Plastome Sequencing of Ten Nonmodel Crop Species Uncovers a Large Insertion of Mitochondrial DNA in Cashew

Samar O. Rabah, Chahee Lee, Nahid H. Hajrah, Rania M. Makki, Hesham F. Alharby, Alawiah M. Alhebshi, Jamal S.M. Sabir, Robert K. Jansen, and Tracey A. Ruhlman*

Abstract
In plant evolution, intracellular gene transfer (IGT) is a prevalent, ongoing process. While nuclear and mitochondrial genomes are known to integrate foreign DNA via IGT and horizontal gene transfer (HGT), plastid genomes (plastomes) have resisted foreign DNA incorporation and only recently has IGT been uncovered in the plastomes of a few land plants. In this study, we completed plastome sequences for 10 crop species and describe a number of structural features including variation in gene and intron content, inversions, and expansion and contraction of the inverted repeat (IR). We identified a putative rpl22 in cinnamon (Cinnamomum verum J. Presl) and other sequenced Lauraceae and an apparent functional transfer of rpl23 to the nucleus of quinoa (Chenopodium quinoa Willd.). In the orchard tree cashew (Anacardium occidentale L.), we report the insertion of an ~6.7-kb fragment of mitochondrial DNA into the plastome IR. BLASTn analyses returned high identity hits to mitogenome sequences including an intact ccmB open reading frame. Using three plastome markers for five species of Anacardium, we generated a phylogeny to investigate the distribution and timing of the insertion. Four species share the insertion, suggesting that this event occurred <20 million yr ago in a single clade in the genus. Our study extends the observation of mitochondrial to plastome IGT to include long-lived tree species. While previous studies have suggested possible mechanisms facilitating IGT to the plastome, more examples of this phenomenon, along with more complete mitogenome sequences, will be required before a common, or variable, mechanism can be elucidated.

Core Ideas
- DNA sequence data provides valuable information for biotechnology and evolutionary studies.
- Plastid genomes (plastomes) of 10 nonmodel crop species were sequenced.
- Inversions, gene divergence and loss, and IR boundary variation were identified.
- Transfer of mitochondrial DNA to the plastome was found in Anacardium (cashew).

The emergence of contemporary genomics has dispelled long-held hypotheses fueled by the Darwinian notion of evolution by vertical decent with modification. Drawing on phenotypic data, early investigators could not have predicted the impact of HGT on both the universality of the genetic code and diversity of organisms.

S.O. Rabah, H.F. Alharby, Dep. of Biological Sciences, Faculty of Science, King Abdulaziz Univ., Jeddah 21589, Saudi Arabia; C. Lee, R.K. Jansen, T.A. Ruhlman, Dep. of Integrative Biology, Univ. of Texas at Austin, Austin, TX 78712; N.H. Hajrah, R.M. Makki, A.M. Alhebshi, J.S.M. Sabir, R.K. Jansen, Genomics and Biotechnology Research Group, Dep. of Biological Sciences, Faculty of Science, King Abdulaziz Univ. (KAU), P.O. Box 80141, Jeddah 21589, Saudi Arabia. S.O. Rabah and C. Lee contributed equally to the work. Received 10 Mar. 2017. Accepted 10 July 2017. *Corresponding author (truhlman@austin.utexas.edu).

Abbreviations: aa, amino acid; ARF, auxin response factor; GC, guanine–cytosine; GSAF, Genome Sequencing and Analysis Facility; HGT, horizontal gene transfer; IGT, intracellular gene transfer; IR, inverted repeat; LSC, large single copy; MAFFT, multiple alignment using fast Fourier transform; MCS, membrane contact sites; ML, maximum likelihood; mtDNA, mitochondrial DNA; ncDNA, nuclear DNA; PCR, polymerase chain reaction; PEG, polyethylene glycol; pLDNA, plastid DNA; SC, single copy; SSC, small single copy; TACC, Texas Advanced Computing Center; UT–Austin, University of Texas–Austin.
found on Earth (Vetsigian et al., 2006). Although first recognized among eubacteria (Tatum and Lederberg, 1947), HGT has been detected across all domains of life and has shifted our views on the phylogeny of organisms from one of bifurcation to a more reticulate, web-like mode of evolution (Soucy et al., 2015).

Just as the sharing of DNA sequence among unrelated organisms has shaped their evolutionary history, so has the transfer of sequences among the genome-bearing compartments of individual cells shaped the evolution of eukaryotic species. Intracellular gene transfer, along with HGT, has played a pivotal role in the evolution of multicellularity and the oxygenation of Earth’s atmosphere, facilitating the evolution of plant and animal life (Timmis et al., 2004). The free-living, single-celled organisms that ultimately became the mitochondria, and later plastids, of eukaryotic cells through endosymbiosis contained the necessary complement of genetic material for survival in the extracellular environment. Once housed within the host cell, much of that genetic material was transferred to the host nuclear genome. This massive transfer of DNA has the transfer of sequences among the genome-bearing compartments of foreign DNA in plastomes been uncovered, including DNA of nonplastome origin in mitogenomes (Ohtani et al., 2015). The transfer of DNA sequence from both plastid and nuclear DNA (ncDNA), although, for the most part, the identity was due to the presence of sequences of plastid or nuclear origin in mitogenomes (Ohtani et al., 2002; Chumley et al., 2006). Only recently have legitimate cases of foreign DNA in plastomes been uncovered, facilitated in part by the availability of complete mitogenome sequences. The first, from Goremykin et al. (2009), was discovered during the sequencing of the grapevine (Vitis vinifera L.) mitogenome. An ~1.5-kb fragment situated in the rps12_3′-trnV-GAC intergenic spacer of the plastome IR of carrot (Daucus carota L. [Apiaceae]) (Ruhlman et al., 2006) contained two regions showing high similarity to grapevine and other published mitogenomes but was absent from sequenced plastomes. Within the larger insertion a fragment (74 bp) was identified with 94 to 95% nucleotide sequence identity to mitochondrial coxl. The subsequent sequencing and analysis of the carrot mitogenome revealed the presence of plastid DNA (ptDNA; Iorzizzo et al., 2012b). The integration of carrot mtDNA in the plastome was confirmed and the insert was identified across the Daucinae clade and in Caulacalis platycarpus L. (Torilidinae) (Iorzizzo et al., 2012b; Spooner et al., 2017). Within Apiaceae, the plastomes of Petroselinum and Crithmum also contain DNA sequence with no identity to angiosperm plastome sequences (Downie and Jansen, 2015). These putative insertions are situated between the boundary of IRAs and the 3′ end of trnH-GUG. BLASTn analysis of the Petroselinum sequence (345 bp) returned a best hit of 122 bp to the cob-apt4 (ORF25) intergenic spacer of the carrot mitogenome, while the Crithmum sequence (1463 bp) showed identity to just 39 bp in the same region.

The possibility that insertions of mtDNA into the plastome could involve the plastome IR was suggested by these reports and reinforced by the detection of a large (2.7 kb) insertion in the trnL-CAU-trnL-CAA spacer in the herbaceous bamboo genus Parijana (Poaceae, Olyreeae). This insertion, the first described from a monocot species, returned hits with high similarity to angiosperm mitochondrial sequences in BLASTn searches. The best scoring hit was to the mitogenome of another bamboo, Ferrocalamus rimosivaginus T.H.Wen (Arundinariae), and covered 97% of the insertion sequence (Ma et al., 2015). Insertion of mtDNA into the plastome IR of monocots has also been reported for both Triticum and Paspalum species. While the presence of putative mitochondrial sequences in two Triticum plastomes (Saarela et al., 2015) was not confirmed, the two cases in Paspalum were supported by read mapping and polymerase chain reaction (PCR) of insertion boundaries (Burke et al., 2016). A recent data-mining experiment suggested that plastomes of three Lamiales species contain mtDNA insertions, two cases at the IR–large single copy (LSC) boundaries and one in the small single copy (SSC) region; however, no confirmation of these putative insertions are reported (Gandini and Sanches-Puerta, 2017).

Thus far, a single instance of mtDNA insertion has been observed for the plastome single-copy (SC) region in a location well removed from the IR and its boundaries. Sequencing of the Catharanthus roseus (L.) G. Don (Apocynaceae) plastome, and comparison to a draft of common milkweed (Asclepias syriaca L. [Apocynaceae, Asclepiadaceae]) (Straub et al., 2011), revealed a large insertion in the rps2-rpoC2 intergenic spacer of A. syriaca that produced best BLAST hits to a region of the tobacco genome.
(Nicotiana tabacum L.) mitogenome containing exon 2 of the mitochondrial rpl2 (Ku et al., 2013). Later the same year, Straub et al. (2013) published both A. syriaca organelle genomes and confirmed the plastome insertion of a 2.4-kb segment of the mitogenome. Using complete plastome sequences and a PCR-based survey, inserts ranging in size from ~2.4 to ~4.7 kb were identified across the tribe Asclepiadeae but not in other Apocynaceae.

Plastome sequencing has far outpaced that of mitogenomes, largely because of the challenge of assembling complete mitochondrial sequences. Despite their relatively conserved gene content, the physical structure of plant mitogenomes appears quite labile, and, unlike their animal counterparts, plant mitogenomes vary widely in size and gene order. Noncoding regions far outweigh genic sequence, and conformational changes facilitated by recombination within and between molecules yield its multipartite structure (Sloan, 2013). Plastomes of photosynthetic plants, on the other hand, have been more amenable to complete assembly of the unit genome, for which a canonical structure has been established. Although often still represented as a circular map, numerous studies have demonstrated that the plastome is more likely present largely as linear, branched molecules in planta (Oldenburg and Bendich, 2015; Ruhlman et al., 2017). Nonetheless, a predominant repeating unit (150–160 kb) of two SC regions separated by a large ~25-kb IR defines the structure of the majority of photosynthetic angiosperm plastomes.

The wealth of information included in plastome sequences, in terms of both structural and nucleotide sequence variation, has been employed in phylogenetic analyses and provided markers for studies examining a range of biological questions in both basic and applied research (Ruhlman and Jansen, 2014). As more nonmodel agricultural species, as well as uncultivated wild accessions added to the canon, variation in plastome sequence and structure has turned up in unrelated lineages across the plant phylogeny, sometimes unexpectedly, when close relatives are highly conserved (Ruhlman and Jansen, 2014). For this study we sequenced and assembled complete plastomes from 10 nonmodel agricultural species from diverse lineages. For the most part, these representatives were highly similar to nearest-relative plastomes; however, we identified a few unique inversions, variation in gene and intron content, IR expansion and contraction and, most interestingly, one lineage that contains a large, unique insertion of mitochondrial DNA. Although previously confirmed in three unrelated families, IGT involving insertion of DNA in the plastome of a tree species has yet to be reported. Here we focus on the timing and phylogeographic distribution of a mitochondrial DNA insertion in the plastomes of cashew and related species and discuss possible routes of plastid entry and mechanisms of the insertion.

Materials and Methods

Selection of Taxa, Growth, Harvest, and DNA Isolation

Underrepresented clades among nonmodel agricultural taxa were selected for complete plastome sequencing; 10 species were included in all: cinnamon, fig (Ficus carica L.), guava (Psidium guajava L.), pomegranate (Punica granatum L.), cashew, mango (Mangifera indica L.), lychee (Litchi chinensis Sonn.), okra, [Abelmoschus esculentus (L.) Moench], quinoa, and basil [Ocimum basilicum L.]. Woody species were purchased as young trees from Top Tropicals (http://toptropicals.com) and were housed in the University of Texas–Austin (UT–Austin) greenhouse. Three species were germinated and grown from seed provided by the USDA–ARS National Plant Germplasm System. Supplemental Table S1 contains accessions and voucher information for all specimens included in the study. Four additional species of Anacardium and an independent accession of cashew were examined. Material for these species was obtained from the University of Texas and Lundell herbaria and specimen information is provided in Supplemental Table S1. For fresh material, newly emergent leaves were harvested, flash frozen, and ground in liquid nitrogen. Total genomic DNA isolation used the cetyl trimethylammonium bromide (CTAB) method of Doyle and Doyle (1987). Following treatment with RNase A (ThermoScientific) and phase separation with chloroform, DNA was recovered by ethanol precipitation, resuspended in DNase-free water, and stored at ~20°C for genome sequencing. Isolated DNA was transferred to the UT–Austin Genome Sequencing and Analysis Facility (GSAF), and no steps to enrich for organelle DNA were included in any protocol. For herbarium specimens, ~20 mg of leaf tissue was homogenized in CTAB isolation buffer, and DNA was extracted using Doyle and Doyle (1987) method. This DNA was used for PCR (see below).

Plastome Sequencing, Assembly, and Annotation

Plastome sequencing and assembly followed our established protocol (Weng et al., 2014). Briefly, total genomic DNA was sheared and size selected for ~800-bp fragments. Paired end library construction using the NEB-Next Ultra DNA Library Prep Kit for Illumina (New England BioLabs) and DNA sequencing were performed at the GSAF on the Illumina HiSeq 2500 platform (Illumina). Approximately 60 million reads of 125 bp were targeted from each paired-end library. The quality-filtered (FastxToolkit; hannonlab.cshl.edu/fastx_toolkit/) reads were assembled using Velvet version 1.2.08 (Zerbino and Birney, 2008) on the Texas Advanced Computing Center (TACC) supercomputer. Multiple de novo assemblies were performed with scaffolding off and variable kmer sizes (81–119), default insert size estimation, and three minimum depths of coverage (200×, 500×, and 1000×). Contigs from all assemblies were imported into Geneious version 9.1.6 (http://www.geneious.com/; Kearse et al., 2012). Plastid contigs from each assembly were identified
by comparison to a database of annotated plastomes from closely related taxa. Plastid contigs from multiple assemblies for each species were evaluated to resolve IR boundaries in addition to ambiguities or differences among contigs and Illumina reads (Supplemental Table S2) were mapped to contigs using Bowtie2 to confirm IR read depth as twice that of single copy (Langmead and Salzberg, 2012). Gene annotation of plastomes was performed in DOGMA (Wyman et al., 2004). Verification of protein coding genes was performed in Geneious using the plastid gene database (above) and transfer RNAs were verified using tRNAscan (Schattner et al., 2005).

**Comparative Genomics**

Publicly available plastome sequences were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/; Supplemental Table S3). Whole-genome alignments were performed to identify inversions using progressiveMauve version 2.3.1 (Darling et al., 2010) in Geneious. Tobacco was used as the reference plastome, and other members of the same angiosperm family were used to identify unique inversions. All plastomes were compared in Mauve without the IRa sequence of the plastome–insert junctions (Supplemental Table S4) were extracted from complete plastomes of five other species. Ten new plastomes from nonmodel agricultural species were included in the study. Among nine species we identified several structural variations including intron loss, inversion, gene duplication, and pseudogenization. In one species, cashew, we identified a large insertion in the plastome IR with high similarity to mitochondrial DNA. Table 1 provides an overview of general plastome features for all 10 taxa.

**Identification and Confirmation of Mitochondrial Insertion**

Whole-plastome alignment of the cashew plastome with progressiveMauve using frankincense (Boswellia sacra Flueck.) as the reference suggested a large insertion of nonplastome DNA in the IR between ycf2 and trnL-CAA. All Illumina reads from cashew sequencing (55,994,274 reads) were mapped to the assembly with Bowtie2 to examine read depth over the inserted region. To define the plastome–insert junctions, all available plastome sequences from the same angiosperm order (Sapindales) were downloaded from Genbank (Supplemental Table S3) and aligned with cashew and mango using multiple alignment using fast Fourier transform (MAFFT) (Katoh and Standley, 2013) in Geneious. The inserted sequence, along with flanking intergenic regions, was used to query the Transposable Elements Platform (http://botserv2.uzh.ch/kelldata/trep-db/index.html) and the Plant Genome and Systems Biology Repeat Element Database (http://pgsb.helmholtz-muenchen.de/plant/recat/index.jsp; Nussbaumer et al., 2013).

Oligonucleotide primers at the mitochondrial insert–plastome junctions (Supplemental Table S4) were designed to evaluate the presence or absence of the mitochondrial insert by PCR of genomic DNA from herbarium material of five species of Anacardium (Supplemental Table S1). Both strands of all amplification products were Sanger sequenced. The mitochondrial insert region was extracted from the cashew plastome assembly and used to query (BLASTn) the NCBI nucleotide database with default parameters. All cashew Illumina reads were used to assemble mitochondrial contigs with Velvet v1.2.08 on TACC (above). Multiple de novo assemblies were performed with scaffolding off, several kmer sizes (81–117), expected coverage values (200, 350, 500), and default insert size estimation. Four complete mitogenomes of Gossypium downloaded from NCBI (Supplemental Table S3) were used to identify mitochondrial contigs in the assembly. All putative mitochondrial contigs were imported into Geneious and annotated using a reference database comprising genes from the four Gossypium mitogenomes.

**Polymerase Chain Reaction, Sequence Alignment, and Phylogenetic Analyses**

Oligonucleotide primers (Supplemental Table S4) were designed to amplify plastid markers with demonstrated phylogenetic utility: 5’strnK_UUU-matK and the trnL-F region from Shaw et al. (2007) and 3’ ndhF from Kim and Jansen (1995). The trnL-F region was included, as it was used previously for Anacardiaceae (Weeks et al., 2014). Amplification products from five species of Anacardium (four plus an independent accession of cashew) were prepared and Sanger sequenced. These same regions were extracted from complete plastomes of five other species, four additional Anacardiaceae (Mangifera, Rhus, two species of Spondias), and one outgroup species from the Burseraceae (Boswellia) (Supplemental Table S1, S3). Sequences were aligned in Geneious using MAFFT v7.222 and concatenated. A maximum likelihood tree was generated in Geneious using PhyML (Guindon and Gascuel, 2003) with the HKY85 substitution model and support for nodes was evaluated with 1000 bootstrap replicates.

**Results**

**Plastome Organization of Ten Nonmodel Crop Plants**

Ten new plastomes from nonmodel agricultural species were included in the study. Among nine species we identified several structural variations including intron loss, inversion, gene duplication, and pseudogenization. In one species, cashew, we identified a large insertion in the plastome IR with high similarity to mitochondrial DNA. Table 1 provides an overview of general plastome features for all 10 taxa.

Two different inversions in the LSC region (Fig. 1) were detected using Mauve alignment of each species
compared with an unrearranged reference plastome (tobacco NC_001879). One inversion between ndhC and accD occurred in quinoa and was unique compared with other available members of Amaranthaceae. Another inversion in mango was between trnE-UUC and trnL-UAA and was not present in any other sequenced Anacardiaceae plastomes. Expansion and contraction of the IR–LSC and IR–SSC boundaries were identified among the included taxa (Table 1). Genome size variation among taxa was predominantly driven by IR expansion, and IR size was particularly variable in cinnamon (~20 kb) and lychee (~30 kb). In lychee, IR expansion into both the LSC and SSC duplicated three genes and extended the truncated ycf1 in IRb to >3100 bp. The opposite trend was
observed in cinnamon, where IR contraction resulted in the pseudogenization of one copy of ycf2, reducing its size to just 3168 bp (from 6972 full length).

The rpl23 coding sequence has insertions and deletions in quinoa causing a frame shift and the introduction of stop codons. The rpl23 region from five Amaranthaceae, one Caryophyllaceae, and tobacco were downloaded from Genbank and aligned with quinoa, demonstrating that mutations and truncating stop codons were shared by all examined members of Caryophyllales (Supplemental Fig. S1). A BLASTn search using the Arabidopsis rpl23 gene against the oneKP quinoa transcriptome returned several hits. The best hits were contained within two large contigs (18,545 and 17,354 bp) and were 100% identical over the rpl23 coding region. Further assessment of these two contigs revealed that they were ptDNA contamination of the transcriptome. Reducing the BLAST stringency did not identify a transcript for a plastid-targeted rpl23. Query of the quinoa translated transcriptome database with the spinach nuclear-encoded chloroplastic L23 (Bubunenko et al., 1994) returned a high amino acid (aa) identity hit (234 aa) comprising both the transit peptide (89 aa; 65.2%) and most of the coding sequence of L23 (117 aa; 88.8%). One smaller overlapping contig (30 aa) was identified that extended the Rpl23 region by 24 residues (82.5%) but did not include the terminal four residues. TargetP prediction supported plastid localization of the nuclear encoded L23 protein in quinoa with a score of 0.978.

Like rpl23, rpl22 has been reported as lost from the plastome in various angiosperm taxa (Ruhlemann and Jansen, 2014). In cinnamon, an insertion–deletion mutation in the rpl22 gene has extended the gene by 93 bp, such that its stop codon lies within the coding sequence of rpl3. The homologous region was extracted from available Lauraceae on Genbank and an alignment to cinnamon, with Arabidopsis sequences as well. The rpl22 coding sequence was not annotated in the plastomes of avocado (Persea americana Mill.), Machilus balansae (Airy Shaw) F.N. Wei & S.C. Tang, and Phoebe omeiensis R.H. Miao; however, all these species share the 93-bp extension, had 99.6% pairwise nucleotide sequence identity over the entire gene, and translate to a 212 amino acid polypeptide (Supplemental Fig. S2).

Intracellular Gene Transfer in Cashew

Alignment of the cashew plastome to other species in Anacardiaceae and other members of the Sapindales (Fig. 2) revealed a highly expanded IR as a result of the insertion of a 6767-bp fragment of nonplastome DNA between ycf2 and trnL-CAA (97,249–104,015 in IRb). No plastome coding sequences were interrupted by the insertion. The guanine–cytosine (GC) content of the insert sequence was 43.5%, while the remainder of the plastome had a GC content of 37%. Illumina reads from cashew mapped to the assembly were uniformly distributed across the region and confirmed plastome insertion in the IR (Fig. 3). Oligonucleotide primers designed to amplify the boundary regions yielded the expected size PCR products from total genomic DNA isolated from herbarium specimens of four Anacardium specimens and identified the insertion in three additional species, A. corymbosum Barb. Rodr., A. humile A. St.-Hil., and A. nanum A. St.-Hil. plus another accession of cashew. The amplified regions aligned with the mtDNA insert boundaries in the cashew assembly over 825 bp (5’ boundary) and 685 bp (3’ boundary) with 99.9% nucleotide identity. Figure 2 (lower) shows 100 bp at the 5’ (ycf2) and 3’ (trnL-CAA) ends of the mtDNA insertion comprising 50 bp each of mtDNA and ptDNA (200 bp total displayed). The PCR results for one additional Anacardium [A. excelsum (Bertero & Balb. ex Kunth) Skeels] included the survey along with mango (Anacardiaceae) indicated that they lacked the large insertion. All Sanger sequencing products are available in GenBank (MF045471-MF045482; MF422640-MF422646).

The 6767 bp insert sequence was used to query all nucleotide sequences in GenBank using BLASTn. Blast results returned multiple hits to both coding and noncoding regions of mitogenomes. Query coverage (100 hits returned) ranged from 11 to 52% with identity ranging from 89 to 96%. The top 10 hits are shown in Supplemental Table S5. No hits to any transposable elements were detected. The plastome intergenic regions within the insert and immediately flanking the insertion site were inspected for repetitive elements. There was no evidence to suggest target site duplication via transposable element activity. Nucleotide alignment of the insert region of cashew revealed that ~523 nt of sequence was missing in this region relative to other examined species of Sapindales (Fig. 2).

A subject database comprising 44 mitochondrial contigs assembled from cashew Illumina reads was queried with the plastome insert (Supplemental Table S6). Several mitochondrial genes and noncoding sequences were identified. High nucleotide identity hits included mitochondrial IG sequences as well as fragments of mitochondrial genes atp1 (69 bp; 91.3%), nad7 (59 bp; 79.9%), and rpl2 (44 bp; 100%). A complete copy of mitochondrial ccmB (612 bp; 98.7%) was also identified. Although the cashew mitogenome contig database was queried with plastome sequences flanking the insert, and no significant hits were returned.

Phylogenetic Distribution of Mitochondrial Insertion

Genomic DNA of taxa surveyed for the IGT plastome insert was also used in PCR amplification of three plastid regions for phylogenetic analysis using maximum likelihood (ML). The PCR products were sequenced and combined with sequences of the same three regions extracted from publicly available plastomes of other Anacardiaceae and an outgroup from the Burseraceae to construct a phylogenetic tree for five of the 10 recognized species of Anacardium (Mitchell and Mori, 1987). The ML tree shown in Fig. 4 was congruent with inferred relationships based on morphological and molecular data.
Fig. 2. Boundary regions of mitochondrial insert in Anacardium plastomes. Complete plastomes of Sapindales species were downloaded from NCBI and aligned to A. occidentale (upper) to identify the boundaries between the plastome and the mtDNA insertion. Mean pairwise identity over all pairs in the column is indicated by the histogram (green, 100%; tan, 30–100%; red, <30%). Thick gray bars indicate identical sequence, and black indicates differences among genera. Thin gray bar indicates gapped region in non-Anacardium, while the purple bar indicates the mtDNA insertion in Anacardium. Ellipses indicate that intervening sequences have been truncated for concision. The labeled blue arrow represents the 523 bp of plastome DNA missing from Anacardium that contain the insert. Genomic DNA from herbarium specimens of three other Anacardium species, plus an independent accession of A. occidentale, was used as template for polymerase chain reaction amplification of the boundary regions. Sanger sequencing confirmed the presence of the insert in all four species of Anacardium (lower). Brackets to the enhanced view of the 5’ (thin bracket) and 3’ (thick bracket) end of the mtDNA insert highlight 50 bp on either side of the boundary. The asterisk indicates the extraction from the A. occidentale complete plastome assembly. The blue triangle indicates the boundary site.

Fig. 3. Plastome insertion of mitochondrial DNA in cashew. All Illumina sequencing reads were mapped to the A. occidentale plastome assembly with Bowtie2. The black thick black line represents the plastome region including inverted repeat (IRb; gray), the mtDNA insert (purple), and selected coding regions (brown) from the IR and both up- and downstream single-copy regions. The green histogram indicates read depth coverage across the entire region, the range of which is indicated at the left (maximum depth of coverage = 11,997×).
The plant genome (Mitchell and Mori, 1987; Weeks et al., 2014) and shows the phylogenetic distribution of the putative mitochondrial insertion in plastid genomes across the genus. Of the five Anacardium species included in the study, the four species that were positive for the IGT event formed a strongly supported clade suggesting that the mitochondrial insertion was a single event that occurred on the branch leading to the clade that includes cashew, A. nanum, A. humile, and A. corymbosum.

**Discussion**

An expanding knowledge base facilitated by technological advancement has illuminated every aspect of biology, and comparative genomics is no exception. It was thought that the plastomes of photosynthetic land plants were highly uniform in structure until relatively recently. Over the last three decades, and particularly since the advent of next-generation sequencing technology and assembly platforms, we have amassed a seemingly ever-increasing panoply of plastome sequences, and while, for the most part, the canonical structure predominates, structural variation has been documented in a number of unrelated angiosperm lineages (Lee et al., 2007; Haberle et al., 2008; Fajardo et al., 2013; Weng et al., 2014; Sabir et al., 2014).

**Plastome Evolution in Ten Nonmodel Crop Species**

In this study we focused on nonmodel species of agricultural interest and identified a number of structural changes including inversions, divergence and loss of coding sequences, and expansion and contraction of the IR boundaries. In addition, we discovered that in one lineage, Anacardium, a large fragment of mtDNA has been inserted into the plastome IR. While these phenomena are not unheard of among land plants, apart from the gene divergences (rpl22 extension, rpl23 pseudogenization, and loss of the rpl2 intron), each is unique in its own respect.

The two inversions (Fig. 1), one each in quinoa and mango, may turn out to be informative markers for phylogenetic resolution in their respective families. Informative inversions in the plastome SC regions have been reported for several lineages (Jansen and Palmer 1987; Raubeson and Jansen 1992; Doyle et al., 1996). On the other hand, inversions facilitated by IR sequences may be limited as phylogenetic markers as they are likely impermanent, with inversion and reversal occurring freely as a function of thermodynamic properties (Kim and Lee, 2005). Given that higher hairpin stability plays a crucial role in the recurrence of inversions (Kim and Lee, 2005; Catalano et al., 2009), it may be that inversions that involve longer loop relative to shorter stem sequences are more stable over time providing a reliable signal for phylogenetic analyses.

The boundaries between the plastome IR and the LSC and SSC vary across land plants (Raubeson and Jansen, 2005) and often account for differences in total plastome size (<1 kb to >88 kb; Jansen and Ruhlman, 2012; Weng et al., 2016). Variation was observed in IR boundaries, the most notable cases were found in cinnamon (~20 kb), okra (~28 kb), and lychee (~30 kb). In cinnamon the contraction of the IRb–LSC boundary by nearly 6 kb truncated one copy of ycf2 (IRb) and the expansion into SSC increased the size of the IRb ycf1 open reading frame (ORF) to 1375 bp. Oppositely, okra experienced a minor contraction of the IR–SSC boundary that further truncated the ycf1 ORF to 959 bp, while expansion at the IRb–LSC boundary extended the IR into the intron of rpl16 duplicating three genes (rps19, rpl22, and rps3). The situation in lychee is different in that IR boundary shifts have extended the IR in both directions to include rpl3 at one end and expand the ycf1 ORF to >3 kb at the other.

The pseudogenization of rpl23 is likely substituted by the nuclear-encoded L23 protein in quinoa. The high identity of the quinoa plastid-targeting peptide (65.2% over 89 aa) to that of the spinach chloroplast L23 (Bubunenko et al., 1994) suggests that the substitution event was shared between the two species. The pseudogenization of plastid encoded rpl23 was shared.
by Amaranthaceae, Caryophyllaceae, and Polygonaceae (Supplemental Fig. S1; Logacheva et al., 2008). As the loss appears to have occurred across the Caryophyllales, it is possible that the substitution occurred on the branch leading to the order. Additional analyses using available nuclear transcriptome data (Yang et al., 2015) may resolve this question.

**Mitochondrial DNA Insertion in the Plastomes of Anacardium**

As more sequences become available for comparative analyses, more diversity has been detected in plastome structure including variants that were, until very recently, thought to be nonexistent. The notion that land plant plastomes could incorporate foreign DNA sequences without biotechnological intervention was unheard of prior to 2008 (Richardson and Palmer, 2007) and remains an unusual observation (Smith, 2011). In sequencing the complete mitogenome of grapevine, Goremykin et al. (2009) identified a segment of the carrot plastome that was absent from other plastomes, but fragments were highly similar (up to 95% identical) to plant mitogenomes available at the time and included a portion of the mitochondrial *cox1* gene. Four years passed before this finding was unambiguously confirmed with the sequencing of the carrot mitogenome (Iorizzo et al., 2012b). Subsequent studies have confirmed plastome insertions of mtDNA in two additional unrelated families of angiosperms (Straub et al., 2013; Ma et al., 2015; Burke et al., 2016) and broadened the extent of the carrot event to include 36 members of Daucinae and *C. platycarpos* (Torilidinae; Spooner et al., 2017). This study brings the total number to four, with the identification of a large mtDNA insertion in the plastome of the orchard tree cashew. Given the distribution of these four families across the phylogeny of land plants, it is likely that, in time, as more plant plastome, mitogenome, and even nuclear genome sequences are elucidated, more instances of IGT involving insertions into the plastome will be identified.

Although they share some characteristics, the instances of mitogenome to plastome IGT reported thus far lack informative common features, making it difficult to hypothesize an overarching mechanism. Placing the confirmed IGT events in a phylogenetic context supports at least five independent events across all land plants, which likely occurred only once within each clade. Of course that view must remain flexible since the groups included are widely distributed across four orders: Apiaceae (asterid II), Gentianales (asterid I), Sapindales (rosid II), and Poales (commelinid). Although reports suggest the possibility of either a wider distribution of the documented IGT event within Apiaceae or other independent events in this lineage conclusive evidence is thus far lacking (Peery 2015; Downie and Jansen, 2015).

Among the confirmed instances of IGT, four plastomes contain insertions of mitochondrial DNA in the IR. However, the Asclepiadaceae were positive for a mitochondrial insertion in the plastome *rps2-rpoC2* spacer in the LSC. It is tempting to speculate that IGT events that inserted extraplastomic DNA in one copy of the IR would be more stable than those in the SC regions, as their presence would more likely be stabilized by recombination and gene conversion between the IR sequences both within and between plastid unit–genome copies. Furthermore, the higher GC content of the plastome IR may present a more favorable environment for transferred mitochondrial sequences given their relatively high GC content (40–48% across angiosperm mitogenomes; https://www.ncbi.nlm.nih.gov/genome/browse/). Among the many copies of the unit genome that constitute the plastome within a single plastid, sequences in the SC regions may also undergo gene conversion, albeit half as often. As demonstrated by transplastomic experiments, foreign DNA may be stably integrated into both the IR and SC regions (Verma et al., 2008). Importantly, it is the action of selection that drives the insertion event toward homoplasy in these experiments. Could there be some advantage, or at least a lack of cost, in maintaining foreign DNA insertions in wild-type plants? A common feature among the four IGT events is insertion into noncoding spacer regions of the plastome. The site of integration may be random or facilitated by homologous recombination; however, any disruption of plastid coding sequences that reduces overall fitness would likely be purged.

In *Anacardium*, and among the previously confirmed examples of mitochondrial to plastid IGT, inserted sequences include intergenic regions of the mitogenome, fragments of mitochondrial genes, pseudogenes of plastid and nuclear origin, and sequences of unknown providence. No unifying theme is apparent here either. The insertion in the lineage that includes carrot lacks homology to plastome sequences. The initial IGT in this lineage likely occurred once at the base of Torilidinae and was considerably larger than what is present in the later diverging *Daucus* plastomes (Lee et al., 2001; Spooner et al., 2017). The larger insertion, documented in *C. platycarpos*, contains an intact *cox1* copy and fragments of sequence with identity to a nuclear auxin response factor (ARF) gene. However, direct insertion of ncDNA into the plastome at the very same locus as mtDNA insertion is implausible compared with its insertion along with the mtDNA as mitogenomes of land plants contain abundant foreign DNA from both IGT and HGT events (Knoop, 2004; Alverson et al., 2010; Park et al., 2014). In particular, an ARF gene (*ARF17*) has been transferred to the mitogenome in several genera of Brassicaceae (Qiu et al., 2014).

Among the mtDNA sequence remaining in the plastome of carrot, two segments (108 and 403 bp) were suggested to contain a putative GAG domain, yet evidence for this was not provided (Iorizzo et al., 2012a), while another (144 bp) shared limited similarity to a nuclear retrotransposable element from unrelated species (Iorizzo et al., 2012b). A mitochondrial *rpl2* pseudogene and noncoding sequences comprise the *Asclepias* insert.
As in carrot, these sequences are no longer contiguous in the mitogenome; however, in Asclepias, the split sequences are adjacent to transferred plastome sequences that appear to be undergoing gene conversion. The authors propose that more recent plastome to mitogenome IGT is playing a role in the conservation of these sequences (Straub et al., 2013). The plastome insertion in bamboo has 98% nucleotide sequence identity to a region of the F. rimosivaginatus mitogenome over ~97% of the insert, but no mitochondrial genes were identified, and no transferred plastome sequences were detected in the adjacent mitogenome regions (Ma et al., 2015).

We assembled contigs from the Anacardium mitogenome and identified sequences homologous to the plastome insertion including a complete copy of mitochondrial ccmB (612 bp; 98.7% nucleotide identity). However, we were unable to detect regions with homology to genic or intergenic sequences flanking the plastome insert, and BLASTn queries of the mitogenome contigs suggested that the ~6.7-kb insert was no longer contiguous in the mitogenome, obfuscating mechanisms involved in the IGT event. The nature and context of transferred sequences in both the plastome and mitogenome could suggest possible mechanisms of incorporation. The similarity to retrotransposon elements in the carrot insert, along with the detection of flanking 6-bp repeat units (CTTGAC) led the authors to suggest non-LTR retrotransposon activity may have been responsible for the plastome insertion. It would be unusual, however, for an instance of transposon insertion via target site recognition to delete sequence from the target region between the repeated sequences. The carrot and cumin (Cuminum cyminum L.) plastomes are missing 337 bp of sequence in the insertion region compared with relatives that lack the insert. It is also worth noting that the 6-bp sequence occurs in direct orientation in the carrot plastome 35 times on the forward strand and 30 times on the reverse strand, including within and just 30 bp upstream of the insert in the same orientation. Apart from the Wendy element in Chlamydomonas plastomes (Fan et al., 1995), signatures or sequences of transposable elements are not known to occur in plastomes.

Missing sequence at the insertion site was also reported for bamboo plastomes (1379 bp). The authors found a large deletion (1542 bp) in a relative of the IGT clade, Oatea glauca L.G.Clark & G.Cortés, and suggested that the deletion could have occurred prior to the insertion of mitochondrial DNA. However, O. glauca is not the nearest relative to the Pariana lineage that contained the insert. According to the phylogeny published in their study, two more closely related bamboos (Olyra latifolia L. and two species of Chusquea) each contained the sequence deleted in Pariana. In Anacardium, species that contain the large mitochondrial insert are missing 523 bp of DNA present in the plastome of those that lack the insert. While there was no indication of sequence deletion in the rps2–rpoC2 spacer in Asclepiadaceae (Straub et al., 2013), the detection of plastome sequences up- and downstream of the transferred sequence in the mitogenome strongly suggest a role for homologous recombination in facilitating plastome integration.

The insertion of mtDNA uncovered in Anacardium is unique in that previous examples of mitochondrial to plastome IGT have all been reported for herbaceous annual, biennial, or perennial plants, but cashew is a long-lived woody tree. Several studies have suggested that generation time plays a role in molecular evolution in plants (Lanfear et al., 2013; Zhong et al., 2014; Bromham et al., 2015). Positive correlations have been drawn between rates of nucleotide substitution and the herbaceous habit because these species tend to be small and have faster generation times. However, the hypothesis that accumulated changes in DNA sequence can be attributed to generation time in plants must consider that, unlike animals, plants do not sequester germ line cells allowing inheritance of somatic mutations (Whittle and Johnston, 2003). Longer-lived taxa could be considered to have more time to incorporate mutations, such as a plastome insertion of foreign DNA in the cell lineages that give rise to gametes in the gametophyte.

A recently uncovered variant of the carrot plastome insertion (DcMP) in the plastome of related C. platycarpos suggests that an intact copy of coxB was included in the ancestral insertion event; this sequence has been truncated in all members of the Daucinae (Spooner et al., 2017). The large mtDNA insertion in the cashew plastome contains an intact coding sequence for a mitochondrial protein. The ccmB gene encodes an inner membrane protein that assembles with nuclear and mitochondrial encoded subunits to form an ABC transporter in plant mitochondria that appears to be involved in heme delivery to the matrix for cytochrome c maturation (Giegé et al., 2008). Like those of the mitochondria, the ABC transporters involved in movement of lipid constituents across the plastid envelopes originated with their bacterial predecessor (Kang et al., 2011). Is it possible that Anacardium plastids have functionalized the ccmB gene leading to its persistence? This tantalizing prospect would require protein-level analyses, as mRNA transcription and functional protein expression in plant plastids is essentially uncoupled (Deng and Gruissem, 1987; Gruissem et al., 1988; Quesada-Vargas et al., 2005; Ruhlman et al., 2010). Except for a few unusual groups, the tendency of plant plastomes is to maintain a highly compact gene-rich structure. Apart from its location in the IR there may be other forces stabilizing the large insertion in Anacardium.

In any case, integration of mtDNA in the plastome could only take place after the foreign DNA had entered the plastid. Once entry is gained, the plastid suite of nuclear encoded proteins involved in homologous recombination may facilitate insertion. Numerous publications report the generation of transplastomic plants via homologous recombination between native plastome sequences and their homologs engineered to flank the foreign DNA. Although most transplastomic experiments...
have employed long flanking sequences (Ruhlman and Jansen, 2014), in situations where the normal recombination surveillance system is disrupted, repeats as short as 5 to 15 bp served as recombination substrates (Marechal et al., 2009). While plastids have been shown to import a eukaryotic mRNA from the cytosol (Nicolai et al., 2007), it is the lack of an active DNA uptake system (Smith, 2011) and the plastid double membrane that have likely precluded widespread HGT and IGT in plant plastomes.

The most common approach for experimental delivery of foreign DNA has been particle bombardment (Bock, 2015). Tungsten and, more commonly, gold particles have been used to carry DNA plasmids into plant cells where contact with the plastid membrane can facilitate DNA entry (Svab et al., 1990; Daniell, 1993; Langbecker et al., 2004; Okuzaki et al., 2013). Another approach involving incubation of protoplasts (plant cells subjected to enzymatic degradation of the cell wall) with polyethylene glycol (PEG) and vector DNA has also been reported to produce plastid transformants (Golds et al., 1993). In this case, it is less obvious how, once taken up by the denuded cells, transforming DNA enters the plastid compartment; the exact mechanism is yet to be defined (Kofer et al., 1998). Incubation with PEG may induce changes in the plasma membrane allowing the transforming DNA to enter the cytoplasm. Meanwhile, PEG draws water from the protoplast and likely by extension from cellular compartments including the plastid. Following PEG incubation osmotic adjustment is made setting up a hypotonic environment in the cell, which could permit transforming DNA to enter the plastid as reported by Cerutti and Jagendorf (1995). Heat shock and incubation of isolated plastids were also shown to permit the movement of foreign DNA inside plastids (Cerutti and Jagendorf, 1995). These experiments found that a greater proportion of the DNA taken up by plastids was in open-circular and linear forms along with smaller fragments. Eukaryotic mRNA (Nicolai et al., 2007) and both coding and noncoding viral RNAs are able to enter plastids (Schoelz and Zaitlin, 1989; Gómez and Pallás, 2010). Free mtDNA molecules in the cell, generated, for example, by selective mitophagy (Veljanovski and Batoko, 2014), could gain entry by similar or other mechanisms, accounting for the presence of mtDNA in plant plastids and, in turn, their plastomes.

Early studies suggesting fusion between the organelles (see Cerutti and Jagendorf, 1995), despite the very different compositions of their membranes, may be gaining contemporary traction. The presence of long plastid protrusions has been repeatedly demonstrated (Kwok and Hanson, 2004; Natesan et al., 2005; Hanson and Sattarzadeh, 2008, 2013; Gray et al., 2012; Caplan et al., 2015). Several reports support their role in providing a physical connection between plastids as well as a channel through which proteins, and perhaps other molecules, may move. The upper size limit for passage of proteins though stromules has not been established (Gray et al., 2012). It is plausible that if florescent polypeptides are able to traverse these channels, then DNA molecules should also be able to make the trip. Recent work is now uncovering the interconnectedness of plant cell organelles and suggests that membrane contact sites (MCS) establish foci for the transfer of molecules between cellular compartments and even between cells via plasmodesmata (Brunkard et al., 2015). Observed stress-inducible MCS include direct interaction between plastids and peroxisomes, the nucleus and mitochondria via stromules, peroxules and mitochondrial protrusions, and mitochondrial–plastid MCS (reviewed in Pérez-Sancho et al., 2016). Contemporary data supporting the observations of early investigators reminds us to broaden our perspective with regard to what may be possible.

According to our phylogenetic analysis, the Anacardium IGT event occurred once in the cashew lineage (Fig. 4). Fossil evidence suggests that Anacardium is at least 47 million yr old (Manchester et al., 2007) and molecular divergence time estimates indicate that the clade that includes species with the insertion diverged <20 million yr ago (Weeks et al., 2014), suggesting that the insertion is relatively recent in the evolution of the genus. There are likely other instances of mitogenome to plastome IGT that remain undiscovered in land plants. Work to completely sequence the organelle genomes of plants has focused more on crop species, and this focus has been nearly complete thus far in terms of plastome sequences. Until more species are available that have genome sequences for both organelles, it will be difficult to propose unifying explanation or delineate alternative mechanisms of IGT in land plants.

Conflicts of Interest Disclosure

The authors declare that there is no conflict of interest.

Supplemental Information

Supplemental information is available for this article.

Acknowledgments

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant no. 35-130-36. The authors, therefore, acknowledge with thanks DSR technical and financial support. The authors also gratefully acknowledge President Prof. Dr. Abdulrahman O. Alyoubi. We also thank the Genome Sequencing and Analysis Facility at the University of Texas at Austin for performing the Illumina sequencing, the Texas Advanced Computing Center at the University of Texas at Austin for access to supercomputers, and TEX-LL for allowing us to remove leaf material from herbarium specimens of Anacardium and serving as a repository for voucher specimens.

References


