Iraqi Journal of Biotechnology, 2022, Vol. 21, No. 2, 439-445

Mapping to Reference: An Efficient Approach to Achieve Sufficient Consensus for Phylogenomic Study of *Oryza* Chloroplast Genome

Raffal A. Saloom , Ali M. Moner

Institute of Genetic Engineering and Biotechnology for Postgraduate Studies, University of Baghdad

Received: 1/6/2022 Accepted: 2/10/2022 Published: December 20, 2022

Abstract: The chloroplast is one of the most important organelles in plants, it plays a significant role in following the heredity of the plant species over millions of years. To achieve chloroplast sequence there are many approaches, some of them easy like mapping, while others are consuming-time and need more computing resources. Here three approaches have been utilized *Oryza sativa* raw data as a model, to compare (multiple rounds of mapping to reference and Denovo) to find the fastest way to produce the most reliable consensus for phylogenomic studies, by comparing variations numbers and variation reduction percentages among all three pipelines (A, B, and C). It has been concluded that simple two rounds of mapping will be enough to construct trustworthy phylogeny.

Keywords: *Oryza* chloroplast, mapping, Iraqi rice cultivar, phylogenomic

Corresponding author: (Email: rafal.ali1200a@ige.uobaghdad.edu.iq).

Introduction

The chloroplast is one of the most important organelles in plant cells. It does a substantial role in photosynthesis and maintains the cell life cycle. Crucial metabolism reactions like respiration and phosphorylation occur in the chloroplast. Therefore any variation in it might reflect positively or negatively on the entire plant (1). The chloroplast genome is one of the most conservative sequences, it has vital information about the evolution history of thousand years ago (2), (3).

Previously sequences were obtained by amplifying several large fragments to complete the whole genome. an example is *Arabidopsis thaliana*; however, this technique consumes time money, and effort. New sequencing technology makes it easier to sequence the entire genome using high-throughput sequencing machines like illumina and PacBio as short and long reads (4). A plant cell has three different genomes (nuclear, mitochondria, and chloroplast) all of them will be present in the extracted DNA, in addition to the repetitive sequence among them makes it even tougher to assemble the correct genome sequence. Chloroplast genome unique structure: mainly large single copy, small single copy, inverted repeat, A, and inverted repeat B, this structure makes it difficult to assemble especially with short reads because it could much multiple regions (5).

Providing useful markers for phylogenetic investigations. The low complexity and high copy number of organelle genomes greatly facilitate their characterization. Owing to these properties and the advent of next-generation sequencing, the pace at which chloroplast genome sequences are being produced has dramatically increased during the last years (6). Therefore, the researcher went to extract just chloroplast DNA and separate it from other nuclear and mitochondrial DNA. This is possible but it needs specific procedure with some expensive equipment and kits (1, 7).

Nevertheless, all these efforts, time, expensive equipment and expenditure; there is simpler approach to achieve the almost accurate sequence if there is a good close relative reference genome. *Oryza sativa* is an economically crucial model plant and effort has been put together to achieve well annotated genome sequence, so it will be utilized as an example to clarify how much effort should be put in chloroplast genome assemble especially for phylogenomic studies.

Materials and methods

Whole genome sequence clean data of five *Oryza sativa* cultivars has been obtained from (ongoing study unpublished data (8)). These data were utilized in different pipeline to assemble the entire chloroplast genome as explained in (figure1). Raw reads apply to two protocols, firstly mapping reads to reference (NC_001320) at two different settings. Strict and relax to make sure that all real variations capture and overcome the mapping bias. Bowtie2 version 2.3.0 has been used with alignment type end to end, highest, and lowest sensitivity (9,10). Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy (9).

Then both consensuses compare, call will be taken on those variants based on reads alignment on both settings. Secondly, Denovo assemble for all reads to obtain contigs. Thereafter, contigs has been mapped to reference the consensus checked and trimmed. Tadpole tool version 37.64 has been used with Kmer length=63 with default setting (11).Variant call was record using Geneious version 11.1.5 embedded tool with minimum coverage 1, maximum variant P-value=10^{-6} and strand-bias P-value= 10^{-5}. phylogenetic tree constructed using PhyML package 3.3.2018 (12) with GTR substitution model, Bootstraps =100 and optimize topology.

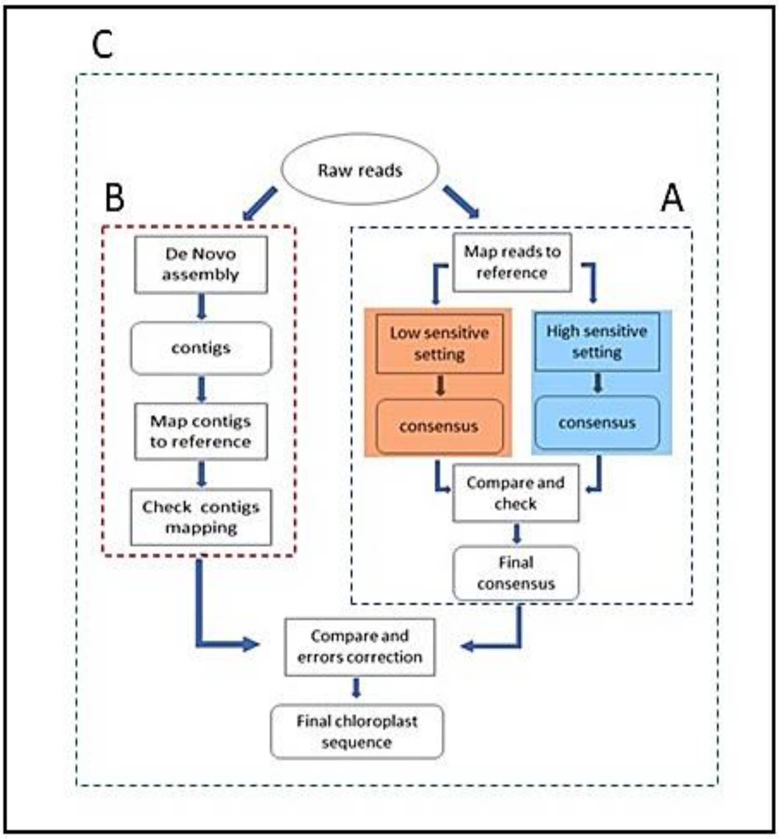


Figure (1): Shows raw reads assembly work flow using mapping to reference and de novo procedures.
A=Mapiing with two different setting Strict and relax, B = Denovo pipeline, C = double pipeline.

Results and discussion

Clean date has been achieved from (ongoing research unpublished (8)), for five *Oryza sativa* Iraqi cultivar namely (Dijla, Ghadir, Baraka, Amber33 and Black rice) Number of reads were ranged from 69.1 to 66.9 million paired end reads. These can provide an estimated average coverage from 27 to 28X of the whole rice genome. This should be more than enough for chloroplast genome assembly assuming that the copy number of chloroplast could range from several to thousand copies per cell depending on tissues type and age. (13). The achieved Chloroplast average coverage was around 3000-9000X with minimum coverage; 27 and maximum 14000 base.

The dual pipelines that have been used for each sample have two rounds of mapping reads to reference with different setting (relax and strict) was very important to produce accurate consensus as much as possible. Both consensuses were compared and inspect carefully. Several variants have been recorded between them, wright call was made based on reads alignment in both setting (5). The final mapping to references compared with the output of Denovo pipeline. More variants were noted. Those were reviewed and corrected according to reads alignment files for both mapping and Denovo. Finally, the most accurate consensuses have been achieved for all samples. The enormous coverage of reads made judging and correcting these variants extremely straightforward process. Mixing both protocols had important role in rearrange contigs and reads on the right orientation and order and decrease assembling errors. Unfortunately, some chloroplast genome has been assembled without considering the repetitive nature of its genome IRA and IRB, which lead to confusing and misassembling for over hundred thousand of SNPs. This could be due to aligning reads or contigs to wrong position especially in the boundary regions of the IRA and IRB or flipping those fragments on wrong direction. Table 1 below shows the different between A and C pipeline and the amount of the improvement in the assembly consensuses by increasing these variants in (IBL, IDJ, IGA and IAN33) varieties when it has been compared sequences between mapping to reference vs Mapping-Denovo to reference. In other words, reduce the mapping bias by forcing reads to align to the reference. Apart from that the accession IBRQ which decrease the variants of Mapping to reference Vs Mapping-Denovo to reference from 921 to 505, which means there were misaligned reads were forced to the reference and caused more variations. Mapping-Denovo vs reference had removed those reads. The distribution of these variants was varied from cultivar to another. Some of them have majority on SNPs while the other was on InDels (insertion or Deletion) as in Table 2. The non-silent SNPs or the functional impact SNP in general were few.

Furthermore, SNPs and indels are also important in the nucleotide sequence evolution analysis of genomes. The calculation of single nucleotide variation may help estimate and comprehend genetic variations of different genome regions (14). Also were improved and decreased in Mapping-Denovo consensuses compare to just mapping to reference consensuses. Therefor it is important to use an appropriate double pipeline to assemble chloroplast genome and double check the fragments orientation and inspect the alignment physically and check the variations that produced from A, B and C pipelines and compare the way that the reads were aligned to reference in all alignment files to outweigh the more likely true aligned reads at each position. Utilizing this procedure will enhance the assembly and remove errors.

However, for phylogenomic study put too much effort to achieve most accurate sequence would not be account or make those big differences. To prove it, phylogenetic tree has been construct using PhyML package (12). PhyML is one of the most widely used phylogenetic tree reconstruction programs (15). and based on entire chloroplast sequence for all samples in addition to reference and five other Oryza sativa chloroplasts as indicate in figure 2 A and B. Trees show same topology and there is no difference between consensuses that produced from both pipelines. Both consensuses from each cultivar were clustered together clearly. This conclude that several differences from mapping pipeline and the accurate sequence from double pipeline would not much impact on the phylogenomic tree especially when conduct it on large sequences like chloroplast.

Table (1): Shows variant call before and after using denovo contigs to correct chloroplast consensus

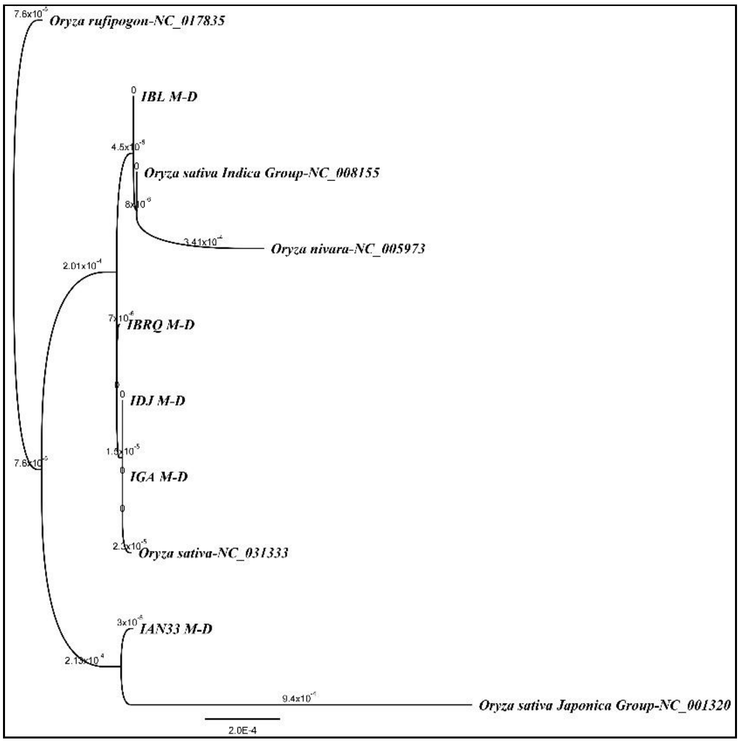
Variety	Mapping Vs Reference	The % of variance to the length of the chloroplast genome	Mapping-Denovo Vs Reference	The % of variance to the length of the chloroplast genome	Mapping Vs Mapping-Denovo	The % of variance to the length of the chloroplast genome
IAN3	599	0.004	874	0.006	275	0.002
IBL	873	0.006	1083	0.008	210	0.001
IBRQ	921	0.006	416	0.003	505	0.003
IDJ	928	0.006	1099	0.008	171	0.001
IGA	311	0.002	1097	0.007	786	0.005

*(IBL= Black rice, IDJ= Dijla, IBRQ= Baraka, IGA= Ghadir, IAN33= Amber33)

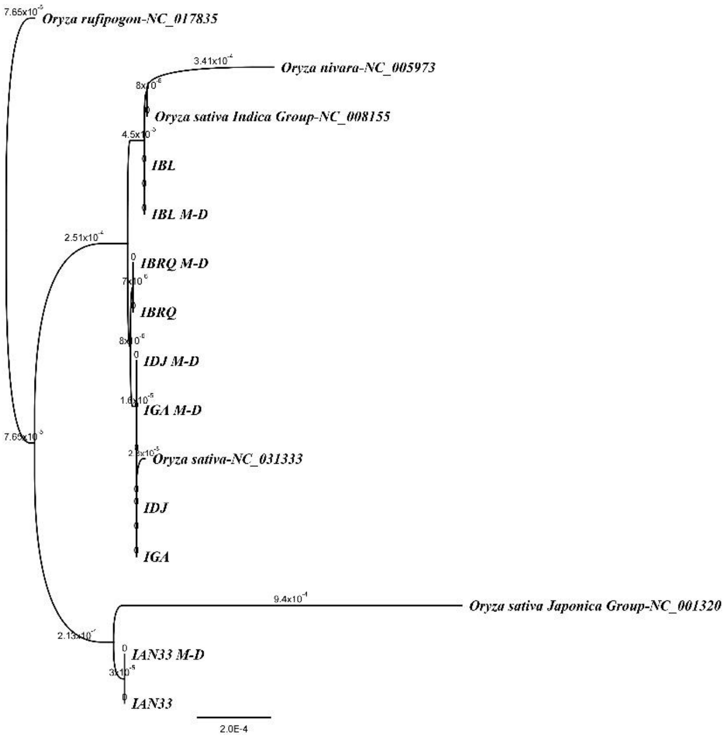
Table (2): Shows the variants distribution on silent and non-silent SNPs and InDels

Variety	InDel (insertion /deletion)	SNP	Non silent SNP
IAN33-M Vs Reference	272	327	36
IAN33-M-D Vs Reference	294	580	36
IBL-M Vs Reference	370	503	45
IBL-M-D Vs Reference	626	457	40
IBRQ-M Vs Reference	397	524	43
IBRQ-M-D Vs Reference	312	104	40
IDJ-M Vs Reference	399	529	44
IDJ-M-D Vs Reference	626	473	40
IGA-M Vs Reference	195	116	43
IGA-M-D Vs Reference	624	473	40

*(M=Mapping, M-D=Mapping-Denovo)



(A)



(B)

Figure (2): Shows chloroplast phylogenomic using PhyML tool. A: phylogenetic tree using C pipeline and B: phylogenetic tree using A pipeline. Both of them have same topology.

References

1. Sang, S.; Mei, D.; Zaman, Q. U.; Liu, J.; Cheng, H.; Fu, L., *et al.* (2020). An efficient approach for obtaining plant organelle genomes. Oil Crop Science, 5(3): 129-135.
2. Daniell, H.; Lin, C. S.; Yu, M. and Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology, 17(1): 1-29.
3. Tong, W.; Kim, T. S. and Park, Y. J. (2016). Rice chloroplast genome variation architecture and phylogenetic dissection in diverse Oryza species assessed by whole-genome resequencing. *Rice*, *9*(1): 1-13.
4. Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications. Genomics, Proteomics and Bioinformatics, 13(5), 278-289.
5. Moner, A. M.; Furtado, A. and Henry, R. J. (2018). Chloroplast phylogeography of AA genome rice species. Molecular Phylogenetics and Evolution, 127: 475-487.
6. Smith, D. R. and Keeling, P. J. (2015). Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. Proceedings of the National Academy of Sciences, 112(33): 10177-10184.
7. Cheng, L.; Nam, J.; Chu, S. H.; Rungnapa, P.; Min, M. H.; Cao, Y., *et al.* (2019). Signatures of differential selection in chloroplast genome between japonica and indica. Rice, 12(1): 1-13.
8. Saloom R. A. and Moner, A. M. (2022). Diversity study of several domesticated rice (local cultivars) cultivated in middle and south of Iraq using NGS technology.
9. Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4): 357-359.
10. Moner, A. M.; Furtado, A. and Henry, R. J. (2020). Two divergent chloroplast genome sequence clades captured in the domesticated rice gene pool may have

significance for rice production. BMC plant Biology, 20(1): 1-9.

11. Bushnell, B.; Rood, J. and Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. PloS One, 12(10):1-15
12. Guindon, S.; Dufayard, J. F.; Lefort, V.; Anisimova, M.; Hordijk, W. and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology, 59(3): 307-321.
13. Moner, A. M.; Furtado, A.; Chivers, I.; Fox, G.; Crayn, D. and Henry, R. J. (2018). Diversity and evolution of rice progenitors in Australia. Ecology and Evolution, 8(8): 4360-4366.
14. Zhang, T. T.; Yang, Y.; Song, X. Y.; Gao, X. Y.; Zhang, X. L.; Zhao, J. J., *et al*. (2021). Novel structural variation and evolutionary characteristics of chloroplast trna in gossypium plants. Genes, 12(6): 822.
15. Torres, M. and Silva, J. O. D. (2018). Parallel solution based on collective communication operations for phylogenetic bootstrapping in PhyML 3.0. In Brazilian Symposium on Bioinformatics (pp. 133-145). Springer, Cham.