A Review of the Prevalence, Utility, and Caveats of Using Chloroplast Simple Sequence Repeats for Studies of Plant Biology

Authors: Gregory L. Wheeler, Hanna E. Dorman, Alenda Buchanan, Lavanya Challagundla, and Lisa E. Wallace

Source: Applications in Plant Sciences, 2(12)
Published By: Botanical Society of America
URL: https://doi.org/10.3732/apps.1400059
A REVIEW OF THE PREVALENCE, UTILITY, AND CAVEATS OF USING CHLOROPLAST SIMPLE SEQUENCE REPEATS FOR STUDIES OF PLANT BIOLOGY

GREGORY L. WHEELER², HANNA E. DORMAN², ALENDA BUCHANAN², LAVANYA CHALLAGUNDLA², AND LISA E. WALLACE²,³

²Department of Biological Sciences, Mississippi State University, P.O. Box GY, Mississippi State, Mississippi 39762 USA

Microsatellites occur in all plant genomes and provide useful markers for studies of genetic diversity and structure. Chloroplast microsatellites (cpSSRs) are frequently targeted because they are more easily isolated than nuclear microsatellites. Here, we quantified the frequency and uses of cpSSRs based on a literature review of over 400 studies published 1995–2013. These markers are an important and economical tool for plant biologists and continue to be used alongside modern genomics approaches to study genetic diversity and structure, evolutionary history, and hybridization in native and agricultural species. Studies using species-specific primers reported a greater number of polymorphic loci than those employing universal primers. A major disadvantage to cpSSRs is fragment size homoplasy; therefore, we documented its occurrence at several cpSSR loci within and between species of Acmispon (Fabaceae). Based on our empirical data set, we recommend targeted sequencing of a subset of samples combined with fragment genotyping as a cost-efficient, data-rich approach to the use of cpSSRs and as a test of homoplasy. The availability of genomic resources for plants aids in the development of primers for new study systems, thereby enhancing the utility of cpSSRs across plant biology.

Key words: chloroplast microsatellite; cpSSR; plastid microsatellite; population genetics; size homoplasy.

Chloroplast genomes (cpDNA) have provided a wealth of data for discerning phylogenetic relationships within and among species (e.g., Olmstead and Palmer, 1994; Parks et al., 2009; Moore et al., 2010; Stech and Quandt, 2010; Drew et al., 2014), for evaluating uniparental patterns of gene flow within and among populations (e.g., Ennos et al., 1999; Wallace et al., 2011; Bai et al., 2014), for barcoding efforts to distinguish taxa (e.g., Hollingsworth, 2009, 2011; Chen et al., 2010), as well as for other studies. The appeal of using cpDNA markers lies in the fact that universal primers are capable of amplifying homologous regions across diverse taxa (Shaw et al., 2005, 2007), the cpDNA genome often exhibits uniparental inheritance (reviewed in Reboud and Zeyl, 1994; Birky, 1995), and the cpDNA genome is generally nonrecombining (Clegg, 1993; but see Ogihara et al., 1988; Marshall et al., 2001). However, instances of biparental inheritance leading to heteroplasy are also reported (e.g., Hansen et al., 2007). Much of the cpDNA genome evolves slowly, which reduces its general utility for evolutionary and population genetic studies, especially at lower taxonomic levels. Therefore, many researchers have sought to use noncoding regions, including introns and intergenic spacers, of the cpDNA genome to characterize genetic variation (Shaw et al., 2005, 2007). Like the nuclear genome, these regions have been found to contain tandem repeat regions or simple sequence repeats (SSRs) (Shaw et al., 2005, 2007; Dong et al., 2012), which often show higher levels of allelic variation than single nucleotide polymorphisms (Haasl and Payseur, 2011).

Since their introduction as a readily usable chloroplast marker by Powell et al. (1995), chloroplast SSRs (cpSSRs) have been increasingly used as genetic markers in a broad range of studies, in both basic plant sciences and applied agricultural research (Ebert and Peakall, 2009). They are commonly used in population genetic studies because of their ability to differentiate recently diverged groups (e.g., Echt et al., 1998), to function as a DNA barcode (e.g., Kuang et al., 2011), to allow for the quick identification of a genetic group or species (e.g., Decroocq et al., 2004), and to study hybridization and introgression (e.g., Wallace et al., 2011). Despite their widespread use in the past two decades, cpSSRs are not failsafe. Among the limitations of using cpSSRs are lack of variation in some species, lack of universal primers, size homoplasy (i.e., the occurrence of underlying sequence variation that does not change the length of the region as determined by genotyping), heteroplasy, and cytoplasmic introgression due to interspecific hybridization. Among these issues, size homoplasy may be the most serious because it can overestimate relatedness among samples and lead to incorrect evolutionary and ecological inferences. Cytoplasmic introgression may also be very serious if
undetected because it reflects the history of the donor genome rather than the full history of the focal species.

Given the continued use of cpSSRs in a wide diversity of systems, reviews of published studies employing them are worthwhile for guiding future research and for addressing outstanding problems with their use in plant biology. The foci of previous reviews of cpSSRs include potential problems associated with homoplasy (Navasucués and Emerson, 2005) and technical resources available for using cpSSRs (Ebert and Peakall, 2009). This latter study is particularly notable as it called attention to the great potential for cpSSRs to be informative for studies of noneconomic species and highlighted methods that could be used to identify variable loci and develop suitable primers (Ebert and Peakall, 2009). Several papers have called attention to the issue of size homoplasy and the importance of examining underlying sequence variation, but it is not clear that these warnings have been realized among published studies. In this review, we document the continued frequency and nature of published studies employing cpSSRs, and discuss the potential for fragment size homoplasy using an empirical data set of cpSSR loci that spans intraspecific and interspecific taxa of the genus *Acmispon* (Fabaceae) as well as reports from published papers. In light of the increased use of genomics techniques, we sought to determine whether researchers are still employing single locus approaches involving cpSSRs and whether there is a shift from using chloroplast markers to genome-wide markers. Finally, we use these data to discuss the strengths and weaknesses of cpSSRs for studies of plant biology, and we make recommendations for developing and using cpSSRs in future studies.

**MATERIALS AND METHODS**

**Review of published studies**—To understand how cpSSRs have been used in plant biology, we conducted a survey of primary literature sources on the subject. This survey was performed using the search engine Scopus (Elsevier B.V., Amsterdam, The Netherlands). We searched papers published between 1960 and 2013 using the following terms in the abstract, title, or key words: chloroplast SSRs, chloroplastic SSRs, plastid SSRs, cpSSRs, chloroplast MSATs, plastidic MSATs, cpMSATs, chloroplast microsatellites, chloroplast microsatellites, plastid microsatellites, chloroplast tandem repeats, chloroplast tandem repeats, plastid tandem repeats, plastid simple sequence repeats, plastid simple sequence repeats, and plastid simple sequence repeats. Only publications with an explicit use or discussion of cpSSR loci were included. Publications that were not available in English-language full-text version were not considered. This review is based on 439 papers (see Appendix S1). From these papers, we recorded the family (or families) of species studied, type of study based on the focal species (agricultural or native), number of loci used, and origin of primers used (e.g., developed de novo or published primer sequences). Studies of agricultural species focused on cultivated and/or economically important species, whereas studies of native species focused on increasing knowledge of species with little economic importance. We further classified by subcategories to develop a better understanding of the types of studies for which cpSSRs have been most useful. Subcategories and their definitions are presented in Table 1.

**Quantifying problems with fragment size homoplasy**—An assessment of whether appropriate methods were used to determine and/or correct for size homoplasy was made for each publication examined in the literature review. Publications in which microsatellite loci were assayed by genotyping (i.e., only fragment sizes were determined) were considered to be “uncorrected” for issues of size homoplasy. We defined “corrected” publications as those in which the authors stated that size homoplasy was examined after fragment genotyping or when they used a direct sequencing approach.

We also evaluated issues of fragment size homoplasy and aberrant mutation motifs (i.e., size changes that deviate from the expected stepwise pattern of microsatellite mutation) in an empirical data set of nine chloroplast microsatellite loci developed for five species of *Acmispon* Raf. (Fabaceae) (Wheeler et al., 2012). Species included were *A. argophyllus* (A. Gray) Brouillet, *A. dendroideus* (Greene) Brouillet, *A. microanthus* (Torr. & A. Gray) Brouillet, *A. glaber* (Vogel) Brouillet, and *A. heermanni* (Durand & Hilg.) Brouillet. Data for four loci (ACcp2, ACcp3, ACcp4, and ACcp5) were derived from a set of direct sequences of 452 *Acmispon* individuals (i.e., 216 from *A. argophyllus*, 166 from *A. dendroideus*, 36 from *A. microanthus*, 22 from *A. glaber*, and 12 from *A. heermanni*). All other loci were sized as fragments, with each unique allele per species also sequenced. Genotyping details are described in Wheeler et al. (2012), and sequencing details are described in Wheeler (2013). Sequences of most alleles were examined in two individuals per species; however, in some cases, particularly when an allele is only present in a single individual of a given species, only one individual was sequenced. As this study does not use sequence data from every individual, nor does it include every species in the genus, the frequencies of motif abnormality and size homoplasy found are likely underestimated. The size of each microsatellite region based on sequence data was compared to size determined by fragment genotyping. Loci were considered to exhibit size homoplasy if any single fragment size corresponded to two or more unique arrangements of sequence gaps (Fig. 1B vs. 1C). A locus was considered to have an abnormal motif if one or more individuals showed a detectable size change that deviated from the expected stepwise pattern of microsatellite expansion or contraction (Fig. 1C). Single-nucleotide mutations that did not change the length of the sequence were ignored.

**RESULTS**

**Prevalence and uses of cpSSRs**—Of the 439 papers considered in this review, 405 contained original empirical data from cpSSR loci; four of these used minisatellites rather than
Applications in Plant Sciences 2014 2(12): 1400059
Wheeler et al.—Review of cpSSRs for plant biology studies

http://www.bioone.org/loi/apps

Fig. 1. Three examples of chloroplast sequence regions containing tandem repeats, with indicated size when genotyped as fragments. (A) A segment of genome sequence used to design primers to be amplified in a number of individuals: (1) is the presumed variable region. (B) The sequence of an individual that meets all the assumptions of a chloroplast microsatellite model, with (2) indicating a 5-step expansion of the repeat region. (C) The sequence of an individual that violates these assumptions because there is a size change due to the addition of a nonrepetitive indel, violating the assumed microsatellite motif (3) and expansion of a shorter secondary T-monomerucotide repeat region (4). When genotyped as fragments, (B) and (C) are indistinguishable, showing an example of size homoplasy.

Microsatellites. Seven papers reused cpSSR data from a previous publication. Eight papers were classified as reviews of microsatellites (n = 3) or molecular markers in plants (n = 5). Nineteen papers included whole chloroplast genome sequences or mRNA sequences and identified SSR regions.

The use of cpSSRs has expanded substantially from 1995, when a single study was published, to 56 publications in 2013 (Fig. 2). cpSSRs have been used as informative genetic characters in a wide range of plant groups, including green algae (n = 2), bryophytes (n = 1), lycophytes (n = 1), gymnosperms (n = 86), magnoliids (n = 9), monocots (n = 79), and eudicots (n = 264). These studies include species from 85 families (Fig. 3). Pinaeae is the most represented family, with 80 studies, followed by several economically important angiosperm families: Poaceae (50 studies), Vitaceae (32 studies), Fabaceae (30 studies), and Brassicaceae (20 studies). However, the majority of families are represented by five or fewer studies employing cpSSRs.

Publications utilizing chloroplast microsatellites in native species (69%) far outweigh those applying cpSSR techniques in agricultural species (31%). In agricultural species, cpSSRs have been used extensively to assess genetic diversity in cultivated taxa (36.4%) and to study the origins or relatedness of cultivars to native species (30.3%; Fig. 4). The other subcategories account for the remaining studies, including identification of cultivars (15.9%), primer notes (8.3%), and various other purposes (9.1%). In native species, cpSSRs have been used most frequently in population genetics studies (59.3%), followed
were used in studies of monocot taxa compared to those focused on eudicot taxa (mean = 7.61 loci), although this difference was not significant (\( t_{0.05[304]} = 0.36, P > 0.05 \)). The other groups did not contain a sufficient number of published studies for comparison. An overwhelming majority of studies used published primers to amplify loci, but more than twice as many loci were used when primers were developed for the taxon of interest (mean = 12.5 loci used) compared to studies using published primers to amplify loci (mean = 5.8 loci used). The most

by primer notes or other foci (11.0% each), systematics (9.0%), hybridization (5.7%), reviews (2.7%), and methodology (1.3%; Fig. 5).

Many different numbers of loci were used across studies, but there are no significant trends in number of loci used and taxonomic category or number of loci used and number of studies published by family. The correlation between number of published studies per family and number of loci used is very weak (\( r^2 = 0.065, P > 0.05 \)). On average, slightly more loci (mean = 8.34) were used in studies of monocot taxa compared to those focused on eudicot taxa (mean = 7.61 loci), although this difference was not significant (\( t_{0.05[304]} = 0.36, P > 0.05 \)). The other groups did not contain a sufficient number of published studies for comparison. An overwhelming majority of studies used published primers to amplify loci, but more than twice as many loci were used when primers were developed for the taxon of interest (mean = 12.5 loci used) compared to studies using published primers to amplify loci (mean = 5.8 loci used). The most

by primer notes or other foci (11.0% each), systematics (9.0%), hybridization (5.7%), reviews (2.7%), and methodology (1.3%; Fig. 5).

Many different numbers of loci were used across studies, but there are no significant trends in number of loci used and taxonomic category or number of loci used and number of studies published by family. The correlation between number of published studies per family and number of loci used is very weak (\( r^2 = 0.065, P > 0.05 \)). On average, slightly more loci (mean = 8.34) were used in studies of monocot taxa compared to those focused on eudicot taxa (mean = 7.61 loci), although this difference was not significant (\( t_{0.05[304]} = 0.36, P > 0.05 \)). The other groups did not contain a sufficient number of published studies for comparison. An overwhelming majority of studies used published primers to amplify loci, but more than twice as many loci were used when primers were developed for the taxon of interest (mean = 12.5 loci used) compared to studies using published primers to amplify loci (mean = 5.8 loci used). The most

by primer notes or other foci (11.0% each), systematics (9.0%), hybridization (5.7%), reviews (2.7%), and methodology (1.3%; Fig. 5).

Many different numbers of loci were used across studies, but there are no significant trends in number of loci used and taxonomic category or number of loci used and number of studies published by family. The correlation between number of published studies per family and number of loci used is very weak (\( r^2 = 0.065, P > 0.05 \)). On average, slightly more loci (mean = 8.34) were used in studies of monocot taxa compared to those focused on eudicot taxa (mean = 7.61 loci), although this difference was not significant (\( t_{0.05[304]} = 0.36, P > 0.05 \)). The other groups did not contain a sufficient number of published studies for comparison. An overwhelming majority of studies used published primers to amplify loci, but more than twice as many loci were used when primers were developed for the taxon of interest (mean = 12.5 loci used) compared to studies using published primers to amplify loci (mean = 5.8 loci used). The most
In the survey of nine Acmispon loci, two-thirds showed size homoplasy or deviation from a standard microsatellite mutation motif. Three loci exhibited size homoplasy, two loci had deviations from an assumed motif, and one locus contained both problems. Departures from expected patterns were found both within and between species, although more alleles exhibited these problems between species (Table 3).

**DISCUSSION**

**Use of cpSSRs across plants**—Since 1995, cpSSRs have been widely used in plant biology to address basic and applied questions, and they continue to be a useful genetic tool (Fig. 2). The ease of PCR amplification and the polymorphic nature of cpSSRs have made them readily available markers for characterizing genetic variation when few genomic resources exist for a study system. For example, the majority of families in our data set were represented by five or fewer studies, suggesting that markers of this type can be developed and put into use for new groups without much a priori genetic information.

Studies of species in certain families have benefitted more than other families from the use of cpSSRs. Among gymnosperms, studies of taxa in Pinaceae are strongly represented (Fig. 3). Most of these studies were conducted within Pinus L. (56 references), which may reflect greater interest in this genus, which is more speciose than other genera in this family, or perhaps may indicate greater success across the genus using loci designed from *P. thunbergii* Parl. (Vendramin et al., 1996), the most commonly reported published primers used in Pinaceae. Common foci of studies within Pinaceae include understanding the effects of Quaternary glaciation on species range distributions (e.g., Gugerli et al., 2001; Gómez et al., 2005; Bucci et al., 2007; Rodríguez-Banderas et al., 2009; Godbout et al., 2010), characterizing genetic structure (e.g., Vendramin et al., 1999; Dyer and Sork, 2001; Viard et al., 2001; Nasri et al., 2008; Scalfi et al., 2009; Wang et al., 2013), testing for hybridization (e.g., 

http://www.bioone.org/loi/apps
TABLE 2. Publications in the data set reporting cpSSR loci from whole chloroplast genome sequences.

<table>
<thead>
<tr>
<th>Family</th>
<th>Species</th>
<th>No. of SSR loci</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Araliaceae</td>
<td><em>Eleutherococcus senticosus</em> (Rupr. ex Maxim.) Maxim.</td>
<td>35</td>
<td>Yi et al. (2012)</td>
</tr>
<tr>
<td>Araliaceae</td>
<td><em>Panax schin-seng</em> Nees</td>
<td>18</td>
<td>Kim and Lee (2004)</td>
</tr>
<tr>
<td>Asteraceae</td>
<td><em>Aremesia frigida</em> Willd.</td>
<td>41</td>
<td>Liu et al. (2013)</td>
</tr>
<tr>
<td>Asteraceae</td>
<td><em>Jacobaea vulgaris</em> Gaertn.</td>
<td>34†</td>
<td>Doorduin et al. (2011)</td>
</tr>
<tr>
<td>Cephalotaxaceae</td>
<td><em>Cephalotaxus oliveri</em> Mast.</td>
<td>24</td>
<td>Yi et al. (2013)</td>
</tr>
<tr>
<td>Cupressaceae</td>
<td><em>Cryptomeria japonica</em> (Thunb. ex L. f.) D. Don</td>
<td>9</td>
<td>Hira et al. (2009)</td>
</tr>
<tr>
<td>Fabaceae</td>
<td><em>Prosopis alba</em> Griseb.</td>
<td>36</td>
<td>Torres et al. (2013)</td>
</tr>
<tr>
<td>Hydrocharitaceae</td>
<td><em>Elodea canadensis</em> Michx.</td>
<td>127</td>
<td>Huotari and Korpelainen (2012)*</td>
</tr>
<tr>
<td>Lamiales</td>
<td><em>Oreganum vulgare</em> L.</td>
<td>56</td>
<td>Lukas and Novak (2013)</td>
</tr>
<tr>
<td>Lamiales</td>
<td><em>Salvia millitiorrhiza</em> Bunge</td>
<td>166</td>
<td>Qian et al. (2013)</td>
</tr>
<tr>
<td>Magnoliaceae</td>
<td><em>Magnolia grandiflora</em> L.</td>
<td>218</td>
<td>Li et al. (2013)</td>
</tr>
<tr>
<td>Magnoliaceae</td>
<td><em>Magnolia kwangsiensis</em> Figlar &amp; Noot.</td>
<td>237</td>
<td>Kuang et al. (2011)</td>
</tr>
<tr>
<td>Pedaliaceae</td>
<td><em>Sesamum indicum</em> L.</td>
<td>28</td>
<td>Yi and Kim (2012)</td>
</tr>
<tr>
<td>Poaceae</td>
<td><em>Brachypodium distachyon</em> (L.) P. Beauv.</td>
<td>17</td>
<td>Sonah et al. (2011)</td>
</tr>
<tr>
<td>Ranunculaceae</td>
<td><em>Megaleranthis saniculifolia</em> Ohwi</td>
<td>54</td>
<td>Kim et al. (2009)</td>
</tr>
<tr>
<td>Rosaceae</td>
<td><em>Pyrus pyrifolia</em> (Burm. f.) Nakai</td>
<td>67</td>
<td>Terakami et al. (2012)</td>
</tr>
<tr>
<td>Solanaceae</td>
<td><em>Solanum L. spp.</em></td>
<td>32 across all taxa</td>
<td>Gargano et al. (2012)</td>
</tr>
<tr>
<td>Trochodendraceae</td>
<td><em>Tetracentron sinense</em> Oliv., <em>Trochodendron aralioides</em> Siebold &amp; Zucc.</td>
<td>154 in both taxa</td>
<td>Sun et al. (2013)</td>
</tr>
</tbody>
</table>

*Ten primer sets were tested in a sample of the focal species.
†This study provides a synthesis of cpSSR regions across many monocots for which whole chloroplast genomes have been sequenced.
‡This study provides a synthesis of cpSSR regions across many species of the Asterid clade.

Fady et al., 2003; Epperson et al., 2009), paternity analysis (e.g., Lambeth et al., 2001), and testing taxonomic hypotheses (e.g., Clark et al., 2000; Ledig et al., 2004; Liu et al., 2012).

Within angiosperms, Poaceae, Vitaceae, and Fabaceae are well represented in published studies employing cpSSRs. This finding likely reflects the great economic importance of these families and availability of fully sequenced chloroplast genomes, which enables development of suitable polymorphic loci. Given the size and diversity of Poaceae and Fabaceae, it is not surprising that studies are varied in their focus and include agricultural and natural species. For Poaceae, notable studies exemplifying these varied foci include documenting movement of herbicide resistance from cultivated rice to weedy rice (Busconi et al., 2012); elucidating evolutionary relatedness among polyploids of the genus *Cynodon* Rich., an important group of grasses used for lawns in temperate areas of the world (Gulsen and Ceylan, 2011); quantifying seed movement in comparison to pollen flow to evaluate the drivers of spatial genetic structure in *Anthoxanthum odoratum* L. (Freeland et al., 2012); and dating the origin of sympatry among three fire-adapted *Triodia* species with varying life history strategies (Armstrong, 2011). Within Fabaceae, notable studies using cpSSRs include evaluating current anthropogenic causes vs. historical fragmentation as a reason for the genetic structure in the endangered *Caesalpinia echinata* Lam. (Lira et al., 2003), a comparison of the ethnobotany of *Phaseolus* L. species among Mexican farmers with genetic differentiation of the species (Soleri et al., 2013), and characterization of genetic diversity in landraces of *P. vulgaris* L. throughout Europe (Angioti et al., 2010).

The use of cpSSR loci in taxa of other diverse families, such as Asteraceae and Orchidaceae, though, is surprisingly uncommon. In all studies reporting empirical data for species of Asteraceae, we found that published primer sequences were used, and the average number of loci was 6.4. For Orchidaceae, a comparably large family, relatively few studies have employed cpSSRs, and in those, six or fewer loci on average were used. This finding may reflect the lack of widespread genomic resources for these families and the difficulty in transferring primer sequences across taxonomic levels, as reported by Ebert and Peakall (2009). Given the usefulness of next-generation

TABLE 3. Deviation from a stepwise mutation motif (i.e., “abnormal motif”) and homoplasy of detected size alleles among cpSSR loci examined in five species of *Acmispon*.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Repeat motif</th>
<th>No. of alleles</th>
<th>Abnormal motif</th>
<th>Within speciesb</th>
<th>Between speciesb</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACep1</td>
<td>(A)10(T)10(A)10(T)10</td>
<td>7</td>
<td>No</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>ACep2</td>
<td>(T)10(T)10</td>
<td>6</td>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACep3</td>
<td>(A)14</td>
<td>6</td>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACep4</td>
<td>(T)14</td>
<td>6</td>
<td>Yes</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ACep5</td>
<td>(A)12</td>
<td>4</td>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACep6</td>
<td>(T)12(A)12</td>
<td>4</td>
<td>No</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ACep7</td>
<td>(T)12(A)</td>
<td>7</td>
<td>No</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>ACep8</td>
<td>(T)12</td>
<td>4</td>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACep9</td>
<td>(ATTTATTCTAT)12(AT)12(T)12</td>
<td>6</td>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

aNumber of alleles showing size homoplasy among conspecific samples and across all species examined.
bNumber of alleles that exhibit similar sizes yet distinct sequences between species.
sequencing (NGS) methods for primer design in nonmodel organisms and the increasing accessibility of these methods (Csencsics et al., 2010; Ekbлом and Galindo, 2011; Guichoux et al., 2011), a wider range of families may see more representation in the future.

The taxonomic scale of studies employing cpSSRs has ranged from conspecific populations to intergeneric comparisons, although the greatest applicability of cpSSRs may be at the intraspecific level, as suggested by the high number of studies in the category of diversity and history for agricultural species (Fig. 4) and population genetics for native species (Fig. 5). Other common uses of cpSSR data are in understanding uniparental genetic structure (e.g., seed or pollen dispersal) and in studies of hybridization. For example, Richardson et al. (2002) were able to compare rates of seed dispersal via birds using mitochondrial markers with rates of pollen dispersal via wind using cpSSRs in whitebark pine and to determine past patterns of species range expansion in association with habitat and climatic changes. Marsico et al. (2009) also leveraged uniparentally inherited cpSSRs with nuclear data to contrast differing patterns of seed and pollen dispersal in Quercus garryana Douglas ex Hook. Bucci et al. (1998) elegantly demonstrated unidirectional introgression in the Pinus halepensis Mill. complex. Heuertz et al. (2006) identified shared chloroplast haplotypes among Fraxinus L. species in Europe and inferred that they had hybridized historically when coming in contact in common refugia or upon species range expansion. The use of cpSSRs across native and agricultural species as well as in studies ranging from population genetics to systematics demonstrates that they can be used to study many different types of questions in plant biology. The chloroplast genome has the potential to reveal much more about evolutionary history across plants than currently realized, and studies of many other taxa could benefit from the development of cpSSR markers.

Variability of cpSSRs—In many ways, cpSSRs are ideal markers for addressing a variety of genetic questions in plant biology. Genotypic variation is easily assessed using fragment-based methods, and the data are easy to interpret due to uniparental inheritance of the chloroplast genome in most species. However, polymorphism of cpSSR loci is highly variable across taxa, making the testing of primers a necessary precursor to data collection. We found that when primers were adapted from another species, many fewer loci were used (5.8) compared to studies employing primers developed for the study system (12.5). This was due to loci not amplifying in the focal samples with similar fragment sizes when they actually have underlying sequence. The severity of this problem will vary by species as well as their discontinued use due to lack of polymorphism. Ebert and Peakall (2009) also found a reduction in amplification success and a decrease in polymorphism among amplified loci with increasing phylogenetic distance.

There is substantial variation in repeat regions of the chloroplast genome across plants. In the extensive review by Ebert and Peakall (2009), the number of cpSSRs of at least eight mononucleotide repeats varied from one in the green alga Nephroselmis olivacea F. Stein to more than 700 in the green alga Chlorella vulgaris Beyerinck [Beijerinck]. Within angiosperms, the number of cpSSRs varied from less than 40 to more than 150 (Ebert and Peakall, 2009). We also found extensive variation in the whole chloroplast genome sequences included in this data set (Table 2). Slightly more loci (mean = 8.34) were used in studies of monocot taxa compared to those focused on eudicot taxa (mean = 7.61). If it is assumed that usable loci were also variable loci, then this finding may indicate a greater number of cpSSR loci in the chloroplast of monocots. However, greater taxonomic coverage of monocot chloroplast genomes is needed to evaluate if this difference is real.

Given that not all chloroplast loci are likely to be equally diverse across plants, it is important to consider a more targeted approach to developing cpSSR loci specific to a study system rather than relying on universal primers. The increasing availability of whole plastome sequences is likely to enhance one’s ability to identify suitable loci in nonmodel species, and is becoming a common pathway to developing variable markers (e.g., Wheeler et al., 2012). The whole chloroplast genome sequences identified in this study as well as those from Ebert and Peakall (2009) could be extremely helpful for the development of primers in novel taxa. Ebert and Peakall (2009) provide an excellent technical review on developing markers utilizing these resources. As Ebert and Peakall (2009) and others have suggested, we also recommend developing markers using sequences from the target species or a closely related species to maximize the number of polymorphic markers available. For taxa lacking genomic resources from a related species, the universal primers of Weising and Gardner (1999), Bryan et al. (1999), and Chung and Staub (2003) may be a good option to finding variable markers.

Fragment size homoplasy—All genetic markers that are assayed using fragment size genotyping are subject to erroneous inferences if the sizes do not reflect similarity of the underlying nucleotide sequence. Size homoplasy is more common in repetitive regions because of the higher mutation rate and complex mutational mechanisms compared to nucleotide substitutions and insertions/deletions involving longer stretches of nucleotides (Provan et al., 2001; Vachon and Freeland, 2011). Previous studies have called attention to the potential for and problems with fragment size homoplasy in cpSSRs (Liepert et al., 2001; Provan et al., 2001; Ebert and Peakall, 2009; Vachon and Freeland, 2011), and others have studied homoplasy in nuclear SSRs (e.g., Estoup et al., 2002; Curtu et al., 2004; Barthe et al., 2012). Despite these calls in the literature to consider fragment size homoplasy of cpSSRs, two-thirds of the studies we reviewed did not indicate that the authors tested for homoplasy. Given that size homoplasy at certain cpSSR loci is expected to be prominent, we urge researchers employing cpSSRs to incorporate tests for homoplasy into their experimental design and data analyses.

Homoplasy can lead to over-estimation of relatedness among samples with similar fragment sizes when they actually have different evolutionary histories that may be detectable in the underlying sequence. The severity of this problem will vary by taxonomic and geographic scale, being greatest at higher taxonomic levels (e.g., among comparisons of congeneric species; Provan et al., 2001) and in large-scale analyses, such as phylogeographic studies (Provan et al., 2001; Vachon and Freeland, 2011). Repeated evolution of similar-sized alleles has been documented in interspecific studies employing cpSSRs, including sympatric orchid taxa (Soliva and Widmer, 1999) and members of the Glycine subg. Glycine J. C. Wendl. (Doyle et al., 1998). Erroneous inferences involving comparisons across taxa may be expected to be more common due to cpSSR homoplasy, but intraspecific comparisons involving populations or subspecific taxa are commonly affected by size homoplasy as well. For example, Ebert and Peakall (2009) found size homoplasy among conspecific individuals sampled from a single population of the orchid Chiloglottis valida D. L. Jones. We found

http://www.bioone.org/loi/apps
similar issues in *Acmispon*, both within and across species (Table 3). The greatest differences in underlying sequences when sizes were similar occurred when interspecific comparisons were made.

Undetected size homoplasy can affect estimates of genetic diversity and genetic distance (e.g., Provan et al., 2001; Navascués and Emerson, 2005), give misleading phylogenetic patterns (e.g., Doyle et al., 1998; Hale et al., 2004), and result in ambiguous ancestor-descendant relationships (e.g., Vachon and Freeland, 2011). For example, Navascués et al. (2006) found that homoplasy in cpSSRs affected inference of demographic scenarios, specifically testing for population expansion using the $F_2$ test (Fu, 1997), because it is based on genetic distance and haplotype diversity. Underestimation of haplotype diversity due to homoplasy reduced their power to detect population expansion in a simulated data set. The greatest distortions occurred in cases of recent population growth because the errors in estimates of genetic distance and number of haplotypes are more unbalanced in recent expansions (Navascués et al., 2006).

The use of cpSSRs in phylogenetic analyses of species has been reported to produce reduced resolution and conflicting results when compared to other data sets. For example, Doyle et al. (1998) compared maximum and minimum estimates of the effects of size homoplasy on a phylogeny of species in *Glycine* subg. *Glycine* by mapping the distribution of cpSSR sizes onto a phylogeny based on restriction fragment length polymorphisms (RFLPs; i.e., maximum estimate) and reconstructing the phylogeny using the cpSSR and RFLP data (i.e., minimum estimate). They found that the addition of the cpSSR slightly reduced resolution of the resulting tree and the cpSSR did not reflect patterns of relationship that were suggested by plastome RFLP profiles (Doyle et al., 1998). In a similar study, Hale et al. (2004) compared phylogenies of *Clusia* L. species based on cpSSRs and chloroplast sequence variation and concluded that the cpSSRs were rich with homoplasy and too variable to be useful in reconstructing evolutionary relationships within the genus. In our data set of *Acmispon*, we have found conflicting phylogenies that are based on cpSSRs compared to nucleotide polymorphisms or other indels from the chloroplast genome (Wallace, unpublished data). Inference of ancestor-descendant relationships based on haplotype networks of cpSSR loci may also be complicated by homoplasy (Vachon and Freeland, 2011) because they increase the probability of loop structures in the network (Saltonstall, 2002; Jimenez et al., 2004).

The presence of size homoplasy in cpSSR loci is troublesome, but this does not void the usefulness of these markers. If the true underlying sequence of a microsatellite locus can be determined, then deviations from the standard assumptions can be taken into account. For example, Liepelt et al. (2001) noted that polymerase error can occur at a higher rate in regions containing multiple variable repeat regions. They suggested that these loci be split into two (or more, if necessary) loci by designing internal primers to increase DNA polymerase fidelity; this solution would also remove issues with size homoplasy by insuring that all variation is produced by a single variable region. Navascués and Emerson (2005) recommended the use of Nei’s unbiased haplotype diversity (Nei, 1978) over the measure developed by Goldstein et al. (1995), which assumes a pairwise mutation model and is affected more severely by homoplasy. The use of greater numbers of loci can also mitigate the effects of homoplasy, as chloroplast haplotypes that appear identical at a small number of sites are more likely to be distinguishable with greater marker coverage (Navascués and Emerson, 2005). Sequencing all chloroplast microsatellite alleles, regardless of method of discovery, can ensure that genotyped fragments reflect both actual length and nucleotide variation (Hale et al., 2004; Ebert and Peakall, 2009). However, Vachon and Freeland (2011) suggested that repetitive regions, whether genotyped or sequenced, should be excluded from certain types of large-scale analyses, such as haplotype networks, because of their high potential for repeated evolution. Instead, they advocate for the incorporation of repetitive regions into other types of analyses that take into account relative mutation rates or that use comparisons of population allele frequencies.

Given the potential for fragment size homoplasy to occur between and within species, we recommend that all studies making use of cpSSRs perform some sequencing of alleles to verify the mutational patterns of underlying sequences. Our approach to sequence at least two representative samples for each allele in each species of *Acmispon* provided a reasonable, cost-effective assessment of size homoplasy in this data set. For studies where deep genetic divergence is expected, further testing may be warranted. If one assumes that all individuals with the same sized allele also have the same underlying sequence, then sequencing cpSSRs also provides one with the ability to convert from fragment sizes to nucleotide sequences, thereby opening up additional types of analyses. Analyzing fragment data assuming a stepwise-mutation model or complete independence of repeats can cause problems in detected divergence. However, by converting fragments to sequence using representative samples, data can be analyzed more accurately, taking all sequence changes into consideration appropriately. Although the cost of this approach varies based on the allelic diversity of the sample and the number of species (or taxa) included, it should always be more cost- and time-efficient than comparable studies using direct sequencing for each individual.

Sequencing of repeat regions can be difficult using Sanger sequencing because of slipped-strand mispairing leading to disassociation of the DNA polymerase from the template. Mononucleotide repeats composed of eight or more nucleotides are the most problematic because the active site of *Taq* DNA polymerase is presumed to be approximately the same length (Eom et al., 1996). Given that most cpSSR loci are composed of mononucleotide A/T repeats, generating clean sequences of alleles may be impossible. Proofreading DNA polymerases can improve the accuracy of PCR and the resulting sequences (Fazekas et al., 2010). For example, Guicking et al. (2008) provided a single nucleotide sequence protocol in which a repeat region is sequenced using one nucleotide instead of all four. Finally, regions that have multiple repeat regions can be broken up and sequenced separately (Liepelt et al., 2001).

Next-generation sequencing methodologies are increasingly used in the development of cpSSR loci (e.g., Doorduin et al., 2011; Kuang et al., 2011) and may provide an alternative to Sanger sequencing for genotyping cpSSRs (Guichoux et al., 2011) in nonmodel organisms (McPherson et al., 2013). NGS approaches experience a different set of issues when facing stretches of repetitive nucleotide sequences. NGS data are produced in the form of many short DNA sequence reads that must be assembled using bioinformatics software (Miller et al., 2010). Sequencing platforms employing shorter reads may pose difficulty in the accurate assembly of repetitive regions because it may be impossible to place reads that lack flanking sequences, which should be distinct for a locus, in a single location with 100% certainty (Treangen and Salzberg, 2012). Given that the majority of microsatellites in the chloroplast genome...
are mononucleotide A/T repeats, divergent flanking regions are especially important for correct assembly. Fortunately, chloroplast microsatellites are often quite short in length, and NGS methods are available that use reads long enough to capture repeat tracts as well as flanking sequences in their entirety. Finally, while polymerase error can potentially occur with NGS methods, computational assembly protocols are designed to handle this issue if read depth is sufficient (Miller et al., 2010; Treangen and Salzberg, 2012).

**Cytoplasmic introgression**—An additional drawback to consider with the use of cpSSRs is the potential for misleading patterns due to cytoplasmic introgression. As more studies have incorporated nuclear and chloroplast markers into the development of phylogenies, many of these have revealed discordant patterns between data sets, which is attributed to cytoplasmic introgression between samples even when there is little morphological or ecological evidence of hybridization. Such a conclusion is supported by Currat et al. (2008), who demonstrated a high likelihood that cytoplasmic introgression will be retained within a lineage from just a single hybridization event. Because the chloroplast genome is uniparentally inherited and nonrecombining in most plants, all chloroplast regions should exhibit introgression when it occurs. As a result, introgression could mask true evolutionary patterns of divergence. On the other hand, cpSSRs can be very helpful when there is a hypothesis of hybridization between species or populations. In our data set, these markers have been used to identify the direction of hybridization in natural populations and in identifying the parents of cultivated species. Because cpSSRs tend to be used more commonly than above the species level and multiple samples are typically examined in such studies, cytoplasmic introgression may be readily detected and accounted for in these studies. However, studies employing cpSSRs above the species level and with the use of few samples per species should exercise caution when making inferences about evolutionary history.

**Heteroplasmy**—Uniparental mode of inheritance is considered a strength of chloroplast markers because the genes involved should not exhibit recombination or heterozygosity. However, some studies have identified biparental inheritance of chloroplasts leading to heteroplasmy (Lee et al., 1988; Johnson and Palmer, 1989; Chat et al., 2002; Frey et al., 2005; Hansen et al., 2007). If heteroplasmic individuals contain multiple variable chloroplast genomes, then genotyping with any chloroplast molecular marker will be problematic. If multiple haplotypes are present in sufficient numbers in the sampled tissue, then heteroplasmy is likely to be detectable in sequence chromatograms via the presence of double peaks or in the presence of more than one allele if genotyping is used. However, if copy number of the individual genomes varies, then it is possible that only the most frequent type will be selected in PCR and sequenced. In this case, a researcher may never know heteroplasmy exists in the study system, and cloning is needed to identify the presence of multiple chloroplast genomes. Given that an estimated one-third of surveyed angiosperms have at least occasional biparental inheritance of plastid DNA (Smith, 1989) and new instances continue to be reported (e.g., Hansen et al., 2007), testing for heteroplasmy prior to widespread use of chloroplast markers may be warranted.

### CONCLUSIONS

This review demonstrates the continued widespread use of cpSSRs in plant biology. These markers have been most used for studies of population genetic structure in native species and in characterizing diversity and the origins of cultivated species. Even in the face of genomic approaches commonly available today, cpSSRs continue to be routinely used. This review demonstrates that researchers who have used primers developed in closely related species have had greater success in finding polymorphic loci. Thus, we believe future studies will benefit from a targeted approach using species-specific primers, rather than adapting published universal primers. The abundance of chloroplast genome sequences should facilitate primer design for many taxa. Impediments to the use of cpSSRs include fragment size homoplasy, introgression, and heteroplasmy. Fragment size homoplasy can be overcome by careful use of these markers, such as targeted sequencing of unique alleles to evaluate potential for homoplasy. We find that problems of size homoplasy are likely to be greatest in studies using cpSSRs for comparisons between species. Tests of homoplasy should be a standard component of all studies employing cpSSRs, regardless of the taxonomic level. Introgression and heteroplasmy can be tested via comparisons of alleles with closely related species and between parent and offspring, respectively. Given the greater prevalence of cpSSRs for intraspecific studies reported in the literature and the lower propensity for size homoplasy, their greatest utility is likely to be below the species level.

### LITERATURE CITED


http://www.bioone.org/loi/appps