Optimized Method of Extracting Rice Chloroplast DNA for High-Quality Plastome Resequencing and *de Novo* Assembly

Takeshi Takamatsu

Doctoral Program in Life and Food Sciences
Graduate School of Science and Technology
Niigata University

Contents

Abbreviations

BAM	binary alignment map
cpDNA	chloroplast DNA
HS	high salt method
InDel	small insertion/deletion
IRs	inverted repeats
LN	liquid nitrogen coupled with sucrose gradient centrifugation method
LSC	large single-copy region
Mt	mitochondria
mtDNA	mitochondrial DNA
MTPTs	mitochondrial plastid like sequences
ncDNA	nuclear DNA
NGS	next-generation sequencing
Pt	plastid
NUPTs	nuclear plastid DNA
PG	percoll gradient centrifugation method
qPCR	quantitative real-time PCR
SNP	single-nucleotide polymorphism
SSC	single copy region
tDNA	total DNA

INTRODUCTION

Chloroplasts, which are important cellular organelles that provide energy to plants, have an independent, circular, double-stranded DNA. The chloroplast genome (plastome), which ranges in size from 110 to 200 kb, consists of a pair of inverted repeats (IRs), a large single-copy region (LSC), and a small single-copy region (SSC). The first whole-chloroplast-genome sequencing of rice (Hirai et al., 1985; Hiratsuka et al., 1989), Arabidopsis (Sato et al., 1999), and maize (Maier et al., 1995), based on Sanger technology, provided a basic understanding of the genome's structure and function. Next-generation sequencing (NGS) technology now offers the promise to further the development of chloroplast genome studies, and has accelerated analyses to the point where more than 1600 accessions of complete chloroplast genome sequences from land plants are available in public databases (https://www.ncbi.nlm.nih.gov/genome/browse/?report=5). The advance of high-throughput, high-resolution analyses has facilitated population genetics and evolutional studies focused on the chloroplast genome (Morris et al., 2011; Tong et al., 2016; Wu et al., 2015). Organelle genomes have less diversity and exert less influence on phenotype than the nuclear genome (Drouin et al., 2008; Kahlau et al., 2006; Wolfe et al., 1987), although some studies have shown that variation in organelle genomes can influence variation in phenotypes (Joseph et al., 2013; Moison et al., 2010; Tang et al., 2013). Joseph et al. (2013) demonstrated that the cytoplasmic genome plays a central role in controlling natural variations in metabolomic networks within a reciprocal Arabidopsis Kas × Tsu recombinant inbred line population. Roux et al., 2016 reported cyto-nuclear co-adaptation by creating a unique series of 56 cytolines resulting from cytoplasmic substitutions among eight natural Arabidopsis species. More recent widespread reports show that interactions between nuclear and cytoplasmic genomes shape natural variation (Greiner and Bock, 2013; Roux et al., 2016; Sloan, 2015). This factor gives rise to new opportunities for using organelle genomes as novel breeding targets (Maliga, 2001; Wang et al., 2008). The plastome thus presents an attractive target for genome engineering and a promising alternative to nuclear transformation (Olejniczak et al., 2016). Indeed, chloroplast genomic studies have crucial translational and biotechnological applications owing to the genome's ability to express >120 foreign genes from different organisms (Daniell et al., 2016). Accordingly, the efficient sequencing of plastid genomes, which requires highly purified plastid DNA, will allow the production of transplastomic plants (Diekmann et al., 2008).

Rice, one of the most important food crops in the world, has important syntenic relationships with the other cereal species and is a model for monocots and grasses. The chloroplast genome of Oryza sativa is 134 525 base pairs long, with 159 unique genes, including 38 tRNA, 8 rRNA, and 108 protein-coding genes (Hiratsuka et al., 1989). The evolutionary transfer of plastid DNA fragments to the nuclear and mitochondrial genomes is frequently found in plants (Lewin, 1984; Martin and Herrmann, 1998; Matsuo et al., 2005). Such transfers to the nuclear genome (nuclear plastid DNA, NUPTs) and to the mitochondrial genome (mitochondrial plastid-like sequences, MTPTs) are more abundant in rice than in other higher plants (Yoshida et al., 2014). As plastome sequencing is frequently based on total DNA, this might reduce the mapping accuracy owing to the difficulty in selecting plastid-derived reads from the whole-genome sequence, which includes NUPT- and MTPT-derived reads, obtained by short read sequencing. In addition, total cellular DNA contains only 1% to 10% cpDNA, and the amount decreases during plant development (Baumgartner et al., 1989; Oldenburg and Bendich, 2004; Oldenburg and J.Bendich, 1991; Shaver et al., 1995). These problems reduce the efficiency of sample multiplex analysis in a single sequencing run on low-output sequencers such as MiSeq, MiniSeq, and PacBio RS II. Therefore, the isolation of high-purity cpDNA will improve the accuracy and cost-efficiency of plastome sequencing. Four procedures for cpDNA isolation have been reported: Percoll density gradient centrifugation to obtain intact chloroplasts free of other organelles (Kaneko et al., 2016; Lang and Burger, 2007); high salt concentration to remove contaminating DNA ionically attached to the chloroplast surface (Shi et al., 2012); the use of DNase I to digest DNA bound to the chloroplast surface (Kolodner and Tewari, 1979); and liquid nitrogen pre-treatment to prevent nuclear breakage (Hirai et al., 1985).

Shi et al. (2012) reported that the DNase I method digested not only the contaminating DNA but also cpDNA within chloroplasts. So we compared Percoll density gradient centrifugation (PG), high salt (HS), and liquid nitrogen coupled with a sucrose gradient (LN) to optimize cpDNA analysis by NGS. We sequenced the highly purified cpDNA to compare the advantage of cpDNA sequencing with whole-genome sequencing. We also assessed the performance of single-nucleotide polymorphism (SNP) / insertion–deletion (indel) calling and de novo assembly on the plastome to evaluate the effect of cpDNA purity on NGS analysis.

MATERIALS AND METHODS

Plant Material

We germinated seeds of several rice (*O. sativa*) cultivars: temperate japonica Nipponbare (Shiga Prefecture Agricultural Research Center) and Koshihikari (Niigata Agricultural Research Institute); tropical japonica Sensho, Urasan, Padi Perak, and Khao Nok (NARO, Genetic Resources Center); aus Chinsurah Boro 2 (Tohoku University) and Kasalath; and indica 93-11 (National Institute of Genetics).

Protocols for Chloroplast Isolation

The three techniques selected for cpDNA isolation are described as fellow: To determine the best method, we first performed three independent experiments to extract the cpDNA from bulk samples of Nipponbare by using each method (LN, HS, PG method). Then, the cpDNA from bulk samples of other cultivars were extracted once or twice by LN method.

Updated Liquid Nitrogen – Sucrose Density Gradient Centrifugation (LN) Method. Rice seeds were sterilized and germinated on moist filter paper in culture dishes in the dark at 28 °C for up to 4 days. The germinated seeds were dispersed onto a layer of water-agarose gel laid to block contamination by fungi over 1/2 MS medium agarose gel. Plants were placed in a growth chamber (28:23 °C, 12:12 h, light:dark; 20 000 lux) for 8 d. cpDNA was isolated as described (Hirai et al., 1985) with some modifications (**Figure 1**). All procedures were performed at 4 °C, and all centrifugations were performed in a CP80NX ultracentrifuge (Hitachi) with a P40ST swing rotor and 13PA tubes. In brief, 50 g of fresh shoot was cut into pieces (~3 cm), frozen in liquid nitrogen for 3 min, and gently ground into a fine powder with a mortar and pestle. The material was suspended in 400 mL of isolation buffer (50 mM Tris·HCl pH 8.0, 0.35 M sucrose, 7 mM EDTA, 5 mM 2-mercaptoethanol, 0.1% BSA) and incubated for 5 min in the dark. The suspension was filtered through two layers of gauze and then two layers of Miracloth (Merck). The filtrate was centrifuged at 1000× *g* for 10 min. The pellet was suspended in 5 mL of isolation buffer, and the suspension was loaded slowly onto a stepwise 20%/45% discontinuous sucrose density gradient in 50 mM Tris·HCl (pH 8.0), 0.3 M sorbitol, and 7 mM EDTA. The gradient was centrifuged at 2000×*g* for 30 min in a swinging bucket rotor. The green band at the 20%/45% sucrose interface was collected, diluted with three volumes of isolation buffer, and centrifuged at 3000×*g* for 10 min in a

swinging bucket rotor. Genomic DNA was isolated from the pellet by the cetyltrimethylammonium bromide method (Shi et al., 2012) or with a DNeasy Plant Mini Kit (Qiagen).

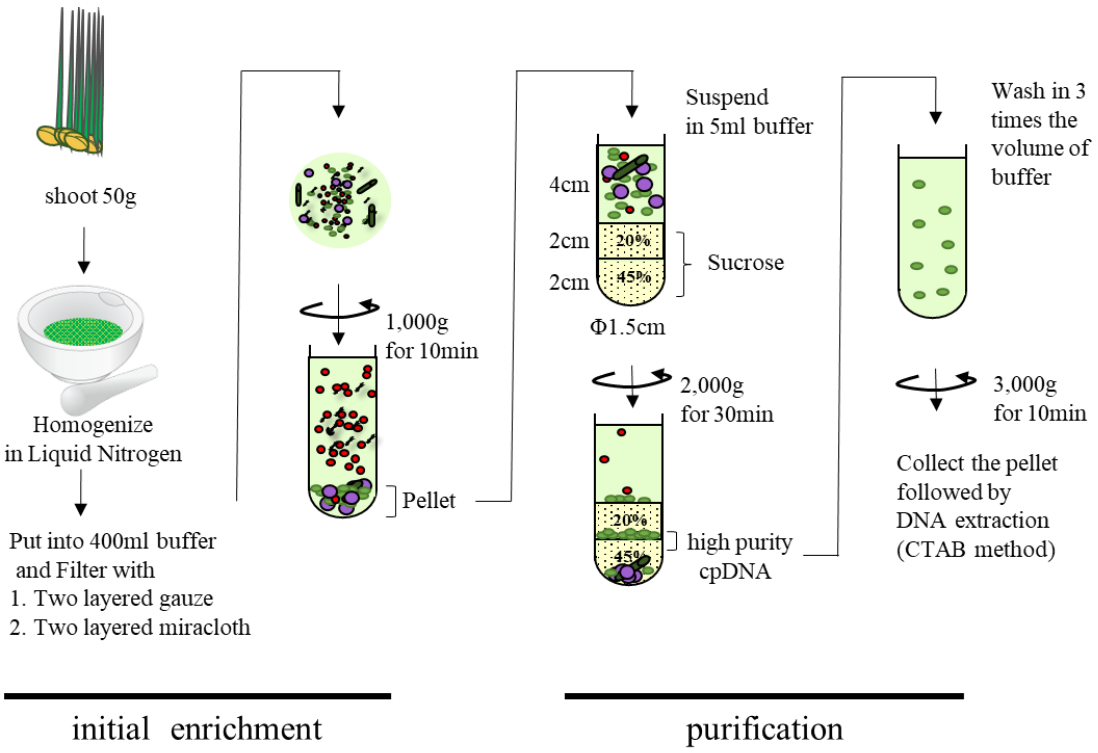


FIGURE 1 | Flowchart of chloroplast DNA isolation using liquid nitrogen coupled with sucrose gradient centrifugation.

Modified High Salt (HS) Method. Sterilized seeds were grown on 0.8% agar in the dark at 28 °C for 7 days and then in natural daylight at 28 °C for 14 days. cpDNA was isolated using the protocol described in Shi et al., 2012 (**Figure 2**). All procedures were performed at 4 °C. In brief, 20 g of fresh leaves was cut into small pieces (~1 cm) and homogenized for 30 s in 400 mL of buffer A (1.25 M NaCl, 0.25 M ascorbic acid, 10 mM sodium metabisulfite, 12.5 mM borax, 50 mM Tris·HCl pH 8.0, 7 mM EDTA, 1% [w/v] PVP-40, 0.1% [w/v] BSA, and 1 mM DTT). The homogenate was filtered through two layers of Miracloth, and the filtrate was centrifuged at 200× *g* for 20 min to remove starch granules, nuclei, tissue debris, and aggregates. The supernatant was centrifuged at 3500× *g* for 20 min, and the pellet was suspended in 250 mL buffer B (1.25 M NaCl, 12.5 M borax, 1% [w/v] PVP-40, 50 mM Tris·HCl pH 8.0, 25 mM EDTA, 0.1% [w/v] BSA, and 1 mM DTT) to increase the purity of the isolated cpDNAs. This step was performed twice. cpDNA was extracted with a DNeasy Plant Mini Kit (Qiagen).

Percoll Gradient (PG) Centrifugation Method. Sterilized seeds were grown on 0.8% agar at 30 °C in the dark for 11 days, and then under continuous light at 28 °C for 3 days for greening. cpDNA was isolated as in our previous report (Kaneko et al., 2016) (**Figure 2**). All procedures were performed at 4 °C. In brief, rice shoots (20 g) were homogenized in 20 mL of isolation buffer (50 mM HEPES-KOH pH 7.5, 0.33 M sorbitol, 5 mM MgCl2, 5 mM MnCl2, 5 mM EDTA, and 50 mM sodium ascorbate). The homogenate was filtered through four layers of gauze and then four layers of Miracloth. The filtrate was layered onto a cushion of 80% (v/v) Percoll (Sigma) in the above isolation buffer (except for the sodium ascorbate) and centrifuged at 2000× *g* for 4 min. The crude chloroplasts on the Percoll surface were collected and diluted with more than twice the volume of isolation buffer, and then layered onto a discontinuous density gradient of 40% and 80% Percoll. The gradient was centrifuged at 4000× *g* for 10 min. Intact chloroplasts enriched around the 40%/80% Percoll interface were collected and centrifuged again as before. Intact chloroplasts were collected, diluted with five volumes of isolation buffer, and centrifuged at 2000× *g* for 4 min. cpDNA was extracted with a DNeasy Plant Mini Kit (Qiagen).

In all three protocols, DNA were eluted from the DNeasy spin column using 80μl (LN and PG method) or 40μl (PG method) of elution buffer. 2μl of the isolated cpDNA was loaded and separated in 0.8% agarose gel and visualized with ethidium bromide.

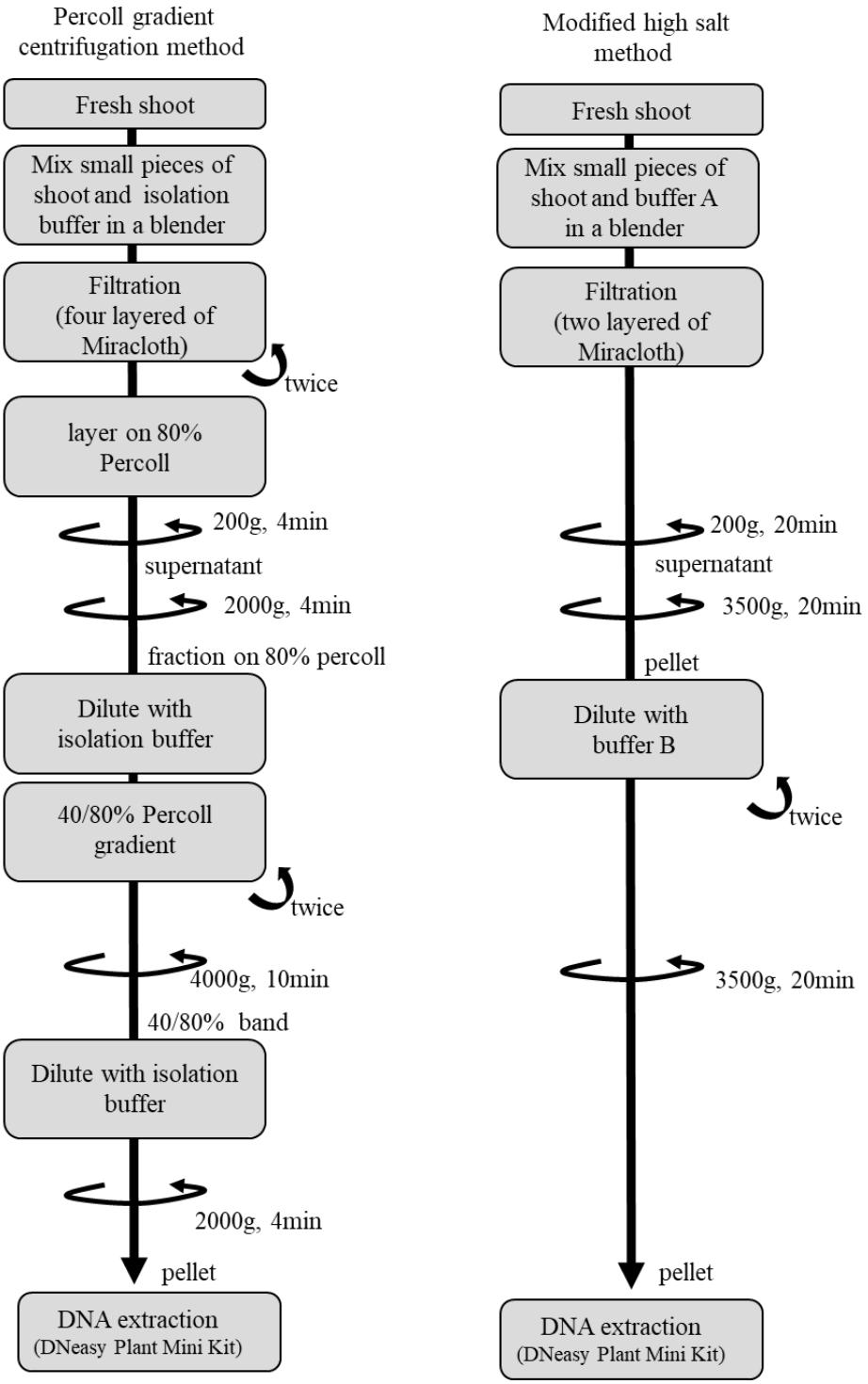


FIGURE 2 | Flowchart for chloroplast DNA isolation using Percoll gradient centrifugation method and Modified high salt method

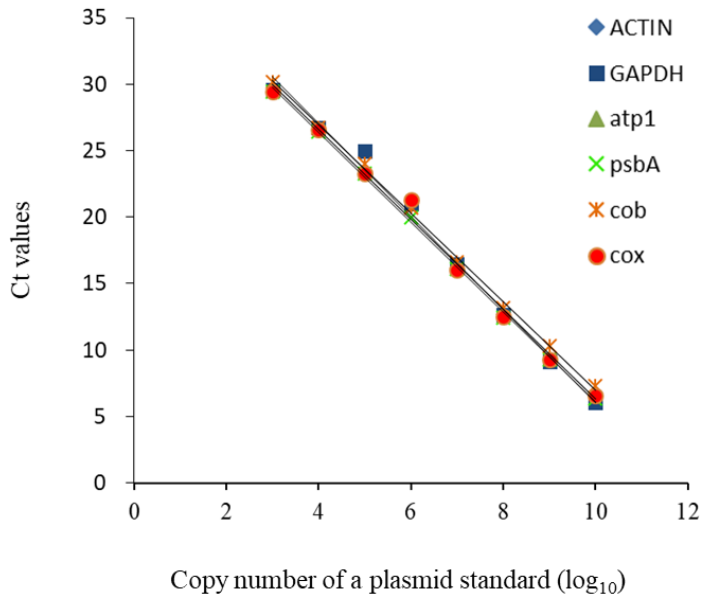
Genome Copy Number Analysis by qPCR

We tested the cpDNA purity of DNA isolates by quantifying the number of copies of chloroplast, mitochondrial, and nuclear genomes. cpDNA, mtDNA, and ncDNA were quantified by quantitative real-time PCR (qPCR) using SsoFast EvaGreen Supermix (Bio-Rad) on a CFX96 real-time PCR system/C1000 Thermal Cycler (Bio-Rad). The thermocycling conditions were denaturation at 98 °C for 2 min, and 39 cycles of 98 °C for 2 s and 60 °C for 5 s. To analyze genome copy number, we designed two sets of primers for each genome (**Table 1**) to improve the accuracy of quantification, and used the mean as the threshold cycle (Ct) value of each genome. First, we calculated the PCR amplification efficiencies of each primer pair with a dilution series of a plasmid standard (10^4 to 10^9 copies of pGEM-T::Actin1::GAPDH::atpI::psbA::cob::coxII). The amplification efficiencies of all six genes were close to 2.0 (1.93–2.03), and R^2 values were between 0.989 and 0.999 (**Figure 3A**). Next, we calculated the copy ratios of cpDNA/ncDNA and cpDNA/mtDNA from the standard curves drawn from the above dilutions and compared these results of absolute quantification with the results of relative quantification using the $2^{-\Delta\Delta ct}$ method (Schmittgen and Livak, 2008). Since the results of relative quantification were almost the same as those of absolute quantification, we used the relative quantification method in the subsequent qPCR analysis, giving priority to efficiency (**Figure 3B**). Samples were assayed in at least three technical replicates, and the average copy ratios were calculated. Then, each genome DNA content in the extracted DNA was estimated from the copy ratio and genome size (plastid, 134 525 bp; mitochondrial, 490 520 bp; nuclear, 373 245 519 bp).

TABLE 1 | Primers used in qPCR for genome copy analysis.

	Gene Symble	ACCETION	Foword Primer	Reverse Primer	amplification test	
					Amplification Efficiency	R^2
Nuclear genome						
	Actin1	Os03g0718100	GGATATGCTCTCCCCCATGC	TCCCTCACAATTTCCCGCTC	1.97	0.993
	GAPDH	Os02g0171100	CGCCAAGCACTGATTTGTTGAAA	GTTCCGCTTGCCCAGGTC	1.93	0.992
Plastid genome						
	atp Ⅰ	CAA33990.1	CCACAAACCATCCCAACCGA	AAAGAGCACCCGACCAGTTC	2.03	0.989
	psbA	CAA34007.1	ACATCGGATGGTTCGGTGTT	GATCGCCGCAGAAGTAGGAA	1.98	0.999
Mitochondrial genome						
	cob	BAC19890.2	TCCTAATGTTTTGGGGCATC	AGAATGGCATGGATCGGTAG	2.00	0.998
	cox Ⅱ	BAC19876.1	GCCAGAAACGGAGAGTTGAG	TCGTATATCGCTCCACCACA	1.97	0.993

A.



B.

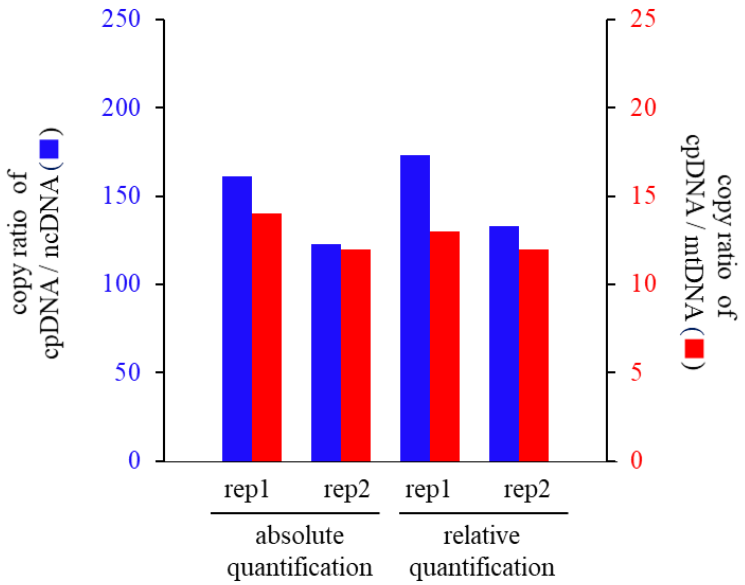


FIGURE 3 | Comparison of qPCR genome copy analysis using absolute and relative quantification. (A) Standard curves of six primers were drawn with a dilution series of a plasmid standard (10^4 to 10^9 copies of pGEM-T::Actin1::GAPDH::atpI::psbA::cob::coxII). (B) Genome copy ratio of leaf total DNAs were assessed by absolute quantification using a standard curve(A) and by relative quantification using the $2^{-\Delta\Delta ct}$ method. Leaf total DNAs were extracted twice (rep1 and rep2).

DNA Library Construction and Next-Generation Sequencing

We used 1 ng of Nipponbare purified cpDNA as input for the Nextera XT DNA library preparation kit and the Nextera XT index kit (Illumina). Constructed libraries were sequenced on an Illumina MiSeq sequencer (300 bp paired-end) with the MiSeq reagent kit v3 (Illumina). All procedures followed the manufacturer's instructions.

Sequence Mapping and Variant Detection

The sequenced paired-end reads from the purified cpDNA and three sets of whole-genome sequencing reads (total DNA [tDNA] 1–3) downloaded from public databases were trimmed in Trimmomatic v. 0.33 software (Bolger et al., 2014) with the following parameters: SLIDINGWINDOW: 8:20; TRAILING: 20; MINLEN: 90 (tDNA 1), 100 (purified cpDNA and tDNA 2), 76 (tDNA 3). The processed reads were aligned to the rice plastid reference genome (X15901.1) or a combined plastid (X15901.1)-mitochondrial (BA000029) reference genome by using the BWA-MEM v. 0.7.15 algorithm (Li and Durbin, 2009) with default parameters. PCR duplicates in binary alignment map (BAM) files were marked with Picard tools v. 1.68 software (http://broadinstitute.github.io/picard/). Then local realignment of reads around indels was done in GATK (Genome Analysis Toolkit) IndelRealigner software (DePristo et al., 2011). To estimate cpDNA purity, we extracted unaligned reads from BAM files in SAMtools software (Li et al., 2009) and re-aligned them on the rice mitochondrial reference genome (BA000029). These unmapped hits were extracted again and realigned on the rice nuclear reference genome (IRGSP-1.0). The coverage depth of each genome was calculated from the number and length of high-quality reads in a 250-nt sliding window. To calculate allele frequency at individual plastid genome positions, we generated wig files describing the base (A/C/G/T) content in a 1-nt sliding window from BAM files in igvtools v. 2.2 software (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). After removal of data neighboring indels because of low reliability, we calculated 1st and 2nd allele frequencies and coverage depths from wig files with a custom Perl script and then visualized them in 3D scatter plots using the scatterplot3d v. 0.3-37 tool of R (Ligges and Mächler, 2003). For variant calling, we used the SAMtools mpileup v. 1.4.1 tool (Li et al., 2009) with default parameters and GATK HaplotypeCaller v. 3.6 software (DePristo et al., 2011) with the '-ploidy 1' parameter to compare SNPs and small indels from BAM files. We filtered out heterozygous and low-quality variants (QUAL < 20) in SAMtools, and low-quality

variants (QUAL < 20) in GATK.

De Novo Assembly

PCR duplicates were removed from paired-end reads using the *k*-mer-based method implemented in a Perl script (https://github.com/linneas/condetri). From paired-reads of total DNA in BAM files, aligned plastid genome reads were extracted for enrichment of plastid reads, and then PCR duplicate reads were filtered out. Contigs were assembled from these reads in SOAP-denovo2 software (Luo et al., 2012) with various sets of *k*-mer parameters (**Table 7**). After assembled scaffolds shorter than 500 bp were filtered out, sequences were compared against the plastid reference genome by NCBI BLAST 2 (Tatusova and Madden, 1999). Alignment results and detected SNPs/indels were visualized by Circos software (Krzywinski et al., 2009).

Accession codes

The sequence data have been deposited in the DDBJ Sequence Read Archive: DRR118684.

RESULTS AND DISCUSSION

Comparison of Chloroplast DNA Isolation Methods for NGS Sequencing

Many areas of chloroplast research require assessment of the quality and quantity of cpDNA. Although assessment several methods have been developed, there is little discussion in the literature of whether they provide similar quality of information (Rowan et al., 2009; Rowan and Bendich, 2011). Large-scale studies of genome variation and evolution, which rely on large quantities of plant material, also require cpDNA isolation. Here we compared three methods of isolating cpDNA (**Table 2**): updated liquid nitrogen – sucrose density gradient centrifugation (LN), high-salt buffer (HS), and Percoll gradient centrifugation (PG). The original LN protocol was improved at various steps (Hirai et al., 1985) associated with gradient centrifugation to isolate cpDNA (**Figure 1**). The main improvements, made to allow the use of general laboratory instruments and to simplify the method, were the use of a mortar and pestle instead of a liquid nitrogen–resistant mixer, and omission of 65% sucrose following a preliminary experiment that showed heavy contamination of nuclear and mitochondrial DNA at the 45%/65% interface. The protocol was further extended to the isolation of cpDNA from several rice accessions from four cultivar groups.

qPCR was used to determine the purity of the cpDNA samples for NGS. The HS method enriched not only cpDNA but also mtDNA (**Figure 4**), which may reduce cpDNA purity. The PG method enriched cpDNA with minimal contamination of mtDNA, but still with significant amounts of ncDNA. On the other hand, the LN technique consistently gave high purity, with cpDNA copy number ratios of 24 643× ncDNA and 155× mtDNA **(Figure 4A**, LN-Nip). A high ratio is critical to reducing misaligned reads on the plastome, such as NUPTs or MTPTs, and to increasing NGS accuracy. Moreover, cpDNA accounted for up to 88% of the total isolated DNA (**Figure 4B**), which is likely to provide meaningful cost-effectiveness of an NGS run. The LN method yielded high-purity cpDNA in all nine cultivars tested. These results show the potential for the LN method to be applied to a wide range of rice cultivars, as we have since confirmed (unpublished data). In the updated LN method, isolated DNA gave a well defined electrophoresis band, which is indicative of undegraded DNA (**Figure 5**). This DNA could be of sufficient quality not only for short-read sequencing, but also for mate-pair and long-read sequencing. This DNA could be of sufficient quality not only for short-read sequencing but also for mate-pair and long-read

sequencing. The high-quality pure cpDNA isolated from LN method would be suitable to facilitate the plastome sequencing using the novel third-generation DNA sequencing technologies such as PacBio RS II and Oxford Nanopore MinION. This process may overcome the hard-to-assembly IRs regions composed of more than 20 000 base pairs long, and can enable new insights such as large indels and structural variations. A large amount of chloroplast pellet was collected, from which abundant cpDNA was extracted: 800 ng of cpDNA from 50 g of shoot (**Table 2**). We also confirmed that as little as 4 g of shoot gave enough cpDNA for our NGS platform (Illumina MiSeq coupled with Nextera XT DNA library preparation) (**Figure 6**). However, the DNA obtained from the HS and PS protocols displayed very weak bands and smeared, indicative of low DNA yield (**Figure 5, Table 2**).

Table 2 | Summary of the materials used and the cpDNA yield in the three different cpDNA isolation methods.

		DNA	Materials	
	Time* (hr)	Yield (ng/g fresh weight)	Tissue	Age
LN method	2.5	16	shoot	12 day-old seedling including germination for 4 days (this study)
HS method	1.8	6	leaf	21day-old seedling including germination for 7 days (Shi et al. 2012)
PG method	3.5	13	shoot	11 day-old seedling (8 day-old etiolated seedling, then greening for 3 days) (Kaneko et al. 2016)

HS, high salt; PG, Percoll gradient centrifugation; LN, liquid nitrogen – sucrose gradient centrifugation.

*The experimental time was estimated from the start of tissue cutting to obtaining cpDNA pellet.

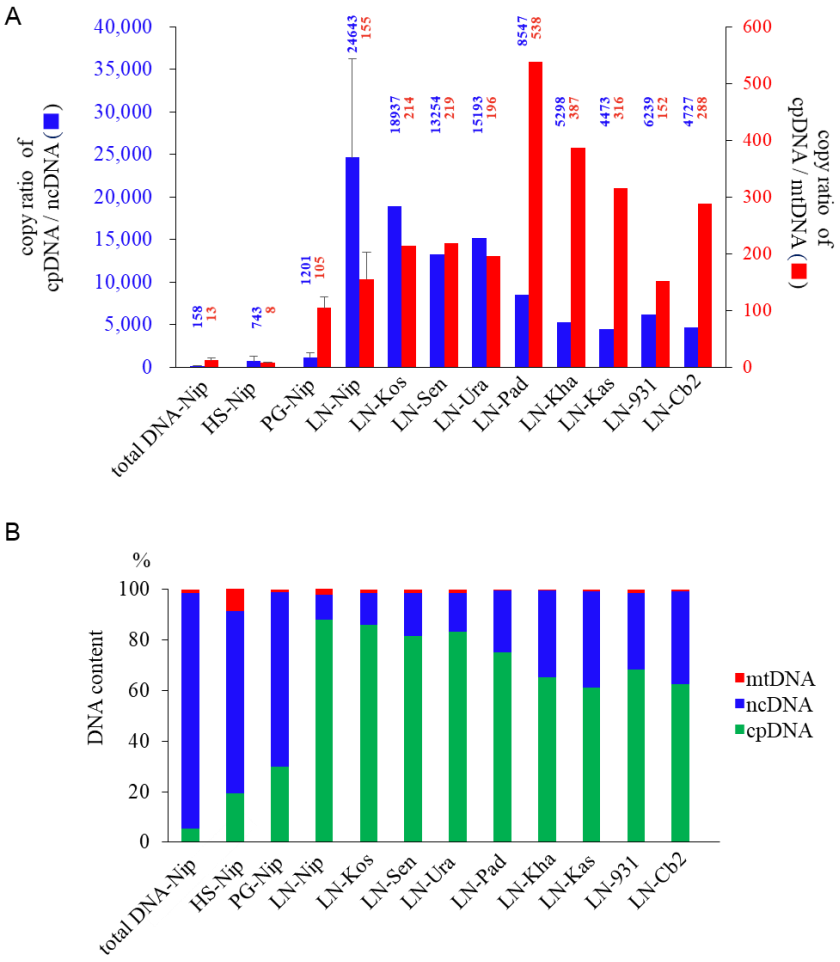


FIGURE 4 | Evaluation of three cpDNA isolation methods. qPCR results of extracted cpDNA: (A) copy number ratio; (B) genomic DNA component rate. HS, high salt; PG, Percoll gradient centrifugation; LN, liquid nitrogen – sucrose gradient centrifugation. Nip, Nipponbare; Kos, Koshihikari; Sen, Sensho; Ura, Urasan; Pad, Padi Perak; Kha, Khao Nok; Kas, Kasalath; 931, *indica* 93-11; Cb2, Chinsurah Boro 2.

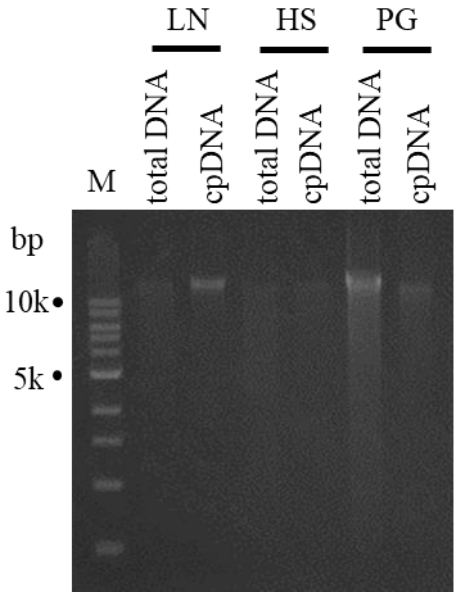


FIGURE 5 | Chloroplast DNA visualization in agarose gel. Total DNA and cpDNA prepared by LN, HS, and PG protocols were subjected to 0.8% TAE agarose gel electrophoresis. M represents DNA size markers.

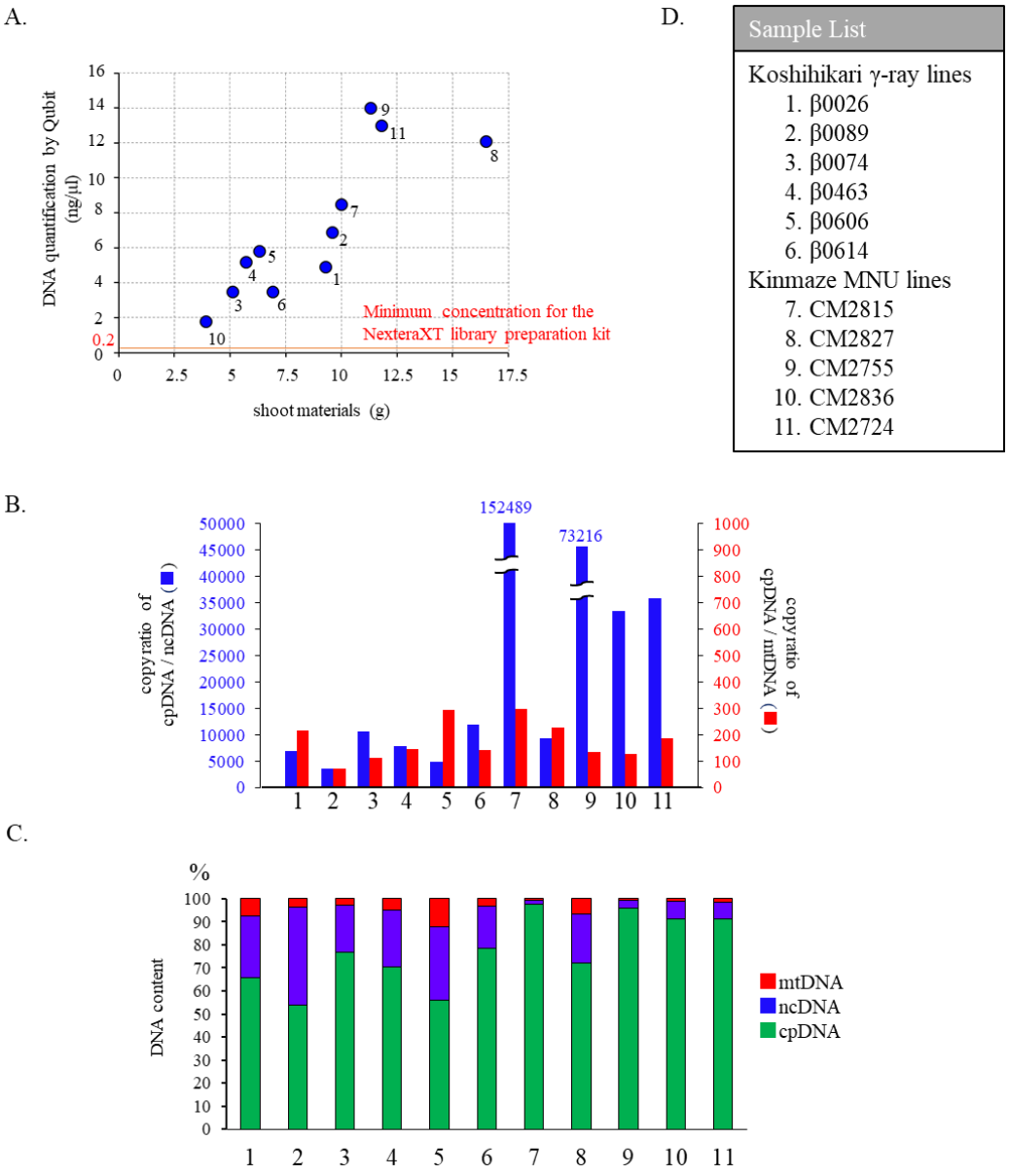


FIGURE 6 | cpDNA yield analysis of the Liquid Nitrogen coupled with sucrose-density-gradient Method. We used DNeasy Plant Mini Kit (Qiagen) for DNA extraction from the cpDNAs pellets, and finally eluted DNA with 30 μ l elution buffer. The relationship between DNA concentration and amounts of shoot materials is shown in scatter plots(A). qPCR results of cpDNA purity exhibit at copy ratio (B) and genomic DNA component rate (C). cpDNA, mtDNA, ncDNA represent chloroplast DNA, mitochondrial DNA, and nuclear DNA, respectively. The plot numbers(A) and column numbers correspond to the sample list(D). MNU represents N-methyl-N-nitrosourea.

Assessments of Resequencing: Mapping and Variant Calling

Illumina sequencing produced 4 105 152 paired-end reads with an average read length of 175 bp and a total of 718 401 600 bases (**Table 3**). We downloaded three sets of whole-genome sequence data (tDNA 1–3) from the public database to assess the influence of cpDNA purity on NGS analysis (**Table 3**). The sequence reads were aligned to the Nipponbare plastid reference genome (X15901.1) to discover putative SNPs and small indels. Over 78% of reads were aligned in the cpDNA purified by the LN method, versus only 1.2% to 10.2% in the tDNA (**Table 3**), corresponding to ratios of 7.7× to 66× tDNA. The plastid gemone (pt) was highly enriched in the cpDNA compared with those in three tDNAs (**Table 4)**.

Massively parallel sequencing gave an increase of at least fivefold in cpDNA purity compared with leaf tDNA (range, <3% to 30%) (Nock et al., 2011). The LN protocol is thus a feasible replacement for the PG and HS methods. The HS method isolated only pea cpDNA (Bookjans et al., 1984), and its improvement (Shi et al., 2012) did not improve purity. Hirao et al., 2008 considered the use of sucrose density gradient centrifugation as the best method for separating ncDNA contamination from cpDNA, which qPCR results (**Figure 4**) and the Illumina sequencing (**Table 4**) strongly supported, indicating high enough yield and purity to perform subsequent resequencing and genome assembly. The small difference between the results of mapping and qPCR analysis is likely due to the calculation methods, since qPCR calculates copy ratio from two target regions on each genome, whereas NGS aligns sequences across the whole genome. Visualization of the read alignment shows that the plastid reference genome is sufficiently covered in each sample (**Figure 7**). While tDNA samples show uniform coverage depth across the reference genome, the purified cpDNA shows irregular depth. The inconsistency would arise from the properties of the Nextera XT library kit, since a high number of PCR enrichment cycles can easily cause PCR-dependent coverage bias (Lan et al., 2015).

Next, we analyzed misaligned NUPT and MTPT reads, which disrupt precise analysis, on the plastid genome. We considered that MTPT reads containing a small number of mismatches against plastid genome could be separated by aligning reads to a combined plastid–mitochondria reference genome. Indeed, tDNA showed differences in coverage depth between the Pt and Pt–Mt reference genomes within regions of high

similarity between the genomes, while purified cpDNA showed approximately similar coverage depth across the whole plastid genome (**Figure 7**; e.g., Region 1). To further detect NUPT- and MTPT-derived noise, we plotted the 1st and 2nd allele frequencies and coverage depth at individual base positions on the plastid genome in 3D graphs (**Figure 8**). In the purified cpDNA, the 1st allele frequencies are close to 1.0 across the genome except in positions of low coverage depth, indicating low contamination by mitochondrial or nuclear DNA. By contrast, tDNA shows several sites where the 1st allele frequency was reduced and one or more other alleles were detected, even in positions with deep coverage depth. This result corroborates the frequency of this tendency and the lower purity of cpDNA (**Table 4**). Using the combined Pt–Mt reference genome improved the percentage of 1st allele frequencies. This result suggests that 1st allele reductions result from contamination by MTPTs, supporting observations of coverage depth differences such as in Region 1 in **Figure 7**. Furthermore, these results demonstrate that aligning on the combined Pt–Mt reference genome enables reduction of mtDNA contamination by a computational approach. However, it seems that the remaining low 1st allele frequencies are subject to noise derived from ncDNA. It is difficult to remove ncDNA-derived contamination by computation because some rice NUPTs have the same sequence as in the complete plastid reference genome and occur in multiple copies in the nuclear genome (Matsuo et al., 2005). The above results indicate that purified cpDNA can lead to high coverage depth of the chloroplast genome with a low number of reads, providing robust mapping and high-throughput sequencing of the rice plastid genome. Using tDNA sequenced on the Illumina platform is not consistently reliable, showing a higher rate of error alignment, which will likely affect later analysis such as plastid genome assembly or variant detection. It is important to highlight here the power of this technique in isolating cpDNA with improved data quality and lowered sequencing costs.

The rice Nipponbare plastid reference genome X15901.1 was released in 1989 (Hiratsuka et al., 1989). Later, Tang et al. (2004) independently published a Nipponbare plastid reference genome, AY522330, noting 79 SNPs and 110 indels of putative sequence errors. To assess the inference of misaligned reads from the NUPTs and MTPTs by detection of these sequencing errors, we used AY522330 as correct data, and processed the BAM files describing the read alignments on X15901.1 to detect variants using two distinct variant callers, SAMtools mpileup and GATK HaplotypeCaller (**Table 5** and **Table 6**). GATK HaplotypeCaller had much higher sensitivity than SAMtools mpileup, as previously reported (Liu et al., 2013; Pirooznia et al., 2014; Yi et al., 2014). SAMtools mpileup failed to detect most indels (**Table 6**), having very low

sensitivity in indel calling (Tian et al., 2016). As sample purity decreases, SAMtools detects more heterozygous variants, which may be false positives, as they don't exist in AY522330. Although it is possible to select highly reliable variants by filtering out heterozygous and low-quality variants, some false-positive variants were found in all samples (e.g., plastid genome sites at 44772, 44775, 70290, and 70291; **Table 6**). On the other hand, all variants identified by GATK HaplotypeCaller are consistent with AY522330 and show a high-quality score tolerant to the filtering process, regardless of sample purity, and no difference was found in the number of variants within the four samples. These results suggest that GATK HaplotypeCaller is superior at detecting mutations with high accuracy, not only in purified cpDNA, but also in total DNA samples, in the resequencing analysis of the plastid genome. However, GATK identified only 129 variants out of 189 reported in AY522330 (Tang et al., 2004). Most of those missing variants lie within the two IRs (**Table 6** and **Figure 9**), since GATK HaplotypeCaller ignores low-mapping-score reads (e.g., multi-mapping reads). Similarly, population genetic analysis of chloroplasts in 383 rice varieties also showed lower SNP/indel density within the IRs than in other plastid genome regions (Tong et al., 2016). Our efforts to detect variants in the IRs by changing several parameters of GATK HaplotypeCaller to allow low-mapping-score reads did not resolve this issue (data not shown). Although GATK HaplotypeCaller is the current gold standard variant caller, our results show that the development of other computation approaches is required for plastome resequencing. The discrepancies between our purified cpDNA and total DNA studies shed light on the importance of plant plastidial studies to thoroughly describe how to map them.

TABLE 3 | Summary of NGS samples and aligned results.

	Library			Reads (After_QC)			Aligned Reads		
DNA	SRA	Instrument	Layout	Read1	Read2	# Reads	# pt	# pt_uniq	% pt
purified cpDNA	this study	Miseq	Paired	204	146	4,105,152	3,233,842	3,230,265	78.8%
total DNA_1	SRR1239746	Hiseq2000	Paired	90	89	10,595,500	1,076,775	1,003,940	10.2%
total DNA_2	SRR1614244	Hiseq2000	Paired	101	101	70,075,112	927,471	898,850	1.3%
total DNA_3	SRR077421 SRR077422 SRR077425	GA Ⅱ	Paired	76	76	55,180,822	673,954	648,676	1.2%

Purified cpDNA: Nipponbare cpDNA obtained by LN method (**Figure 4A, B**, column 4). Total DNAs: published Nipponbare whole-genome sequencing reads downloaded from public database. Reads 1 and 2 indicate mean lengths of reads.

TABLE 4 | cpDNA purity: coverage depths and copy ratio of plastid (pt), mitochondrial (mt), and nuclear (nc) genomes of the purified chloroplast DNA (cpDNA) and three total genomic DNAs.

	Depth			Copy ratio	
	pt	mt	nc	pt/nc	pt/mt
purified cpDNA	3071	67	0.30	10172	45.7
total DNA_1	668	44	2.22	301	15.3
total DNA_2	1094	187	18.3	60	5.8
total DNA_3	366	66	11	33	5.5

cpDNA purity was calculated from coverage depth of each genome. After alignment on the plastid genome, unmapped reads were realigned on the mitochondrial genome, then the unmapped reads on the mitochondrial genome were realigned on the nuclear genome.

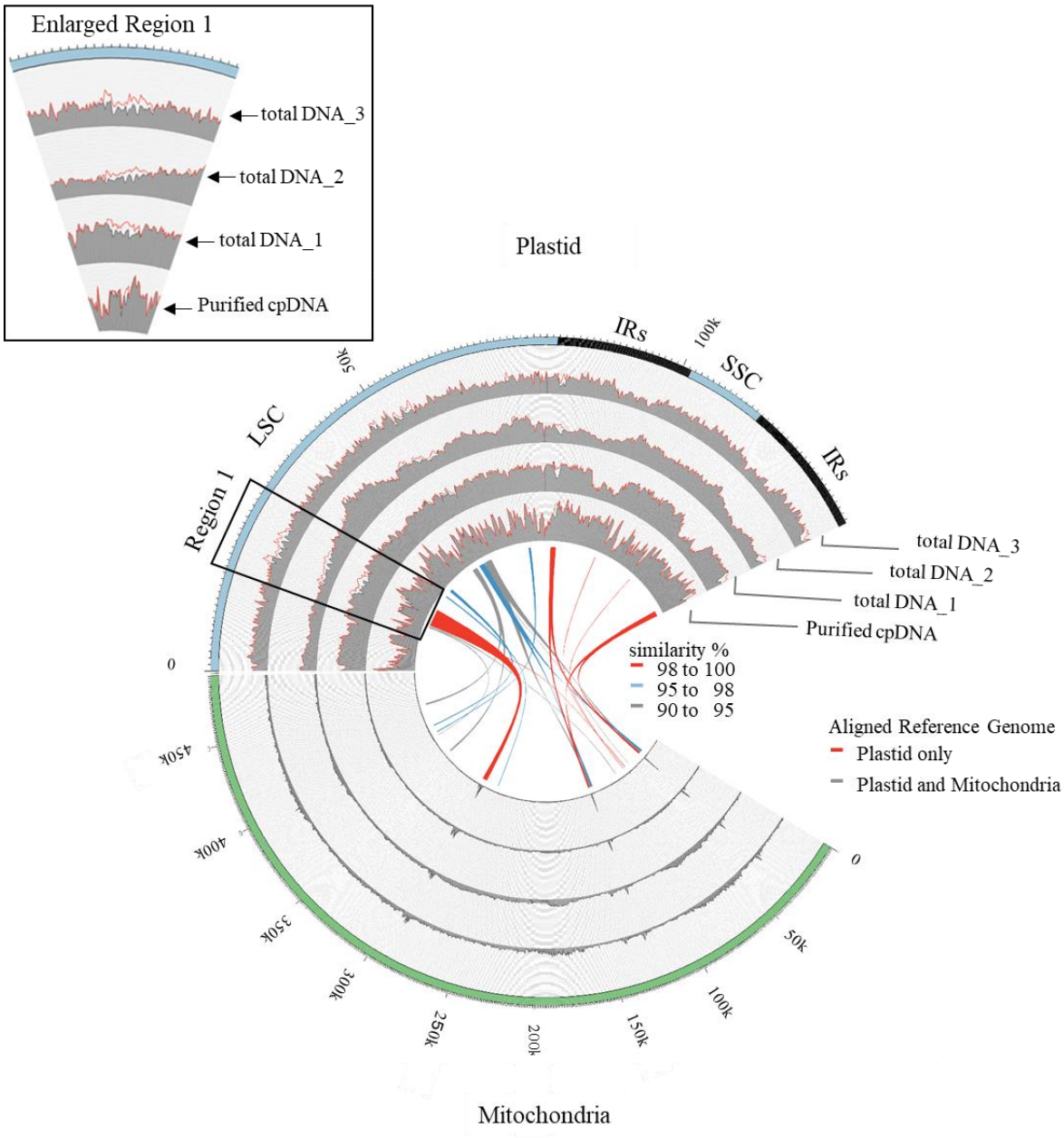


FIGURE 7 | Visualization of read alignment on plastid reference genome. FASTQ reads were aligned on plastid reference genome only (red line) or combined plastid–mitochondrial reference genomes (gray area) followed by visualization of coverage depth in a 250-nt sliding window. Outermost circle shows common plastid genome structures: LSC, large single-copy; SSC, small single-copy; IRs, pair of inverted repeats. Inner colored ribbons mark BLASTN-identified PTMTs; width represents homology; color represents similarity (%).

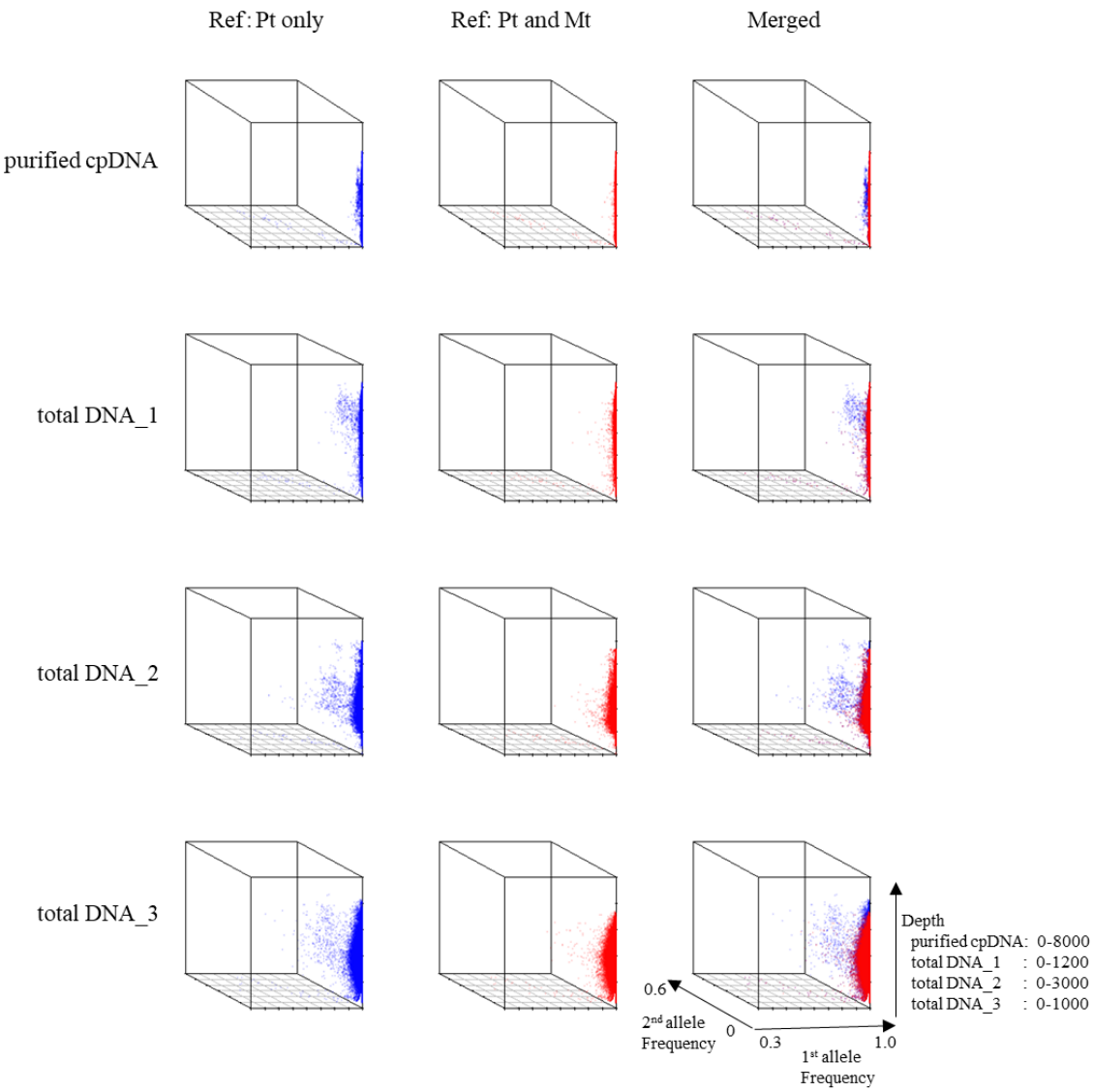


FIGURE 8 | 3D plots of allele frequency at individual base positions of the plastid genome of purified cpDNA and tDNA. Axes: *x*, 1st allele frequency; *y*, 2nd allele frequency; *z*, coverage depth. Blue, plastid (Pt) reference genome; red, combined plastid–mitochondria (Mt) reference genome.

FIGURE 9
(continued on next page)

FIGURE 9 | Graphic summary of de novo assembly of whole plastid genome. The assembled scaffolds (>500 bp) were aligned to the plastid reference genome by NCBI BLAST 2, and the hit regions are indicated by central colored ribbons and bars. Scaffold hits in the opposite direction to the reference genome are represented with lightest color in the bar on the side marked "Plastid". Detected SNPs (red tick marks) and indels (blue tick marks) are plotted in the green (GATK, **Table 6**) and yellow (*de novo* assembly) tracks (**Table 6**). Regions of undetermined base positions ('N'; green tick marks) and sequencing artifacts such as insertions of mitochondrial homology sequence (red tick marks) and duplication of plastid sequence (orange tick marks) are plotted in the gray track (**Table 8**). LSC, large single-copy; SSC, small single-copy; IRs, pair of inverted repeats.

Figure 9A.

(purified cpDNA)

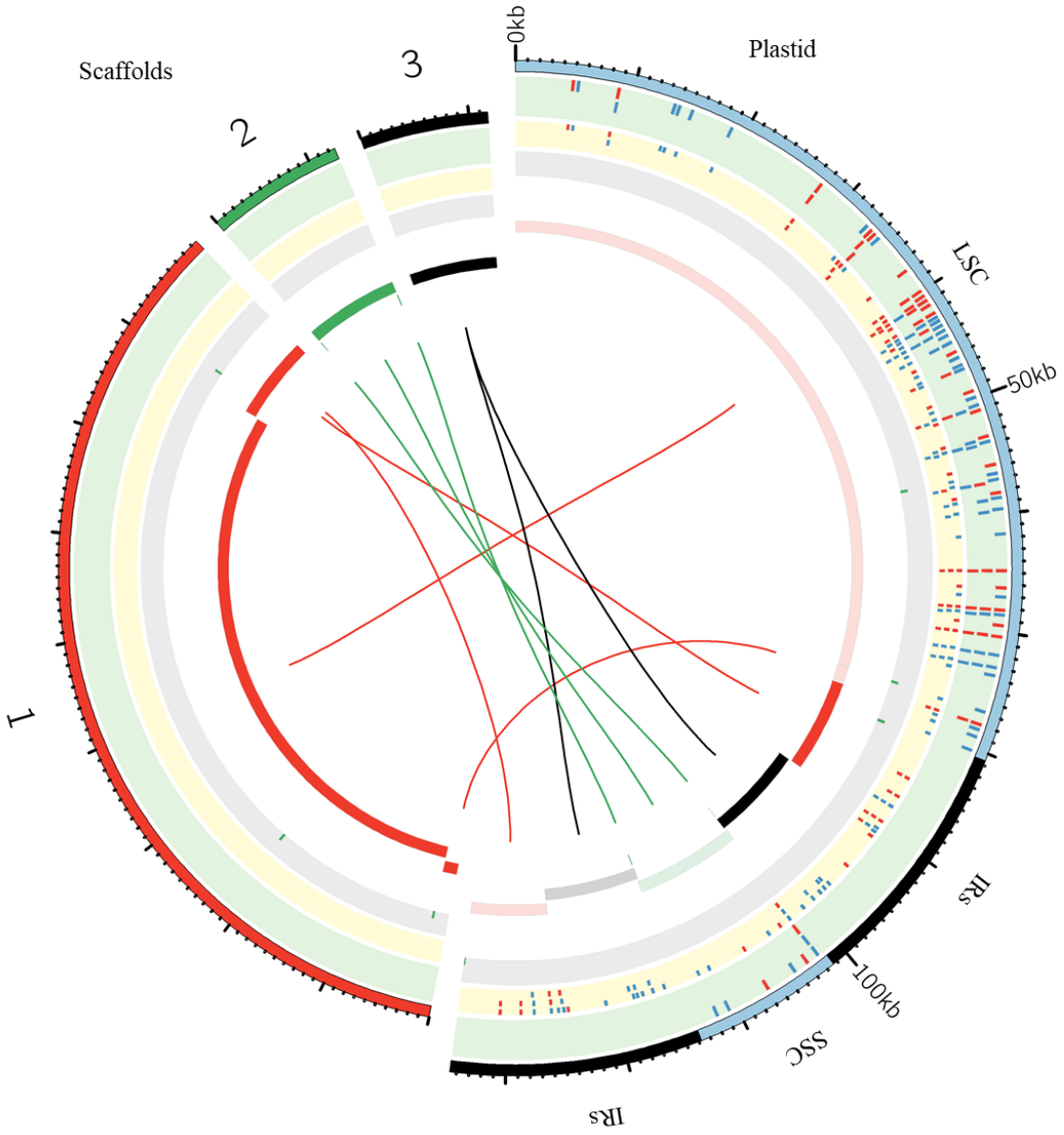


Figure 9B.

(total DNA_1)

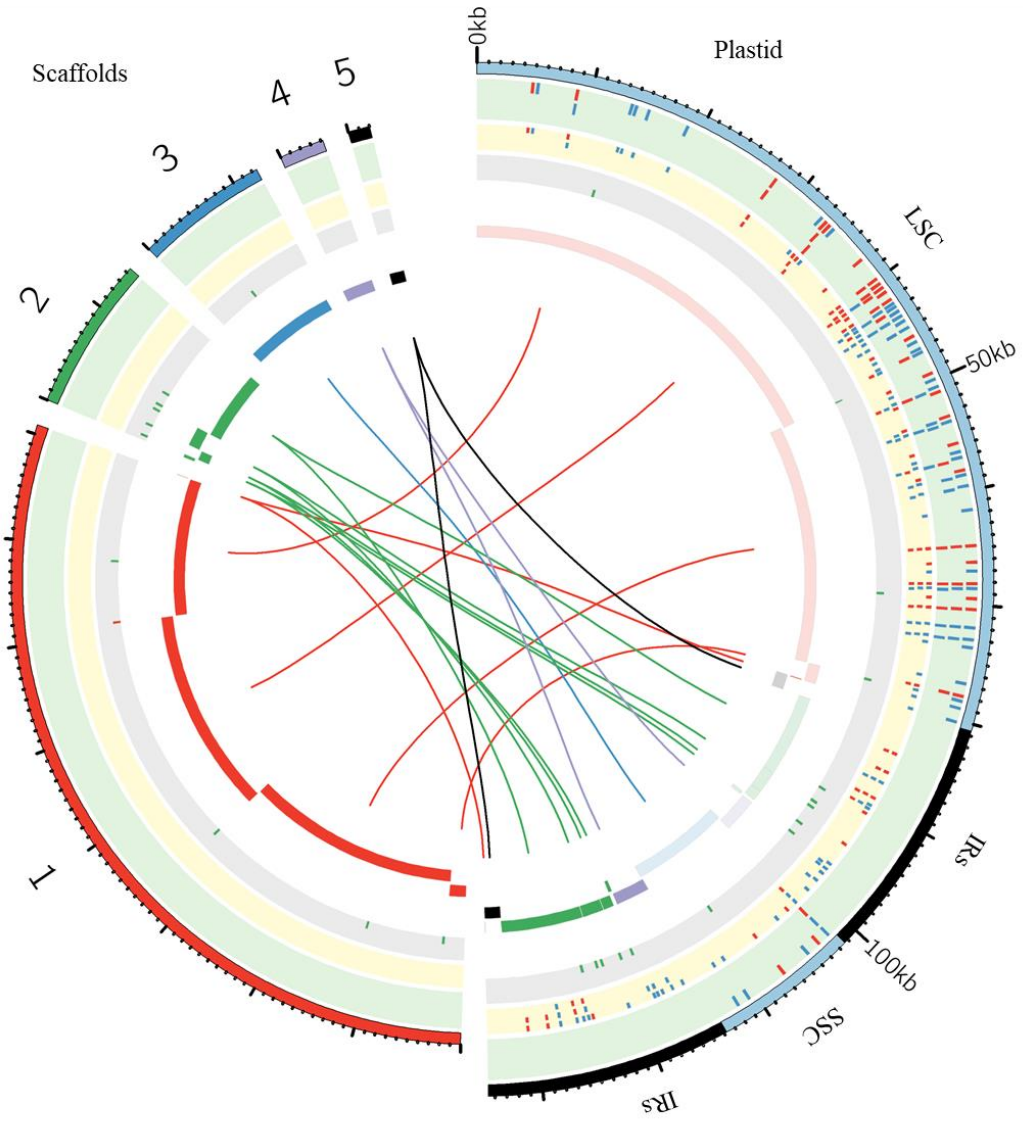


Figure 9C.

(total DNA_2)

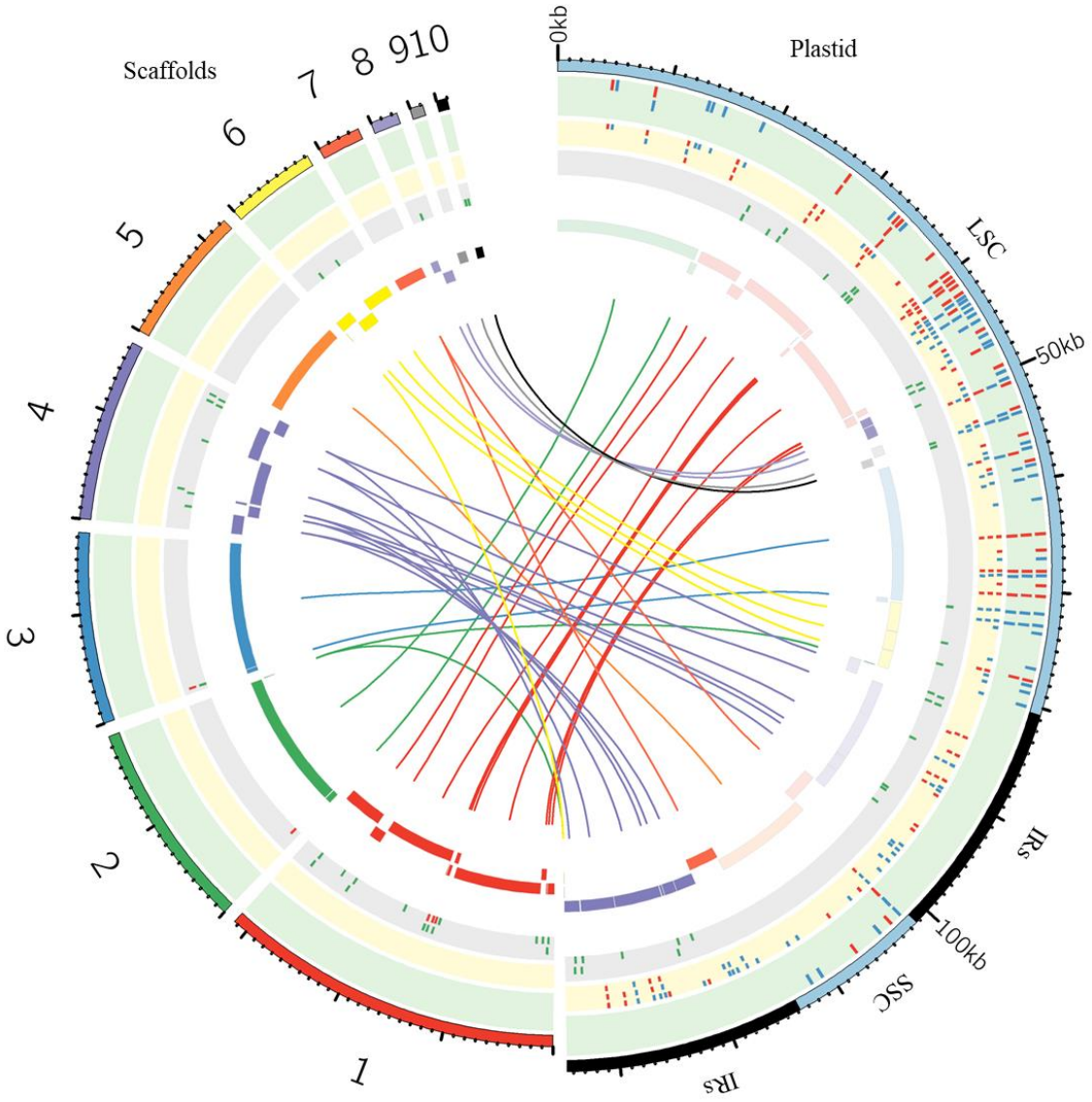


Figure 9D.

(total DNA_3)

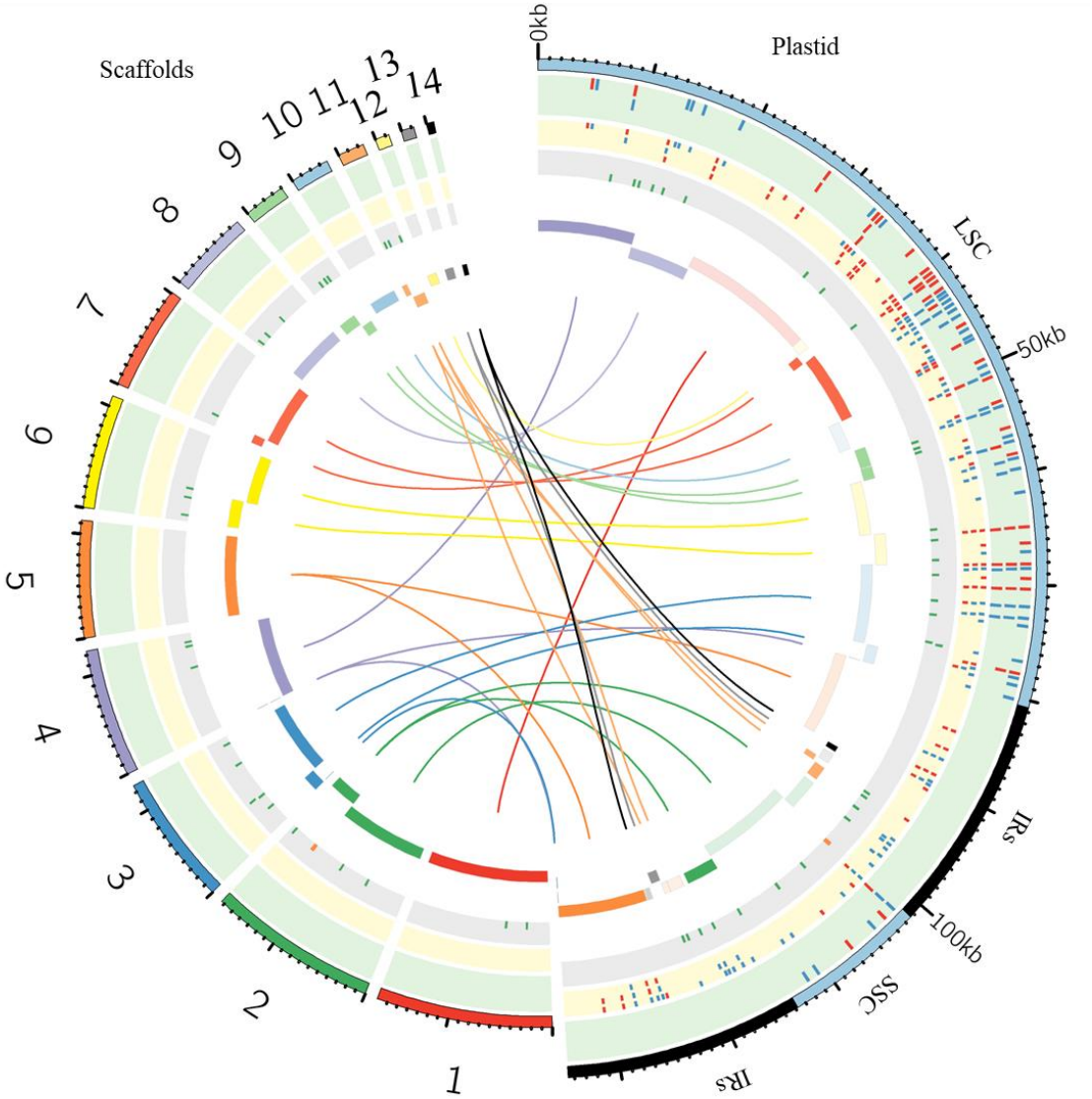


TABLE 5 | Plastid variant call results from SAMtools mpileup versus GATK HalotypeCaller using the purified cpDNA and the three total genomic DNA samples.

	SAMtools mpileup				GATK HaplotypeCaller			
Aligned (ref)	Pt only		Pt and Mt		Pt only		Pt and Mt	
Filter	Total	Pass	Total	Pass	Total	Pass	Total	Pass
purified cpDNA	86	64	86	64	129	129	129	129
total DNA_1	86	66	86	67	129	129	129	129
total DNA_2	131	64	87	65	129	129	129	129
total DNA_3	128	69	99	69	129	129	129	129

Processed reads were aligned on the plastid reference genome (Pt only) or on the combined plastid–mitochondrial reference genome (Pt and Mt). Heterozygous and low-quality variants (QUAL < 20) were filtered out in SAMtools, and low-quality variants (QUAL < 20) were filtered out in GATK.

TABLE 6 | Summary of SNP/Indel detection. Plastid reference genome,AY522330, was used as correct data. The same base as the ALT column is rebresented by '✓'. In addition, in the columns of SAMtools and GATK, filtered out variants are colored in red (Heterozygote), blue (Low QUAL score) and green (Heterozygote and Low QUAL score). IRs regions are represented in grey rows**.**

(Continued on page 49)

Assessments of Plastome *de novo* Assembly: Scaffolds and Variant Calling

We generated *de novo* assembled plastid genomes for assessment of the purified cpDNA and the three tDNAs (**Figure 9** and **Tables 6, 7, 8**). Comparison of the scaffold alignments showed that the plastid genome was almost fully covered by the assemblies of all samples, but the tDNAs generated numerous shorter scaffolds. Repeat sequences are well known to hinder long contig formation (Kim et al., 2015). Our results also show that the contig elongations stop around the boundaries of the evolutionally conserved pair of large IRs. Since it is difficult to construct a complete plastid genome from a single library, supplementation with libraries of different insert sizes and mate pairs, long read sequencing, and PCR analysis is also necessary (Naito et al., 2013; Ferrarini et al., 2013).

The genome assembled *de novo* from the purified cpDNA has long contiguous sequences which were joined into 3 scaffolds in the correct order (**Figure 9A**). This is potentially another advantage of the LN approach. In contrast, *de novo* assembly using tDNA produced 5, 10, and 14 scaffolds (**Figure 9B–D**). A simulation study demonstrated that de Bruijn Graph Assemblers such as SOAPdenovo2, which we used, can improve the assembly of longer contigs from longer reads (Knudsen et al., 2010). Certainly, among tDNAs 2 and 3, the plastid DNA ratio was approximately the same, but the longer read length of tDNA 3 resulted in the assembly of longer contigs (**Table 3** and **Figure 9C, D**). Moreover, high heterozygosity increases the complexity of the de Bruijn graph structure, leading to small contigs and base call errors (Kajitani et al., 2014). Other reports also suggest that sequence error and GC bias create ‘dropouts’—multiple gaps in assemblies—and hence small contigs and scaffolds, even in small genomes such as those of plastids (Li et al., 2010; Minoche et al., 2011). In the *de novo* plastome assembly, widespread NUPTs and MTPTs are likely to behave like heterozygous sites and sequence errors, and thus to interrupt contig formation. In fact, our results show a clear tendency for scaffolds to be longer with purer plastid DNA (**Table 3** and **Figure 9**). Additionally, scaffolds in tDNAs contain multiple low-quality regions composed of both small and large gaps of consecutive undetermined (‘N’) bases and sequencing artifacts such as the insertion of regions with high homology to mitochondrial sequences and the duplication of plastid sequences (**Figure 9** and **Table 8**). By contrast, we obtained scaffolds of purified cpDNA across the whole plastid genome with very few ‘N’ sequences.

As a result of SNP/indel detection, *de novo* sequencing identified 187 or 188 variants out of 189 reported in AY522330, with a higher sensitivity than resequencing

analysis by GATK HaplotypeCaller (**Figure 9** and **Table 6**). Despite this high sensitivity, tDNA 2 and 3 returned 41 and 58 additional variants, indicative of high false-positive rates, but the purified cpDNA did not return other variants. Taken together, these results reveal specificity and robustness in the identification of SNPs/indels by *de novo* sequencing with high-purity cpDNA.

Overall, read length and cpDNA purity are key to the successful *de novo* assembly of plastid genomes. Increasing cpDNA purity compensates for the low yield of MiSeq and enables the best use of its 300-bp paired-end sequencing, which is the longest read length of current Illumina next-generation sequencers. High-purity cpDNA is crucial for *de novo* assembly and SNP/indel calling without sequencing artifacts. It's worth noting that *de novo* sequencing allowed the identification of SNPs/indels within the two IRs where GATK HaplotypeCaller ignored variant calling. This result indicates that *de novo* sequencing using high-purity cpDNA could be an effective method for detecting variants within IRs.

TABLE 7 | Lists of assembly parameters and scaffolds/contigs. All scaffolds/contigs obtained by SOAPdenovo2 were listed. For enrichment of plastid reads, we aligned fastq reads of total DNAs on the plastid reference genome (X15901.1), and then extracted the mapped reads from bam files (pt reads enrichment: processed). SOAPdenovo2 was performed at various k-mers (tested k-mers), and the best k-mer was selected. The scaffolds/contigs were arranged in order of length, and the most homologous genome was shown in ‘homology (blast)’ column. We filtered out less than 500 bp and mitochondrial genome homologous scaffolds/contigs (marked in grey).

(Continued on next page)

purified cpDNA

name	scaffold number in Fig.6	length of scaffold (bp)	homology (blast)
scaffold8	1	89604	pt
C117	2	12597	pt
C115	3	11742	pt
scaffold4		1751	mt
scaffold7		1564	mt
scaffold5		1553	mt
scaffold3		1483	mt
scaffold6		1011	mt
C107		919	mt
scaffold2		692	mt
scaffold1		688	mt
C93		571	mt
C89		562	mt
C87		527	mt
C83		512	mt
C85		512	mt
C81		497	mt
C79		493	mt
C75		484	mt
C73		477	mt
C71		471	mt
C69		469	mt
C67		466	mt
C65		437	mt
C63		434	mt
C57		431	mt
C59		431	mt
C55		428	mt
C51		422	mt
C45		418	mt
C41		414	mt
C35		411	mt
C37		411	mt
C33		410	mt
C31		409	mt
C29		408	mt
C27		403	mt
C23		400	mt
C19		396	mt
C17		393	mt
C13		387	mt
C11		386	mt
C9		385	mt
C7		384	mt
C3		246	pt
C1		245	pt

pt reads enrichment: not processed
tested k-mers: 21, 31, 41, 51, 61, 71, 81, 91, 101, 111, 127
the best k-mer: 127

total DNA_1

name	scaffold number in Fig.6	length of scaffold (bp)	homology (blast)
scaffold1	1	80853	pt
scaffold2	2	14466	pt
scaffold3	3	12485	pt
C31	4	4218	pt
C25	5	2050	pt
C13		234	mt
C11		216	mt
C9		215	mt
C1		143	pt
C3		143	mt
C5		143	mt
C7		143	pt

pt reads enrichment: processed
tested k-mers: 21, 31, 41, 51, 61, 71, 81
the best k-mer: 71

total DNA_2

name	scaffold number in Fig.6	length of scaffold (bp)	homology (blast)
scaffold2	1	31387	pt
scaffold10	2	18983	pt
scaffold9	3	17425	pt
scaffold1	4	16555	pt
C231	5	12485	pt
scaffold12	6	8250	pt
C221	7	3946	pt
scaffold11	8	2485	pt
C199	9	1220	pt
scaffold3	10	1022	pt
scaffold6		762	mt
C193		751	mt
scaffold8		714	mt
scaffold4		691	mt
scaffold5		655	mt
C185		582	mt
C179		525	mt
C173		481	mt
scaffold7		476	mt
C171		454	mt
C167		411	mt
C161		385	mt
C157		371	pt
C153		368	pt
C151		359	mt
C147		337	mt
C143		327	mt
C139		298	mt
C129		286	pt
C125		280	pt
C121		278	mt
C115		265	mt
C109		248	mt
C107		247	mt
C101		238	mt
C97		227	mt
C87		216	pt
C85		211	mt
C77		194	mt
C73		179	mt
C67		176	mt
C65		175	pt
C61		172	mt
C53		141	mt
C51		140	mt
C49		139	pt
C47		138	mt
C39		136	pt
C41		136	pt
C43		136	pt
C45		136	pt
C27		135	pt
C29		135	pt
C31		135	pt
C33		135	pt
C35		135	pt
C37		135	pt
C19		134	pt
C21		134	pt
C23		134	pt
C25		134	pt
C15		133	pt
C17		133	pt

pt reads enrichment: processed
tested k-mers: 21, 31, 41, 51, 61, 71, 81, 91
the best k-mer: 71

total DNA_3

name	scaffold number in Fig.6	length of scaffold (bp)	homology (blast)
scaffold9	1	16991	pt
scaffold4	2	16338	pt
scaffold1	3	12654	pt
scaffold7	4	12376	pt
C91	5	11273	pt
scaffold6	6	10993	pt
scaffold8	7	10141	pt
scaffold3	8	7957	pt
scaffold5	9	4046	pt
C75	10	3780	pt
scaffold2	11	2583	pt
C49	12	1351	pt
C41	13	1255	pt
C29	14	716	pt
C19		426	pt
C17		425	pt
C13		419	pt
C7		240	pt
C1		219	mt

pt reads enrichment: processed
tested k-mers: 21, 31, 41, 51, 61, 71
the best k-mer: 71

TABLE 8 | Summary of large indels, and undetermined (N) base positions in de novo sequencing.

sample	Scaffold	REF_position	scaffold_position	REF_sequence	Scaffold_Sequence
cpDNA	1	56398	21428	AAAAATCAA	ANNNNNNN-
cpDNA	1	76460	1604	G-(and 52nt)	GNNNN[(-) x48]
cpDNA	1	80719; 134398	80662	C	CN
total_DNA1	1	12482	68298	CC	CN
total_DNA1	1	19673	61007	G-(and 63nt)	GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNT[(-) x13]
total_DNA1	1	47952	32872	TG	TN-(171nt of mitochondrial sequence)-NNNNG
total_DNA1	1	69067	11561	AGTTATG	A-----N
total_DNA1	1	78394	2324	TATGTTTATTTATCTATATTCTATATTTAATGTAATTC	GTNNNNNNNT[(-) x29]
total_DNA1	2	91080; 124012	6033	GTTTTTTCTATTTT	GN----------
total_DNA1	2	92701; 122367	4489	GAGGG-(and 45nt)	GTAAGNNNNNNNNNNNNN[(-) x36]
total_DNA1	2	93347; 121763	3811	CACTC	C---N
total_DNA1	2	95437; 119676	1829	A-(and 65nt)	ANNNNNNNNNNNNNNNNNNNNNNNNNNNNN[(-) x35]
total_DNA1	2	96767; 118354	436	CTCGAGCCGAGGTCCGAGTACCAAGCGCTGCAGC	N[(-) x34]
total_DNA1	3	108802	5002	AG	AN
total_DNA2	2	18201	619-888	-	269nt insertion of mitochondrial sequence
total_DNA2	1	20823	29119	C	CN
total_DNA2	1	20955	28986	T	TN
total_DNA2	1	24053	25715	G---G	GANNN
total_DNA2	1	25429	24334	G	GNNNNNN
total_DNA2	1	25572	24057	A	AN
total_DNA2	1	32071	17536	T----------CTTGCTTCT----------	TNNNNNNNNNNNCTTGCTTCTNNNNNNNNNNN
total_DNA2	1	34865	14742	T	TN
total_DNA2	1	34971	14635	A[(-) x130]CGGCATACTTTAATAA-------	AN-(99nt insertion of mitochondrial sequence)-(N x33)-CGGCATACTTTAATAA-(N x9)
total_DNA2	1	35154	14315	T	TNNNNNNNNNNNNNNNNNN
total_DNA2	1	35154	14307	TCTGCTTTAC	TNNNNNNNNNNNNNNNNNN
total_DNA2	1	35379	13853	C----------GCG	CNNNNNNNNNNNNNNN
total_DNA2	1	35383	14051	-	N -(159nt of mitochondrial sequence)-NNNNNNNNNNNNNNG
total_DNA2	1	35535	13684	-	N- (213nt of mitochondrial sequence)
total_DNA2	1	35541	13686	AAGAGGA	AN-----
total_DNA2	1	35763	13249	G	GN
total_DNA2	1	47219	1865	GT-(and 62nt)	NT[(-) x62]
total_DNA2	1	47437	1130	TGG	TN-
total_DNA2	1	47951	615	CA	CANNNNNNNT
total_DNA2	1	48053	501	C	CNNNNNNNNNNNN
total_DNA2	8	49510	1031	CA-(and 62nt)	CN[(-) x62]
total_DNA2	9	53916	534	T	TN
total_DNA2	9	54327	181	T	TN
total_DNA2	3	71976	398	GA	GCNNNNNNNN
total_DNA2	3	72177	400-603	-	200nt insertion of mitocondrial sequence
total_DNA2	6	75956	4639	CA-(and 62nt)	CN[(-) x62]
total_DNA2	6	78352	2255	A-(and 63nt)	ANNNNNNN[(-) x56]
total_DNA2	4	81628; 133491	15520	TT	TN
total_DNA2	4	81751; 133358	15402	GAGTGGATA	GN-------
total_DNA2	4	81864; 133253	15270	G	GNNNNNNNNNNNNNNNNNNN
total_DNA2	4	82484; 132633	14649	G	GN
total_DNA2	4	82562; 132450	14570	G-(and 103nt)	GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN[(-) x74]
total_DNA2	4	86919; 128100	10253	AA-(and 63nt)	AN[(-) x63]
total_DNA2	4	92934;122121	4339	TA-(and 61nt)	TN[(-) x61]
total_DNA2	4	93309; 121805	4025	GG-(and 62nt)	GN[(-) x62]
total_DNA2	4	94890; 120228	2507	TT-(and 63nt)	TN[(-) x63]
total_DNA3	4	7858	7930	ACAA	ANNNNNN
total_DNA3	4	10500	10576	AAAA	AN--
total_DNA3	4	11052	11126	AAA	ANNN
total_DNA3	8	12596	298	CAGCAA	CN----
total_DNA3	8	13796	1495	AG	AN
total_DNA3	8	16409	4108	TT	TN
total_DNA3	1	32079	5106	GGA	GNNNNN
total_DNA3	1	34613	2569	CC	CNNN
total_DNA3	7	39861	1220	GACCTAAAGCACTCATGGTATCATTATGAATGTATAAA	GNNNNNN[(-) x31]
total_DNA3	9	54273	1292	GAATC	GN---
total_DNA3	9	54985	2001	T	TN
total_DNA3	9	55470	2487	G-(and 54nt)	GNNNNNNNNNNNNNNNNNNNNNN[(-) x31]
total_DNA3	6	64110	3904	AATCAAATCCTTTTCTACTCCTAATGTGTCTC	ANN[(-) x29]
total_DNA3	6	65264	2780	GC	GN
total_DNA3	6	67451	595	AGCATT	AN----
total_DNA3	3	69082	11555	GC	GN
total_DNA3	3	71895	8742	CTT	CNN
total_DNA3	3	76500	4134	CT	CN
total_DNA3	3	76650	3984	TTT	TN-
total_DNA3	3	73453	2280	TCTAG-(and 56nt)	GGATNNNNNNNNNNNNNNNNNNNNNNNNNN-ATTT[(-) x28]
total_DNA3	11	94492; 120625	240	GAG	GN-
total_DNA3	11	95022; 120096	769	GT-(and 60nt)	GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN[(-) x28]
total_DNA3	11	96515; 118603	2233	TA	TN
total_DNA3	2	98079; 117020	15684	CCCAAGATGAGTGCTCTCTCC	CN[(-) x19]
total_DNA3	2	101321; 113798	12458	CT	CN-(and 252nt duplicate sequence of Plastid:113545-113797)
total_DNA3	2	105135	8387	TT	TN
total_DNA3	2	108740	4782	GATAC	GN---

In REF_sequence columns of the IRs region, the sequence of the first IRs are shown as a representative.

CONCLUSIONS

The updated LN technique permits the extraction of enriched cpDNA, allowing the investigation of plastid genomes in a more cost-effective, time-saving manner, with huge increases in sequence throughput. We demonstrated that it is possible to obtain high-quality cpDNA with which to perform functional analysis to widen the scope for high-throughput sequencing and gain new insights for other genetic studies. Collectively, our analyses strongly support that the LN protocol increases the depth of coverage with a low output of short-read sequencing, allowing the large-scale bioinformatic/computational analysis of data. Using this protocol, we generated highly accurate plastid genome sequences without sequencing artifacts. The application of NGS followed by read mapping analysis to highly purified cpDNA would allow efficient detection of SNPs and indels within a plant population and accessions. This improvement in chloroplast sequencing technologies may help the rapid advancement of the chloroplast genomics field, the understanding of plastid genome replication and repair, the high-resolution analysis of heteroplasmy, and the development of technologies for chloroplast transformation.

FUTURE PERCEPECTIVE

This high quality and high content of cpDNA is suitable for not only Miseq in this study but also various sequencing platform including third generation sequencers. Moreover, its high purity could enable high resolution analysis such as low frequency variations maybe caused by DNA lesions, natural mutations, mutagens. In plastid genome homeostasis, it is still unclear what is the molecular mechanism responsible for extremely slow evolutionary rate, quite low transformation efficiency and tolerant to mutagens. This method enables us to use various sequence platforms and applications, so we hope to capture these frontier studies from a new perspective and contribute to the development of chloroplast genomics.

Summary

Chloroplasts, which are important cellular organelles that provide energy to plants, have an independent, circular DNA. Chloroplast genome (plastome), which ranges in size from 110 to 200 kb, exists dozens to thousands of copies per cell. Organelle genomes have been considered less influence on phenotype due to less diversity and coding a small number of genes. However, recent studies demonstrated the cytoplasmic genome effects and/or cyto-nuclear genome interaction play a more important role in controlling phenotypes than previously thought (Joseph et al., 2013; Roux et al., 2016). These results suggest the potential values in the organelle genome. Therefore, we focus on the rice chloroplast genome as a novel breeding target, and are developing fundamental technology and basic research of rice chloroplast breeding. The evolutionary transfer of plastid DNA fragments to the nuclear and mitochondrial genomes is frequently found in plants. Such transfers to the nuclear genome (nuclear plastid DNA, NUPTs) and to the mitochondrial genome (mitochondrial plastid-like sequences, MTPTs) are more abundant in rice than in other higher plants. As current studies of plastome sequencing are frequently based on total DNA, these might reduce the mapping accuracy owing to the difficulty in selecting plastid-derived reads from the whole-genome sequence, which includes NUPT- and MTPT-derived reads, obtained by short read sequencing. In this thesis, for high-quality next-generation sequencing of rice plastome, we first optimized rice chloroplast DNA method, and then demonstrated high purity chloroplast DNA (cpDNA) was crucial for high accuracy analysis using next-generation sequencing.

To optimize rice chloroplast DNA methods for next-generation sequencing, we compared cpDNA extraction by three extraction protocols: liquid nitrogen coupled with sucrose density gradient centrifugation (LN method), high-salt buffer, and Percoll gradient centrifugation. The original LN method was improved at various steps to allow the use of general laboratory instruments and to simplify the method. Extracted cpDNA were assessed by qPCR and only LN methods succeeded in the removal of nuclear and mitochondrial DNA. The isolated DNA gave a well-defined electrophoresis band, which is indicative of undegraded DNA. In addition, the LN method yielded high-purity cpDNA in all ten cultivars from temperate Japonica, tropical Japonica, indica, aus, showing the potential for the LN method to be applied to a wide range of rice cultivars.

Next, to assess cpDNA quality for plastome sequencing accuracy, we performed resequencing and de novo sequencing. We generated fastq reads of purified chloroplast DNA isolated from LN methods (purified cpDNA) by using Illumina Miseq, and downloaded three sets of fastq reads of whole-genome sequencing (tDNAs) from

the public database as controls. In the mapping process of resequencing, we analyzed misaligned NUPT and MTPT reads, which disrupt precise analysis, on the plastid genome. We considered that MTPT reads containing a small number of mismatches against plastid genome could be separated by aligning reads to a combined plastid–mitochondria reference genome. Indeed, tDNA showed differences in coverage depth between the Pt and Pt–Mt reference genomes within regions of high similarity between the genomes, while purified cpDNA showed approximately similar coverage depth across the whole plastid genome. To further detect NUPT- and MTPT-derived noise, we calculated allele frequencies at individual base positions on the plastid genome. In purified cpDNA, most positions consisted of a single allele, indicating low contamination by mitochondrial or nuclear DNA. By contrast, in tDNAs, one or more other alleles were detected in many sites, and Pt-Mt combine reference genome reduced such alternative allele frequency. This result corroborates the frequency of this tendency and the lower purity of cpDNA. The above results indicate accurate mapping of purified cpDNA. Using tDNA sequenced on the Illumina platform is not consistently reliable, showing a higher rate of error alignment, which will likely affect later analysis such as plastid genome assembly or variant detection.

To assess the inference of misaligned reads from the NUPTs and MTPTs by detecting variants, we processed the BAM files describing the read alignments on plastid reference genome (X15901.1), which includes 189 SNPs/InDels of sequencing errors, to detect variants using two distinct variant callers, SAMtools mpileup and GATK HaplotypeCaller. Identified variants were compared with the correct plastid reference genome, AY522330. GATK HaplotypeCaller had much higher sensitivity than SAMtools mpileup, and SAMtools mpileup failed to detect most InDels. As sample purity decreases, SAMtools detects more false positive variants, as they don’t exist in AY522330. Furthermore, some false-positive variants still remained in all samples after filtering. On the other hand, all variants identified by GATK HaplotypeCaller are consistent with AY522330 and show a high-quality score tolerant to the filtering process, regardless of sample purity, and no difference was found in the number of variants within the four samples. These results suggest that GATK HaplotypeCaller is superior at detecting mutations with high accuracy, not only in purified cpDNA, but also in total DNA samples, in the resequencing analysis of the plastid genome. However, GATK HaplotypeCaller failed to detect variants within the two inverted repeats, since this algorism ignores low-mapping-score reads such as multi-mapping reads. Although GATK HaplotypeCaller is the current gold standard variant caller, our results show that the development of other computation approaches is required for plastome resequencing.

Next, we generated *de novo* assembled plastid genomes for assessment of the purified cpDNA and the three tDNAs by using SOAPdenovo2. Comparison of the scaffold alignments showed that the plastid genome was almost fully covered by the assemblies of all samples, but the tDNAs generated numerous shorter scaffolds. The genome assembled *de novo* from the purified cpDNA has long contiguous sequences which were joined into 3 scaffolds in the correct order whereas *de novo* assembly using tDNA produced 5, 10, and 14 scaffolds. Longer reads improve the assembly of longer contigs. Increasing cpDNA purity compensates for the low yield of MiSeq and enables the best use of its 300-bp paired-end sequencing, which is the longest read length of current Illumina next-generation sequencers, contributing to long scaffolds. Furthermore, these results show a clear tendency for scaffolds to be longer with purer plastid DNA. It is considered that widespread NUPTs and MTPTs are likely to behave like heterozygous sites and sequence errors, and thus to interrupt contig formation. Additionally, scaffolds in tDNAs contain multiple low-quality regions composed of both small and large gaps of consecutive undetermined ('N') bases and sequencing artifacts such as the insertion of regions with high homology to mitochondrial sequences, the duplication of plastid sequences, and many false-positive variants. By contrast, we obtained scaffolds of purified cpDNA with very few 'N' sequences. Furthermore, in SNPs/InDels detection, *de novo* sequencing using purified cpDNA showed no false-positive variants and higher sensitivity than resequencing analysis by GATK HaplotypeCalle. It's worth noting that *de novo* sequencing allowed the identification of SNPs/InDels within the two inverted repeats where GATK HaplotypeCaller ignored variant calling. Taken together, these results reveal high-accuracy of *de novo* sequencing with high-purity cpDNA, and *de novo* sequencing using high-purity cpDNA could be an effective method for detecting variants within IRs.

In conclusions, we optimized high-quality chloroplast DNA isolation methods and demonstrated high-accurate plastid genome sequencing using purified cpDNA. This high-purity and high-quality cpDNA can be used for not only for short-read sequencing but also for various next-generation sequencing platforms such as the long-read sequencer, PacBio RS II. This improvement technologies may help the rapid advancement of the chloroplast genomics field, the understanding of plastid genome replication and repair, the high-resolution analysis of heteroplasmy, and the development of technologies for chloroplast transformation.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to Professor Toshiaki Mitsui, Professor Kimiko Ito, and Assistant Professor Marouane Baslam for their valuable suggestions and discussions through this study. I owe my deepest gratitude to Associate Professor Takashi Abe, Faculty of Engineering, for his great lecture on bioinformatics. I would also like to express the deepest appreciation to Dr. Kazusato Oikawa, RIKEN, and PhD student Takuya Inomata for their supporting my experiments. I would also like to express the deepest appreciation to Professor Tetsu Kinoshita, Yokohama City University, and Associate Professor Takayuki Ohnishi, Utsunomiya University, for their suggestions and providing rice materials. I would also like to express the deepest appreciation to the members of laboratory of biological chemistry for their help. Finally, I would like to give my special thanks to my family for understanding and kind support.

本研究を遂行し学位論文をまとめるにあたり、以下の方々に多大なご助力を頂き、ここに感謝の意を表します。三ツ井敏明教授におきましては研究全般でご指導頂き、何よりも、特筆した技術も知識も持たなかった私を研究支援員として招き入れ、再び学術研究の道を歩む機会を与えて頂きました。社会人博士後期課程と合わせてこの５年間、私の常識では計り知れないご理解とご支援を賜りましたことに対して、適切な感謝の言葉が見つかりません。また、伊藤紀美子教授からはジェネティクスや水稲栽培管理において専門的なご指導を頂き、Marouane Baslam 特任助教には論文執筆に多大なご尽力をいただき内容を磨き上げることが出来ました。ご両名の親身なご指導ご鞭撻に深く感謝致します。バイオインフォマティクス解析は、専門家である工学部・阿部貴志准教授の半期に渡るマンツーマン講義が無ければ成しえなかった事をここに記してお礼申し上げます。本研究を始める契機を頂いた横浜市立大学の木下哲教授ならびに大西孝幸特任助教(現：宇都宮大学・特任准教授)には、実験サンプルを提供頂き、幾度となく有益な情報交換をさせて頂きました。さらに、実験でご助力いただいた及川和聡特任助教(現：理化学研究所)ならびに同輩の猪俣拓也氏、多方面にわたり惜しみないご協力をいただきました生化学研究室の方々、以上の皆様に心より感謝申し上げます。最後に、回り道を決断して歩む私に、いつでも温かい理解を示し応援してくれる両親に心より感謝します。

2018 年 8 月　　高松　壮

REFERENCES

Baumgartner, B. J., Rapp, J. C., and Mullet, J. E. (1989). Plastid transcription activity and DNA copy number increase early in barley chloroplast development. *Plant Physiol.* 89, 1011–8. doi:10.1104/PP.89.3.1011.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bookjans, G., Stummann, B. M., and Henningsen, K. W. (1984). Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic strength. *Anal. Biochem.* 141, 244–247. doi:10.1016/0003-2697(84)90452-4.

Daniell, H., Lin, C.-S., Yu, M., and Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 134. doi:10.1186/s13059-016-1004-2.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi:10.1038/ng.806.

Diekmann, K., Hodkinson, T. R., Fricke, E., and Barth, S. (2008). An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. *PLoS One* 3, e2813. doi:10.1371/journal.pone.0002813.

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49, 827–31. doi:10.1016/j.ympev.2008.09.009.

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., et al. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* 14, 670. doi:10.1186/1471-2164-14-670.

Greiner, S., and Bock, R. (2013). Tuning a ménage à trois: Co-evolution and co-adaptation of nuclear and organellar genomes in plants. *BioEssays* 35, 354–365. doi:10.1002/bies.201200137.

Hirai, A., Ishibashi, T., Morikami, A., Iwatsuki, N., Shinozaki, K., and Sugiura, M. (1985). Rice chloroplast DNA: a physical map and the location of the genes for the large subunit of ribulose 1,5-bisphosphate carboxylase and the 32 KD photosystem II reaction center protein. *Theor. Appl. Genet.* 70, 117–22. doi:10.1007/BF00275309.

Hirao, T., Watanabe, A., Kurita, M., Kondo, T., and Takata, K. (2008). Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biol.* 8, 70. doi:10.1186/1471-2229-8-70.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol Gen Genet* 217, 185–194.

Joseph, B., Corwin, J. A., Li, B., Atwell, S., and Kliebenstein, D. J. (2013). Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. *Elife* 2, e00776. doi:10.7554/eLife.00776.

Kahlau, S., Aspinall, S., Gray, J. C., and Bock, R. (2006). Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. *J. Mol. Evol.* 63, 194–207. doi:10.1007/s00239-005-0254-5.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–95. doi:10.1101/gr.170720.113.

Kaneko, K., Takamatsu, T., Inomata, T., Oikawa, K., Itoh, K., Hirose, K., et al. (2016). *N*-glycomic and microscopic subcellular localization analyses of NPP1, 2 and 6 strongly indicate that trans-Golgi compartments participate in the Golgi to plastid traffic of nucleotide pyrophosphatase/phosphodiesterases in rice. *Plant Cell Physiol.* 57, 1610–28. doi:10.1093/pcp/pcw089.

Kim, K., Lee, S.-C., Lee, J., Yu, Y., Yang, K., Choi, B.-S., et al. (2015). Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. *Sci. Rep.* 5, 15655. doi:10.1038/srep15655.

Knudsen, B., Forsberg, R., and Miyamoto, M. M. (2010). A computer simulator for assessing different challenges and strategies of de novo sequence assembly. *Genes (Basel).* 1, 263–82. doi:10.3390/genes1020263.

Kolodner, R., and Tewari, K. K. (1979). Inverted repeats in chloroplast DNA from higher plants. *Proc. Natl. Acad. Sci. U. S. A.* 76, 41–5.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–45. doi:10.1101/gr.092759.109.

Lan, J. H., Yin, Y., Reed, E. F., Moua, K., Thomas, K., and Zhang, Q. (2015). Impact of

three Illumina library construction methods on GC bias and HLA genotype calling. *Hum. Immunol.* 76, 166–75. doi:10.1016/j.humimm.2014.12.016.

Lang, B. F., and Burger, G. (2007). Purification of mitochondrial and plastid DNA. *Nat. Protoc.* 2, 652–660. doi:10.1038/nprot.2007.58.

Lewin, R. (1984). No Genome Barriers to Promiscuous DNA: The movement of DNA sequences between mitochondrial, chloroplast and nuclear genomes is even more prolific than had been expected. *Science* 224, 970–1. doi:10.1126/science.224.4652.970.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi:10.1101/gr.097261.109.

Ligges, U., and Mächler, M. (2003). Scatterplot3d – an R package for visualizing multivariate. *J. Stat. Softw.* 8, 1–20.

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.-Z. (2013). Variant callers for next-generation sequencing data: A comparison study. *PLoS One* 8, e75619. doi:10.1371/journal.pone.0075619.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1, 18. doi:10.1186/2047-217X-1-18.

Maier, R. M., Neckermann, K., Igloi, G. L., and Kössel, H. (1995). Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* 251, 614–628. doi:10.1006/jmbi.1995.0460.

Maliga, P. (2001). Plastid engineering bears fruit. *Nat. Biotechnol.* 19, 826–827. doi:10.1038/nbt0901-826.

Martin, W., and Herrmann, R. G. (1998). Gene transfer from organelles to the nucleus: how much, what happens, and Why? *Plant Physiol.* 118, 9–17. doi:10.1104/PP.118.1.9.

Matsuo, M., Ito, Y., Yamauchi, R., and Obokata, J. (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause

chloroplast-nuclear DNA flux. *Plant Cell* 17, 665–75. doi:10.1105/tpc.104.027706.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* 12, R112. doi:10.1186/gb-2011-12-11-r112.

Moison, M., Roux, F., Quadrado, M., Duval, R., Ekovich, M., Lê, D.-H., et al. (2010). Cytoplasmic phylogeny and evidence of cyto-nuclear co-adaptation in Arabidopsis thaliana. *Plant J.* 63, 728–738. doi:10.1111/j.1365-313X.2010.04275.x.

Morris, G. P., Grabowski, P. P., and Borevitz, J. O. (2011). Genomic diversity in switchgrass (Panicum virgatum): from the continental scale to a dune landscape. *Mol. Ecol.* 20, 4938–4952. doi:10.1111/j.1365-294X.2011.05335.x.

Naito, K., Kaga, A., Tomooka, N., and Kawase, M. (2013). De novo assembly of the complete organelle genome sequences of azuki bean (*Vigna angularis*) using next-generation sequencers. *Breed. Sci.* 63, 176–82. doi:10.1270/jsbbs.63.176.

Nock, C. J., Waters, D. L. E., Edwards, M. A., Bowen, S. G., Rice, N., Cordeiro, G. M., et al. (2011). Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* 9, 328–333. doi:10.1111/j.1467-7652.2010.00558.x.

Oldenburg, D. J., and Bendich, A. J. (2004). Changes in the structure of DNA molecules and the amount of DNA per plastid during chloroplast development in maize. *J. Mol. Biol.* 344, 1311–1330. doi:10.1016/J.JMB.2004.10.001.

Oldenburg, D. J., and J.Bendich, A. (1991). Degradation of chloroplast DNA in second leaves of rice (*Oryza sativa*) before leaf yellowing. *Protoplasma* 160, 89–98. doi:10.1007/BF01539960.

Olejniczak, S. A., Łojewska, E., Kowalczyk, T., and Sakowicz, T. (2016). Chloroplasts: state of research and practical applications of plastome sequencing. *Planta* 244, 517–27. doi:10.1007/s00425-016-2551-1.

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., et al. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* 8, 14. doi:10.1186/1479-7364-8-14.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–6. doi:10.1038/nbt.1754.

Roux, F., Mary-Huard, T., Barillot, E., Wenes, E., Botran, L., Durand, S., et al. (2016). Cytonuclear interactions affect adaptive traits of the annual plant *Arabidopsis thaliana* in the field. *Proc. Natl. Acad. Sci. U. S. A.* 113, 3687–92. doi:10.1073/pnas.1520687113.

Rowan, B. A., and Bendich, A. J. (2011). “Isolation, Quantification, and Analysis of Chloroplast DNA,” in (Humana Press), 151–170. doi:10.1007/978-1-61779-234-2_10.

Rowan, B. A., Oldenburg, D. J., and Bendich, A. J. (2009). A multiple-method approach reveals a declining amount of chloroplast DNA during development in Arabidopsis. *BMC Plant Biol.* 9, 3. doi:10.1186/1471-2229-9-3.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6, 283–290. doi:10.1093/dnares/6.5.283.

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* 3, 1101–8.

Shaver, J. M., Oldenburg, D. J., and Bendich, A. J. (1995). Differential transcription of Pea Chloroplast Genes during Light-Induced Leaf Development (Continuous far-red light activates chloroplast transcription). *Plant Physiol.* 109, 105–112. doi:10.1104/pp.89.3.1011.

Shi, C., Hu, N., Huang, H., Gao, J., Zhao, Y.-J., and Gao, L.-Z. (2012). An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS One* 7, e31468. doi:10.1371/journal.pone.0031468.

Sloan, D. B. (2015). Using plants to elucidate the mechanisms of cytonuclear co-evolution. *New Phytol.* 205, 1040–1046. doi:10.1111/nph.12835.

Tang, J., Xia, H., Cao, M., Zhang, X., Zeng, W., Hu, S., et al. (2004). A comparison of rice chloroplast genomes. *Plant Physiol.* 135, 412–420. doi:10.1104/pp.103.031245.

Tang, Z., Yang, Z., Hu, Z., Zhang, D., Lu, X., Jia, B., et al. (2013). Cytonuclear epistatic quantitative trait locus mapping for plant height and ear height in maize. *Mol. Breed.* 31, 1–14. doi:10.1007/s11032-012-9762-3.

Tatusova, T. A., and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–50.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–92. doi:10.1093/bib/bbs017.

Tian, S., Yan, H., Neuhauser, C., and Slager, S. L. (2016). An analytical workflow for accurate variant discovery in highly divergent regions. *BMC Genomics* 17, 703. doi:10.1186/s12864-016-3045-z.

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 (N. Y).* 9, 57. doi:10.1186/s12284-016-0129-y.

Wang, S., Liu, J., Feng, Y., Niu, X., Giovannoni, J., and Liu, Y. (2008). Altered plastid levels and potential for improved fruit nutrient content by downregulation of the tomato DDB1-interacting protein CUL4. *Plant J.* 55, 89–103. doi:10.1111/j.1365-313X.2008.03489.x.

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.* 84, 9054–8.

Wu, C.-C., Ho, C.-K., and Chang, S.-H. (2015). The complete chloroplast genome of *Cinnamomum kanehirae* Hayata (Lauraceae). *Mitochondrial DNA* 27, 1–2. doi:10.3109/19401736.2015.1043541.

Yi, M., Zhao, Y., Jia, L., He, M., Kebebew, E., and Stephens, R. M. (2014). Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res.* 42, e101–e101. doi:10.1093/nar/gku392.

Yoshida, T., Furihata, H. Y., and Kawabe, A. (2014). Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. *DNA Res.* 21, 127–40. doi:10.1093/dnares/dst045.

TABLE 6 | Summary of SNP/Indel detection.

X15901.1				SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3					
Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	cpDNA	tDNA_1	tDNA2	tDNA3
True Positive and False Negative Base Call																							
4726	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5172	C	CT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8127	A	G	✓																	A	A	A	A
8622	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8666	T	TC	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13502	T	TG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13897	C	CT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15055	GC	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
18640	TC	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
27516	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
27517	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
32686	TA	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
33207	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
33408	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
33633	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
33639	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
34109	T	TG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
37674	A	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
39771	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
39772	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40251	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40482	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40684	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40687	A	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40688	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40689	T	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40851	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40839	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
41145	G	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
41920	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
41921	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42329	C	CT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42360	CT	C	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42582	AC	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42655	GC	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42674	TC	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42678	TC	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42698	TC	T	✓							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
42896	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43129	TA	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43607	G	GGT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43624	GT	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(continued)

TABLE 6. (continued)

X15901.1				SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3					
Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	cpDNA	tDNA_1	tDNA2	tDNA3
True Positive and False Negative Base Call																							
43640	AC	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43650	CT	C	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43658	GT	G	✓							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
43928	TC	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
44108	CA	C	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
44768	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
44774	GA	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
45634	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
46175	AT	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
46850	C	CA	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
47132	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
47196	T	TC	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	N	✓
49212	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
49787	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50464	AT	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50641	T	TGC	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50643	G	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50932	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53201	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53390	CA	C	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53670	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53707	TA	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53815	AT	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53828	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53835	A	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53851	G	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
53857	T	TCGAATTCCTATAGTA	✓				✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
55656	T	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
56419	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
56926	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
57112	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
57121	AG	A	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
58052	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
58475	A	AG	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
58579	C	CA	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
59428	GC	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
59582	CA	C	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
61549	TG	T	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
64660	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
64661	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
64688	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
64689	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
66104	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
66878	GC	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
67982	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
68001	T	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(continued)

TABLE 6. (continued)

	X15901.1			SAMtools mpileup								GATK Haplotypecaller								de novc sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
	Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	P:-Mt				
True Positive and False Negative Base Call																							
	68008	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68021	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68028	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68073	A	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68110	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68388	CG	C	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68429	AG	A	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68510	T	TG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	68610	A	AAT	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	69275	T	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70224	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70225	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70256	A	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70278	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70281	A	T	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70289	AC	A	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70292	T	TC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70307	A	G	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	70308	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	71793	TC	T	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	71841	A	AT	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	71874	T	TG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	71920	T	TG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	72845	GA	G	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	72909	A	AT	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	72961	A	AG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	73024	A	AG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	73099	T	TG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	73530	CT	C	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	76691	C	CG	✓						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	77793	C	G	✓	✓	✓	✓	✓	✓	G,A	G,A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	77794	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	78230	GC	G	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	78872	T	TG	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	80074	TC	T	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	84653	C	G	✓	✓	✓	✓	✓	✓											✓	✓	✓	✓
	84654	G	C	✓	✓	✓	✓	✓	✓											✓	✓	✓	✓
	86574	G	C	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	86575	C	G	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
IRs	87665	CG	C	✓																✓	✓	✓	✓
	87670	AG	A	✓																✓	✓	✓	✓
	87687	AA	A	✓																✓	✓	✓	✓
	87693	TT	T	✓																✓	✓	✓	✓
	87696	GA	G	✓																✓	✓	✓	✓

(continued)

TABLE 6. (continued)

	X15901.1			AY522330	SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
	Position	REFERENCE	ALT		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
					Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt				
True Positive and False Negative Base Call																								
	87798	C	CG	✓																	✓	✓	✓	✓
	87806	AT	A	✓																	✓	✓	✓	✓
	89318	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	89339	A	C	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	89395	T	C	✓	✓	✓	✓	✓	✓	✓		✓									✓	✓	✓	✓
	89400	CT	C	✓																	✓	✓	✓	✓
	90072	G	GC	✓																	✓	✓	✓	✓
	90299	G	GG	✓																	✓	✓	✓	✓
IRs	90343	T	G	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	90448	A	AA	✓																	✓	✓	✓	✓
	90563	A	-	✓																	✓	✓	✓	✓
	90968	G	T	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	94502	G	-	✓																	✓	✓	✓	✓
	96863	T	TC	✓																	✓	✓	✓	✓
	97151	G	GG	✓																	✓	✓	✓	✓
	97443	C	CCC	✓																	✓	✓	✓	✓
	97511	CA	A	✓																	✓	✓	✓	✓
	98009	C	CC	✓																	✓	✓	✓	✓
	98867	G	GGG	✓																	✓	✓	✓	✓
	99004	G	GG	✓																	✓	✓	✓	✓
	100107	G	GC	✓																	✓	✓	✓	✓
	101752	T	TC	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	101796	GC	G	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	101871	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	102839	T	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	104127	C	CT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	106801	G	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	110509	C	CATAACT	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	111607	T	TC	✓									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	115012	T	TG	✓																	✓	✓	✓	✓
	116113	C	CC	✓																	✓	✓	✓	✓
	116251	C	CCC	✓																	✓	✓	✓	✓
	117109	G	GG	✓																	✓	✓	✓	✓
	117604	CT	C	✓							✓	✓									✓	✓	✓	✓
IRs	117673	G	GGG	✓																	✓	✓	✓	✓
	117967	C	CC	✓																	✓	✓	✓	✓
	118256	C	CG	✓																	✓	✓	✓	✓
	120615	AC	A	✓							✓	✓									✓	✓	✓	✓
	124150	C	A	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	124553	CT	C	✓																	✓	✓	✓	✓
	124668	T	TT	✓																	✓	✓	✓	✓
	124775	A	C	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	124819	C	CC	✓																	✓	✓	✓	✓
	125047	A	AG	✓																	✓	✓	✓	✓

(continued)

TABLE 6. (continued)

	Position	REFERENCE	ALT	AY522330	SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
					cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
					Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt				
True Positive and False Negative Base Call																								
IRs	125714	GA	G	✓																	✓	✓	✓	✓
	125723	A	G	✓	✓	✓	✓		✓	✓	✓	✓									✓	✓	✓	✓
	125779	T	G	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	125800	A	G	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	127310	AA	A	✓																	✓	✓	✓	✓
	127321	C	CC	✓																	✓	✓	✓	✓
	127420	TT	T	✓																	✓	✓	✓	✓
	127427	AAT	A	✓																	✓	✓	✓	✓
	127446	CC	C	✓																	✓	✓	✓	✓
	127451	TC	T	✓																	✓	✓	✓	✓
	128543	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	128544	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	130464	C	G	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
	130465	G	C	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	✓	✓	✓
artifacts: novel variants not existing in AY52230																								
	412	T	C						✓		✓	✓												
	6282	T	C						✓															
	7134	T	C						✓	✓														
	12788	C	G																				✓	✓
	12832	CTGCTAAGTGC GGAGA	ATGGAGCGGGTAG																				✓	✓
	12955	G	C																				✓	✓
	12999	ATGGAGCGGG TAG	CTGCTAAGTGCGGAGA																				✓	✓
	17895	T	G																					✓
	17915	T	C						✓														✓	
	17927	C	T																					✓
	17944	G	T						✓		✓												✓	
	17961	C	A						✓		✓												✓	
	17972	A	G						✓														✓	
	17977	T	A						✓		✓												✓	
	18008	G	A																					✓
	18010	A	G								✓													✓
	18016	G	T																					✓

(continued)

TABLE 6. (continued)

X15901.1				SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt				
artifacts: novel variants not existing in AY52230																							
18479	T	G						✓															
18634	G	T						✓															
18649	C	A						✓		✓													
18655	G	A						✓															
18676	C	T						✓															
18742	A	C						✓															
18746	A	C						✓															
19011	G	C						✓															
19515	A	G								✓													
20587	T	A						✓		✓													
21589	A	G								✓													
23942	A	G																					✓
23997	G	A																					✓
25497	G	A																					✓
26635	A	G																				✓	
26637	G	A																				✓	
26639	C	T																				✓	
26668	AAAGACTCAG AGA	GAGGAATCCGAGG																				✓	
26694	C	T																				✓	
26700	A	G																				✓	
26704	G	A																				✓	
26713	A	C																				✓	
26836	G	A																				✓	
26838	A	G																				✓	
26840	T	C																				✓	
26869	GAGGAATCCG AGG	AAAGACTCAGAGA																				✓	
26889	T	C																				✓	
26895	T	C																				✓	
26901	G	A																				✓	
26905	A	G																				✓	
26914	C	A																				✓	
29113	A	G								✓	✓												
32173	A	T						✓															
32324	C	T						✓															
35059	C	T								✓													
35232	A	G																					✓
35274	G	A																					✓

(continued)

TABLE 6. (continued)

X15901.1				SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3					
Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	cpDNA	tDNA_1	tDNA2	tDNA3
artifacts: novel variants not existing in AY52230																							
35298	C	T																					✓
35358	A	T																					✓
35382	G	A								✓													
35613	T	C																					✓
35635	A	C																					✓
35700	A	G																					✓
35731	T	A																					✓
35732	A	C																					✓
35742	C	T																					✓
43131	A	T		✓	✓																		
44772	A	G		✓	✓	✓	✓	✓	✓	✓	✓												
44775	A	G		✓	✓	✓	✓	✓	✓	✓	✓												
44921	C	A						✓															
44957	A	T						✓															
45021	G	A						✓															
45098	C	T						✓															
45122	T	C						✓															
46099	TTATAT	T								✓	✓												
47506	AG	A																					✓
47514	T	A																					✓
47537	A	T																					✓
47538	C	A																					✓
47549	G	A																					✓
47570	T	A																					✓
47578	A	C																					✓
47592	T	A																					✓
47677	ACAA	A																					✓
47691	TA	T																					✓
47697	G	T																					✓
47716	A	C						✓		✓	✓											✓	
47727	T	C																					✓
47737	A	G																					✓
47779	A	G																					✓
47792	C	T																					✓
47796	C	T																					✓
47846	T	C																					✓
47857	A	G																					✓
47858	A	AC																					✓
47872	T	A																					✓
47883	A	C																					✓

(continued)

TABLE 6. (continued)

X15901.1				SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
				cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
Position	REFERENCE	ALT	AY522330	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt:-Mt	Pt	Pt-Mt				
artifacts: novel variants not existing in AY52230																							
47884	G	C																					✓
47888	G	A																					✓
47907	ATCTAGG	TAGACGA																					✓
47920	G	A																					✓
47924	A	C																					✓
47943	C	A																					✓
48125	A	T								✓													
48150	A	C								✓													
48685	T	G								✓													
48713	G	A								✓													
48735	T	C								✓													
48786	C	T								✓													
48848	G	A								✓													
49856	C	T								✓													
50670	G	T								✓													
50695	A	G								✓													
50703	T	G								✓													
50954	C	T						✓															
51466	C	G						✓		✓													
53213	A	G								✓													
53285	C	T						✓		✓													
53355	G	T						✓															
53521	C	T						✓		✓													
53562	T	C						✓															
53591	T	C						✓															
53735	A	G						✓															
53763	T	C						✓															
54340	C	A								✓													
54376	T	A								✓													
54640	A	G						✓															
54676	C	T						✓		✓													
54776	T	G						✓															
54850	A	G						✓		✓													
54913	G	A						✓		✓													
55475	A	G								✓													
55577	A	C						✓															
55788	G	T						✓		✓													
57070	A	ACTATA								✓	✓												
64251	T	C						✓															
64321	A	G						✓															
68219	A	C						✓															
68223	G	A						✓															

(continued)

TABLE 6. (continued)

	X15901.1			AY522330	SAMtools mpileup								GATK Haplotypecaller								de novo sequencing			
					cpDNA		tDNA1		tDNA2		tDNA3		cpDNA		tDNA1		tDNA2		tDNA3		cpDNA	tDNA_1	tDNA2	tDNA3
	Position	REFERENCE	ALT		Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt	Pt	Pt-Mt				
artifacts: novel variants not existing in AY52230																								
	68389	G	T						✓	✓														
	70290	C	G		✓	✓	✓	✓	✓	✓	✓	✓												
	70291	G	T		✓	✓	✓	✓	✓	✓	✓	✓												
	71847	G	T																					✓
	71867	GAAATTCT	CTG																					✓
	71882	C	A																					✓
	71915	C	T																					✓
	71920	TG	T																					✓
	71927	G	GAAGTT																					✓
	71937	CTT	C																					✓
	76692	G	GA								✓	✓												
	78232	C	A		✓	✓																		
	78411	CTTTTTTTTT	C								✓	✓												✓
	78418	T	C																					✓
	79470	C	T						✓	✓	✓	✓												
	80076	C	T		✓	✓	✓	✓	✓	✓	✓	✓												
	84808	A	G																			✓		
	84844	C	A																			✓		
	84846	T	G																			✓		
IRs	84858	G	A																			✓		
	84894	A	C																			✓		
	84896	G	T																			✓		
	95020	T	G																			✓		
	97150	A	AG								✓	✓												
	109142	A	G								✓	✓												
	116112	A	AC								✓	✓												
	117672	A	AGG				✓	✓				✓												
	118255	A	AG				✓	✓																
	120098	A	C																			✓		
IRs	130222	C	A																			✓		
	130224	T	G																			✓		
	130260	C	T																			✓		
	130272	A	C																			✓		
	130274	G	T																			✓		
	130310	T	C																			✓		

Plastid reference genome,AY522330, was used as correct data. The same base as the ALT column is rebresented by '✓'. In addition, in the columns of SAMtools and GATK, filtered out variants are colored in red (Heterozygote), blue (Low QUAL score) and green (Heterozygote and Low QUAL score). IRs regions are represented in grey rows.