TITLE: DEVELOPMENT OF A PIPELINE TO ASSEMBLE SHORT READS SEQUENCED BY ILLUMINA HISEQ

AUTHORS: MIRANDA, F.M.[1]; SOUSA, T.J.[1]; KATO, R.B.[1]; CAVALCANTE, A. L.Q.[2]; AZEVEDO, V.A.[2]; SILVA, A.L.C.[1,2]; RAMOS, R.T.J.[1,2].

INSTITUTION: 1. INSTITUTE OF BIOLOGICAL SCIENCES, FEDERAL UNIVERSITY OF MINAS GERAIS, BELO HORIZONTE, MINAS GERAIS, BRAZIL; 2. LABORATORY OF GENOMIC AND BIOINFORMATICS, CENTER OF GENOMICS AND SYSTEM BIOLOGY, FEDERAL UNIVERSITY OF PARÁ, BELÉM, PARÁ, BRAZIL.

ABSTRACT:

The Illumina platform is still the most used worldwide, accounting for about 90% of the sequencing data generated. Although there are new sequencing technologies that produce longer reads, they can reach up to 15% error rate. Thus, either the Illumina platform continues to be used in order to combine its accurate data with the latest sequencing technologies or it is the preferred standalone choice when project costs are tight. However, obtaining a finished genome only from short reads is still a difficult task. Therefore, in this work we present a new pipeline that exploits previous advances in genomic assembly, in order to optimize *de novo* assembly of bacterial data sequenced by Illumina HiSeq. The proposed pipeline automatically installs the dependencies, then merges overlapping paired-end reads, cleans the data by removing adapters and trimming low quality sequences with AdapterRemoval v2, followed by k-mer prediction with KmerStream and assembly with Edena and SPAdes. Experimental results on 3 public datasets – *Mycobacteroides abscessus* (*M. abscessus*), *Staphylococcus aureus* (*S. aureus*) and *Vibrio cholerae* (*V. cholerae*) – obtained from GAGE-B showed that our pipeline achieved similar or better results when compared with QUAST using reference genomes to another pipeline known as Unicycler. The genome fraction obtained by the proposed pipeline was higher in all cases (99.189% for *M. abscessus*, 92.78% for *S. aureus* and 96.95% for *V. cholerae*) when compared to Unicycler (96.94%, 92.709% and 95.981%, respectively). Our pipeline also produced less misassemblies, reporting 6 misassemblies for *M. abscessus*, 39 for *S. aureus* and 3 for *V. cholerae*, while Unicycler produced 10 misassemblies for *M. abscessus*, 40 for *S. aureus* and 11 for *V. cholerae*. The number of whole genomic features represented on Unicycler's assembly was better only for *S. aureus* (5176 vs. 5125), while our pipeline represented more features for the *M. abscessus* and *V. cholerae* assemblies (9887 vs. 9725 and 7109 vs. 7058, respectively). These preliminary results show that the proposed pipeline can improve bacterial assemblies for Illumina HiSeq datasets and produce contigs with more biological meaning, which can positively impact downstream analysis. In future work, we plan to add new features to the pipeline in order to perform hybrid assemblies using both Illumina short reads and long reads sequenced by new platforms such as PacBio and Oxford Nanopore.

Keywords: genome assembly, sequencing, bacteria

Development Agency: CNPq, CAPES