Mutational Dynamics of Aroid Chloroplast Genomes

Ibrar Ahmed[1,2,*], Patrick J. Biggs[3], Peter J. Matthews[4], Lesley J. Collins[5], Michael D. Hendy[6], and Peter J. Lockhart[1,5]

[1]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand

[2]Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan

[3]Institute of Veterinary, Animal, and Biomedical Sciences, Massey University, Palmerston North, New Zealand

[4]Department of Social Research, National Museum of Ethnology, Osaka, Japan

[5]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[6]Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

*Corresponding author: E-mail: I.Ahmed@massey.ac.nz, iaqureshi_qau@yahoo.com.

Accepted: November 25, 2012

Data deposition: Chloroplast genome sequences of two morphotypes of taro (*Colocasia esculenta* var. GP and *Colocasia esculenta* var. RR) have been deposited to GenBank under accession numbers JN105689 and JN105690, respectively. The sequence alignments generated for this work will be provided on request.

Abstract

A characteristic feature of eukaryote and prokaryote genomes is the co-occurrence of nucleotide substitution and insertion/deletion (indel) mutations. Although similar observations have also been made for chloroplast DNA, genome-wide associations have not been reported. We determined the chloroplast genome sequences for two morphotypes of taro (*Colocasia esculenta*; family Araceae) and compared these with four publicly available aroid chloroplast genomes. Here, we report the extent of genome-wide association between direct and inverted repeats, indels, and substitutions in these aroid chloroplast genomes. We suggest that alternative but not mutually exclusive hypotheses explain the mutational dynamics of chloroplast genome evolution.

Key words: Araceae, indels, phylogeny, repeats, substitution mutations, taro.

Introduction

Comparative studies of chloroplast genome sequences have investigated divergences spanning an enormous range of evolutionary times. These have included studies of intraspecific variation in domesticated plants (Yamane et al. 2003), studies of early land plant evolution (Kugita et al. 2003) and also the earliest events of oxygenic photosynthesis (Martin et al. 2002). This range of comparisons has been possible because of the conservative nature of chloroplast (cp) genome evolution (Palmer 1985), which involves relatively slow rates of sequence evolution in some parts of the cp genome (Sammut and Huttley 2011) and elevated rates in other parts (Magee et al. 2010; Sammut and Huttley 2011).

Molecular evolution of the cp genome sequences is typically modeled as a time reversible substitution process, in which changes at any one site are independent of changes at any other site (Liò and Goldman 1998; Drouin et al. 2008). However, observations have suggested more complex processes of evolution in which both lineage-specific and nonrandom spatial patterns of substitution occur (Liò and Goldman 1998; Lee et al. 2007; Gruenheit et al. 2008; Magee et al. 2010; Wu et al. 2011; Zhong et al. 2011). Such observations have practical significance for understanding the limitations of cp genomes in phylogenetic analyses of highly diverged lineages (Gruenheit et al. 2008), and for understanding the mutational dynamics of "hotspot" regions studied in comparisons of closely related taxa (Shaw et al. 2007; Worberg et al. 2007).

In prokaryotes and eukaryotes, analyses of DNA sequence alignments show that indels commonly occur in regions that are hotspots for nucleotide substitutions. Alternative hypotheses have been proposed to explain this co-occurrence. It has been suggested that certain genome regions are predisposed to mutational events such as substitutions and insertion/deletions—"the regional difference hypothesis" (Silva and Kondrashov 2002; Hardison et al. 2003). A second hypothesis explaining the association between indels and substitutions is that certain (large) indels act to induce substitutions through a DNA repair process that recruits error-prone DNA polymerases—"the indel-induced mutation hypothesis" (Tian et al. 2008; Zhu et al. 2009). A third and related hypothesis is that it is the presence of repeat sequences rather than indels per se, that actually promotes replication fork arrest, causing the recruitment of the error-prone DNA polymerases, and in doing so generates nucleotide substitutions (McDonald et al. 2011).

These hypotheses have not been explicitly investigated in cp genomes yet these genomes are known to contain very high densities of direct and inverted oligonucleotide repeats. Associations between repeats, indels, and substitutions have also been reported in cp genomes (McLenachan et al. 2000; Lockhart et al. 2001 and references cited therein). Cp genome repeats include simple sequence repeats (SSRs, also known as microsatellites) and other moderate to long (8–48 bp) repeats. Contraction and expansion of the SSR units, caused by slipped strand mispairing during DNA replication (Levinson and Gutman 1987), frequently produces short indels at these SSR loci (Masood et al. 2004). The moderate-to-long repeats have also been suggested to cause indels (Kawata et al. 1997) and inversions (Kim and Lee 2005; Whitlock et al. 2010). Most angiosperms also contain two large inverted repeat (IR) regions, commonly known as IRa and IRb (5–76 kb; Palmer 1991).

Here, we report the cp genome sequences of two morphotypes of taro (*Colocasia esculenta*; var. RR and var. GP; Matthews 1985) and examine the genome wide association of repeats (excluding IRa and IRb), indels and substitutions in the cp genomes of these taro morphotypes and four other distantly related aroids in the duckweed (Lemnoideae) subfamily.

The *Colocasia esculenta* cp Genome

Colocasia esculenta (L.) Schott, commonly known as taro, is an ancient root crop in subfamily Aroideae of the monocot family Araceae. This species is distributed in the tropical to subtropical and some temperate regions of the world (Bown 1988).

Gene arrangement and other features of the *C. esculenta* cp genome are shown in figure 1. Size of the cp genome was 162,546 bp (GC content: 36.1%) in var. RR, and 162,424 bp (GC content: 36.2%) in var. GP. The GC content varied from 42.4% in IRs to 34.4% in the large single copy (LSC) and only 28.4% in the small single copy (SSC) regions of the taro cp genomes. Higher GC content in the IR regions corresponded to the presence of the ribosomal DNA locus. Pair-wise sequence alignment between the taro cp genomes revealed 99.5% identical sequence, 241 substitutions, and 92 indels. The LSC region contained 141 (58.6%) substitutions and 65 (71%) indels, the SSC region contained 83 (34.4%) substitutions and 25 (27%) indels, whereas the IRa and IRb regions collectively contained only 17 (7%) substitutions and 2 (2%) indels, indicating that the IR was the most evolutionarily stable region. Prominent differences between the two taro cp genomes were found at the IRb–SSC boundary (numerous indels making up a 91 bp difference in size), and at the SSC–IRa boundary (a shift of 64 bp in the repeat boundary without causing indels). Thus, the IR boundaries at both ends of the SSC region were polymorphic at intraspecific level in taro. Polymorphism between the two taro cp genomes included 59 substitutions in 29 protein coding genes. Among these, the most polymorphic gene was *ycf1* even when normalized for its size, showing 16 substitutions between the two genomes. Some protein coding genes (including *atpH*, *psbM*, and *psbZ)* and tRNA genes (including *trnH*, *trnG*, and *trnW*) in particular showed a relatively high density of substitutions and indels within 20 bp upstream of their respective coding regions. Whether this observation has functional significance needs to be further explored. A set of 30 functional tRNA genes covering all 20 amino acids required for protein synthesis was present in the taro cp genome.

The overall gene arrangement was similar between taro (*C. esculenta*) and the duckweed (*Lemna minor*; Mardanov et al. 2008) cp genomes. However, notable differences were as follows:

(a) *trnH* gene is reported in the LSC region in duckweed, whereas the 5′-end of this gene extended into the IRa region in taro.
(b) *infA* gene is completely missing in duckweed, but a pseudo-copy of this gene with internal stop codons was observed in taro.
(c) A single functional *rpl2* gene spanning the IRb–LSC boundary is reported in duckweed, whereas two functional copies of this gene were found in taro, one in each of the IR regions.
(d) A pseudo-copy of *ycf68* gene is reported in duckweed; however, a functional copy of this gene was observed in each IR region in taro.
(e) Duckweed has *ycf1* and *rps15* genes within its IR regions, whereas these genes were placed within the SSC region in taro.

The *infA* gene is considered to be among the most mobile cp genes. Multiple independent gene transfers from cp to nuclear genomes are thought to have occurred during angiosperm evolution (Millen et al. 2001). The *ycf68* gene is present in a range of plant families as a functional or a pseudo-gene, and may have functional significance even in its noncoding form (Raubeson et al. 2007). Other genes showing variation in comparison with *L. minor* include *trnH, rpl2, ycf1,* and *rps15*. These are located at or near the boundaries of IRs with single copy regions. These boundaries are well known to exhibit expansion and contraction in angiosperms (Whitlock et al. 2010) as well as in gymnosperms (Lin et al. 2012). A comparison of

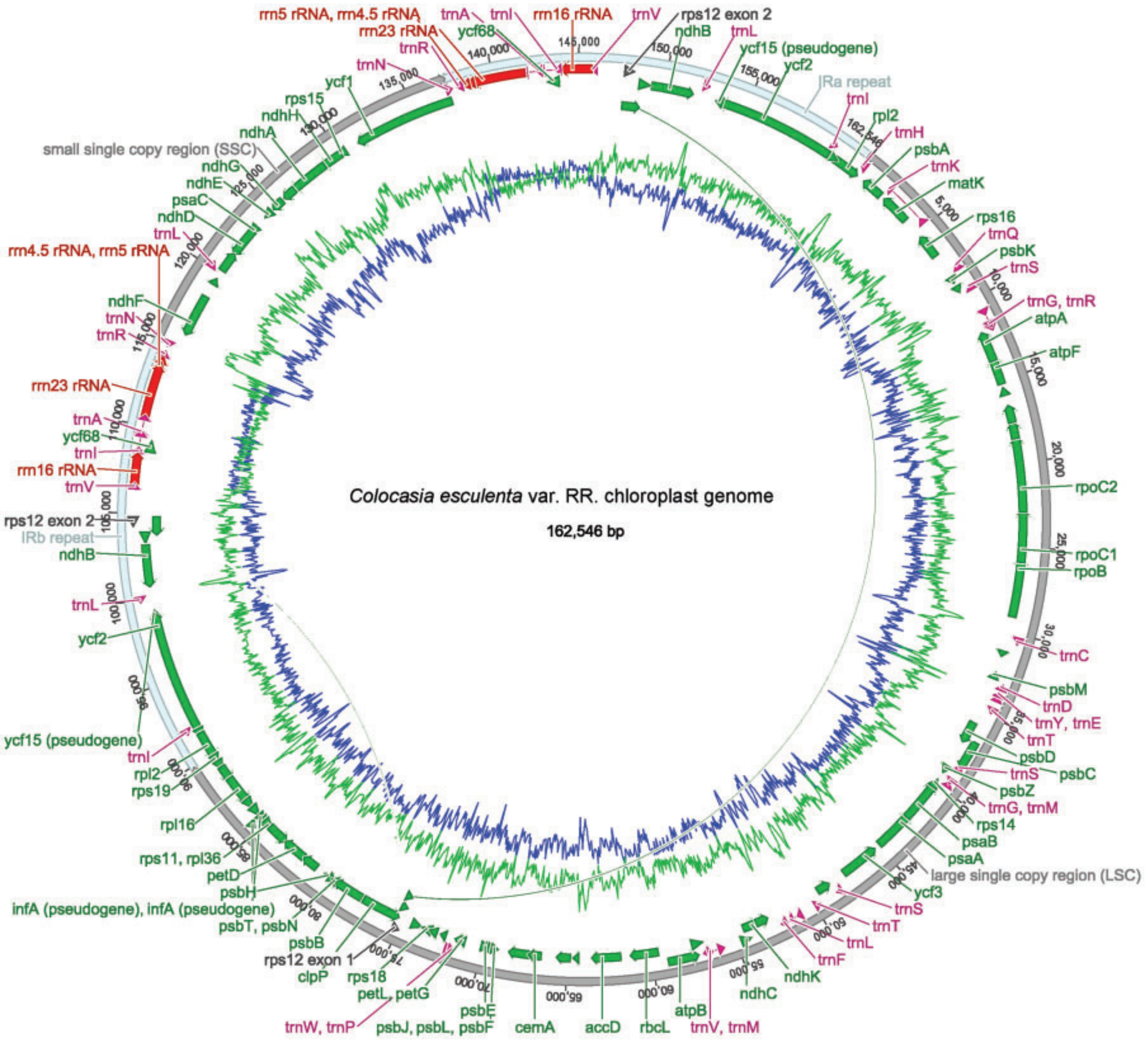

Fig. 1.—*Colocasia esculenta* var. RR chloroplast genome (GenBank accession: JN105690). Brown lines in the outer circle represent the LSC and SSC regions, cyan lines represent the IRs, whereas inner green lines show AT and blue lines show GC percentage throughout the cp genome.

Table 1
Comparison among Total Size (bp) and Sizes of the LSC, SSC, and Two IR Regions in Taro and Other Aroid Chloroplast Genomes

Species	GenBank ID	Genome Size	LSC	SSC	IR
Colocasia esculenta var. GP	JN105689	162,424	89,670 (55.21)	22,208 (13.67)	25,273 (31.12)
C. esculenta var. RR	JN105690	162,546	89,817 (55.26)	22,075 (13.58)	25,327 (31.16)
Lemna minor	NC010109	165,955	89,906 (54.17)	13,603 (8.20)	31,223 (37.63)
Spirodela polyrhiza	JN160603	168,788	91,222 (54.04)	14,056 (8.33)	31,755 (37.63)
Wolffiella lingulata	JN160604	169,337	92,015 (54.34)	13,956 (8.24)	31,683 (37.42)
Wolffia australiana	JN160605	168,704	91,454 (54.21)	13,394 (7.94)	31,930 (37.85)

Note.—Percentage proportions of the LSC, SSC, and IRs are given in parenthesis.

the size and percentage proportions of LSC, SSC, and IR regions in taro and other aroid cp genomes is given in table 1. Characterization of these boundaries is likely to provide useful insights into the dynamics of single copy—IR boundary shifts in *Colocasia* and other aroid cp genomes.

Correlations among Repeats, Indels, and Substitutions in Aroid cp Genomes

We have visualized the extent to which indel and substitution mutations are nonrandomly distributed between taro and other aroid cp genomes, using a Circos

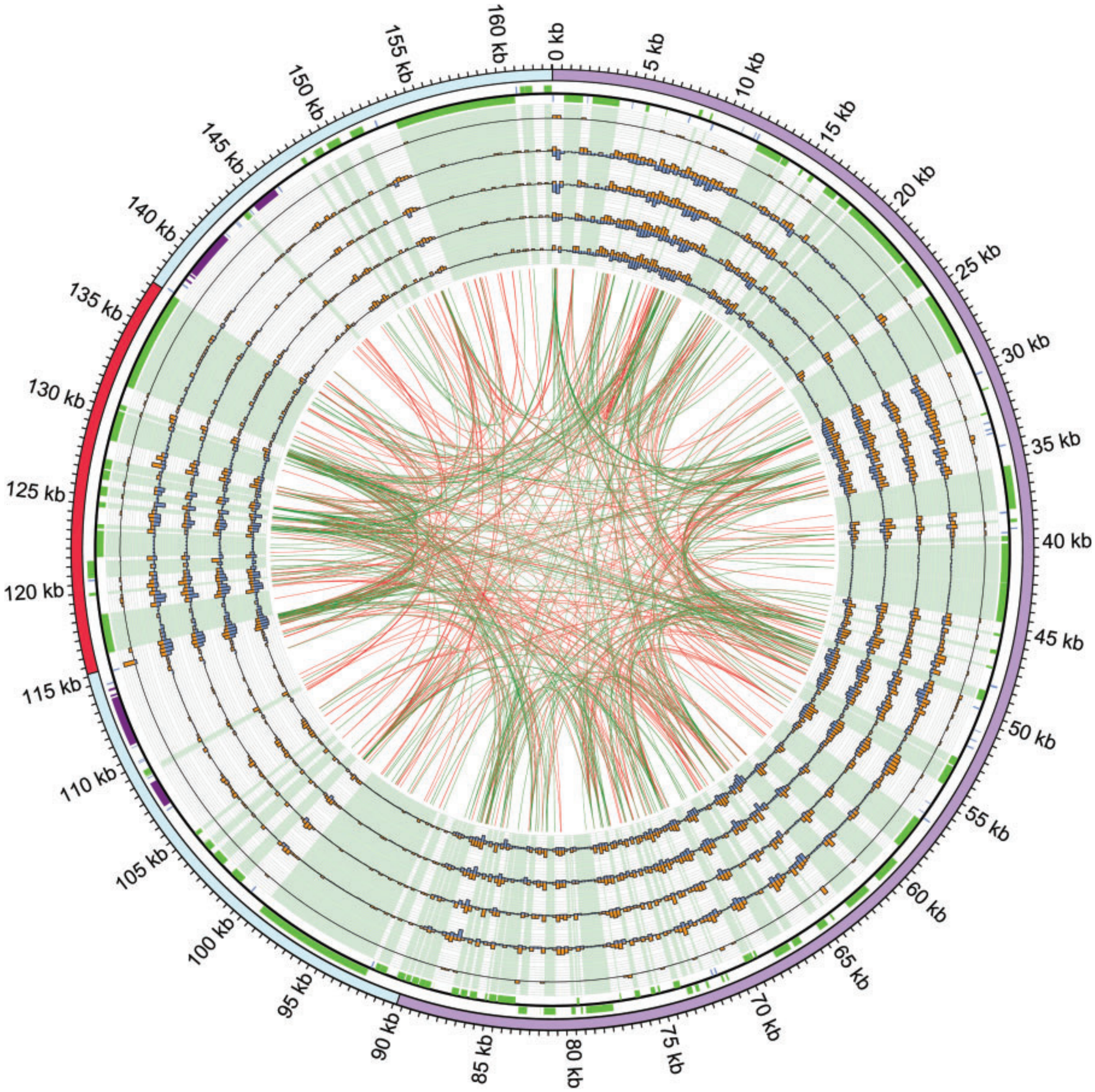

FIG. 2.—Circos plot of taro (*Colocasia esculenta*) var. RR showing the relationship between short repeats within the chloroplast genome and distribution of indels and substitutions in pairwise comparisons of taro var. RR cp genome with other aroid cp genomes. All data in the histogram tracks are shown in nonoverlapping 250 bp bins, with the taro var. RR genome taken as a reference for the coordinate space. Tracks from the outermost to innermost show: taro var. RR chloroplast ideogram (LSC in purple, SSC in red, and IRs in cyan); genome annotation on the positive and negative strand (genes in green; tRNAs in blue and rRNAs in purple); five circular plots showing comparisons between *C. esculenta* var. RR with 1) *C. esculenta* var. GP, 2) *Lemna minor*, 3) *Spirodela polyrhiza*, 4) *Wolffia australiana,* and 5) *Wolffiella lingulata.* For each genome comparison, the number of indels in each 250 bp bin is shown in orange (scale of 0–10), and the number of substitutions is shown in blue (scale of 0–160). Across these five plots, the light green coloring indicates the coding regions. The figure center shows the results of Reputer mapping using the taro var. RR chloroplast genome. Two ends of a red line mark the two locations of the forward (direct) repeats, whereas those of a green line mark the two locations of the reverse (inverted) repeats on the genome. In this part of the figure, the large IRs are not plotted, as they would obscure a large part of the figure. Number of repeats shown in the diagram is 667, with a size range from 15 to 48 bp (average repeat size: 16 bp).

(Krzywinski et al. 2009) plot as given in figure 2. This plot shows that substitutions are very closely correlated in their distribution with moderate (15 bp) to long (48 bp) repeat sequences mainly found in noncoding regions. Correlation (*r*) and related values for these data are given in table 2. Correlations were highly significant in comparisons of three types of mutations, including 1) repeats and substitutions, 2) substitutions and indels, and 3) repeats and indels. In a pairwise comparison of the two closely related taro genomes, the strength of correlations was greatest for

Table 2

Comparisons among the Pairwise Alignments (*Colocasia esculenta* var. RR taken as a Reference) to Calculate the Correlations between 1) Repeats and Substitutions, 2) Insertion-Deletions (indels) and Substitutions, and 3) Repeats and Indels

Comparison	*C. esculenta* var. GP	*Wolffiella lingulata*	*Wolffia australiana*	*Lemna minor*	*Spirodela polyrhiza*
Repeats and substitutions					
Correlation between repeats and substitutions (*r*)	0.245	0.391	0.416	0.424	0.491
Significance of correlation (*t*)	6.44***	10.81***	11.657***	11.92***	14.37***
Coefficient of determination (r^2)	0.060	0.152	0.173	0.180	0.241
Insertion–deletions (indels) and substitutions					
Correlation between indels and substitutions (*r*)	0.391	0.220	0.245	0.323	0.387
Significance of correlation (*t*)	10.82***	5.75***	6.43***	8.71***	10.69***
Coefficient of determination (r^2)	0.153	0.048	0.060	0.105	0.150
Repeats and indels					
Correlation between repeats and indels (*r*)	0.640	0.168	0.178	0.224	0.212
Significance of correlation (*t*)	21.20***	4.33***	4.59***	5.87***	5.51***
Coefficient of determination (r^2)	0.409	0.028	0.032	0.050	0.045

NOTE.—The alignments compared closely related (var. RR to var. GP) and distantly related (var. RR to *W. lingulata*, *W. australiana*, *L. minor*, and *S. polyrhiza*) aroid chloroplast genomes. The alignments were partitioned into 651 nonoverlapping bins of 250 bp size each to calculate these correlations.

***All correlations were highly significant at 0.001α and 649 degree of freedom.

"repeats and indels" followed by "substitutions and indels" and then "repeats and substitutions." In contrast, when pairwise comparison was made between a taro genome and a more distantly related aroid genome, the strength of correlations reversed. The strongest correlation was for "repeats and substitutions" followed by "substitutions and indels" and then "repeats and indels" (table 2). The strongest correlation value observed was for "repeats and indels" in comparison of the two taro genomes. Similar observations have previously been reported in prokaryotes and eukaryotes (Kawata et al. 1997; McDonald et al. 2011) and have led to a hypothesis that repeat sequences play a pivotal role in generation of indel and substitution mutations (McDonald et al. 2011).

Since Tian et al. (2008) proposed that moderate-to-large–sized indels induce substitutions in their surrounding sequences, we also investigated this relationship in a multiple sequence alignment (parental alignment) of all six aroid cp genomes. From the this parental alignment, we extracted data partitions containing distinct indel location points (ILPs) to make mutually exclusive partitions with respect to locations of the ILPs. Partition A contained ILPs associated with SSR indels in both coding and noncoding regions. Partition B contained ILPs associated with large (oligonucleotide long, non-SSR) indels in both coding and noncoding regions. Partition C contained ILPs in noncoding regions, associated with both SSR indels and large indels. Partition D contained ILPs in coding regions, associated with both SSR and large indels. The density of substitutions in all partitions was highly dependent upon inverse of distance from the ILPs (r^2 ranged from 0.85 to 0.97 for all bin sizes; supplementary fig. S1, Supplementary Material online). Higher substitution density in bins closer to the ILPs was a general trend in all five comparisons above, including the partition in which coding regions were removed (partition C); however, in this case, distance from the ILPs was relatively shorter than in the other four comparisons. The indel-induced mutation hypothesis was further explored in a comparison including the parental alignment and partitions A and B, as shown in figure 3. From this comparison, it is evident that the partition B (containing ILPs associated with large indels) displayed a higher density of substitutions closer to ILPs, and the density of substitutions decreased with an increase in distance from the ILPs. In contrast, the partition A (containing ILPs associated with SSRs) exhibited a low density of substitutions close to ILPs, and the density of substitutions showed a net increase with increase in distance from the ILPs. These observations are consistent with the indel-induced mutation hypothesis suggested for diploid eukaryote (Tian et al. 2008) as well as bacterial genomes (Zhu et al. 2009).

It is well known that certain regions of the chloroplast genome show different rates of mutations (Lee et al. 2007; Gruenheit et al. 2008; CBOL Plant Working Group 2009; Zhong et al. 2011). These are observations consistent with a regional difference hypothesis (Silva and Kondrashov 2002; Hardison et al. 2003) and the suggestion that purifying selection operates at both coding and noncoding regions (Petersen et al. 2011). However, these explanations are alone insufficient to explain substitution and indel patterns of the chloroplast genome. The extent of genome wide correlations reported here for indels, repeats, and substitution provides further support for the hypothesis by McDonald et al. (2011), which emphasizes the evolutionary importance of the repeats in causing mutations. In addition, our observations on substitution densities also provide support for an indel-induced mutation hypothesis (Tian et al. 2008;

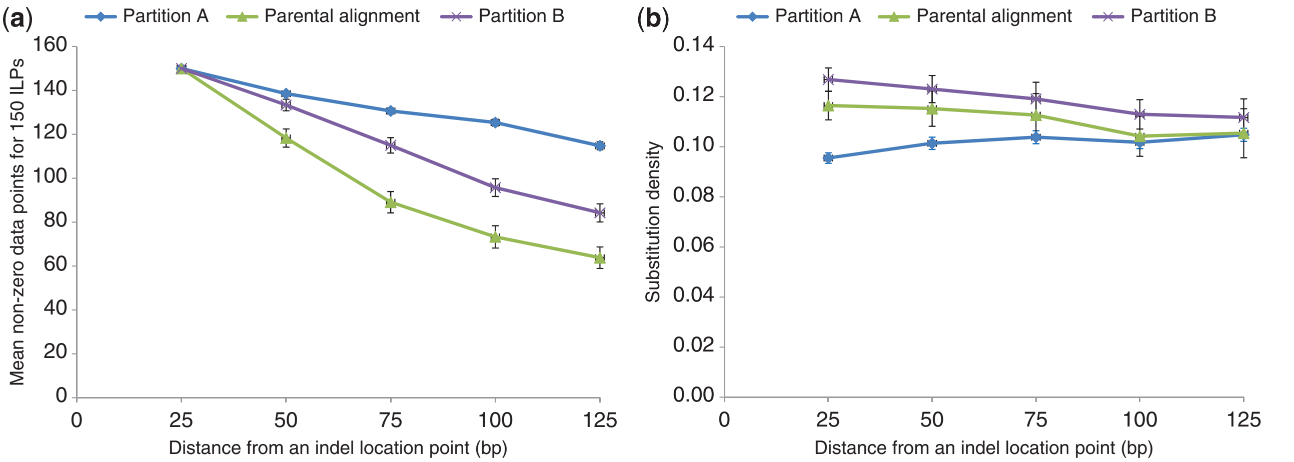

Fig. 3.—Results showing (*a*) the number of mean non-zero data points used to calculate the substitution density and (*b*) the values of substitution density in 125 bp sequence adjacent to the ILPs for the parental alignment as well as two of its partitions, A and B (Partition A contains ILPs associated with SSR indels in coding and noncoding regions, while partition B contains ILPs associated with large indels in coding and noncoding regions). Lower than 150 values for non-zero data points at >25 bp distance in (*a*) represents that taking an average for 1,000 random iterations, lesser than 150 ILPs are 125 bp apart from each other in all three types of comparisons.

Zhu et al. 2009) and further our understanding for the sometimes poor fit between time reversible substitution models and chloroplast sequence data. Perhaps, most interestingly, the relationship between repeats, substitutions, and indels implies that, if the distribution of repeat sequences in a chloroplast genome is determined, there is a possibility to predict the mutational hotspot regions and other sequences that are most appropriate for population genetic, phylogeographic, and phylogenetic analyses.

Materials and Methods

Taro plants (*C. esculenta* var RR; voucher number MPN:46548, and var GP; voucher number MPN:46549 in the Dame Ella Campbell Herbarium, Massey University, New Zealand) were obtained from the University of Auckland campus. Chloroplasts were enriched following procedure given in Atherton et al. (2010). DNA was extracted using a DNeasy Plant Mini Kit (Qiagen, USA) and quantified using a Qubit Fluorometer (Invitrogen) and Quant-iT-ds DNA HS Assay kit (Invitrogen). Illumina sequence reads were generated using the GAIIx platform at the Massey Genome Service, Massey University, New Zealand. Illumina sequencing produced 33 million reads of 75 base long (16.5 million paired-end reads) for var. RR, and 26.4 million reads of 75 base long (13.2 million paired-end reads) for var. GP. The reads were mapped to the duckweed cp genome (*L. minor*; Mardanov et al. 2008) using BWA mapping tool (Li and Durbin 2009). Mapping results were visualized using Tablet (Milne et al. 2010). The reads from var. RR were de novo assembled into contiguous sequences ("contigs") of variable lengths using Velvet (v.0.7.60; Zerbino and Birney 2008), as described elsewhere (Collins et al. 2008). These contigs were BLAST-searched (Altschul et al. 1997) to determine homology to the duckweed cp genome. The contigs of cp origin were assembled in Geneious Pro (Drummond et al. 2009) to deduce the cp genome of the taro var. RR morphotype. The two IRs were distinguished by visual inspection of the boundaries between the repeat and single copy regions. Genome annotation was carried out using Dual Organellar GenoMe Annotator (DOGMA; Wyman et al. 2004) and also by direct comparison with the duckweed cp genome. Contigs were generated similarly for the var. GP morphotype. The completed var. RR cp genome was then used as our reference genome to help assemble the var. GP cp genome. To verify integrity of the de novo assembly process, the original 75 base long reads from both taro samples were mapped back to their respective, assembled cp genomes. Summary statistics for the BWA mapping of 75 base long reads to the *L. minor* cp genome, as well as to their respective assembled var. RR and var. GP genomes are given in table 3.

The var. RR cp genome was pairwise aligned to the var. GP cp genome, as well as to four aroid cp genomes from the Lemnoideae subfamily, using DIALIGN alignment (Morgenstern 2004). The four aroid cp genomes included *L. minor* (GenBank ID: NC010109; Mardanov et al. 2008), *Spirodela polyrhiza* (GenBank ID: JN160603), *Wolffiella lingulata* (GenBank ID: JN160604), and *Wolffia australiana* (GenBank ID: JN160605; Wang and Messing 2011). Selecting *C. esculenta* var. RR cp genome as a reference for the coordinate positions, indels, and substitutions were counted in pairwise comparisons in nonoverlapping bins of 250 bp through the entire length of the aligned cp genomes (partitioning each of the five alignments into 651 bins). For the

Table 3

Summary Statistics for BWA Mapping of 75 Base, Paired-End Reads Obtained from the *Colocasia esculenta* var. RR and var. GP Morphotypes to the *Lemna minor* Chloroplast Genome and to Their Assembled Chloroplast Genomes

Parameter	*L. minor*						*C. esculenta* var. RR			*C. esculenta* var. GP		
	RR1	RR2	RPE	GR1	GR2	GPE	RR1	RR2	RPE	GR1	GR2	GPE
Genome coverage (%)	68.5	68.1	85	68.5	68.7	85.6	99.99	100	100	100	100	100
Average coverage depth	129	128	337	319	317	825	296	294	593	665	659	1,338
Maximum coverage depth	674	623	1,531	1,984	1,853	5,194	641	656	1,020	1,940	2,021	3,304

Note.—The acronyms RR1, RR2, and RPE represent mapping with the read 1, read 2, and paired-end (reads 1 and 2 taken together) reads obtained from the var. RR morphotype. Similarly, GR1, GR2, and GPE represent mapping with the read 1, read 2, and paired-end reads obtained from the var. GP morphotype.

substitution count, indels in the var. RR cp genome were deleted from the alignments to preserve the coordinate positions. Similar patterns of indel and substitution counts were obtained using a MAFFT alignment (Katoh et al. 2005; results not shown). A total of 5,000 forward (direct) and reverse (inverted) repeats with a minimum size of 14 bp, a maximum size of 48 bp, and a maximum of three nucleotide mismatch between the two repeat copies in the taro var. RR cp genome were calculated using Reputer (Kurtz et al. 2001). Of these 5,000 repeats, 667 locations of the forward and reverse in var. RR (minimum size: 15 bp; zero mismatch between the two copies), as well as polymorphic sites (indels and substitutions) in all five pairwise comparisons with respect to the var. RR cp genome were plotted as a circular diagram using Circos (Krzywinski et al. 2009). Correlations (r) were calculated between numbers of 1) repeats and substitutions, 2) substitutions and indels, and 3) repeats and indels. This was done for comparisons of closely related (two taro genomes) and distantly related (taro with other Lemnoideae) cp genomes. The correlation values (r) were used to determine the significance of correlation (t) and the coefficient of determination (r^2), according to Lowry (2012).

To further investigate the relationships between substitutions and indels, a multiple sequence alignment of all six aroid cp genomes was generated using DIALIGN alignment (Morgenstern 2004). Hyper variable regions causing problems in the alignment were removed to ensure conservative estimates. This 122-kb long parental alignment contained 457 ILPs. This parental alignment was used to generate mutually exclusive alignment combinations with respect to locations of the ILPs, to include ILPs associated with coding and noncoding regions and SSR indels (171 ILPs; partition A) and coding and noncoding regions and large indels (286 ILPs; partition B). The parental alignment was also used to generate two further mutually exclusive alignment combinations to include ILPs associated with SSR indels and large indels in noncoding regions (376 ILPs; partition C) and SSR indels and large indels in coding regions (81 ILPs; partition D). Using a Perl script, we counted the number and positions of substitutions with respect to the ILPs, and plotted the substitution density as a function of distance from the ILPs in nonoverlapping bins of 50, 100, 150, 200, and 250 bp each for the parental alignment as well as partitions A, B, and D; and 10, 20, 30, 40, and 50 bp for the partition C. The effect of large indels in causing substitutions was further explored by comparing first three alignment combinations (parental alignment along with partitions A and B) and plotting the substitution density as a function of distance from the ILPs in 125 bp sequence adjacent to the ILPs. For this purpose, a jacknifing approach was used to randomly select 150 ILPs from each of these three partitions with 1,000 random iterations to count substitutions within the 125 bp distance, divided into five nonoverlapping bins of 25 bp in size. Plots showing the relationship between substitutions and ILPs were generated using MS Excel 2010 worksheets.

Supplementary Material

Supplementary figure S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge the financial assistance of the New Zealand Royal Society (The New Zealand Marsden Fund and James Cook Fellowship scheme) and the Higher Education Commission, Government of Pakistan.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Atherton RA, et al. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 6:22.

Bown D. 1988. Aroids: plants of the arum family. London: Century Hutchinson.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proc Natl Acad Sci U S A. 106:12794–12797.

Collins LJ, Biggs PJ, Voelckel C. 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform. 21:3–14.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast, and nuclear genomes of seed plants. Mol Phylogenet Evol. 49:827–831.

Drummond AJ, et al. 2009. Geneious. Available from: http://www.geneious.com/ (last accessed December 10, 2012).

Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol. 25: 1512–1520.

Hardison RC, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518.

Kawata M, Harada T, Shimamoto Y, Oono K, Takaiwa F. 1997. Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). Curr Genet. 31:179–184.

Kim K-J, Lee H-L. 2005. Widespread occurrence of small inversions in the chloroplast genomes of land plants. Mol Cells. 19:104–113.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Kugita M, et al. 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucleic Acids Res. 31:716–721.

Kurtz S, et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29:4633–4642.

Lee H-L, Jansen RK, Chumley TW, Kim K-J. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol. 24:1161–1180.

Levinson G, Gutman G. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4:203–221.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Lin CP, Wu CS, Huang YY, Chaw SM. 2012. The complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. Genome Biol Evol. 4:374–381.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8:1233–1244.

Lockhart PJ, et al. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. Ann MO Bot Gard. 88:458–477.

Lowry R. 2012. Concepts & applications of inferential statistics. Available from: http://vassarstats.net/textbook/index.html (last accessed August 14, 2012).

Magee AM, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700–1710.

Mardanov AV, et al. 2008. Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. J Mol Evol. 66:555–564.

Martin W, et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A. 99: 12246–12251.

Masood MS, et al. 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.

Matthews PJ. 1985. Nga taro o Aotearoa. J Polynesian Soc. 94:253–272.

McDonald MJ, Wang W-C, Huang H-D, Leu J-Y. 2011. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biol. 9:e1000622.

McLenachan PA, et al. 2000. Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. Mol Ecol. 9:1899–1903.

Millen RS, et al. 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645–658.

Milne I, et al. 2010. Tablet—next generation sequence assembly visualization. Bioinformatics 26:401–402.

Morgenstern B. 2004. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res. 32:W33–W36.

Palmer JD. 1985. Chloroplast DNA and molecular phylogeny. Bioessays 2: 263–267.

Palmer JD. 1991. Plastid chromosomes: structure and evolution. In: Vasil IK, Bogorad L, editors. Cell culture and somatic cell genetics in plants. Vol. 7A: The molecular biology of plastids. San Diego (CA): Academic Press. p. 5–53.

Petersen K, Schöttler MA, Karcher D, Thiele W, Bock R. 2011. Elimination of a group II intron from a plastid gene causes a mutant phenotype. Nucleic Acids Res. 39:5181–5192.

Raubeson LA, et al. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 8:174.

Sammut R, Huttley G. 2011. Regional context in the alignment of biological sequence pairs. J Mol Evol. 72:147–159.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am J Bot. 94:275–288.

Silva JC, Kondrashov AS. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends Genet. 18: 544–547.

Tian D, et al. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455:105–108.

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

Whitlock BA, Hale AM, Groff PA. 2010. Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. PLoS One 5: e11533.

Worberg A, et al. 2007. Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. Organisms Divers Evol. 7: 55–77.

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. 3: 1284–1295.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Yamane K, Yasui Y, Ohnishi O. 2003. Intraspecific cpDNA variations of diploid and tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). Am J Bot. 90:339–346.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhong B, et al. 2011. Systematic error in seed plant phylogenomics. Genome Biol Evol. 3:1340–1348.

Zhu L, Wang Q, Tang P, Araki H, Tian D. 2009. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. Mol Biol Evol. 26:2353–2361.

Associate editor: Shu-Miaw Chaw