TITLE: COMPARISON OF THE *de novo* ASSEMBLY PROCESS USING A BASIC COMPUTER APPLIED TO NGS DATA OF PROKARYOTES

AUTHORS: SÁ, P. H. C. G.[1]; VERAS, A. A. O.[2]

INSTITUTION: 1. UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA (UFRA), TOMÉ-AÇU, PA (RODOVIA PA 140, 2428, CEP 68680-000, TOMÉ-AÇU - PA, BRASIL); 2. UNIVERSIDADE DA AMAZÔNIA (UNAMABR), ANANINDEUA, PA (RODOVIA BR 316, KM 3, CEP 67113-901, ANANINDEUA - PA, BRASIL).

ABSTRACT:

Among the various bioinformatics analyzes that can be performed using NGS data, one of the main and most frequently is *de novo* genome assembly. The *de novo* assembly process has a high computational cost that is directly proportional to the amount of reads used in the assembly. However, most of the research labs, especially the small and medium ones, do not have the financial conditions to obtain a high performance computer, so it is interesting to evaluate if the *de novo* assembly can be done using ordinary computers. Thus, the purpose of this work is to perform a comparison of the *de novo* assembly process applied to microorganisms using a basic computer, which can be found in different types of laboratories more easily. The Velvet assembler and the SPAdes assembler will be used to carry out the assembly process. The sequencing data of four microorganisms available in the Sequence Read Archive database (SRA) will be used, each organism obtained by a different sequencing platform. *Corynebacterium pseudotuberculosis* 262 (SRX1669990) generated by the Ion Torrent PGM. *Escherichia coli* P12b (SRX1012260) sequenced by the Illumina MiSeq. *Rhodococcus equi* ATCC 33707 (SRX037194) obtained by the 454 GS FLX Titanium. And *Nocardia brasiliensis* (SRX2717972) sequenced by the Illumina HiSeq 4000. The raw data of the organisms will be assembled by the two assemblers using a notebook with Debian 8 64-bit, Core i7-4510u 2.0GHz processor and 16 GB of RAM, dedicated only for this task. As a result, the only organism that was not assembled was *Nocardia brasiliensis*, in which for both assemblers the process was terminated due to lack of memory. In terms of time Velvet was faster than SPAdes for *C. pseudotuberculosis* 262, *E. coli* P12b and *R. equi* ATCC 33707. However, about the amount of contigs, for the three organisms the SPAdes assembler generated a result with a smaller number of contigs. Comparing the total assembly of each organism with the reference size, Velvet and SPAdes obtained similar results for *E. coli* P12b, with Velvet yielding a better result for *R. equi* ATCC 33707 while SPAdes obtained a better result for *C. pseudotuberculosis* 262. Although Velvet performs the assembly in less time, SPAdes generates an assembly with fewer contigs. The ideal here would be to combine the results of both assemblers to generate more complete assemblies. Thus, it was possible to compare how much *de novo* assembly can be performed using only a basic computer.

Keywords: NGS; de novo assembly; basic computer; genome assembly; comparison