Plastid Genome Assembly Using Long-read data (ptGAUL)

Wenbin Zhou[1], Carolina Armijos[2], Chaehee Lee[3], Ruisen Lu[4], Jeremy Wang[1], Tracey Ruhlman[5], Robert Jansen[6], Alan Jones[1], and Corbin Jones[1]

[1]The University of North Carolina at Chapel Hill
[2]Universidad San Francisco de Quito
[3]University of California Davis
[4]Institute of Botany Jiangsu Province and Chinese Academy of Sciences
[5]University of Texas at Austin
[6]The University of Texas at Austin

December 2, 2022

Abstract

Although plastid genome (plastome) structure is highly conserved across most seed plants, investigations during the past two decades have revealed several disparately related lineages that have experienced substantial rearrangements. Most plastomes have two inverted repeat regions and two single-copy regions with few dispersed repeats. However, the plastomes of some taxa do harbor long repeat sequences (>300 bp). These long repeats make it difficult to assemble complete plastomes using short read data, leading to misassemblies and consensus sequences that have spurious rearrangements. Long read sequencing can potentially overcome these challenges. However, there is no consensus as to the most effective method for accurately assembling plastomes using long read data. Here, we generated a pipeline, plastid Genome Assembly Using Long-read data (ptGAUL) to address the problem of assembling of plastomes using long read data from Oxford Nanopore Technologies (ONT) or Pacific Biosciences (Pacbio) platforms. We demonstrated the efficacy of the ptGAUL pipeline using 16 published long read datasets. We showed that ptGAUL produces accurate and unbiased assemblies. Additionally, we applied ptGAUL to assemble four Juncus (Juncaceae) plastomes using ONT long reads. Our results revealed many long repeats and rearrangements in Juncus plastomes compared with basal lineages of Poales.

Hosted file

`chloroplast_genome_ms draft_1182022_BEAN-tr-bob.doc` available at https://authorea.com/users/562144/articles/609397-plastid-genome-assembly-using-long-read-data-ptgaul