De Novo assembly of complete chloroplast genomes from non-model species based on a K-mer frequency-based selection of chloroplast reads from total DNA sequences

ABSTRACT

Whole Genome Shotgun (WGS) sequences of plant species often contain an abundance of reads that are derived from the chloroplast genome. Up to now these reads have generally been identified and assembled into chloroplast genomes based on homology to chloroplasts from related species. This re-sequencing approach may select against structural differences between the genomes especially in non-model species for which no close relatives have been sequenced before. The alternative approach is to de novo assemble the chloroplast genome from total genomic DNA sequences. In this study, we used k-mer frequency tables to identify and extract the chloroplast reads from the WGS reads and assemble these using a highly integrated and automated custom pipeline. Our strategy includes steps aimed at optimizing assemblies and filling gaps which are left due to coverage variation in the WGS dataset. We have successfully de novo assembled three complete chloroplast genomes from plant species with a range of nuclear genome sizes to demonstrate the universality of our approach: Solanum lycopersicum (0.9 Gb), Aegilops tauschii (4 Gb) and Paphiopedilum henryanum (25 Gb). We also highlight the need to optimize the choice of k and the amount of data used. This new and cost-effective method for de novo short read assembly will facilitate the study of complete chloroplast genomes with more accurate analyses and inferences, especially in non-model plant genomes.

Keyword: Chloroplast genome; De novo assembly; Solanum; Aegilops; Paphiopedilum; DNA sequencing; Whole genome shotgun sequencing; k-mer analysis