Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae)

Andrew J. Alverson,*1 XiaoXin Wei,†1 Danny W. Rice,1 David B. Stern,2 Kerrie Barry,3 and Jeffrey D. Palmer1
1Department of Biology, Indiana University
2Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY
3DOE Joint Genome Institute, Walnut Creek, CA
†Present address: State Key Laboratory of Systematics and Evolutionary Botany, Institute of Botany, The Chinese Academy of Sciences, Xiangshan, Beijing 100093, China.

*Corresponding author: E-mail: andy.alverson@gmail.com.
Associate editor: Charles Delwiche

Abstract

The mitochondrial genomes of seed plants are unusually large and vary in size by at least an order of magnitude. Much of this variation occurs within a single family, the Cucurbitaceae, whose genomes range from an estimated 390 to 2,900 kb in size. We sequenced the mitochondrial genomes of Citrullus lanatus (watermelon: 379,236 nt) and Cucurbita pepo (zucchini: 982,833 nt)—the two smallest characterized cucurbit mitochondrial genomes—and determined their RNA editing content. The relatively compact Citrullus mitochondrial genome actually contains more and longer genes and introns, longer segmental duplications, and more discernibly nuclear-derived DNA. The large size of the Cucurbita mitochondrial genome reflects the accumulation of unprecedented amounts of both chloroplast sequences (>113 kb) and short repeated sequences (>370 kb). A low mutation rate has been hypothesized to underlie increases in both genome size and RNA editing frequency in plant mitochondria. However, despite its much larger genome, Cucurbita has a significantly higher synonymous substitution rate (and presumably mutation rate) than Citrullus but comparable levels of RNA editing. The evolution of mutation rate, genome size, and RNA editing are apparently decoupled in Cucurbitaceae, reflecting either simple stochastic variation or governance by different factors.

Key words: Cucurbitaceae, genome size, mitochondria, mutation rate, RNA editing.

Introduction

Seed plant mitochondrial genomes are exceptional for their generally very low mutation rate (Wolfe et al. 1987; Palmer and Herbon 1989), relatively high incidence of RNA editing and trans-splicing of coding sequences (Hiesel et al. 1989; Knoop 2004), frequent uptake of foreign DNA by intracellular and horizontal gene transfer (Stern and Lonsdale 1982; Richardson and Palmer 2007), dynamic structure (Lonsdale et al. 1988; Palmer and Herbon 1989), and, historically first and perhaps foremost, their extraordinarily large and highly variable sizes (Quetier and Vedel 1977; Ward et al. 1981; Hsu and Mullin 1989). Seed plants house by far the largest known mitochondrial genomes, with sizes ranging from 222 to 773 kb among the 16 genomes sequenced to date (http://www.ncbi.nlm.nih.gov/Genomes/). The full range of sizes well exceeds an order of magnitude, however, and most of this variation occurs within a single family, the Cucurbitaceae. In a landmark study, Ward et al. (1981) showed, based on analysis of reassociation kinetics, that cucurbit mitochondrial genomes vary in size from an estimated 390 kb in Citrullus lanatus (watermelon) to an astounding 2.9 Mb in Cucumis melo (muskmelon) (fig. 1). For perspective, the muskmelon mitochondrial genome is bigger than the genomes of many free-living bacteria (Moran 2002).

The limited number of follow-up studies to Ward et al. (1981) has largely come up short in identifying the sources of the “extra” DNA in the largest cucurbit genomes. For example, size differences do not appear to reflect large-scale genome duplications (Havey et al. 1998) or major differences in gene content (Stern and Newton 1985; Adams et al. 2002). Although chloroplast-derived sequences are common in plant mitochondrial genomes, the larger cucurbit genomes do not appear to contain disproportionate amounts of captured chloroplast DNA either (Stern et al. 1983; Havey et al. 1998). Small repetitive sequences can also comprise a substantial portion of plant mitochondrial genomes (André et al. 1992), including in cucurbits (Ward et al. 1981; Lilly and Havey 2001). Indeed, a handful of small repetitive motifs account for as much as 13% of the large (~1.8 Mb) Cucumis sativus (cucumber) mitochondrial genome (Lilly and Havey 2001), but reassociation kinetics did not show a positive correlation between genome size and proportion of repetitive DNA across Cucurbitaceae (Ward et al. 1981). Thus, the sources underlying major expansions in mitochondrial genome size in cucurbits, and

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

across all seed plants for that matter, are largely unknown. In fact, most plant mitochondrial DNA (as much as 90%) is noncoding (Kubo and Mikami 2007), and the origins of most of this sequence are unknown. Clearly more data—preferably whole-genome sequences—are necessary to understand the ebb and flow of these remarkable genomes.

In this study, we present complete mitochondrial genome sequences and RNA editing data for *Ci. lanatus* (watermelon) and *Cucurbita pepo* (zucchini). With size estimates of 390 kb and 1.0 Mb, *Citrullus* and *Cucurbita* have the smallest characterized cucurbit mitochondrial genomes (Ward et al. 1981). Even so, the nearly 1-Mb (as determined in this study) mitochondrial genome of *Cucurbita* is the largest organelle genome sequenced to date, exceeded only by the slightly larger 1.02-Mb genome of the freshwater amoeboid, *Paulinella* (Nowack et al. 2008). The early diverging phylogenetic positions of *Cucurbita* and *Citrullus* within the Cucurbitaceae are such that these two species offer an important glimpse into the ancestral features of cucurbit mitochondrial genomes, that is, before the events that led to even greater size expansions in *Cucumis* (fig. 1). Analysis of these two genomes reveals a number of surprising features and important insights into the evolution of genome size in plant mitochondria.

**Materials and Methods**

**Mitochondrial DNA Isolation, Genome Sequencing, and Assembly**

Mitochondria were isolated from etiolated seedlings of *Ci. lanatus* (cultivar [cv.] Florida Giant) and *Cu. pepo* (cv. Dark Green Zucchini) using the DNase I procedure (Kolodner and Tewari 1972), and mitochondrial DNA was purified from lysed mitochondria by CsCl centrifugation (Palmer 1982). We chose these cultivars rather than those used by Ward et al. (1981) because of the availability of almost-ready-purified mitochondrial DNA. Relative proportions of mitochondrial DNA and contaminant chloroplast and nuclear DNA were assessed by Southern hybridization of conserved mitochondrial (cob), chloroplast (rbcL), and nuclear (SSU ribosomal DNA) probes using a chemiluminescent detection protocol (Nakazato and Gastony 2006).

A single 3-kb library was made for each of *Citrullus* and *Cucurbita*. Library construction, cloning, and Sanger sequencing were carried out by the US DOE Joint Genome Institute in Walnut Creek, CA. Detailed protocols are available at http://www.jgi.doe.gov/sequencing/protocols/prots_production.html.

Sequence reads initially were assembled with CAP3 (Huang and Madan 1999). Consed (Gordon et al. 1998) was then used to calculate library statistics on the initial assembly, and CAP3 was run again with forward–reverse constraints. These steps were repeated with refined forward–reverse constraints until the CAP3 assembly no longer improved. For both genomes, CAP3 assembled the vast majority of reads into a single contig. Consed was then used to visualize and validate the final assemblies and to design polymerase chain reaction (PCR) primers for filling gaps and augmenting regions of low sequence coverage. Annotated genome sequences are available from GenBank (accession numbers GQ856147 and GQ856148).

**Gene Annotation**

We made amino acid databases for protein-coding genes and nucleotide databases for ribosomal RNA (rRNA) and transfer RNA (tRNA) genes, compiled from all previously sequenced seed plant mitochondrial genomes. NCBI-BlastX and -BlastN searches of the genomes against these databases were performed to find protein and structural RNA genes, respectively. We also used tRNAscan-SE (Lowe and Eddy 1997) to corroborate tRNA boundaries identified by BlastN. Blast and tRNAscan output were converted into an HTML display format similar to that used for the annotation of chloroplast and animal mitochondrial genomes (Wyman et al. 2004), which allowed clear visualization of gene and intron boundaries. Relevant annotation data were entered into a web-based form and written to a Sequin-formatted table file with a set of Perl and CGI scripts.

**Analysis of Intergenic Sequences**

Each genome was searched against a database of all previously sequenced seed plant mitochondrial genomes with NCBI-BlastN. All BlastN searches used the following settings unless stated otherwise: *r* = 5, *q* = −4, *G* = 8, *E* = 6, and *W* = 7. Visual inspection of BlastN hits of each genome to a database of all fully sequenced seed plant mitochondrial genomes identified regions encompassing and extending beyond coding regions that were conserved across eudicots, angiosperms, or all seed plants. The strong syntenic and sequence-level conservation suggests that these regions might contain trans-spliced introns, promoters, untranslated regions, or otherwise important sequences. Boundaries of these putatively functional “conserved syntenic regions” were manually determined and...
annotated for each gene (supplementary figs. 1 and 2, Supplementary Material online).

Chloroplast-like sequences were identified with NCBI-BlastN searches of mitochondrial genomes against a database of representative angiosperm chloroplast genomes. The remaining intergenic regions were extracted and searched against the following databases maintained by the National Center for Biotechnology Information (NCBI): the nonredundant nucleotide and protein databases, the whole-genome shotgun database, and the est_others database. BlastN hits to plant mitochondrial genomes were precluded with settings “all[filter] NOT (viridiplantae[ORGN] AND mitochondrion[filter]).” Nuclear-derived insertions are common in plant mitochondrial genomes, and transposable elements are often the hallmark of these sequences (Satoh et al. 2006). With this in mind, each genome was also searched against the Repbase repetitive element database (version 13.05; Jurka 2000).

Analysis of Repeated Sequences

Short conserved repeats with mismatches and indels were found by searching each genome against itself using Washington University (WU)-Blast with the following settings: \( M = 1, N = 3, Q = 3, \) and \( R = 3, \) kap, span, \( B = 1 \times 10^3, \) and \( W = 7. \) With these settings, WU-Blast detected perfect repeats of minimum length 19 nt and imperfect repeats of minimum length 23 nt. The minimum percent identity for imperfect repeats was 78.6%, for repeats of 154 nt and larger. All Blast hits with an expect value \( \leq 1 \) were considered repeats.

RNA Editing

Total RNA was isolated from fresh young leaves. RNA preparation, cDNA synthesis, and experimental safeguards to identify potential contaminating genomic DNA followed Mower and Palmer (2006). Genomic sequences were used to design PCR primers that immediately flanked genes (supplementary table 1, Supplementary Material online), based on the assumption that these regions comprised parts of the 5’ and 3’ untranslated regions. In some cases, these primers failed to amplify any product, so primers within the coding sequence were used, resulting in cDNA sequences that were incomplete at one or both ends or, in some cases, the middle of the gene (supplementary table 3, Supplementary Material online). Thus, our data represent minimal estimates for both the number of total edits per gene and the number of shared edits between the two species. Gene sequences were screened with PREP-Mt to ensure that internal primers were not located in regions with predicted editing sites (Mower and Palmer 2006; Mower 2009). PCR conditions were as follows: 36 cycles of (45 s at 94 °C, 45 s at 48–55 °C, and 2–2.5 min at 72 °C), with an initial step of 3 min at 94 °C and a final step of 10 min at 72 °C. Amplicons were purified with ExoSAP-IT (USB Corporation, Cleveland, OH) and directly sequenced with an ABI 3730 (Applied Biosystems, Foster City, CA). RNA editing sites were determined by comparing cDNA and genomic sequences. Sites with both T and C chromatogram peaks clearly above background in a majority of sequenced strands were considered partially edited. If any ambiguity existed about whether the site was fully or partially edited, it was scored as fully edited.

Analysis of Nucleotide Substitution Rates

We extracted nucleotide sequences for all protein genes (excluding introns) from all 18 completely sequenced seed plant mitochondrial genomes. Sequences were aligned with MUSCLE (version 3.6: Edgar 2004), and ambiguously aligned regions were manually adjusted with MacClade (version 4.07). We excluded the following regions from subsequent analyses: 1) codons with known RNA editing in Arabidopsis, Brassica, Beta, Oryza, Citrullus, or Cucurbita, 2) regions for which positional homology could not be determined with confidence, and 3) regions for which data were missing from a majority of taxa. We also excluded several genes that are missing from many of the sequenced genomes (sdh3, sdh4, rpl2, rpl5, rpl6, rps1, rps2, rps8, rps10, rps11, rps14, and rps19), resulting in a final alignment of 30 concatenated genes and 25,553 positions.

We constrained the tree topology to reflect known phylogenetic relationships (Kellogg and Birchler 1993; Soltis et al. 2000; Barker et al. 2001; Mathews et al. 2002) and estimated rates of synonymous (\( d_s \)) and nonsynonymous (\( d_a \)) substitutions across the tree using the MG94W9 codon model, allowing for independent estimates of \( d_s \) and \( d_a \) on each branch. We also performed relative rate tests using the same model, separately constraining both \( d_s \) and \( d_a \) in Citrullus and Cucurbita, with Carica as the outgroup. All analyses were performed with HyPhy (version 0.9920070130 for Macintosh).

Results and Discussion

Genome Size and Characteristics

The Citrullus and Cucurbita mitochondrial genomes assembled into single circular-mapping (Bendich 1996) molecules of lengths 379,236 nt and 982,833 nt, respectively (table 1). The sizes of these two genomes are remarkably close to the size estimates based on reassociation kinetics (~390 kb and ~1 Mb, respectively) (Ward et al. 1981). Although the estimates of Ward et al. (1981) are conventionally cited as 330 kb for Citrullus and 800 kb for Cucurbita (e.g., Lilly and Havey 2001), these values are problematic owing to 1) rounding issues, 2) uncertainty in converting the megadalton estimates of Ward et al. (1981) to kilobases, and 3) the use of a reassociation-kinetics size standard (Bacillus subtilis) of uncertain size. A proper accounting of these gives size estimates of 391, 1014, 1837, and 2936 kb for the mitochondrial genomes of Citrullus, Cucurbita, Cu. sativus, and Cu. melo, respectively (fig. 1). Still, some uncertainty remains in comparing these estimates with the more precise sizes determined in this study. The estimates of Ward et al. (1981) used different cultivars than did the present study, and the genome size of the B. subtilis strain (#746) used by Ward et al. (1981) has not yet been determined. The limited size variation (4,187–4,293 kb) among
the five B. subtilis strains so far sequenced suggests, however, that this is unlikely to contribute significant error to the above estimates.

Although difficult to reconstruct with so few genomes, the size disparity between the two species appears to reflect a dynamic history of expansion and possibly contraction (fig. 1). Like Citrullus, the common ancestor of cucurbits might have had a relatively compact genome, with a series of independent expansions leading to the large genomes in Cucurbita and Cucumis (fig. 1). Alternatively, the common ancestor of cucurbits might have possessed an unusually large mitochondrial genome, with a contraction resulting in the relatively small mitochondrial genome of Citrullus (fig. 1). Clearly more data are necessary to distinguish among the possible scenarios. No correlation is seen between the mitochondrial and nuclear genome sizes of these four species (Ren et al. 2009).

Using a Blast expect cutoff of 1 $\times$ 10$^{-6}$, Citrullus and Cucurbita share $\sim$240 kb of genomic sequence, both coding and noncoding, which translates to roughly 63% and 25% genomic coverage, respectively. This is somewhat more than the amount of shared sequence between two other confamilial species, Arabidopsis and Brassica, which were found to share 143 kb of genomic sequence (Handa 2003). The relatively small size of the Brassica genome (222 kb) does, however, predict a lower amount of shared sequence between it and Arabidopsis. Coverage by mitochondrial-like sequence increases to 74% for Citrullus and 38% for Cucurbita when all seed plant mitochondrial genomes are considered. So although most previously sequenced plant mitochondrial DNA is species specific (Kubo and Newton 2008), a large fraction of the modestly sized Citrullus mitochondrial genome is not unique to this genome.

Genomic coverage by genes and introns totals $\sim$70 kb for each of the two species (table 1). Conserved syntenic regions—genes, introns, and conserved flanking sequences (see Materials and Methods)—likely include most of the functional sequence in the genome. These regions total $\sim$102 and $\sim$94 kb in Citrullus and Cucurbita, respectively (table 1; supplementary figs. 1 and 2, Supplementary Material online). Therefore, as in other seed plants, most of the sequence in these genomes is noncoding and probably nonfunctional.

### Gene Complement and Synteny

Both genomes share the same core set of 37 intact protein genes and 3 rRNA genes. Gene content in Citrullus and Cucurbita is consistent with the results of a Southern hybridization survey of mitochondrial gene content across 280 diverse angiosperms, which included the closely related cucurbit species, Cu. sativus (Adams et al. 2002). The one exception is that Cu. sativus has apparently very recently lost the rps19 gene (Adams et al. 2002), which is present twice in both Citrullus and Cucurbita. One of the rps19 genes is part of the rpl2–rps19–rps3–rpl16 arrangement conserved as far back as liverworts (Takemura et al. 1992), whereas the second copy is part of an rps19–rps10–cox1 cluster unique to these two cucurbits (supplementary figs. 1 and 2, Supplementary Material online). Like other eudicots (Adams et al. 2002), Citrullus and Cucurbita lack the rps2 and rps11 genes, which among sequenced seed plants are found only in Cycas and grasses (rps2) or just Cycas (rps11). Finally, with an additional sdh3 gene and slightly more coding sequence across shared genes (table 1), the smaller Citrullus genome has more overall coding sequence than does Cucurbita.

A total of 14 syntenic gene clusters (defined as two or more colinear and identically oriented genes) are shared between the two genomes. Many of the syntenic clusters represent maintenance of well-characterized, highly conserved arrangements and cotranscription units (e.g., Takemura et al. 1992; Perrotta et al. 1996; Quiñones et al. 1996; Hoffmann et al. 1999; Placido et al. 2006), whereas others are present across a more restricted phylogenetic range (supplementary figs. 1 and 2, Supplementary Material online). With the exception of clusters 12 and 13, the arrangement of the clusters is essentially scrambled between the two genomes (fig. 2). This high level of rearrangement is entirely expected (Palmer and Herbon 1989; Satoh et al. 2006; Allen et al. 2007).

Most genes show the expected, highly conserved level of sequence and structural conservation. For example, two previously characterized sets of overlapping genes, rps3–rpl16 and cox3–sdh4, were found in both genomes (Takemura et al. 1992; Giegé et al. 1998). As in other angiosperms, the rpl16 genes of both species likely use a GTG start codon (Bock et al. 1994; Sakamoto et al. 1997), and the “t-element” (likely modified from a chloroplast-derived trnI gene) immediately downstream from the ccmC gene probably facilitates formation of the 3’ terminus of the transcript in both cucurbits, just as it does in Arabidopsis (Forner et al. 2007). Finally, the rps14 gene in both genomes is likely nonfunctional due to numerous indels that disrupt the reading frame. Blast searches to expressed sequence tag databases of Citrullus, Cu. sativus, and Cu. melo found an intact and full-length mitochondrial rps14 homolog in Cu. melo that showed the expected, high level of sequence

### Table 1. Genome Coverage (number of nonredundant nucleotide positions in the genome) by Coding and Noncoding Features in the Citrullus and Cucurbita Mitochondrial Genomes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Feature</th>
<th>Citrullus (%)</th>
<th>Cucurbita (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>Total Size</td>
<td>379,236</td>
<td>982,833</td>
</tr>
<tr>
<td></td>
<td>Protein exons</td>
<td>32,370 (8.5)</td>
<td>32,032 (3.3)</td>
</tr>
<tr>
<td></td>
<td>Cis-spliced introns</td>
<td>32,476 (8.6)</td>
<td>30,557 (3.1)</td>
</tr>
<tr>
<td></td>
<td>tRNA</td>
<td>5,148 (1.4)</td>
<td>5,109 (0.5)</td>
</tr>
<tr>
<td></td>
<td>tRNA</td>
<td>1,358 (0.4)</td>
<td>966 (0.1)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>379,236</td>
<td>982,833</td>
</tr>
<tr>
<td>Noncoding</td>
<td>Mitochondrial-like</td>
<td>159,032 (41.9)</td>
<td>180,008 (18.3)</td>
</tr>
<tr>
<td></td>
<td>Chloroplast-like</td>
<td>22,779 (6.0)</td>
<td>113,347 (11.5)</td>
</tr>
<tr>
<td></td>
<td>Nuclear-like</td>
<td>20,914 (5.5)</td>
<td>17,820 (1.8)</td>
</tr>
<tr>
<td></td>
<td>Transposable elements</td>
<td>3,438 (0.9)</td>
<td>2,818 (0.3)</td>
</tr>
</tbody>
</table>

*a*  Includes all genes, cis- and trans-spliced introns, and the highly conserved (putatively functional) sequences immediately flanking them.

*b*  Intergenic sequences with similarity to DNA in at least one other seed plant mitochondrial genome, excluding chloroplast-like sequences.
Introns

The smaller *Citrullus* genome actually contains more and longer cis-spliced introns (table 1). The two species share 19 cis- and 5 trans-spliced group II introns, fully 15 of which are longer in *Citrullus*. *Citrullus* also contains the well-characterized *cox1* group I intron, which has spread widely across angiosperms by horizontal transfer (Sanchez-Puerta et al. 2008). The *cox1* intron is also known from three *Cucumis* species, indicating its gain in the *Citrullus–Cucumis* lineage some 20–30 Ma, following the split from *Cucurbita* (fig. 1) (Sanchez-Puerta et al. 2008). Altogether, the *Citrullus* genome contains nearly 2 kb of additional intronic sequence compared with the larger *Cucurbita* genome (table 1). The ~1.8-Mb genome of their relative, *Cu. sativus*, contains the largest known plant mitochondrial introns (sometimes two to three times larger than homologous introns in other land plants) for three surveyed genes.
Transfer RNAs
Although both genomes use all 64 codons and show highly similar patterns of codon usage (not shown), tRNA complement differs between the two species. We classified tRNAs based on origin (chloroplast or mitochondrial) and whether they were embedded within larger tracts of captured chloroplast DNA (supplementary table 2, Supplementary Material online). The Citrullus and Cucurbita mitochondrial genomes encode 18 and 13 intact and putatively functional tRNAs, respectively, that lie outside larger segments of chloroplast-derived segments (supplementary table 2, Supplementary Material online). In both genomes, three of these tRNAs (trnH-GTG, trnM-CAT, and trnN-GTT) are nevertheless chloroplast in origin. Both genomes lack several tRNAs that are present in either bryophytes (trnA and trnT) or bryophytes and Cycas (trnR and trnL) (Li et al. 2009) but are commonly missing from angiosperm mitochondrial genomes. Codons for these amino acids are abundant in both genomes, so missing tRNAs are likely encoded in the nucleus (Dietrich et al. 1996). Citrullus has duplicate copies of three tRNAs (trnC-GCA, trnG-GCC, and trnQ-TTG), and Cucurbita has lost both native trnS variants that are otherwise universally present across sequenced seed plant mitochondrial genomes. The mitochondrial genome of the hornwort, Megaceros aequatricus, is the only other land plant known to lack native trnS genes (Li et al. 2009). So altogether, the smaller Citrullus genome contains five more tRNAs than Cucurbita.

The Citrullus and Cucurbita mitochondrial genomes contain substantial amounts of chloroplast-derived sequences (see next section), many of which contain the expected tRNAs. Citrullus and Cucurbita have 8 and 24 such tRNAs in their mitochondrial genomes, respectively (supplementary table 2, Supplementary Material online). Of these, 7 and 15 are intact and potentially functional, respectively, with the rest appearing to have degenerated to the point of being nonfunctional. In some cases, the same syntenic tract of chloroplast sequence contains both intact and degenerate tRNAs. The apparently differential constraints on the embedded chloroplast tRNAs provide circumstantial evidence that some of them might be functional. For example, of the 5 chloroplast-derived trnS genes in Cucurbita, 3 remain intact and 2 of these recognize the same codons as their notably absent mitochondrial homologs (see above), making them candidates for unusually recent, functional replacement of native copies.

Noncoding and Promiscuous Sequences
Most of the sequence in both genomes—73% in Citrullus and 90% in Cucurbita—is intergenic, lying outside of conserved syntenic regions (table 1). A large fraction of these intergenic sequences, 159–180 kb, shows similarity to previously sequenced seed plant mitochondrial DNA (table 1), excluding chloroplast-like sequences. Chloroplast-derived DNA accounts for 1–9% of sequenced seed plant mitochondrial genomes (Kubo and Mikami 2007; Goremykin et al. 2009), so in this respect, Citrullus resembles the typical plant mitochondrial genome, containing 23 kb (6% coverage) of chloroplast-derived sequence distributed among 20 distinct regions in the genome (fig. 2 and table 1; supplementary fig. 1, Supplementary Material online). Cucurbita, on the other hand, has a remarkable 113 kb of chloroplast-derived sequence—fully 1.7–29 times more than other fully sequenced seed plant mitochondrial genomes. Put another way, the Cucurbita mitochondrial genome contains more chloroplast DNA than it does mitochondrial genes and introns combined (table 1). Chloroplast sequences are divided among 29 distinct regions, ranging from 92 to 18,534 nt in length (fig. 2; supplementary fig. 2, Supplementary Material online). The regions are relatively large (median length = 2.3 kb), with nine exceeding 5 kb and two exceeding 15 kb in length. At 16.6 and 18.5 kb, the latter two fragments are among the largest contiguous stretches of chloroplast-derived DNA so far characterized in plant mitochondria, though much of the 25 kb of chloroplast DNA in maize likely arrived as a single segment that was subsequently fragmented inside the mitochondrial genome (Clifton et al. 2004; Allen et al. 2007). Counting twice those regions that map entirely within both copies of the large chloroplast inverted duplication, the 29 regions cover 79% of the Cu. sativus chloroplast genome. Some regions of the chloroplast genome are represented more than once in the mitochondrial genome, reflecting either multiple independent transfers or single transfers that were subsequently duplicated inside the mitochondrial genome.

Plant mitochondrial genomes typically house some small fraction of discernibly nuclear-derived sequences, most commonly identified as transposable elements. The mitochondrial genome of the lycophyte, Isoetes engelmannii, is exceptional in that it contains degenerate intergenic sequences matching an auxin-responsive transcription factor and a phytochrome gene, both of which are encoded in the nucleus (Grewe et al. 2009). Although nuclear sequences are generally more difficult to detect, detailed studies of a few species have nevertheless shown that 5% of their mitochondrial DNA can be traced to the nucleus (Knoop et al. 1996; Unseld et al. 1997; Notsu et al. 2002). The Citrullus and Cucurbita mitochondrial genomes contain 24 kb (6.4%) and 21 kb (2.1%), respectively, of clearly identifiable nuclear-derived sequences, most of which resemble copia- and gypsy-like retrotransposons (table 1). Both genomes also contain regions with strong matches to nuclear protein-coding genes. Citrullus and Cucurbita each contain sequences with similarity to an (R)-mandelonitrile lyase gene and a lectin protein kinase gene. In both cases, the gene fragments cover large and similar tracts of their cognate nuclear copies. A close homolog of the mandelonitrile lyase gene in the nuclear genome of Arabidopsis (GenBank GI: 15238300) has two introns, and whereas virtually the entire length of the Cucurbita fragment is from exon 2, the longer Citrullus fragment
covers much of exons 2 and 3 along with the intervening intron, which indicates that the transfer did not involve an RNA intermediate. For both species, the lectin protein kinase fragments fall within the large second exon of a homolog in the nuclear genome of *Populus* (GenBank GI: 116256320). The lectin protein kinase fragment, which is nearly full length in *Citrullus*, is divided between two distantly spaced fragments in *Cucurbita*, an apparent consequence of intramolecular recombination following the transfer.

Given that most plant mitochondrial DNA (as much as 80–90%) shows no similarity to known sequences, one hypothesis is that much of the variation in genome size reflects different amounts of DNA acquired (and retained) from the large and mostly noncoding plant nuclear genome (Palmer 1990). It therefore came as a surprise that the smaller *Citrullus* mitochondrial genome contains more identifiably nuclear DNA than does the *Cucurbita* genome. The unsequenced nuclear genomes of *Citrullus* (430 Mb) and *Cucurbita* (539 Mb) (Ren et al. 2009) represent huge reservoirs of unexamined sequence, some of which could have found its way into the mitochondrial genome. In addition to the present availability of only a few nuclear genome sequences from relatively distantly related plants, the challenge of identifying putative nuclear sequences is further complicated by the possibility that the ancestral nuclear genomes of these species could have been much larger. A large fraction of the sequence in these two mitochondrial genomes—21% in *Citrullus* and 58% in *Cucurbita*—shows weak or no similarity to known sequences, so the possibility remains that they might contain substantial amounts of additional nuclear-derived DNA.

**Repeats**

Reassociation kinetics suggested that 5–10% of the sequence in the *Citrullus* and *Cucurbita* mitochondrial genomes consists of low-complexity repetitive DNA (Ward et al. 1981). Consistent with this estimate, *Citrullus* has 1,154 repeats that cover 10% of the genome, based on our Blast settings and an expect cutoff of 1. The largest repeat—a 7.3-kb inverted repeat—creates duplicate copies of the *sdh3*, *trnQ*, and *trnN* genes. The short 3-kb clone library used to sequence the *Citrullus* genome (see Materials and Methods) could not provide insights into whether this repeat engages in high-frequency recombination, as would be expected for a repeat of this size (Lonsdale et al. 1988; Palmer and Herbon 1989). All remaining repeats are <400 nt in length, and most of these (900 of 1,154) are only 19–40 nt in length (table 2). Repeat coverage is not simply a reflection of duplicated genes, as the majority of repeat coverage (81%) in *Citrullus* lies outside of genes and introns.

Despite similar distributions of repeat lengths, *Cucurbita* has nearly 50 times more repeats (54,783 vs. 1,154) than *Citrullus* (table 2), based on our Blast settings and an expect cutoff of 1. In fact, much of the genome expansion in *Cucurbita* owes to an accumulation of repeats in intergenic regions, with total genomic coverage by repeats (371 kb, or 38% of the genome) summing to nearly the entire length of the *Citrullus* genome. Like *Citrullus*, the overwhelming majority of repeats in *Cucurbita* are short; a total of 38,724 of them (71%) are just 19–40 nt in length, accounting for >272 kb (28%) of the total genome coverage. Most of the repeats lie within intergenic regions, with <2% of repeated sequences occurring within genes, introns, and RNA genes. A larger fraction of the total repeat coverage (11%) overlaps with chloroplast-derived regions, pointing to multiple introductions of the same chloroplast region and/or duplications of chloroplast-like sequences within the mitochondrial genome. Tandem repeats account for a virtually negligible fraction of the repeat coverage in both genomes.

**Table 2. Frequency Distributions of Repeat Lengths in the Mitochondrial Genomes of *Citrullus* and *Cucurbita***

<table>
<thead>
<tr>
<th>Repeat Length (nt)</th>
<th>Citrullus Number of Repeats (% coverage)</th>
<th>Cucurbita Number of Repeats (% coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19–20</td>
<td>95 (0.47)</td>
<td>4,331 (7.09)</td>
</tr>
<tr>
<td>21–40</td>
<td>805 (3.44)</td>
<td>34,393 (26.92)</td>
</tr>
<tr>
<td>41–60</td>
<td>134 (1.34)</td>
<td>8,552 (15.08)</td>
</tr>
<tr>
<td>61–80</td>
<td>39 (0.66)</td>
<td>3,417 (6.67)</td>
</tr>
<tr>
<td>81–100</td>
<td>23 (0.46)</td>
<td>1,591 (6.94)</td>
</tr>
<tr>
<td>101–120</td>
<td>14 (0.31)</td>
<td>821 (5.38)</td>
</tr>
<tr>
<td>121–140</td>
<td>15 (0.48)</td>
<td>510 (4.24)</td>
</tr>
<tr>
<td>141–160</td>
<td>7 (0.23)</td>
<td>362 (3.57)</td>
</tr>
<tr>
<td>161–180</td>
<td>6 (0.27)</td>
<td>246 (2.45)</td>
</tr>
<tr>
<td>181–200</td>
<td>6 (0.30)</td>
<td>173 (2.09)</td>
</tr>
<tr>
<td>201–220</td>
<td>0 (0.00)</td>
<td>96 (1.47)</td>
</tr>
<tr>
<td>221–240</td>
<td>0 (0.00)</td>
<td>67 (1.16)</td>
</tr>
<tr>
<td>241–260</td>
<td>2 (0.14)</td>
<td>72 (1.49)</td>
</tr>
<tr>
<td>261–280</td>
<td>0 (0.00)</td>
<td>34 (0.79)</td>
</tr>
<tr>
<td>281–300</td>
<td>0 (0.00)</td>
<td>34 (0.90)</td>
</tr>
<tr>
<td>301–400</td>
<td>6 (0.55)</td>
<td>56 (1.64)</td>
</tr>
<tr>
<td>401–500</td>
<td>0 (0.00)</td>
<td>20 (0.63)</td>
</tr>
<tr>
<td>501–600</td>
<td>0 (0.00)</td>
<td>4 (0.22)</td>
</tr>
<tr>
<td>601–700</td>
<td>0 (0.00)</td>
<td>4 (0.25)</td>
</tr>
<tr>
<td>≥7,286</td>
<td>2 (3.84)</td>
<td>0 (0.00)</td>
</tr>
</tbody>
</table>

**Note:**—Repeats were identified by Blasting each genome to itself (see Materials and Methods) and considering all hits with a Blast expect value <1. Repeats are defined by their begin and end coordinates in the genome. In tallying the total number of repeats, those with identical begin and end coordinates are counted only once. Percent coverage is the percentage of nucleotide positions in the genome occupied by repeats of given length category (column 1), calculated without respect to repeats in other length categories. Whereas the number of repeats is additive among rows, the coverage values are not because repeats can (and do) overlap.

Large repeats are common in seed plant mitochondrial genomes, with lengths of maximal perfect repeats ranging from 897 nt in the large (773 kb) *Vitis* genome to 87 kb in the 501-kg *Beta* (Owen-CMS) genome (Satoh et al. 2004; Goremykin et al. 2009). In general, genome size tends to scale positively with genomic coverage by large (>500 nt) repeats. The trend appears to hold within species (Allen et al. 2007), within genera (Palmer and Herbon 1989), and—with some notable exceptions (e.g., *Vitis*)—across the whole of angiosperms. Moreover, large sequence duplications can cause major and rapid increases in genome size, as is the case for five maize mitochondrial genomes, which have virtually identical sequence complexities but nevertheless range from 536 to 740 kb in size (Allen et al. 2007). It therefore came as a surprise that the largest
repeat in the nearly 1-Mb \textit{Cucurbita} genome is quite small, just 621 nt in length. Thus, both \textit{Cucurbita} and \textit{Vitis} are exceptional in combining an unusually large genome with a scarcity of large repeats.

As highlighted by the \textit{Cucurbita} genome, small repeated sequences are an important determinant of plant mitochondrial genome size. As sites of active recombination, repeats have important impacts on the structure of plant mitochondrial genomes as well (Palmer and Herbon 1989; André et al. 1992). At least 38% of the \textit{Cucurbita} genome is repetitive DNA, considerably more than the 5–10% estimate based on reassociation kinetics (Ward et al. 1981). This discrepancy probably reflects the very short size of the repeats, which are likely to fail to reassociate with repetitive kinetics. The repetitive fraction alone accounts for >60% of the absolute size difference between \textit{Citrullus} and \textit{Cucurbita}. Limited sequencing and hybridization analyses led, by extrapolation, to the estimate that seven short (30–53 nt) motifs account for some 13% of the large (~1.8 Mb) mitochondrial genome of the close relative, \textit{Cu. sativus} (fig. 1) (Lilly and Havey 2001), providing an early indication that repetitive sequences might be a more important factor in the growth of cucurbit genomes than originally thought (Ward et al. 1981). Blast searches of the 7 dominant repetitive motifs from \textit{Cu. sativus} (Lilly and Havey 2001) confirmed their absence from both \textit{Citrullus} and \textit{Cucurbita}, indicating that entirely different suites of small repeats underlie the genome expansions in \textit{Cucurbita} and \textit{Cu. sativus}. One caveat to this conclusion is that, along with point mutations, recombination across repeats can in principle effectively shuffle their sequences, erasing any signal of homology over time (André et al. 1992). In \textit{Cycas}, the proliferation of a single, short (36 nt), and apparently self-replicating repeat (termed “Bpu sequence”) accounts for 5% coverage of its 415-kb mitochondrial genome (Chaw et al. 2008). The Bpu sequence was not found in either \textit{Citrullus} or \textit{Cucurbita}, nor does it appear that the large repetitive fraction of the \textit{Cucurbita} genome reflects the proliferation of just one or a few motifs. Finally, although less common, small repeats can be a major source of expansion in other organelle genomes. Despite having a relatively reduced gene complement, the green alga \textit{Chlamydomonas} has an unusually large, 203-kb chloroplast genome, ~20% of which comprises short sequence repeats located in intergenic regions (Maul et al. 2002). Another green alga, \textit{Volvox carteri}, has an extraordinarily large (≥420 kb) chloroplast genome, more than 60% of which consists of short, presumably self-replicating palindromic repeats located primarily in intergenic regions (Smith and Lee 2009b). The same repeat element has led to major expansion of the \textit{Volvox} mitochondrial genome as well (Smith and Lee 2009b).

**RNA Editing**

Sequencing of full or nearly full-length cDNAs for 37 mitochondrial protein genes identified a similar set of RNA editing sites in the two genomes. RNA editing data are summarized in supplementary table 3 (Supplementary Material online), and the locations of editing sites are available in the GenBank records (GQ856147 and GQ856148). The \textit{Citrullus} and \textit{Cucurbita} mitochondrial genomes contain a minimum of 463 and 444 sites with C-to-U editing, respectively. We found no evidence of U-to-C editing. Considering only one copy of each of the duplicated genes, the average density of editing sites is also higher in \textit{Citrullus} (1.57 edits/100 nt) than in \textit{Cucurbita} (1.51 edits/100 nt), so the absolute difference in the number of edits does not reflect minor variation in gene length or cDNA sequence coverage between the two species (fig. 3). \textit{Citrullus} does, however, have one additional source of five edited sites in the form of a partial \textit{atp9} pseudogene.

The \textit{Citrullus} and \textit{Cucurbita} totals (463 and 444 edits, respectively, in 37 genes) are comparable with those in the four other angiosperms for which comprehensive cDNA sequencing has been done: \textit{Arabidopsis} (441 edits in 36 genes), \textit{Brassica} (427 edits in 34 genes), \textit{Beta} (357 edits in 31 genes), and \textit{Oryza} (491 edits in 34 genes) (Giegé and Brennicke 1999; Notsu et al. 2002; Handa 2003; Mower and Palmer 2006). \textit{Citrullus} and \textit{Cucurbita} have 437 and 422 editing sites, respectively, in those regions of the genes with cDNA coverage in both species (27,881 nt). A total of 394 of these sites are shared between the two species, which translates to 90% of the sites in \textit{Citrullus} and 93% of the sites in \textit{Cucurbita}. By comparison, two other confamilial species with comprehensive editing data, \textit{Arabidopsis} and \textit{Brassica} (Brassicaceae), share 81% and 84% of edited sites, respectively (Handa 2003); it is not clear, however, if or how these calculations accounted for missing and/or unshared cDNA coverage.

Both cucurbit species showed the typical pattern of relative editing levels across genes. For example, ribosomal proteins tend to have fewer edits than other genes, and the \textit{mttB}, \textit{ccmB}, and \textit{ccmFn} genes are highly edited. As with other...
angiosperms, the majority of edits (>92%) alter the amino acid, and most of these sites are fully edited. Although the set of fully edited, nonsynonymous sites is generally highly conserved between the two species, there are several striking differences between them in the nonsynonymous edits for six genes. Most of the edited sites in the ccmFc, cob, matR, and mttB genes are nonsynonymous edits that are fully edited in *Cucurbita* but partially edited in *Citrullus*. One notable example is the mttB gene, in which 18 of the 20 nonsynonymous edit sites are fully edited in *Citrullus* whereas just 3 of 18 nonsynonymous edits are fully edited in *Citrullus*. In the other direction, most of the nonsynonymous edits in *nad9* and *rps4* are fully edited in *Citrullus* but partially edited in *Cucurbita*. For example, *Citrullus* and *Cucurbita* each have eight edit sites in the *rps4* gene, all of which are nonsynonymous and shared between the two species. Whereas all eight of the sites are fully edited in *Citrullus*, just one of them is fully edited in *Cucurbita*.

For both species, RNA editing creates start codons for the *nad1*, *nad4L*, and *rps10* genes and stop codons for the *atp9* and *rps10* genes. We had insufficient cDNA coverage to determine whether several other putative start (i.e., ACG) and stop (i.e., CAA) codons that would seem to require editing are in fact edited in either or both species. In addition, canonical start codons were not detected for the *matR* gene in *Citrullus* or for the *mttB* gene in *Cucurbita*, nor was a canonical stop codon detected for the *Citrullus sdh3* gene. The cDNA sequence of the latter showed that neither the CGA nor the CAA codon in the 129 nt immediately downstream of *sdh3* is edited to produce a stop codon.

**Nucleotide Substitution Rates and the Mutation Pressure Hypothesis**

As exemplified by these two genomes, seed plants are notorious for their extremely large and variably sized mitochondrial genomes. Rates of synonymous substitution, which presumably reflect the underlying mutation rate, show similarly dramatic fluctuations across seed plants as well (Mower et al. 2007). Notwithstanding this variation, the mutation rate is generally extremely low, some 3–5 times lower than in the chloroplast and 40–100 times lower than in animal mitochondria (Wolfe et al. 1987), whose genomes are comparatively miniscule in size (typically 14–20 kb) (Lynch et al. 2006). Thus, a seemingly strong negative correlation between organelle genome size and mutation rate is seen across several groups of eukaryotes (Lynch et al. 2006). This pattern, along with the apparent lack of correlation between the effective number of genes per locus in organelles (N_e) and their genome sizes, led to the hypothesis that mutation rate is the primary determinant of organelle genome size (Lynch et al. 2006). According to this hypothesis, the generally low mitochondrial mutation rate of plants facilitates the accumulation of noncoding sequences and hence the overall growth of their mitochondrial genomes. The elevated mitochondrial mutation rate likewise maintains (indirectly) the small, streamlined mitochondrial genomes of animals. In essence, the superfluous noncoding DNA carries less potential burden (Lynch 2006; Lynch et al. 2006) in the low mutational environment present in most plant mitochondria (Wolfe et al. 1987; Mower et al. 2007). Like genes and introns, sites of RNA editing require up- and downstream sequence conservation to be properly processed (Choury et al. 2004; Mulligan et al. 2007), which, again, are thought to be more easily preserved in a low mutational background (Lynch et al. 2006). The mutation pressure hypothesis therefore predicts that both RNA editing frequency and genome size should be negatively correlated with the mutation rate (Lynch et al. 2006).

We made a concatenated alignment of 30 mitochondrial genes from the now 18 fully sequenced seed plant mitochondrial genomes and used it to estimate rates of synonymous (d_S) and nonsynonymous (d_N) substitution in *Citrullus* and *Cucurbita*. The d_S and d_N trees show similar patterns of divergence. Namely, *Cucurbita* falls on relatively long branches in both d_S and d_N trees (fig. 4). Relative rate tests confirm that both d_S and d_N are significantly higher in *Cucurbita* (P < 0.001). Compared with *Citrullus*, *Cucurbita* has a 4-fold higher d_S and an 8-fold higher d_N (fig. 4). That most of the total variation is at silent sites implicates mutation rate as the underlying cause of the observed rate increase in *Cucurbita* (Kimura 1983).

Based on the previous knowledge of a nearly 3-fold larger genome size than *Citrullus*, the mutation pressure

---

**Fig. 4.** Nucleotide substitution rates in 17 angiosperms with fully sequenced mitochondrial genomes. Branch lengths are proportional to rates of synonymous (left panel) and nonsynonymous (right panel) substitutions, based on a concatenated alignment of 30 protein genes (25,553 positions), and with the tree topology constrained as shown (see Materials and Methods). Relative rate tests show significantly higher d_S and d_N in *Cucurbita* compared with *Citrullus*. *Cycas* was included in the analysis to root the basal angiosperm divergence but, for simplicity, is not shown.
hypothesis (Lynch et al. 2006) predicts that Cucurbita should have a lower mutation rate and more RNA editing sites. However, since its split from Citrullus some 30 Ma, Cucurbita has actually experienced a 4-fold higher mutation rate than Citrullus, whereas its density of RNA editing is only slightly lower (97% of the level of Citrullus; determining the polarity—the gain and loss—of editing site differences is difficult with these data alone). So whereas RNA editing frequencies are weakly consistent with predictions of the mutation pressure hypothesis, genome sizes clearly are not. Although a quantitative assessment of any correlations cannot be measured from these two data points alone, the elevated synonymous substitution rate (and presumably the underlying mutation rate) in Cucurbita was nonetheless surprising to discover in light of its long-established larger genome size.

The mutation pressure hypothesis was based on broad-scale patterns of organelle mutation rate, \( N_g \) and genome size across a diverse set of eukaryotes, so it is unclear how or when these same factors are manifest in the genomes of taxa spanning a much narrower phylogenetic range (see discussions in Lynch and Conery 2004; Vinogradov 2004). That is, many large-scale mitochondrial genomic changes clearly have occurred over the course of angiosperm—even Cucurbitaceae—evolution, but it is unclear whether the magnitude of mutation rate variation and the timescales over which those shifts have occurred are sufficient to drive the predicted changes in genome size and RNA editing frequency. The stochastic nature of both \( N_g \) (for which we have no data) and mutation rate predicts local departures from the overall trend (Lynch and Conery 2004), making the hypothesis difficult to reject at more local phylogenetic scales (Vinogradov 2004). That is, many large-scale mitochondrial genomic changes clearly have occurred over the course of angiosperm—event Cucurbitaceae—evolution, but it is unclear whether the magnitude of mutation rate variation and the timescales over which those shifts have occurred are sufficient to drive the predicted changes in genome size and RNA editing frequency. The stochastic nature of both \( N_g \) (for which we have no data) and mutation rate predicts local departures from the overall trend (Lynch and Conery 2004), making the hypothesis difficult to reject at more local phylogenetic scales (Vinogradov 2004), within Cucurbitaceae for example. Limited data from this and other studies of organelle genome evolution (Smith and Lee 2008, 2009a) show mixed support for the mutation pressure hypothesis. We do note, however, the long \( d_s \) branch leading to the common ancestor of Arabidopsis and Brassica (fig. 4), both of which have relatively low editing levels and, by angiosperm standards, compact mitochondrial genomes. Additional plant mitochondrial genome sequences will no doubt shed valuable light on the extent to which \( N_g \) and mutation rate drive organelle genome evolution. One important and likely difficult challenge will be finding markers from the slowly evolving plant mitochondrial genomes with enough variability to allow estimation of \( N_g \) from the standing level of neutral variation. Such data might show, for example, that the large and variably sized cucurbit mitochondrial genomes reflect stochastic variation resulting from a generally low \( N_g \) across Cucurbitaceae.

**Perspectives on Genome Size Evolution**

The genome size estimates of Ward et al. (1981) for four species of Cucurbitaceae provided very early indications that plant mitochondrial genome size is extremely fluid—a finding that has been validated repeatedly in the ensuing years. The discovery of 8-fold variation in genome size, and organelle genomes that measured in the megabases, sparked years of discussion, speculation, and study about the origin of the "extra" DNA in the largest genomes. Data from this study show that growth of the nearly 1-Mb Cucurbita genome is largely attributable to the uptake and retention of an usually large quantity of exogenous chloroplast sequences and the virtually unprecedented proliferation of small repeats, the origins of which remain unclear. Indeed, subtracting both of these components yields an average-sized plant mitochondrial genome.

Major and rapid changes in genome size are common in the mitochondrial and nuclear genomes of seed plants, whereas chloroplast genome size is much more conserved. Paradoxically, though, the forces underlying genome size evolution have greater overlap between the two organelle genomes than between the dynamic genomes of the mitochondrion and nucleus. In plant nuclear genomes, rapid growth is usually the result of polyploidization and/or the proliferation of transposable elements (Hawkins et al. 2008). Polyploidization is a uniquely nuclear process, and active mobile elements have not been identified in any plant mitochondrial genome, with the possible exception of the Cycas genome, 5% of which consists of a family of very short (36 nt), putatively self-replicating elements (Chaw et al. 2008). The chloroplast genomes of photosynthetic seed plants vary by less than 2-fold in size, with most angiosperm genomes occupying an even narrower size range (135–160 kb; Ravi et al. 2008). Most of this size range is the result of expansion or contraction of a large (0–76 kb) inverted repeat (e.g., Chumley et al. 2006).

A diverse and fairly lineage-specific constellation of factors appears to underlie rapid and major size fluctuations in the mitochondrial genomes of seed plants. Whereas differences in length and number of large duplicated (sometimes triplicated) regions are, as in chloroplast genomes, the paramount drivers among related species and genetic lines of Zea and Beta (Satoh et al. 2004; Allen et al. 2007), the primary force driving genome expansion in Cucurbita is the proliferation of short dispersed repeats, with the accumulation of chloroplast sequences playing an important secondary role. The large mitochondrial genome of Cucurbita is surprisingly devoid of large segmental duplications. Like Cucurbita, the next largest sequenced mitochondrial genome (Vitis—773 kb) contains a large quantity of chloroplast sequences and no major segmental duplications but differs in having relatively few short repeats (Goremykin et al. 2009). The gain of foreign sequences from more-or-less distantly related species by horizontal gene transfer is relatively common in plant mitochondrial genomes (Richardson and Palmer 2007) and is a major force underlying genome expansion in the flowering plant Amborella trichopoda (Bergthorsson et al. 2004; Rice DW, Palmer JD, unpublished data). Horizontal transfer does not, however, appear to have been an important factor in the overall evolution of the now 18 fully sequenced seed plant mitochondrial genomes. Insofar as can be determined, the uptake of nuclear sequences seems to have been of relatively minor importance in the completely sequenced genomes as well (but see above for caveats concerning this conclusion).
It is increasingly clear that two forces that sometimes play a major role in genome size evolution in other mitochondrial lineages, gene loss (e.g., in some green algae—Nedelcu 1998) and intron gain and loss (e.g., in fungi—Cummings et al. 1990), have been of negligible importance in the remarkable expansion and contraction of seed plant mitochondrial genomes. This is particularly striking in comparing the Citrullus and Cucurbita genomes, which are so different in size but nearly identical in gene and intron content. Current efforts to sequence and characterize the extraordinarily large Cucumis genomes have largely confirmed the early size estimates of Ward et al. (1981) and will undoubtedly turn up additional insights about the interrelationships of mutation rate, RNA editing, and genome size in organellar genomes.

Supplementary Material

Supplementary figures 1 and 2 and tables 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Arnold Bendich, Weilong Hao, Dan Sloan, and two anonymous reviewers for commenting on earlier versions of the manuscript. We thank Stacia Wyman for providing the DOGMA source code and for advice on developing the gene annotation scripts. This work was supported by a National Institutes of Health (NIH) Ruth L. Kirschstein National Research Service Award Postdoctoral Fellowship (1F32GM080079-01A1) to A.J.A., an NIH research grant RO1-GM-70612, and the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc., to J.D.P. This work was performed under the auspices of the US Department of Energy’s Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract no. DE-AC02-06NA25396.

References


Handa H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.); comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res. 31:5907–5916.

extensive rearrangements and the origin of the mitotype-unique regions. Theor Appl Genet. 113:477–484.