difficulty of the students in relation to the main bioinformatics analyzes. In addition, this study also allowed us to identify how much reads a low cost computational server can process in different analyzes.

Keywords: Bioinformatics; Education; NGS; Reads; Sequencing;