Efficient gap-closure of eukaryotic genome sequence assemblies by third-generation sequencing

Izinara Rosse da Cruz, Juliana Assis Geraldo, Laura Rabelo Leite, Adhemar Zerlotini, Flávio Araújo, Marcos Vinícius G.B. Silva, Roney Santos Coimbra, Maria Raquel Santos Carvalho, Guilherme Oliveira

Universidade Federal de Minas Gerais-UFMG, Centro de Excelência em Bioinformática-CEBio/FIOCRUZ-Minas, Centro de Excelência em Bioinformática-CEBio/FIOCRUZ-Minas, EMBRAPA Informática Agropecuária, FIOCRUZ-Minas, EMBRAPA Gado de Leite, Centro de Excelência em Bioinformática-CEBio/FIOCRUZ-Minas, EMBRAPA Gado de Leite, Centro de Excelência em Bioinformática-CEBIO/FIOCRUZ-Minas

Recently, Pacific Bioscience released to scientific community their first commercial third-generation sequencing instrument, the PacBio RS, a single-molecule real-time sequencing, which is capable of generating significantly longer reads and with low composition bias. This technology has been able to resolve complex repeats and to close gaps in draft genomes, contributing towards the goal of "one chromosome, one contig". Recently, we produced the first draft assemblies of Guzerá and Gir genomes. Two Zebu (Bos indicus) breeds adapted to tropical climate and under genetic evaluation to milk purposes in Brazil. The average depth of coverage was 26X, but about 13% of the consensus sequence contained gaps compared to the reference genome of Bos taurus (UMD 3.1). In order to further improve the genome drafts, we performed a hybrid de novo assembly with MIRA3 using as input: 1) the contigs, larger than 1 Kb, previously generated by de novo assembly of SOLiD reads using SOAPdenovo and Graph Constructor (Convey hybrid-core system); 2) pseudo-contigs from chromosome 18 derived from the consensus of SOLiD reads mapped against the reference genome using LifeScope (Life Technologies); 3) and the contigs and singletons resulting from a de novo assembly using continuous long-reads from PacBio RS using MIRA3. In this proof of concept, we limited our analysis to the chromosome 18 of Guzerá. As a result, from a total of 400,045 contigs submitted to the hybrid assembly, 18,843 were assembled in new and longer 2,000 supercontigs. Comparing the sequence of these new contigs with their respective regions in the consensus from the mapping step against the reference genome, we observed that, 86% of the "Ns" and 2,673 gaps (99%) (≥4Ns) were closed in the hybrid assembly. Considering only supercontigs containing PacBio data, 91% of "Ns" and 481 (99%) gaps were closed. This is the first study to demonstrate that hybrid assembly strategies using short reads from SOLiD and large reads from PacBio may improve the assembling of eukaryotic genomes. Funding: SECTES/FAPEMIG(485/2009, CBB-1181/0, TCT 12.093/10, REDE-56/11), CAPES, CNPq, NIH-USA (TW007012), CAPES/CDTS-FIOCRUZ, FIOCRUZ-Minas

Keywords: zebu genome sequencing, PacBio RS sequencer, SOLiD platforms, improve assembly genome

Concentration area: Genomics Evolution