Unprecedented Heterogeneity in the Synonymous Substitution Rate within a Plant Genome

Andan Zhu,¹,² Wenhu Guo,¹,³ Kanika Jain,¹,³ and Jeffrey P. Mower*,¹,²

¹Center for Plant Science Innovation, University of Nebraska-Lincoln
²Department of Agronomy and Horticulture, University of Nebraska-Lincoln
³School of Biological Sciences, University of Nebraska-Lincoln

*Corresponding author: E-mail: jpmower@unl.edu.
Associate editor: Brandon Gaut

Abstract

The synonymous substitution rate varies widely among species, but it is generally quite stable within a genome due to the absence of strong selective pressures. In plants, plastid genes tend to evolve faster than mitochondrial genes, rate variation among species generally correlates between the mitochondrial and plastid genomes, and few examples of intragenomic rate heterogeneity exist. To study the extent of substitution rate variation between and within plant organellar genomes, we sequenced the complete mitochondrial and plastid genomes from the bugleweed, Ajuga reptans, which was previously shown to exhibit rate heterogeneity for several mitochondrial genes. Substitution rates were accelerated specifically in the mitochondrial genome, which contrasts with correlated plastid and mitochondrial rate changes in most other angiosperms. Strikingly, we uncovered a 340-fold range of synonymous substitution rate variation among Ajuga mitochondrial genes. This is by far the largest amount of synonymous rate heterogeneity ever reported for a genome, but the evolutionary forces driving this phenomenon are unclear. Selective effects on synonymous sites in plant mitochondria are generally weak and thus unlikely to generate such unprecedented intragenomic rate heterogeneity. Quickly evolving genes are not clustered in the genome, arguing against localized hypermutation, although it is possible that they were clustered ancestrally given the high rate of genomic rearrangement in plant mitochondria. Mutagenic retroprocessing, involving error-prone reverse transcription and genomic integration of mature transcripts, is hypothesized as another potential explanation.

Key words: Ajuga reptans, mitochondrial genome, plastid genome, intragenomic rate heterogeneity, nonstandard start codons.

Introduction

One of the fundamental principles of the neutral theory of molecular evolution is that the substitution rate of a sequence evolving in the absence of selection should approximate the underlying mutation rate (Kimura 1984). Because synonymous sites are free from selection at the protein level, they are widely used to infer mutation rates. Of course, selection can act on synonymous sites in other ways, such as to affect translational efficiency, to maintain regulatory motifs, or to produce localized nucleotide biases (Chamary et al. 2006; Shabalina et al. 2013). Nevertheless, in many genomes, the effects of selection on synonymous sites are generally quite limited, especially in those species with small populations (Chamary et al. 2006).

When the strength of selection on synonymous sites is weak or absent in a genome, the synonymous substitution rate should be fairly constant among genes. This is generally true in yeast nuclear genomes (Fay and Benavides 2005; Fox et al. 2008), in mammalian mitochondrial genomes (Pesole et al. 1999), and, to a lesser extent, in mammalian and insect nuclear genomes (Fox et al. 2008). Minor variations in synonymous rate within a genome have been detected in human (Chuan and Li 2004), rat (Gaffney and Keightley 2005), Caenorhabditis (Stein et al. 2003), and Arabidopsis (Yang and Gaut 2011). Synonymous rate variation can also be region specific. For example, rates in X-chromosome-linked genes were decreased approximately 40% in mammals (Waterston et al. 2002; Malcom et al. 2003) and 30% in Drosophila (Singh et al. 2009) compared with autosomal genes, while Y-linked genes generally evolve several times faster than X-linked genes and autosomal genes (Ellegren and Fridolfsson 1997; Makova and Li 2002). Several more dramatic examples of intragenomic rate variation also exist, including 65-fold variation in synonymous rates among Chlamydomonas nuclear genes (Popescu et al. 2006) and 50-fold variation among genes in Escherichia coli (Martincarena et al. 2012). Potential evolutionary explanations of substitution rate heterogeneity include context compositional mutagenesis (e.g., CpG hypermutability) (Smith et al. 2002; Berglund et al. 2009), mutagenic recombination (Hellmann et al. 2003), transcription-coupled repair efficiency (Martincarena et al. 2012), gene-specific optimization (Martincarena et al. 2012), and natural selection (Chuang and Li 2004).

In angiosperms, mitochondrial genomes (mitogenomes) are more variable than plastid genomes ( plastomes) in terms of size, structure, and content (Jansen and Ruhlman 2012; Mower, Sloan et al. 2012), whereas synonymous
substitution rates are typically several times slower in mitochondria genes compared with plastid genes (Wolfe et al. 1987; Drouin et al. 2008; Richardson et al. 2013). Rate changes, when they have occurred, tend to correlate between the mitogenome and plastome (Eyre-Walker and Gaut 1997; Drouin et al. 2008; Huang et al. 2012). However, several angiosperm lineages (Geraniaceae, Plantago, and Silene) have independently experienced a massive increase in the mitochondrial synonymous substitution rate (Cho et al. 2004; Parkinson et al. 2005; Mower et al. 2007; Sloan, Alverson, Chuckalovcak et al. 2012). Mitochondrial synonymous rates have episodically increased >100-fold in each of these lineages, with equally remarkable rate reversions approaching typical angiosperm levels in some Plantago and Geraniaceae lineages (Cho et al. 2004; Parkinson et al. 2005; Mower et al. 2007). The underlying mechanisms for the dramatic rate accelerations remain uncertain, although defects in DNA replication, repair, or recombination have been suggested (Cho et al. 2004; Parkinson et al. 2005; Mower et al. 2007; Sloan, Alverson, Chuckalovcak et al. 2012). Synonymous substitution rates can also vary within a genome, but few examples are known for plant organellar genomes. A 2- to 5-fold higher rate was identified in the plastid single-copy region compared with the large inverted repeat, perhaps due to biased gene conversion or some other copy-correction process between duplicate genes (Wolfe et al. 1987; Perry and Wolfe 2002).

The 149,963 bp plastome has a 1,190-fold mean coverage depth with the large inverted repeat, perhaps due to biased gene conversion or some other copy-correction process between duplicate genes (Magee et al. 2010). A broad comparison of rates for particular organellar rate variation (fig. 1). Unexpectedly, the Ajuga atp8 and rps12 genes provide clear evidence for usage of alternative start codons (fig. 1). Both genes appear functional because they are intact (no premature stop codons or frameshifting indels) and transcribed (based on reverse transcriptase-polymerase chain reaction [RT-PCR] analysis), except that there is no viable upstream or downstream ATG codon in either gene and no ACG codon that is RNA edited to AUG in the transcript. However, at the evolutionarily conserved start position in other angiosperms, the Ajuga atp8 gene has a GTG codon and the rps12 gene has an ATA codon. Previous studies have suggested that GTG and ATA codons may initiate translation of several plant mitochondrial genes including cob, mtfB (previously orfX), and rpl16 (Bock et al. 1994; Sunkel et al. 1994; Siculella et al. 1996). These findings indicate that the mitochondrial atp8 and rps12 genes can also be initiated on alternative start codons in plants.

A Mitochondrion-Specific Increase in Substitution Rate

We used concatenated data sets of 27 mitochondrial genes and 78 plastid genes to examine organellar rate variation (fig. 2A) among angiosperms (supplementary table S5, Supplementary Material online). We discovered higher synonymous and nonsynonymous divergence in Ajuga mitochondrial genes in comparison with other sampled species. The synonymous and nonsynonymous branches for Ajuga are 5.6 times and 4.5 times longer than for its closest sampled relative, Boea hygrometrica. No increased divergence was observed in Ajuga’s plastid genes, indicating that the accelerated substitution rates are restricted to the mitogenome in Ajuga.
We also recovered higher mitochondrial sequence divergence in two *Silene* species, *Silene conica* and *S. noctiflora*, as previously reported (Mower et al. 2007; Sloan, Alverson, Chuckalovcak et al. 2012), which appears to be much more pronounced in the mitogenome compared with the plastome (Sloan, Alverson, Wu et al. 2012).

The uncoupling of substitution rates in the *Ajuga* and some *Silene* organellar genomes contrasts sharply with correlated organellar rate variation observed for most other angiosperms (fig. 2B). Indeed, linear regression of mitochondrial vs. plastid substitution rates for all branches in the phylogenetic tree (excluding the *Ajuga* and *Silene* branches with mitochondrial-specific rate accelerations) shows a significant correlation. After excluding *Ajuga, S. conica*, and *S. noctiflora*, all remaining angiosperms have plastid rates that are on average 4.9 times faster for synonymous substitutions and 1.8 times faster for nonsynonymous substitutions compared with mitochondrial rates, which is consistent with previous studies (Wolfe et al. 1987; Drouin et al. 2008; Richardson et al. 2013). *Ajuga* organelles show an opposite pattern, in which the mitochondrial rate is 1.5 times higher at synonymous sites and 4.0 times higher at nonsynonymous sites than its plastid rates.

**Unprecedented Synonymous Rate Heterogeneity in the Mitochondrial Genome**

Examination of synonymous divergence of individual mitochondrial genes revealed that the extent of rate increase in *Ajuga* is not consistent among genes (fig. 3). Representative trees show extreme diversity in synonymous rates among genes, ranging from no increase in *nad2* to highly accelerated rates in *atp6* and *rps3* (fig. 3A). A complete assessment of this mitochondrial rate variation (fig. 3B) showed that genes such as *atp9, rps3,* and *rps12* have high levels of synonymous sequence divergence, whereas all *nad* genes and several other genes have low levels of synonymous divergence similar to the typically low levels in most other angiosperms (represented here by *B. hygrometrica*). Synonymous rate heterogeneity among genes varies 340-fold in *Ajuga*, from 0.00359 substitutions/site in *nad7* to 1.22 substitutions/site in *atp9*. Given an organismal divergence time of 78 Ma between *Ajuga* and *Boea* (Bremer et al. 2004), absolute synonymous rates vary from $4.6 \times 10^{-11}$ to $1.6 \times 10^{-8}$ substitutions/site/year in *Ajuga* mitochondrial genes. These values are both unusually slow and extremely fast compared with other angiosperms, whose synonymous rates are typically around $0.2–1.0 \times 10^{-9}$ substitutions/site/year (Wolfe et al. 1987; Mower et al. 2007).

Mapping the synonymous substitution rates onto the genome shows that the quickly evolving genes are scattered throughout the genome and mostly interspersed among slowly evolving genes (fig. 3C).

Rate variation, to a lesser extent, was also observed at nonsynonymous sites and among rRNA genes, but introns, which are only found in slowly evolving genes, appear to be slowly evolving (fig. 3B). In the *Ajuga* plastome, genes exhibit

**Table 1. Summary of *Ajuga* Mitochondrial and Plastid Genomes.**

<table>
<thead>
<tr>
<th></th>
<th>Plastome</th>
<th>Mitogenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>149,963</td>
<td>352,069</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>38.3</td>
<td>45.1</td>
</tr>
<tr>
<td>Repeat content (%)</td>
<td>17.2</td>
<td>1.5</td>
</tr>
<tr>
<td>Coding content (%)</td>
<td>66.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Protein genes</td>
<td>79</td>
<td>27</td>
</tr>
<tr>
<td>rRNA genes</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td>Introns</td>
<td>21</td>
<td>20</td>
</tr>
</tbody>
</table>

**Fig. 1.** Alternative mitochondrial start codons. Protein sequence alignments of (A) *atp8* and (B) *rps12*. *Ajuga* protein sequences translated from DNA and RNA sequences are shown. Nonstandard codons in the *Ajuga* sequences are boxed when they occur at the conserved start codon positions in other species (marked with arrows). Downstream ATG and ACG codons in the *Ajuga* sequences are shaded in gray. The first in-frame stop codon upstream of the nonstandard start codon for each *Ajuga* sequence is shown.
very little synonymous rate variation (supplementary fig. S3, Supplementary Material online). However, it is curious that the one plastid gene with slightly accelerated rates is \( \text{atpH} \). This gene is homologous to \( \text{atp9} \), which has the fastest synonymous rate in the \( \text{Ajuga} \) mitogenome. The potential significance of this apparent rate correlation between the \( \text{atpH} \) and \( \text{atp9} \) homologs is not clear.

\section*{Discussion}

In most genomes, the synonymous substitution rate is fairly constant among genes. This is because the strength of selection on synonymous sites is generally weak or absent, and thus the synonymous rate is widely used in molecular evolution as an approximation of the underlying genomic mutation rate. In this study, we have shown that substitution rates are extraordinarily variable within the mitogenome of \( \text{A. reptans} \). The 340-fold range of synonymous rate variation among \( \text{Ajuga} \) mitochondrial genes is the largest range of heterogeneity for any genome sequenced to date. The \( \text{Ajuga} \) mitogenome also exhibits several other atypical characteristics including a reduction in gene, intron, repeat, and MIPT content as well as the usage of nonstandard start codons. In contrast, the \( \text{Ajuga} \) plastome is typical of other photosynthetic plants, with a standard set of genes, introns, and repeats and no major synonymous rate variation among genes.

Our findings raise fundamental questions about the evolutionary processes driving gene-to-gene variation in the synonymous substitution rate. What is the underlying cause of such unprecedented mitochondrial rate heterogeneity, and is the evolution of the other unusual genomic characteristics causally connected? The fast-evolving genes are unlikely to be pseudogenes because they are intact and \( d_N / d_S < 1 \), indicating that they are under selective constraint and probably encode functional proteins. Errors arising from our use of topological constraints during the rate calculations can also be

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Mitochondrial and plastid sequence divergence among selected angiosperms. (A) Mitochondrial (Mt) and plastid (Pt) phylograms of concatenated genes depicting synonymous (\( d_S \)) or nonsynonymous (\( d_N \)) sequence divergence. All trees were drawn to the same scale. (B) Scatter plot of synonymous (\( d_S \)) or nonsynonymous (\( d_N \)) sequence divergence between each shared branch in the mitochondrial and plastid phylograms shown in figure 2A. Linear regression analyses included all data points except for the \( \text{Ajuga} \) terminal branch value (triangle), the \( \text{Silene conica} \) and \( \text{S. noctiflora} \) terminal branch values (open circles), and the value for the internal branch leading to \( \text{S. conica} \) and \( \text{S. noctiflora} \) (open circle labeled Scon/Snoc). Significance of fit was evaluated by a Pearson correlation test in the R package.}
\end{figure}
discounted because unconstrained phylogenetic analyses firmly place *Ajuga* as sister to *Boea* and within the asterid clade, as expected (supplementary fig. S4, Supplementary Material online). The unconstrained phylogenetic results also exclude the possibility that the fast-evolving *Ajuga* genes were acquired by horizontal transfer from an organism with fast substitution rates. The lack of a plastid rate increase indicates that organismal effects (such as differences in
and those genes that are clustered (e.g., because the divergent genes are not clustered in the genome, it would have to act on the scale of individual genes or smaller hypothesized to be at work in legume plastomes (Magee et al. 2006; Islam et al. 2013). Across angiosperms, biased usage of preferential codons is implied by the lower GC content at third codon positions in plant mitochondrial genes compared with the genome average (Sloan and Taylor 2010), suggesting at least some selection to maintain higher AT content at synonymous sites. In Ajuga, divergent mitochondrial genes have a higher GC content at synonymous sites than conserved genes (supplementary fig. S5A, Supplementary Material online). The increased synonymous GC content in Ajuga may be due to a slight relaxation of selection for codon usage, causing GC content to drift closer to the genomewide average and resulting in a minor increase in the effective number of codons in quickly evolving genes relative to slowly evolving genes (supplementary fig. S5B, Supplementary Material online). Plant mitochondrial genes also contain regulatory motifs for binding of RNA editing and intron-splicing factors. In Ajuga, divergent genes lack introns (supplementary fig. SSC, Supplementary Material online) and tend to contain fewer edit sites than conserved genes (supplementary fig. SSD, Supplementary Material online), suggesting that divergent genes are under less selective constraint because they have fewer binding sites for splicing and editing factors. However, it has been shown that selection on synonymous sites appears to be quite weak in plant mitochondrial genes (Sloan and Taylor 2010), and it is difficult to imagine how a relaxation of these relatively minor selective effects could generate the 340-fold difference in synonymous rates in the Ajuga mitogenome.

The intragenomic synonymous rate variation may also result from a localized variation in mutation rate, which could be explained by at least two factors: localized hypermutation and mutagenic retroprocessing. Localized hypermutation due to frequent double-strand break repair was hypothesized to be at work in legume plastomes (Magee et al. 2010). Mutagenic retroprocessing (Parkinson et al. 2005) describes a process involving the frequent reverse transcription and genomic integration of mitochondrial transcripts. Because RNA polymerases and reverse-transcriptases are more error-prone than DNA polymerases, repeated rounds of retroprocessing would lead to increased mutation accumulation in the gene sequence. If either of these processes are the cause of increased rates in Ajuga mitogenomes, it would have to act on the scale of individual genes or smaller because the divergent genes are not clustered in the genome, and those genes that are clustered (e.g., nad4L + atp4; nad6 + rps3; atp8 + nad5) exhibit very different synonymous substitution rates. For localized hypermutation, the dispersed distribution of fast-evolving genes would require a large number of independent hypermutation hotspots, which seems unlikely. However, given the high rearrangement frequency of plant mitochondrial genomes (Arrieta-Montiel and Mackenzie 2011), it is conceivable that the genes were ancestrally more clustered in the genome, requiring fewer hypermutation hotspots. For mutagenic retroprocessing, the rate differences among genes could be tied to transcriptional levels if highly transcribed genes are retroprocessed more often. Intriguingly, mitochondrial transcriptome analysis in a grass and a legume show that atp9, rps3, and rps12 genes are highly expressed (Kazakoff et al. 2012; Islam et al. 2013); analysis of the Ajuga mitochondrial transcriptome will be necessary to more fully explore the possible link between transcription level and substitution rate. Alternatively, the gene-specific rate variation may mean that particular genes are more susceptible to retroprocessing. For example, transcripts with introns and frequent RNA editing will have splicing and editing factors bound along the transcript, which may provide some protection against retroprocessing by blocking reverse transcriptase binding and processivity. Consistent with this idea is the fact that divergent genes have no introns and fewer edit sites than conserved genes in Ajuga (supplementary figs. SSC and D, Supplementary Material online).

Finally, our results draw some interesting parallels, but also important distinctions, to the highly accelerated substitution rates observed in other plant lineages, including Plantago, Silene, and Geraniaceae. First, as we observed for quickly evolving genes in Ajuga, it is notable that there is a reduction in RNA editing frequency in the other three fast-evolving lineages as well (Cho et al. 2004; Parkinson et al. 2005; Sloan et al. 2010). This similarity lends further support to the idea that retroprocessing is a mutagenic process in Ajuga and the other species. Second, similar to Ajuga, mitochondrial genes in Silene show some gene-to-gene variation (fig. 3B; Sloan, Alverson, Chuckalovcak et al. 2012). However, other than atp9, which appears to be fast-evolving in many species (including the generally slowly evolving plants B. hygrometrica and S. latifolia), there is not a clear correlation of the faster and slower genes in Ajuga and S. noctiflora. For example, atp1, nad4L, nad9, and rps12 are some of the fastest evolving genes in S. noctiflora, but they are quite slowly evolving in Ajuga. This indicates that a gene’s identity is not a major determinant of substitution rate variation, suggesting that locus-specific rates are not based on gene or protein function.

In summary, the underlying cause of the extreme synonymous rate heterogeneity in the Ajuga mitochondrial genome is unclear from the present data, although locus-specific variation caused by mutational hotspots or mRNA retroprocessing seems to be most likely. To better address this question, it will be necessary to sequence additional genomes from close relatives to reconstruct the ancestral genomic state, identify the phylogenetic point of origin of the accelerated substitution rates, and correlate the evolutionary progression of the substitution rate variation with intron and RNA editing content. It will also be essential to get an estimate of relative
expression levels to determine whether transcription level correlates with substitution rate.

**Materials and Methods**

**DNA and RNA Sequencing**

*Ajuga reptans* individuals were obtained from a local nursery (Lincoln, Nebraska). We prepared organelle-enriched DNA as described previously (Grewe et al. 2013) and then sequenced 2 Gb of 100-bp paired-end reads from an 800-bp library at BGI (Shenzhen, China) using the Illumina HiSeq2000 platform. RNA isolation and RT-PCR were performed as described previously (Mower and Bonen 2009; Hepburn et al. 2012).

**Genome Assembly**

We ran multiple iterations of Velvet version 1.1.06 (Zerbino and Birney 2008) using different pairwise combinations of Kmer values (51, 61, 71, 81, 91) and expected coverage values (50, 100, 200, 500, 1,000, 2,000). For each run, minimum coverage was set to 10% of expected coverage and scaffolding was turned off. Mitochondrial and plastid contigs were identified in each assembly by using BlastN with known organellar protein-coding genes from related asterids as query sequences. For each genome, contigs from the three best assemblies (that maximized total mitochondrial or plastid length and minimized the number of mitochondrial or plastid contigs) were aligned manually, and the alignment consensus was taken as the final sequence. To assess depth of coverage, read pairs were mapped to the genome with Bowtie version 2 (Langmead and Salzberg 2012) with default parameters.

**Genome Annotation**

Protein and rRNA genes were annotated by using BlastN with homologous sequences from sequenced angiosperms, while tRNA genes were identified using tRNAscan-SE (Lowe and Eddy 1997). Unusual features such as alternative start codons and premature stop codons were rechecked by mapping the raw reads onto each genome using bowtie2 (Langmead and Salzberg 2012). Tandem repeats in the genomes were identified by using the Tandem Repeats Finder program (http://tandem.bu.edu/trf/trfhtml, last accessed March 3, 2014) with the following parameters: match = 2, mismatch = 7, indel = 7, minscore = 50, MaxPeriod = 2,000. Dispersed repeats were identified by using BlastN with an e-value = 1 and requiring > 20 bp matches. MIPTs were identified by using BlastN with an e-value = 1 and requiring > 100 bp matches. For the *Ajuga* atp8 and rps12 genes, RNA edit sites were determined experimentally by comparing the genome sequence with the organelle’s sequences in birds.

**Substitution Rate Estimation**

For phyllogenetic analysis, we selected angiosperms with sequenced mitochondrial and plastid genomes (supplementary table S5, Supplementary Material online). All 27 mitochondrial and 78 of 79 plastid protein-coding genes in *Ajuga reptans* were selected for rate analysis, excluding plastid ycf1 due to its weak conservation among angiosperms. Individual gene alignments were generated with clustalW2 (Larkin et al. 2007), manually adjusted, and then used to constrain nucleotide alignments with PAL2NAL (Suyama et al. 2006). Poorly aligned regions were removed by using Gblocks (Castresana 2000) with parameters t=c b1=11 b2=16 b3=8 b4=5 b5=none. synonymous and nonsynonymous substitution rates were calculated in HyPhy version 2.1 (Pond et al. 2005) with a local MG94xHK85 codon model and a topology constrained according to the Angiosperm Phylogeny website (http://www.mobot.org/MOBOT/research/APweb/, last accessed March 3, 2014). For rRNA genes and introns, sequences were aligned with clustalW2, manually adjusted, filtered through Gblocks, and then substitution rates were calculated with HyPhy using the HKY85 model.

**Supplementary Material**

Supplementary figures S1–S5 and tables S1–S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank Yizhong Zhang for performing the organelle-enriched DNA extractions, Dan Sloan for helpful discussion, and Jeffrey Palmer for critically reading the manuscript. This research was supported in part by National Science Foundation awards IOS-1027529 and MCB-1125386 and by startup funds from the University of Nebraska–Lincoln (to J.P.M.).

**References**


