New algorithms for accurate and efficient de-novo genome assembly from long DNA sequencing reads

David Guevara-Barrientos, Laura Gonzalez (ln.gonzalez138@uniandes.edu.co), Daniela Lozano, Juanita Gil, Maria Camila Hoyos, Christian Chavarro, Natalia Guayazan, Luis Alberto Chica, Maria Camila Buitrago, Edwin Bautista, Juan Camilo Bojacá, Miller Trujillo, Jorge Duitama
Departamento de ingeniería de sistemas y computación, Universidad de los Andes, Bogotá, Colombia.

Producing high-quality de-novo genome assemblies for complex genomes is possible thanks to the development of long read DNA sequencing technologies. New algorithms have been developed to achieve contiguous assemblies of complex genomes; however, new algorithmic techniques have the potential to further improve the accuracy and computational efficiency to build both haploid and diploid genome assemblies.

We present here the implementation of a new algorithm for assembly of large DNA sequencing reads, following the overlap-layout-consensus (OLC) process. This algorithm builds an undirected graph having two vertices for each read representing the start (5'-end) and the end (3'-end) of the read. A minimizers table is constructed from the reads to identify overlaps in linear time. K-mer hash codes are calculated based on rankings relative to the mode of the k-mer counts distribution. For each candidate overlap, matching k-mers are clustered following an efficient dynamic programming algorithm. Statistics collected during this process include an overlap estimation, a coverage of shared k-mers weighted by the repetitiveness of each k-mer (WCSK), and the percentage of the predicted overlap supported by k-mers. These statistics are used as features to build layout paths following a Naive Bayesian machine learning approach in which selecting path edges can be thought of as a binary classification problem. Expected distributions of the different features are inferred from a selected subset of "safe" edges which are reciprocal best in both overlap and WCSK.

We ran the implemented algorithms on HiFi PacBio sequencing data taken from haploid or low heterozygosity diploid samples of rice, maize and the human cell line CHM13, comparing our assemblies with those of other currently used software tools. Our algorithm ranked first on the rice sample, achieving a median contig length of 30 Mbp, and a complete assembly of 7 of the 12 chromosomes. Statistics of numbers of misassemblies also indicate that our assembly has the second best precision for this dataset. The more complex genomes of maize and the human cell line were sequenced at lower read depth, making the assembly obtained with our solution rank below the two most contiguous assemblies. However, the obtained genomes are still highly contiguous and accurate, with N50 values above 10 Mbp.

We also integrated our previous work on single individual haplotyping to perform phased assemblies of diploid samples. Our algorithm builds first a contiguous haploid assembly which is used to identify heterozygous sites within each contig and to perform single individual haplotyping. The phasing procedure is used to remove edges connecting reads assigned to different haplotypes and obtain a phased assembly running the layout algorithm on the filtered graph.

We expect that this new development would be useful for research groups developing different strategies for genome assembly, as well as researchers currently building genome assemblies for different species.