A Strategy to Recover a High-Quality, Complete Plastid Sequence from Low-Coverage Whole-Genome Sequencing

Authors: Garaycochea, Silvia, Speranza, Pablo, and Alvarez-Valin, Fernando

Source: Applications in Plant Sciences, 3(10)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1500022

Applications in Plant Sciences

Applications in Plant Sciences 2015 3(10): 1500022

Application Article

A strategy to recover a high-quality, complete plastid sequence from low-coverage whole-genome sequencing[1]

Silvia Garaycochea[2,5], Pablo Speranza[3], and Fernando Alvarez-Valin[4]

[2]Unidad de Biotecnología, Instituto Nacional de Investigación Agropecuaria (INIA), Rincón del Colorado, Canelones, Uruguay; [3]Departamento de Biología Vegetal, Facultad de Agronomía, Universidad de la República, Montevideo, Uruguay; and [4]Sección Biomatemática, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

- *Premise of the study:* We developed a bioinformatic strategy to recover and assemble a chloroplast genome using data derived from low-coverage 454 GS FLX/Roche whole-genome sequencing.
- *Methods:* A comparative genomics approach was applied to obtain the complete chloroplast genome from a weedy biotype of rice from Uruguay. We also applied appropriate filters to discriminate reads representing novel DNA transfer events between the chloroplast and nuclear genomes.
- *Results:* From a set of 295,159 reads (96 Mb data), we assembled the chloroplast genome into two contigs. This weedy rice was classified based on 23 polymorphic regions identified by comparison with reference chloroplast genomes. We detected recent and past events of genetic material transfer between the chloroplast and nuclear genomes and estimated their occurrence frequency.
- *Discussion:* We obtained a high-quality complete chloroplast genome sequence from low-coverage sequencing data. Intergenome DNA transfer appears to be more frequent than previously thought.

Key words: bioinformatic methods; chloroplast genome; next-generation sequencing; weedy rice.

With the advent of new sequencing technologies, deep sequencing is becoming the standard approach to obtain complete genomes (Zhang et al., 2011). In particular, there has been a rapid increase in the number of sequenced organelle genomes that have been widely used for evolutionary and phylogenetic studies. Today, there are 858 plastid genomes and 137 mitochondrial genomes publicly available for plants (National Center for Biotechnology Information [NCBI] database, accessed 16 July 2015). Advances in DNA sequencing technologies are providing a new cost-effective option not only for genome comparisons at a large scale but also for the study of interactions between organelle and nuclear genomes in plants.

The plastid genome is haploid, but there are several copies per organelle that do not recombine. In angiosperms, they usually exhibit uniparental (maternal) inheritance and considerable sequence and structural conservation (Birky, 1978). This structure consists of a quadripartite organization with two single-copy regions, a longer (long single copy [LSC], 80–90 kb) and a shorter one (short single copy [SSC], 16–27 kb), and two inverted repeat regions (IR; 12–25 kb) (Yang et al., 2010). Despite their high degree of structural and sequence conservation, chloroplast genomes usually display enough variation to perform inter- and intraspecific variability studies using whole chloroplast genome comparisons (Neale et al., 1988; Provan et al., 2001; Matsuoka et al., 2002; Cronn et al., 2008; Kumagai et al., 2010).

It is also known that several fragments, and in some cases almost entire copies of the chloroplast genome, may be found in the nuclear genomes of plants. However, few studies have described these sequences and their evolutionary processes in detail (Martin, 2003; Richly and Leister, 2004). These nuclear copies of the organelle genomes are the product of a continuous process of transference of plastid sequences to the nucleus. After their insertion into the nuclear genome, plastid sequences exhibit a high rate of fragmentation and accumulation of single nucleotide substitutions (Huang et al., 2005). Specifically, these authors presented evidence indicating that there is at least a 10-fold increase in the nucleotide substitution rate in nuclear-inserted plastid DNA when compared with their counterparts that remain in the chloroplast genome. On the other hand, the fragmentation of these DNA segments in the nucleus is expected to render many sequencing reads that will exhibit a chimeric matching pattern (discussed below). Therefore, depending on the time of DNA transfer events, these sequences may retain different degrees of similarity to the original plastid genome, and consequently introduce noise that poses additional complications to chloroplast sequence assembly from whole genome data sets. As a consequence, these nuclear DNA segments of chloroplast origin should be taken into account for chloroplast genome recovery from total

[1]Manuscript received 5 March 2015; revision accepted 28 August 2015.

The authors thank Fabián Capdeville for his important support at the beginning of this work. This project received financial support from the Instituto Nacional de Investigación Agropecuaria (INIA), Uruguay. S.G. was the recipient of a Master's degree fellowship awarded by Agencia Nacional de Investigación e Innovación (ANII), Uruguay. F.A.-V. and P.S. are researchers from Sistema Nacional de Investigadores (ANII, Uruguay).

[5]Author for correspondence: sgaraycochea@inia.org.uy

doi:10.3732/apps.1500022

DNA sequence data, not only because they can introduce distortions in the assembly, but also because they can provide valuable evolutionary information.

Different strategies to obtain whole chloroplast sequences have been reported, and they often involve prior plastid DNA isolation (Dong et al., 2013; McPherson et al., 2013; Vieira et al., 2014) or plastid DNA enrichment (Cronn et al., 2008; Dong et al., 2013; Stull et al., 2013; Kaya et al., 2014). Although successful for a range of less-studied species, these approaches can be time consuming and costly (McPherson et al., 2013). An alternative strategy consists of sequencing the total genomic DNA of a plant and subsequently isolating the chloroplast sequences using in silico approaches (Nock et al., 2011; Wang and Messing, 2011; Zhang et al., 2011; Kane et al., 2012). Such methods may require both a reference genome and resequencing (Nock et al., 2011; Wang and Messing, 2011), or the use of paired-end or mate-pair libraries to recover whole chloroplast genome sequences without using reference genomes (Zhang et al., 2011; Kane et al., 2012). Owing to the well-documented difficulty of removing all plastid DNA, even when nuclear DNA enrichment protocols are used (Atherton et al., 2010; Zhang et al., 2011), raw read samples, produced by projects aimed at obtaining nuclear genomes, also contain plastid-derived reads that are usually in sufficient amounts to assemble their corresponding genomes. Therefore, developing strategies to efficiently recover and analyze this type of data deposited in public repositories is highly desirable. In this study, we describe an approach to recover high-quality complete chloroplast genome sequences from a whole plant DNA single-read data set produced on a 454 FLX Titanium platform (454 Life Sciences, a Roche Company, Branford, Connecticut, USA).

In this study, we used weedy rice (*Oryza sativa* L.) as a model plant. This choice allowed us both to obtain a chloroplast sequence of interest for research, and to take advantage of the wealth of available information to validate our results. Publicly available information on rice genomes includes two complete nuclear genomes representing the two main domesticated subspecies, *O. sativa* subsp. *japonica* S. Kato (Goff et al., 2002) and *O. sativa* subsp. *indica* S. Kato (Yu et al., 2002); five chloroplast genomes representing different taxa within the *O. sativa* complex: the two cultivated subspecies *O. sativa* subsp. *japonica* and *O. sativa* subsp. *indica* (Tang et al., 2004), and three wild species, *O. nivara* Sharma & Shastry (Shahid Masood et al., 2004), *O. rufipogon* Griff., and *O. meridionalis* N. Q. Ng (Waters et al., 2012); and one mitochondrial genome from *O. sativa* subsp. *japonica* (Notsu et al., 2002).

Weedy rice, also called "red rice" because of its colored endosperm, is a clear example of a conspecific weed that is a major problem for the irrigated rice production system. Like many domesticated plants, rice occurs as part of a crop-weed-wild complex (Basu et al., 2004; Warwick and Stewart, 2005). Complete plastid genomes may provide a wealth of information and genetic markers that can be applied to evolutionary studies. A better understanding of the evolution of these weeds can contribute to unraveling the genetic basis underlying their ecological success (Basu et al., 2004).

MATERIALS AND METHODS

***Plant material*—**Red rice biotypes were collected on a farm in Cerro Largo, Uruguay (31°46′S, 54°26′W), and maintained in a greenhouse (approximately 25°C) with regular irrigation. After two weeks of growth, fresh green leaves from one of these individuals (AM356-8) were collected and genomic DNA was extracted. For DNA extraction, 0.2–0.4 g of fresh green plant tissue was ground with liquid nitrogen. Then 700 µL of cetyltrimethylammonium bromide (CTAB) extraction buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA [pH 8], 100 mM Tris [pH 8], PVP 2% β-mercaptoethanol 0.125%) was added, and the mix was incubated at 65°C for 20 min. After incubation, 700 µL of chloroform : isoamyl alcohol (24 : 1) was added and samples were centrifuged at 12,000 × *g* for 20 min at 4°C. The aqueous phase was precipitated with 0.7 volumes of isopropanol, and the precipitate was washed with 70% ethanol. The pellet was dissolved in 100 µL of bidistilled water.

***Library construction and sequencing*—**We used 5 µg of purified DNA to construct the sequencing genomic libraries using the GS FLX Titanium Rapid Library with Multiplex Identifier (MID) 5 adapters for barcoding (454 Life Sciences, a Roche Company). Briefly, genomic DNA was mechanically sheared to obtain 400–1000-bp fragments, ligated to the A and B adapters, and amplified using adapter-specific primers. We used Agencourt AMPure XP beads (Beckman Coulter, Brea, California, USA) to discard fragments smaller than 350 bp. A TBS 380 Fluorometer (Turner BioSystems, Sunnyvale, California, USA) was used to adjust aliquot concentrations to 1×10^7 molecules/µL. Emulsion PCR (emPCR) was performed with the GS FLX Titanium SV emPCR Kit (Lib-L) (454 Life Sciences, a Roche Company) for 50 amplification cycles as follows: 30 s at 94°C, 4.5 min at 58°C, and 30 s at 68°C. For sequencing, we used the GS Titanium Sequencing XLR70 kit (454 Life Sciences, a Roche Company) for 1/4 of a GS Titanium PicoTiterPlate (PTP) 70 × 75 in a 454 Genome Sequencer FLX System (454 Life Sciences, a Roche Company). The raw reads obtained were deposited in the NCBI Sequence Read Archive (SRA) public repository (Bioproject ID PRJNA284786).

***Representation of the three plant cell genomes in the sequence data*—**For the nuclear genome, sequence data were mapped against the nuclear genome of *O. sativa* subsp. *japonica* cv. Nipponbare with Newbler (454 Life Sciences, a Roche Company) using optimized mapping conditions for eukaryotic genomes. Both organelle sequences were aligned against the cv. Nipponbare mitochondrial (NC_011033.1) and chloroplast (AY522330.1) genomes using BLASTN algorithms.

***Identification and classification of chloroplast sequences and de novo assembly*—**The identification of chloroplast reads was performed by comparative analysis with BLAST (Altschul et al., 1990) against the three reference plastid genomes available in March 2011 (*O. nivara* [AP006728], *O. sativa* subsp. *indica* cv. 9311 [AY522329.1], and *O. sativa* subsp. *japonica* cv. Nipponbare [AY522330.1]) from the NCBI (ftp server http://www.ncbi.nlm.nih.gov/genome/browse/?report=5). Two different filters were applied on the results: one on alignment length (>100 nucleotides) and a second one on the overlap percentage >99% (alignment overlap, from now on referred to as O%). The resulting set of reads was identified as the set of reads with complete alignment (RC). One set of reads was generated this way for each of the three reference genomes used at this stage (RC *japonica*, RC *indica*, RC *nivara*). Reads exhibiting incomplete alignment, namely those with overlap percentages <90%, were called the RI sets (RI *japonica*, RI *indica*, RI *nivara*). Subsequent analyses were performed on RC *japonica* and RI *japonica*.

Reads with an overlap higher than 99% with the chloroplast genome (RC *japonica*) were assembled with Newbler. The parameters used were those recommended by the manufacturer for genomic data of noncomplex organisms (i.e., minimum overlap length = 50 bases and minimum identity of overlapping regions = 90%). The chloroplast genome sequence we obtained for the individual AM356-8 was annotated with BLAST (Altschul et al., 1990) before submission to GenBank (GenBank accession KP878280).

***Search for divergent regions between public reference genomes and chloroplast AM356-8 in silico classification*—**Identification of divergent regions among rice chloroplast genomes was made by comparison with four chloroplast genomes: *O. nivara*, *O. rufipogon* (NC017835.1), *O. sativa* subsp. *indica*, and *O. sativa* subsp. *japonica*. *Oryza meridionalis* was disregarded from this analysis because the species is a distant relative from cultivated rice, its genetic distance with *O. sativa* (subsp. *indica* and subsp. *japonica*) being 20× as much as that between *O. sativa* and Asian *O. rufipogon* (Waters et al., 2012). This alignment was carried out with Whole Genome VISTA Tools (Zambon et al., 2005). We performed a visual inspection of the alignment using a sliding window 600 nucleotides in length to identify regions containing variable sites such as

single-nucleotide polymorphisms (SNPs) and indels (Kumagai et al., 2010). A 600-bp region containing each informative indel was used for comparative analyses. These regions were compared to the RC *japonica* set with BLASTN with an *E*-value of 1×10^{-10} and flag -FF to keep the low-complexity sequences. We selected reads with sequence identity (ID%) >90% and alignment lengths >100 bases. Finally, the selected reads were aligned to the reference regions with CLUSTALW (Thompson et al., 1994) to confirm the presence of variants in the AM356-8 chloroplast read set.

***Strategy to identify chloroplast-nucleus DNA transfer events*—**The read sets with complete and incomplete alignment to the *O. sativa* subsp. *japonica* chloroplast genome (RC *japonica* and RI *japonica*) were compared to the *O. sativa* subsp. *japonica* nuclear genome to identify segments of chloroplast origin inserted in the nucleus. We then classified these segments on the basis of a tentative estimation of the age of insertion into the nuclear genome. To help understand how different types of reads were identified, we present a schematic representation of the evolutionary process that the inserted chloroplast DNA segments undergo after their insertion in the nuclear genome (Fig. 1). As shown in this figure, it is evident that one should consider the degree of overlap (O%) between the read and both genomes as well as their degree of sequence identity (ID%) with both the nuclear and chloroplast genomes. Modern inserts (recently transferred) are somewhat difficult to identify because they are expected to show a high sequence identity with both the nuclear and chloroplast genomes. Consequently, reads derived from internal parts of the transferred DNA (reads type 1A in Fig. 1) are indistinguishable from reads derived from the chloroplast genome. However, when sequencing reads derived from recent transfers include the insertion edges (boundary of insertion), they can be readily identified. This type of read (represented by 1B in Fig. 1) will partially align with the chloroplast genome (with a high degree of identity), but they will align completely and with very high identity with the nuclear genome. Therefore, we identified these sequences by using the following filtering criterion: O% < 90% with the chloroplast genome and O% ≥ 99% with the nuclear genome together with the ID% set to ≥99% to both genomes.

Reads derived from older transfers (represented by 2A and 2B in Fig. 1) can be identified by using the sequence identity level. Specifically, they are expected to exhibit a very high nucleotide identity with the nuclear genome and noticeably less identity with the chloroplast genome. Reads of type 2B can also be identified using the O%. Consequently, we looked for reads with an identity percentage threshold lower than 98% (ID% < 98%) with the chloroplast genome and an ID% ≥ 99% with the nuclear genome. More ancient transfers (represented by 1C and 2C in Fig. 1) were identified using the same filtering criterion and were not treated separately.

Finally, we searched for transfers specific to the AM356-8 biotype, namely those that occurred after the separation of this specific biotype from the reference *O. sativa* subsp. *japonica* genome. In this case, it is possible to identify only those reads encompassing an insertion boundary as in case 1B from Fig. 1. In contrast with 1B reads, reads representing new insertions are not expected to match over their entire length with the nuclear genome that is being used as a reference, because these insertions are not present in that genome (*O. sativa* subsp. *japonica* cv. Nipponbare). Instead, these reads are expected to have a chimeric alignment pattern with both genomes (nuclear and chloroplast), one

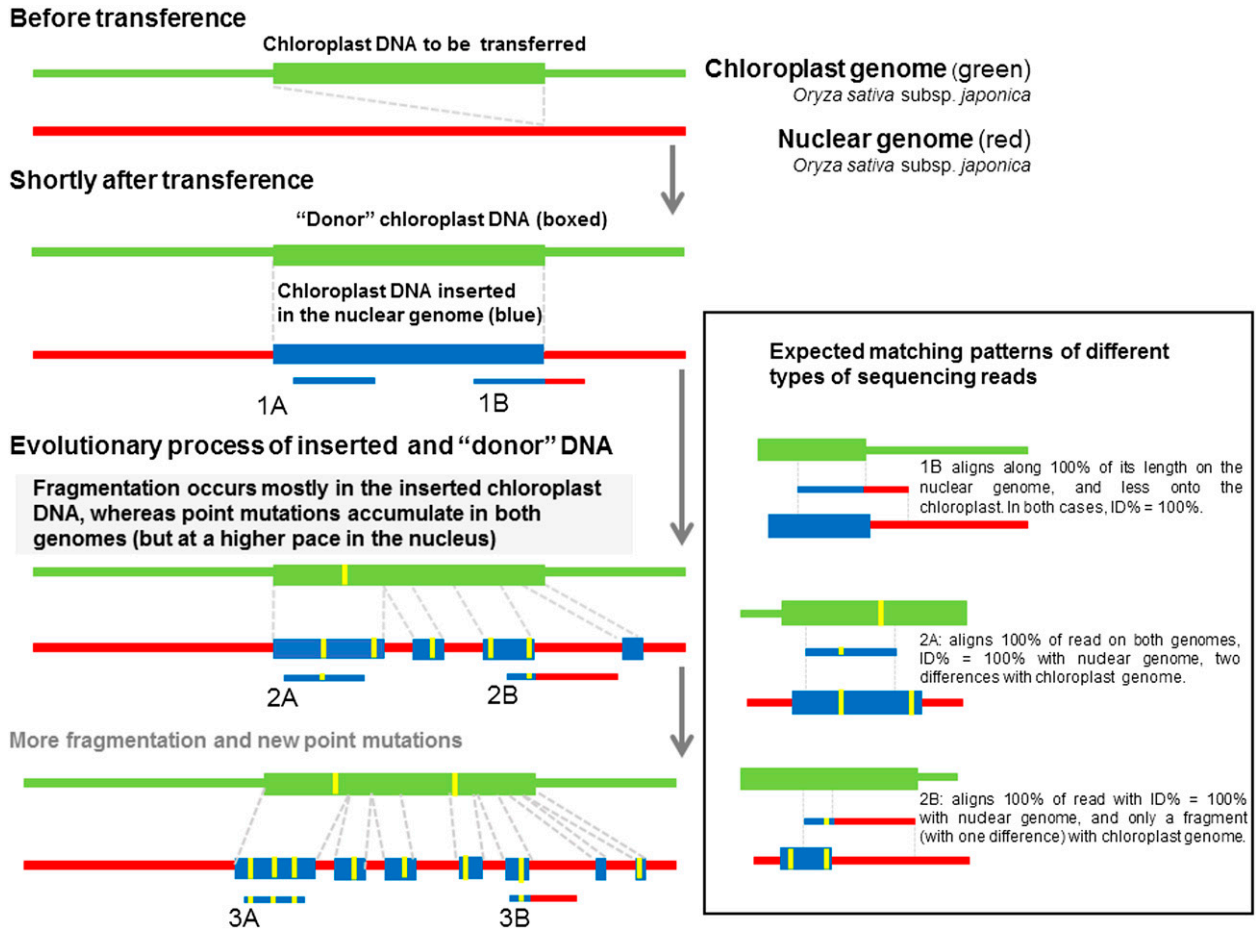

Fig. 1. Schematic representation of the evolutionary process that occurs after the insertion of a chloroplast DNA segment in the nuclear genome. The inserted fragment is represented by a blue box in the nuclear genome, whereas the homolog fragment that remains in the chloroplast (referred to as "donor" DNA) is represented by a green box. The main evolutionary events are depicted: accumulation of point mutations in both genomes (represented by yellow vertical lines) and fragmentation of inserts in the nuclear genome. The different predicted types of sequencing reads (1A, 1B, 2A, 2B, 3A, and 3B) and how they are expected to match with both genomes are also schematized.

segment aligning with 100% identity with the nuclear genome (*O. sativa* subsp. *japonica* cv. Nipponbare) and the remaining part of the read aligning with 100% identity with the chloroplast genome (Fig. 2). We looked for reads exhibiting this alignment pattern in the RI *japonica* data set using the following filtering criterion: O% < 80% with the chloroplast and nuclear genomes and ID% > 99% with both genomes (further details in Fig. 2).

RESULTS

***Sequence data*—**A set of 295,159 single reads of red rice biotype AM356-8 was generated with a mean length of 277 bp (96 Mb data) from a 1/4 run on a 454 GS FLX (454 Life Sciences, a Roche Company). Read quality was satisfactory, with a low ratio of duplicates (9.6%). The representation of the three plant cell genomes in the data were as follows: 177,920 reads were mapped to the nuclear genome with a coverage level of 0.13×, 17,003 reads showed similarity with the mitochondrial genome (coverage level 10×), and 47,817 reads showed similarity with the chloroplast genome (coverage level 106×).

***Identification and classification of chloroplast sequences*—**The identification of chloroplast sequences within the complete sequence data set was performed by following the comparative genomics strategy shown in Fig. 3. On average, 47,800 reads were identified with similarity to at least one of the chloroplast reference genomes, which represent 16.2% of the data. After applying the two previously defined filters (alignment length >100 nucleotides; overlap percentage >99%), we obtained three new sequence data sets: RC *japonica* with 34,091 reads, RC *indica* with 33,888 reads, and RC *nivara* with 33,925 reads (Fig. 3).

***Assembly of the chloroplast genome*—**The de novo assembly of the chloroplast sequence of AM356-8 was obtained with Newbler software using default parameters. The assembler generated two contigs: a larger one of 101,363 bp in length and a shorter one of 12,637 bp, with a total length of 114 kb corresponding to 85% of the expected length of the chloroplast genome. Because Newbler collapses repeated sequences, we interpreted this difference as the result of collapsing the two IR regions into only one sequence. To confirm this, the two contigs were aligned against the *O. sativa* subsp. *japonica* chloroplast genome with BLAST/ACT (Artemis Comparison Tool; Carver et al., 2005). The alignment of the two contigs against the *O. sativa* subsp. *japonica* chloroplast genome confirmed this interpretation because the longer contig had a very high sequence identity to the LSC region and the inverted repeat, whereas the shorter contig showed high similarity with the SSC region (Fig. 4).

***Search for divergent regions among public reference genomes and AM356-8 chloroplast in silico classification*—**We identified 47 indels and 111 SNPs among the four *Oryza* public chloroplast genomes. All variable sites with the exception of one of the indels were found in the single copy regions. This spatial distribution of SNPs and indels is congruent with the divergence rates reported for chloroplast genomes, as previous studies showed that the IR region has a slower rate of divergence than the LSC and SSC regions (Wolfe et al., 1987; Shahid Masood et al., 2004). This observation is in line with the well-established concept that evolutionarily divergent regions exhibit higher intraspecific variability; the link between both levels of variability is given by the degree of functional constraints (less constrained regions evolve faster and have higher polymorphism levels, see for instance Tajima, 1989). Indels showed the following interspecific distribution: 14 indels were exclusive to the *O. sativa* subsp. *japonica* chloroplast genome, 12 to *O. rufipogon*, seven to *O. sativa* subsp. *indica*, and nine to *O. nivara.* Three indels were shared between *O. sativa* subsp. *japonica* and *O. rufipogon*, and another three were shared between *O. sativa* subsp. *indica* and *O. nivara*. We did not identify indels shared between both *O. sativa* subspecies, nor between the two wild species.

Indels encompassing at least two nucleotides were searched for their presence in the data set used for the AM356-8 chloroplast assembly (RC *japonica* set). Eight of these indels were identified in the AM356-8 chloroplast DNA, among which three correspond to those shared by the *O. sativa* subsp. *japonica* and *O. rufipogon* chloroplast sequences and the remaining five were only shared with the *O. sativa* subsp. *japonica* chloroplast sequence (Table 1).

***Identification of DNA transfer from the chloroplast to the nucleus*—**We identified 30 "internal" reads derived from ancient DNA transfers to the nucleus (represented as 2A and 3A in Fig. 1) in the RC *japonica* set. Additionally, in the RI *japonica* set, we

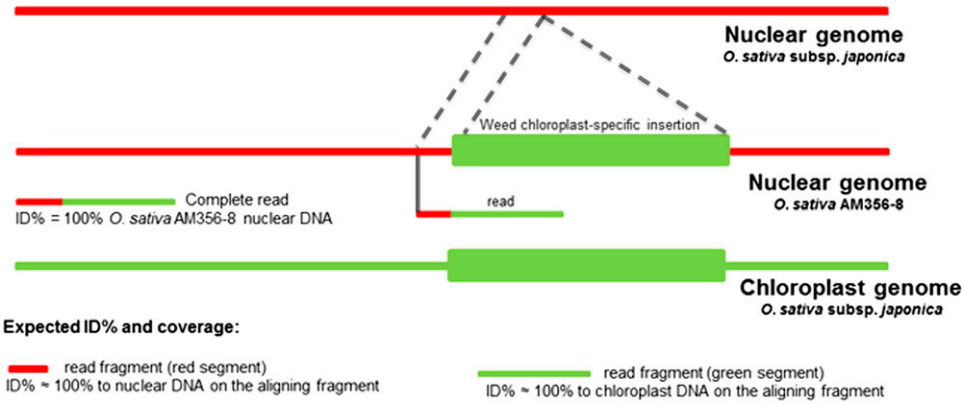

Fig. 2. Strategy to identify AM356-8-specific chloroplast DNA insertions into the nuclear genome. Read-filtering criteria: partial, non-overlapping alignment with both genomes (O% < 80 with both genomes) and 100% identity with both genomes on the respective segments (ID% ≈ 100%).

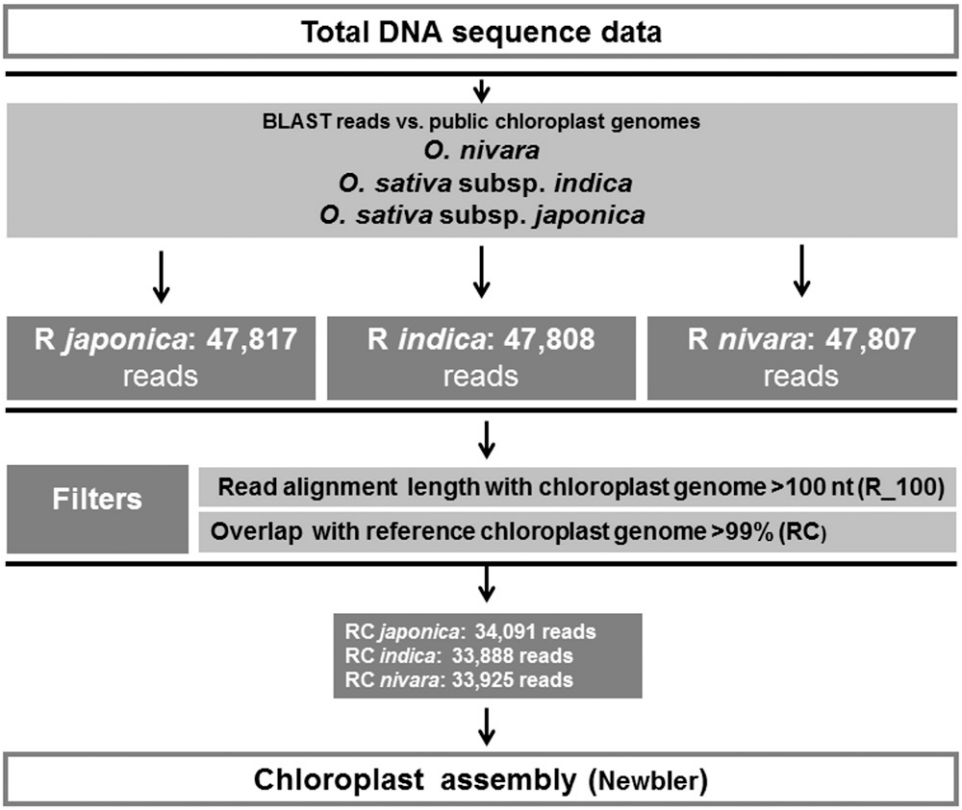

Fig. 3. The strategy followed for the identification of chloroplast reads from whole genome data. R *japonica*, R *indica*, R *nivara* = reads aligned with each respective chloroplast reference genome; RC = set of reads with more than 99% overlap with the chloroplast reference genome.

identified 50 reads representing either edges of old transfers or short chloroplast DNA inserts (types 2B and 3B in Fig. 1). These reads can potentially introduce noise and were consequently discarded in the assembly step.

As far as recent transfers are concerned, an important point to consider is that almost all of the reads contained in the RC *japonica* set showed high identity with both genomes, indicating that sequences representing the complete chloroplast genome are found among the *O. sativa* subsp. *japonica* nuclear sequences. Reads matching both genomes along their entire length (type 1A), cannot be assigned to actual chloroplast sequences or recent nuclear-inserted plastid DNA. Nevertheless,

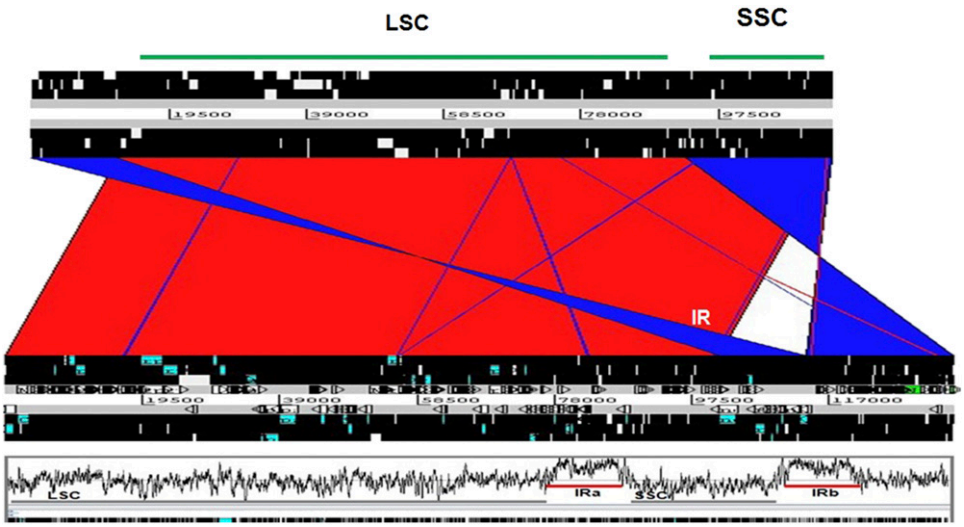

Fig. 4. Overview of alignment between contigs recovered for de novo AM356-8 chloroplast and the reference chloroplast genome. The alignment was made with Artemis Comparison Tool (ACT; Carver et al., 2005). LSC = long single copy; SSC = short single copy.

TABLE 1. Indel comparison among the four reference genomes and reads from AM356-8. Only indels in the AM356-8 chloroplast genome are shown.

Indel position (kb)[a]	Variant genome	AM356-8 (no. of reads)	Length (bp)	Variable sequence	Annotation
8	*japonica*/*rufipogon*	23	69	GAATCCTATTTTTGTTCTTATACCCATGCA ATAGAGAGGAGTGGGAAAAGGGAGGT TACTTTTTTTCA	Nonannotated predicted ORF
12	*japonica*	14	4	AGGG	Intergenic
14	*japonica*	1	2	AC	Intergenic
46	*japonica*	8	5	TATAT	Intergenic
57	*japonica*/*rufipogon*	10	16	TTTTTTAGAATACTAA	Intergenic
60	*japonica*	32	5	deletion: TATTG	Intergenic
65	*japonica*	23	2	TT	Intergenic
77	*japonica*/*rufipogon*	43	3	deletion: TGG	Intergenic

[a] Position of the indel in the alignment of the AM356-8 genome with the four reference genomes.

this does not pose a serious problem because type 1A reads would not introduce much noise in the assembly other than increasing sequencing depth in some parts. Type 1B reads, however, can certainly induce assembly distortions. Twenty-five type 1B reads were identified, which were also excluded in the assembly step. This group of reads, on the other hand, can be used to estimate the amount of recent transfer events. Considering that only about 13% of the nuclear genome was represented in our sequence data set, it follows that there may be about 192 (25/0.13) insertion borders in the nuclear genome, which in turn would represent approximately 96 recent transfer events. It is important to stress that only reads derived from recent transfers (1B) should be used for these estimations, because in the case of more ancient transfers (2B and 3B) it is not possible to determine whether the boundary-like pattern has arisen as a result of a new insertion or by fragmentation of a previously inserted fragment.

The last group of transfers investigated included those presumably present only in the weedy rice nuclear genome. We found 11 reads that aligned with both the nuclear and chloroplast reference genomes without completely overlapping with either of them (see Fig. 2 for details on how to interpret the aligning pattern of this type of read). The alignment details of each of the 11 reads identified (listed in Table 2) reflect the predicted pattern. For instance, the first read listed in Table 2 (GCFF90V2GW6GW) aligns from position 1 to 270 with Chromosome 2 (ID% = 99.63) and from base 271 to 515 with the plastid genome (ID% = 99.59). Equivalent alignment patterns can be observed in the remaining 10 reads. These 11 reads mapped to six of the 12 rice chromosomes, with no evident distribution pattern. The segments aligned with the chloroplast genome of all but one of these reads showed sequence identities close to 100%. The remaining read (GCFF90V2H8EV7), which

TABLE 2. Reads representing AM356-8-specific nuclear insertion of chloroplast DNA.

Read ID[a]	Matching genome	ID%	Aln length (bp)	Read length (bp)	Aln start (read)	Aln end (read)	Start DB (genome)	End DB (genome)
2GW6GW	Plastid[b]	99.59	246	517	271	515	54,081	54,326
	Chr. 2	99.63	270	517	1	270	1,879,451	1,879,182
02F9BJE	Plastid	100	116	382	267	382	111,458	111,573
	Chr. 2	98.13	268	382	1	268	15,462,658	15,462,391
2IEQIE	Plastid	100	115	516	1	115	30,704	30,818
	Chr. 4	99.76	411	516	106	516	32,858,510	32,858,919
2JK097	Plastid	100	199	421	223	421	43,044	42,846
	Chr. 5	99.56	228	421	1	228	2,460,207	2,460,433
2HC0XG	Plastid	100	131	437	307	437	45,371	45,241
	Chr. 5	98.04	306	437	1	306	3,748,044	3,748,344
2GJK3F	Plastid	100	134	359	226	359	46,482	46,615
	Chr. 5	99.12	228	359	1	228	113,635,156	113,634,929
2G6KGA	Plastid	98.70	154	351	198	351	67,685	67,533
	Chr. 5	100	205	351	1	205	26,830,841	26,831,045
2H8EV7	Plastid	89.36	141	488	1	134	71,978	71,844
	Chr. 8	96.87	319	488	113	430	25,340,556	25,340,238
2JD9YU	Plastid	100	185	397	1	185	88,217	88,401
	Chr. 12	100	218	397	180	397	23,930,784	23,930,567
2GBG4O	Plastid	100	123	281	159	281	77,439	77,561
	Chr. 4	100	162	281	1	162	16,565,916	16,566,077
2F55JZ	Plastid	100	185	397	1	185	88,217	88,401
	Chr. 12	100	218	397	180	397	23,930,784	23,930,567

Note: Read ID = read name; ID% = percentage of nucleotide identity; Aln length = alignment length in base pairs (between read and genome); Aln start = coordinate on the read where the alignment starts; Aln end = coordinate on the read where the alignment ends; Start DB = coordinate on the database (genome) where the alignment starts; End DB = coordinate on the genome where the alignment ends.

[a] All read names start with the sequence GCFF90V.

[b] Plastid stands for *Oryza sativa* subsp. *japonica* (cv. Nipponbare) chloroplast genome, accession number: AY522330.1. Chromosomes 1–12 are the chromosomes from the same cultivar.

showed higher identity with the nuclear genome, may represent a more ancient insertion.

DISCUSSION

Several studies have reported different methods to obtain organelle genome sequences without prior isolation or enrichment steps (Nock et al., 2011; Wang and Messing, 2011; Zhang et al., 2011; Kane et al., 2012); however, those methods involve additional library construction steps or the use of reference genomes and resequencing. We achieved the de novo assembly of a chloroplast genome into only two contigs from the data set produced by a 1/4 run of 454 FLX Titanium standard single reads without using a reference genome. Although the 454 platform has been discontinued, this strategy can be applied to unexplored data sets deposited in public repositories that may contain valuable plastid read data. The strategy can also be applied to data sets produced with other sequencing technologies achieving similar read lengths with lower costs per run than 454 such as the current Illumina platforms (2 × 150 paired end). In our case, reference genomes were used only for read filtering. It is worth noting that although we used the closest available genomes, which in this case belonged to the same species, any public set of more distantly related chloroplast genomes may have been used for filtering chloroplast reads due to the structure and sequence conservation of plastid genomes (Zhang et al., 2011). Very recently, we applied a similar strategy (using 2 × 100 and 2 × 150 paired-end Illumina reads) for obtaining very high-quality mitochondrial genomes from *Trypanosoma vivax*, using as a reference another *Trypanosoma* species (*T. brucei*) that is relatively divergent (Greif et al., 2015). To validate our results, we showed that the sequence we obtained de novo can be completely aligned to the publicly available *O. sativa* subsp. *japonica* chloroplast genome, indicating that the RC *japonica* set contained a whole chloroplast genome despite consisting of two contigs. Performing the assembly without the assistance of reference genomes eliminates the probability of introducing biases caused by favoring any particular contig arrangement.

Characterization of phylogenetically informative genome variants—We tested the chloroplast sequence of AM356-8 for the presence or absence of 47 indels and identified eight of the variable indels found among the other *Oryza* genomes in this chloroplast genome. The specific distribution of these indels among the genomes analyzed group *O. sativa* subsp. *japonica* together with *O. rufipogon* and *O. sativa* subsp. *indica* with *O. nivara* in agreement with previous studies (Huang et al., 2005). Five of these indels were shared with the haplotype observed in the *O. sativa* subsp. *japonica* chloroplast genome, and the remaining three indels present in the red rice chloroplast genome were shared with the *O. sativa* subsp. *japonica* and *O. rufipogon* chloroplast haplotypes. The AM356-8 chloroplast genome can then be classified as an *O. sativa* subsp. *japonica* type. This is consistent both with the history of the crop in Uruguay and the particular field where this biotype was found. Several evolutionary processes may explain the presence of weedy rice in areas where native wild rice relatives are absent. The results of this work support the hypothesis that weedy rice could have originated by introgression from originally contaminated germplasm. Subsequent selection and rehybridization may have led to the weedy biotypes found today (Chen et al., 1993).

As in other cases, the indels identified may be useful tools for population or phylogenetic studies in the genus *Oryza*, as well as to trace the origin of other weedy rice populations (Chen et al., 1993; Tang et al., 2004; Kumagai et al., 2010). This combination of indels represents a haplotype that can be clearly differentiated from other subspecies of *O. sativa*, as well as from the wild haplotypes of *O. nivara* and *O. rufipogon*.

Chloroplast sequence transfers to the nucleus—The sequence data obtained in this study originated from a mix of three genomes present in a plant cell (nuclear, chloroplast, and mitochondria). An additional degree of difficulty is given by the fact that the existing plant nuclear genomes are the outcome of the balance between ongoing processes of integration and elimination of inserted plastid DNA by genome shuffling (Matsuo et al., 2005). Therefore, to identify the source of the sequences and the reads that derive from them, it was necessary to take into account molecular evolutionary concepts. A very informative element is the disparity in evolutionary rates between the nuclear and plastid genomes. Specifically, it is expected that these inserts will diverge from the donor sequence by the accumulation of point mutations at very high rates. The intensity of the signal left by these processes is expected to be directly proportional to the time elapsed since the insertion of the chloroplast DNA into the nuclear genome (Martin, 2003; Leister, 2005; Matsuo et al., 2005). In the case of old transfers, we could clearly see evidence of these evolutionary processes of DNA chloroplast inserts into the nuclear genome. It should be noted that when we sought for reads representing insertion edges (2B and 3B), we found a twofold amount compared with that found when searching for internal reads with complete alignment with both genomes (2A and 3A). This excess of edge-containing reads is in agreement with the reported fragmentation process because it is expected that the length of the insert will decay with time (Leister, 2005; Matsuo et al., 2005). The fragmentation mechanism of organelle inserts is not clear, but it may be caused by insertions of transposable elements into nuclear copies of chloroplast fragments (Noutsos et al., 2007). In fact, most long chloroplast inserts appear to begin to decrease in size within the first million years (Matsuo et al., 2005). In *Arabidosis thaliana* (L.) Heynh. and *O. sativa*, two kinds of inserts were observed, one collinear with the original genome and the other formed by mosaic organelle DNA, often of both mitochondrial and chloroplast origin (Noutsos et al., 2007).

In the case of modern transfers, because sequence divergence has not taken place yet, it is only possible to identify those reads aligning with less overlap with the chloroplast than with the nuclear genome (1B reads). Fragmentation is also expected to be minimal; consequently, the edges identified by this criterion (1B) will mostly correspond to those produced by new insertions instead of fragmentation. We found 25 reads matching this criterion, which yields an estimate of around 100 "new" insertions. Although these data alone do not allow estimation of an insertion rate, they do indicate that this rate is very high. In effect, if one considers both that these insertions must have taken place before single nucleotide substitutions occurred and that the nucleotide substitution rate in inserted chloroplast DNA is at least 10-fold higher than that of the source chloroplast genomes (Huang et al., 2005), it follows that these inserts must be very recent.

Reads that represent transfers exclusive to the specific weed biotype AM356-8 and are not present in the reference rice nuclear genomes were also identified. These are reads from the

nuclear genome that correspond to chloroplast genome insertions that may have occurred after the separation of this particular weedy rice lineage from the available cultivated *O. sativa* subsp. *japonica* sequence. However, we cannot exclude the possibility that some of these reads may actually correspond to regions not included in the original assembly of the available reference nuclear genome of the *O. sativa* subsp. *japonica* cv. Nipponbare due to their high similarity to the chloroplast genome. In either case, this group of reads is expected to produce a mosaic alignment with both genomes.

Based on our results, we can make a rough calculation of the number of transfers that occurred since the split between this weedy rice biotype lineage and cultivated *O. sativa* subsp. *japonica* rice, as well as a putative transfer rate, assuming that AM356-8-specific chloroplast DNA insertions are indeed not absent from the available public reference genome due to assembly incompleteness of specific genomic regions. We estimated that our data set contained a 0.13× coverage of the nuclear genome, and 11 reads were identified as biotype-specific transfers. It follows that about 49 new insertions occurred since the divergence of this nuclear lineage from the reference *O. sativa* subsp. *japonica* rice sequence.

It has been suggested that weedy rice could be the result of hybridization between the crop and a related wild species (Olsen et al., 2007; Londo and Schaal, 2007), a proposal that was supported by studies on genome-wide variation (Londo and Schaal, 2007; Gealy et al., 2009; Gross and Olsen, 2010; Reagon et al., 2010). Based on this well-grounded hypothesis, we can safely assume that the AM356-8 biotype probably evolved by hybridization and introgression with cultivated *O. sativa* subsp. *japonica*. Very recent estimations suggested an upper bound of this event of 10,000 years, namely at the beginning of domestication (Subudhi et al., 2014). This yields an estimate of the transfer rate as high as one insertion arising and becoming fixed in the population every 200 years. Laboratory assessment of transfer rate between chloroplast and nucleus estimated that one out of 16,000 pollen grains carries a new insertion (Huang et al., 2003). Assuming that these insertions are neutral (i.e., they do not affect the function), this will produce one established insertion (fixed in the population) every 16,000 years. However, many of these insertions will be disruptive of function (e.g., those occurring within coding regions). Consequently, fixation rates are expected to be substantially lower. The discrepancies between these estimates and the results presented here may deserve further consideration. Although the literature on DNA flow between plastid genomes and the cell nucleus is not currently abundant, we can foresee an accumulation of data based on the new sequencing technologies in the near future that will provide greater insight into these mechanisms and their evolutionary significance.

In conclusion, in this work we presented a data analysis approach that can recover a whole chloroplast sequence genome from whole genome sequences even from a low-coverage data set and a simple sequencing strategy. The deliberate identification of reads that represent chloroplast DNA inserts into the nuclear genome allowed us to refine our read sets to attain a higher-quality chloroplast genome assembly in a time- and cost-effective way. Finally, we obtained a full sequence of a weedy rice biotype from Uruguay; although an exhaustive comparative analysis of this genome was not the focus of this work, we have provided information that will contribute to the understanding of the evolutionary processes that have shaped the *O. sativa* complex.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Atherton, R. A., B. J. McComish, L. D. Shepherd, L. A. Berry, N. W. Albert, and P. J. Lockhart. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods* 6: 22–28.

Basu, C., M. Halfhill, T. Mueller, and C. Stewart. 2004. Weed genomics: New tools to understand weed biology. *Trends in Plant Science* 9: 391–398.

Birky, C. W. 1978. Transmission genetics of mitochondria and chloroplasts. *Annual Review of Genetics* 12: 471–512.

Carver, T., K. Rutherford, M. Berriman, M. Rajandream, B. Barrell, and J. Parkhill. 2005. ACT: The Artemis Comparison Tool. *Bioinformatics (Oxford, England)* 21: 3422–3423.

Chen, W., I. Nakamura, Y. Sato, and H. Nakai. 1993. Distribution of deletion type in cpDNA of cultivated and wild rice. *Japanese Journal of Genetics* 68: 597–603.

Cronn, R., A. Liston, M. Parks, D. S. Gernand, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

Dong, W., C. H. Xu, T. Cheng, K. Lin, and S. Zhou. 2013. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biology and Evolution* 5: 989–997.

Gealy, D. R., H. A. Agrama, and G. C. Eizenga. 2009. Exploring genetic and spatial structure of U.S. weedy red rice (*Oryza sativa*) in relation to rice relatives worldwide. *Weed Science* 57: 627–643.

Goff, S., D. Ricke, T. H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100.

Greif, G., M. Rodriguez, A. Reyna-Mello, C. Robello, and F. Alvarez-Valin. 2015. Kinetoplast adaptations in American strains from *Trypanosoma vivax. Mutation Research* 773: 69–82.

Gross, B. L., and K. M. Olsen. 2010. Genetic perspectives on crop domestication. *Trends in Plant Science* 15: 529–537.

Huang, C., M. Ayliffe, and J. Timmis. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* 422: 72–76.

Huang, C. Y., N. Grünheit, N. Ahmadinejad, J. N. Timmis, and W. Martin. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology* 138: 1723–1733.

Kane, N., S. Sveinsson, H. Dempewolf, J. Y. Yang, D. Zhang, J. M. M. Engels, and Q. Cronk. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* 99: 320–329.

Kaya, B., O. Erol, L. Sik, A. Gedik, H. Ozkan, and M. Tanyolac. 2014. Chloroplast genome sequencing of 7 *Crocus* species through Illumina platform from total DNA. Plant and Animal Genome XXII, San Diego, California, USA [online abstract]. Website https://pag.confex.com/pag/xxii/webprogram/Paper9684.html.

Kumagai, M., L. Wang, and S. Ueda. 2010. Genetic diversity and evolutionary relationships in genus *Oryza* revealed by using highly variable regions of chloroplast DNA. *Gene* 462: 44–51.

Leister, D. 2005. Origin, evolution and genetic effect of nuclear insertion of organelle DNA. *Trends in Genetics* 21: 655–663.

Londo, J. P., and B. A. Schaal. 2007. Origins and population genetics of weedy red rice in the USA. *Molecular Ecology* 16: 4523–4535.

Martin, W. 2003. Gene transfer from organelles to the nucleus: Frequent and big chunks. *Proceedings of the National Academy of Sciences, USA* 100: 8612–8614.

Matsuo, M., Y. Ito, R. Yamauchi, and J. Obokata. 2005. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. *Plant Cell* 17: 665–675.

Matsuoka, Y., Y. Yamazaki, Y. Ogihara, and K. Tsunewaki. 2002. Whole chloroplast genome comparison of rice, maize and wheat: Implication for chloroplast gene diversification and phylogeny of cereals. *Molecular Biology and Evolution* 19: 2084–2091.

McPherson, H., M. van der Merwe, S. K. Delaney, M. A. Edwards, R. J. Henry, E. McIntosh, P. D. Rymer, et al. 2013. Capturing chloroplast variation for molecular ecology studies: A simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology* 13: 8–19.

Neale, D., M. Saghai-Maroof, R. Allard, Q. Zhang, and R. Jorgensen. 1988. Chloroplast DNA diversity in populations of wild and cultivated barley. *Genetics* 120: 1105–1110.

Nock, C. J., D. L. E. Waters, M. A. Edwards, S. G. Bowen, N. Rice, G. M. Cordeiro, and R. J. Henry. 2011. Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* 9: 328–333.

Notsu, Y., S. Masood, T. Nishikawa, N. Kubo, G. Akiduki, M. Nakazono, A. Hirai, et al. 2002. The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: Frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Molecular Genetics and Genomics* 268: 434–445.

Noutsos, C., T. Kleine, U. Armbruster, G. DalCorso, and D. Leister. 2007. Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends in Genetics* 23: 597–601.

Olsen, K., A. Caicedo, and Y. Jia. 2007. Evolutionary genomics of weedy rice in the USA. *Journal of Integrative Plant Biology* 49: 811–816.

Provan, J., W. Powell, and P. M. Hollingsworth. 2001. Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution* 16: 142–147.

Reagon, M., C. S. Thurber, B. L. Gross, K. M. Olsen, Y. Jia, and A. L. Caicedo. 2010. Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. *BMC Evolutionary Biology* 10: 180–191.

Richly, E., and D. Leister. 2004. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular Biology and Evolution* 21: 1972–1980.

Shahid Masood, M., T. Nishikawa, S. Fukuoka, P. Njenga, T. Tsudzuki, and K. Kadowaki. 2004. The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: First genome wide comparative sequence analysis of wild and cultivated rice. *Gene* 340: 133–139.

Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H.-R. Kates, X. Qi, S. F. Brockington, et al. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1: 1200497.

Subudhi, P. K., P. K. Singh, T. DeLeon, A. Parco, R. Karan, H. Biradar, M. A. Cohn, and T. Sasaki. 2014. Mapping of seed shattering loci provides insights into origin of weedy rice and rice domestication. *Journal of Heredity* 105: 276–287.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

Tang, J., H. Xia, M. Cao, X. Zhang, W. Zeng, and S. Hu. 2004. A comparison of rice chloroplast genomes. *Plant Physiology* 135: 412–420.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673–4680.

Vieira, L. d. N., H. Faoro, M. Rogalski, H. Pacheco de Freitas Fraga, R. Alves Cardoso, E. Maltempi de Souza, F. De Oliveira Pedrosa, et al. 2014. The complete chloroplast genome sequence of *Podocarpus lambertii*: Genome structure, evolutionary aspects, gene content and SSR detection. *PLoS One* 9: e90618.

Wang, W., and J. Messing. 2011. High-throughput sequencing of three *Lemnoideae* (duckweeds) chloroplast genomes from total DNA. *PLoS One* 6: e24670.

Warwick, S., and N. Stewart. 2005. Crops come from wild plants: How domestication, transgenes, and linkage together shape ferality. *In* J. Gressel [ed.], Crop ferality and volunteerism, 9–25. CRC Press, Boca Raton, Florida, USA.

Waters, D. L. E., C. J. Nock, R. Ishikawa, N. Rice, and R. J. Henry. 2012. Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. *Ecology and Evolution* 2: 211–217.

Wolfe, K. H., W. H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.

Yang, M., X. Zhang, G. Liu, Y. Yin, K. Chen, Q. Yun, D. Zhao, et al. 2010. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS One* 5: e12762.

Yu, J., S. Hu, J. Wang, G. Wong, S. Li, B. Liu, Y. Deng, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.

Zambon, A. C., L. Zhang, S. Minovitsky, J. R. Kanter, S. Prabhakar, N. Salomonis, K. Vranizan, et al. 2005. Gene expression patterns define key transcriptional events in cell-cycle regulation by cAMP and protein kinase A. *Proceedings of the National Academy of Sciences, USA* 102: 8561–8566.

Zhang, T., X. Zhang, S. Hu, and J. Yu. 2011. An efficient procedure for plant organelle genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods* 7: 38–46.