Complete plastome sequences from Glycine syndetika, and six additional perennial wild relatives of soybean.

Sue Sherman-Broyles*, Aureliano Bombarely*, Jane Grimwood §, Jeremy Schmutz §, and Jeff Doyle*

*Cornell University, Department of Plant Biology, Ithaca, NY 14853
§ HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806

GenBank PopSet: 514252878 KC893632-KC893640
ABSTRACT

Organelle sequences have a long history of utility in phylogenetic analyses. Chloroplast sequences when combined with nuclear data can help resolve relationships among flowering plant genera, and within genera incongruence can point to reticulate evolution. Plastome sequences are becoming plentiful because they are increasingly easier to obtain. Complete plastome sequences allow us to detect rare rearrangements and test the tempo of sequence evolution. Chloroplast sequences are generally considered a nuisance to be kept to a minimum in bacterial artificial chromosome libraries. Here we sequenced two bacterial artificial chromosomes per species to generate complete plastome sequences from seven species. The plastome sequences from Glycine syndetika and six other perennial Glycine species are similar in arrangement and gene content to the previously published soybean plastome. Repetitive sequences were detected in high frequencies as in soybean but further analysis showed that repeat sequence numbers are inflated. Previous chloroplast based phylogenetic trees for perennial Glycine were incongruent with nuclear gene based
phylogenetic trees. We tested whether the hypothesis of introgression was supported by the complete plastomes. Alignment of complete plastome sequences and Bayesian analysis allowed us to date putative hybridization events supporting the hypothesis of introgression and chloroplast “capture”.

INTRODUCTION

The flowering plant (angiosperm) chloroplast genome (plastome) has a wide range of uses in plant biology. Much of this is due to its behavior as a single phylogenetic unit of tightly linked genes, typically comprising 125 and 160 kilobases (kb). In the majority of angiosperms, the plastome is uniparentally, generally maternally, inherited (Corriveau and Coleman 1988). Plastome sequences are useful for species identification (Nock et al. 2011; Kane et al. 2012) and biotechnology applications (Sabir et al. 2014; Ruhlman and Jansen 2014), and have a long history of utility for phylogenetic inference (Palmer et al. 1988b; Jansen et al. 2007; Ruhfel et al. 2014). Plastome sequence data combined with genes from the mitochondrion and nucleus can provide the ability to resolve ancient, higher order relationships among genera that have remained unresolved until adequate data became available (Soltis et al. 2011). At the genus level and below, uniparentally inherited plastome data can indicate reticulate evolution when incongruence between plastome derived trees do not agree with other data such as morphology or nuclear loci (Rieseberg and Soltis 1991).

Angiosperm chloroplast sequences generally evolve more slowly than angiosperm nuclear sequences (Palmer 1990). Organelle size, gene content and gene order have remained quite conserved among land plants as compared to mitochondrial
genome sequences. The typical chloroplast genome, comprising around 150 kb, is composed of four regions, a large and small single copy region (LSC and SSC, respectively), separated by a pair of large inverted repeats (IR). Rearrangements in chloroplast gene order are generally found in taxa that have one of the following qualities: changes in the size of the IRs or complete loss of one copy of the repeat; a high frequency of small dispersed repeats; biparental chloroplast inheritance; or complete or near-complete absence of photosynthesis (Wicke et al. 2011). Although inversions are not common in angiosperm chloroplasts, there are several families that are known for having a number of chloroplast genome inversions: Geraniaceae (Weng et al. 2014; Guisinger et al. 2011), Onagraceae (Greiner et al. 2008), Campanulaceae (Haberle et al. 2008), Asteraceae (Kim et al. 2005), and Leguminosae (Cai et al. 2008).

The Leguminosae or legume family is the third largest land plant family. It has been divided into three subfamilies, only two of which, Mimosideae and Papilionoideae, are monophyletic (LPWG: et al. 2013). Research interests centered on the papilionoids are driven by the wide array of grain, oil and forage legumes in the subfamily, encompassing at least 24 economically important genera from peanuts to beans (Cannon et al. 2009). Prior to the current study, there were seventeen complete plastome sequences published from fourteen papilionoid legume genera (Figure 1). One of the largest clades of the Papilionoideae is characterized by the loss of one copy of the IR, and is thus called the “Inverted Repeat Loss (or Lacking) Clade” (Wojciechowski et al. 2000). One genistoid chloroplast genome has recently been published, Lupinus luteus, European yellow lupine (Martin et al. 2014). Chloroplast genomes from species representing eight genera in the IRLC have been sequenced:
Trifolium aureum, T. grandiflorum, T. repens and T. subterraneum (clover), Medicago truncatula (barrel medic), Cicer arietinum (chickpea), Pisum sativum (pea), Lathyrus sativus (grass pea), Glycyrrhiza glabra (licorice), Lens culinaris (lentil) and Vicia faba (broad bean) (Sabir et al. 2014; Saski et al. 2005; Cai et al. 2008; Jansen et al. 2008; Magee et al. 2010). The IRLC is sister to the clade that includes the model legume Lotus, which retains both IRs (Kato et al. 2000); together these clades comprise the Hologalegina (Figure 1). Sister to Hologalegina is the millettioid clade (Wojciechowski et al. 2004), from which complete chloroplast genome sequences have been obtained from Milletia pinnata (= Pongamia pinnata, pongam oiltree), as well as three phaseoloid genera, Vigna radiata (mung bean), Phaseolus vulgaris (common bean) and Glycine max (soybean) (Saski et al. 2005; Tangphatsornruang et al. 2010; Guo et al. 2007; Kazakoff et al. 2012).

In the legume family, a 51 kb inversion is shared by most members of subfamily Papilionoideae (Doyle et al. 1996; Doyle et al. 1990b; Palmer et al. 1988a). This inversion is present in Glycine chloroplast genomes and occurs in the LSC region, changing the gene order between trnK and accD. Three additional inversions have been reported, a newly described 36 kb inversion within the 51 kb inversion is present in Lupinus and other genistoids (Martin et al. 2014), a 78 kb inversion shared by several closely related genera including Phaseolus and Vigna (Guo et al. 2007; Tangphatsornruang et al. 2010; Bruneau et al. 1990), and a 5.6kb inversion reported in Milletia (Kazakoff et al. 2012). The 78 kb inversion spans almost the entire LSC between trnH and rps19, returning the genes in the 51 kb inversion to the order found in most land plants (Guo et al. 2007). This change in gene order has also been attributed
to the expansion and contraction of the IR, leaving the gene order as described in papilionoids with the 51 kb inversion but changing the genes bordering the IR (Tangphatsornruang et al. 2010; Perry et al. 2002).

Legume chloroplasts also show variation in the presence and absence of genes. All legumes are missing two chloroplast encoded genes, infA and rpl22 (Doyle et al. 1995), both have nuclear copies targeted to the chloroplast (Gantt et al. 1991; Millen et al. 2001). Loss of rps16 from the chloroplast has been reported in a number of legume lineages excluding Glycine. The mitochondrial copy is dual targeted to both the mitochondria and chloroplast (Jansen et al. 2008; Sabir et al. 2014)). Intron losses in clpP and rps12, found in IRLC lineage (Jansen et al. 2008) are not detected in Glycine (Saski et al. 2005).

The genus Glycine includes at least 28 species divided into two subgenera. The annuals include G. soja and the cultivated soybean, G. max, which are native to eastern Asia, whereas the majority of species are perennials found in Australia. Early investigations grouped Glycine species into “genome groups” (designated by letters, A-I) based on the fertility of artificially produced hybrids and the degree to which meiotic chromosomes paired (Singh and Hymowitz 1985). These data, combined with isozyme data and sequences of two nuclear loci (the nuclear ribosomal gene cistron internal transcribed spacer; nrDNA ITS, and the low copy gene histone H3D), have been used to delimit nine genome groups as reviewed by Ratnaparkhe (Ratnaparkhe et al. 2011).

Chloroplast data from annual and perennial Glycine species have been used in genetic diversity studies and in phylogenetic studies (Doyle et al. 1990b; Doyle et al. 1990c; Sakai et al. 2003; Xu et al. 2002; Xu et al. 2001; Xu et al. 2000) including
investigations of neopolyploidy in perennial taxa (Doyle et al. 2004a; Doyle et al. 1990a). For the perennial subgenus as a whole, Doyle et al. (1990b) identified three major clades, termed “plastome groups,” which showed varying degrees of agreement with nuclear genome groups. The B- and C-plastomes were found to characterize species of the B- and C-genome groups, respectively, though later work revealed incongruence between nuclear and chloroplast phylogenies within the B-genome group itself (Doyle et al. 1999). The A-plastome group included chloroplast genomes from all of the remaining species in the subgenus. The majority of these species belonged to the histone H3D clade comprising the A-, D-, E-, H-, and I-genomes, a result broadly congruent between chloroplast and nuclear data; however, subsequent studies suggested little agreement between genome groups and groupings within the A-plastome clade (Rauscher, J.T., A.H.D. Brown and J.J. Doyle, unpublished). The most obvious disagreement between nuclear and chloroplast phylogenies was the placement of *G. falcata*, the sole species of the F-genome, which in the histone H3D phylogeny was sister to all other perennials, but in the chloroplast phylogeny was strongly supported as part of the A-plastome clade. A sister relationship between *G. falcata* and other perennial *Glycine* agrees with the distinctiveness of this species (Doyle et al. 1996; Doyle et al. 2004a).

In this study we describe complete plastome sequences from seven perennial *Glycine* species, and compare these results to *G. max* (G-genome). We describe the *G. syndetika* plastome in detail as a representative genome of the perennial *Glycine* species. *Glycine syndetika* is currently the perennial species with the most extensive nuclear sequence available (Bombarely, A., J. Schmutz, J. Grimwood, S.A. Jackson...
and J.J. Doyle, unpublished). The perennial *Glycine* chloroplast genomes are compared to the annual soybean, *G. max* (Saski *et al.* 2005) and *Phaseolus vulgaris* (Guo *et al.* 2007) as well as other closely related legumes, *Milletia pinnata* and *Vigna radiata* (Kazakoff *et al.* 2012; Tangphatsornruang *et al.* 2010).

**MATERIALS AND METHODS**

Five plants of a single accession each from seven perennial *Glycine* species representing five of the nine genome groups (Table 1) were grown in the greenhouse at Cornell. Plants were placed in the dark for 48 hours prior to collecting leaf tissue. Tissue was frozen and shipped to Arizona Genomics Institute (AGI) where Bacterial Artificial Chromosome (BAC) libraries were prepared from *HindIII* digests of genomic DNA. BAC ends were sequenced for all BACs. Two BACs per species were selected from the BAC libraries based on BLAST matches of perennial BAC END sequence data to the *G. max* chloroplast sequence (Supplemental Table 1). BAC DNA was sheared to 3-4kb using Adaptive Focused Acoustics technology (Covaris, Woburn, MA, USA) cloned into the plasmid vector pIK96, sequenced with Sanger technology to an average depth of 9x and then assembled (using Phrap Version 0.990319) and finished as previously described (Ferris *et al.* 2010). Additional BACs were selected as needed to close the chloroplast circles based on BLAST to the finished sequences.

**DOGMA** (Wyman *et al.* 2004) was used with default parameters for preliminary annotation of each sequence. The complete chloroplast sequences were arranged so that each sequence started with *trnH*. The SSC regions were arranged so that the complete *ycf1* gene followed IRb, as in the *G. max* orientation depicted in Saski *et al.*
2005. Sequences were then aligned using default parameters in Mulan (Ovcharenko et al. 2005) with minor adjustments to the alignment made manually. *Glycine max* gene features were downloaded from NCBI (DQ317523) and aligned against the eight sequence alignment using Sequencher (Gene Codes). BioEdit (Hall 2011) was used for manual alignment and confirmation of annotations. Analysis of the total chloroplast alignment for GC content and codon bias were performed using DNAsp (Librado and Rozas 2009).

Previously published primers were used for amplification of *trnL-trnF* (Shaw et al. 2005) from 11 *G. falcata* accessions. The resulting sequences were aligned by MUSCLE (Edgar 2004) and edited in BioEdit (Hall 1999). This alignment was used to determine pairwise nucleotide diversity (\(\pi\)) between individuals. Levels of nucleotide polymorphism were calculated using DNAsp (Librado and Rozas 2009).

Direct and palindromic repeated sequences from each chloroplast sequence were identified using REPuter (Kurtz et al. 2001). The number of repeats identified was limited by searching for repeats greater than 30 bp in length and with a sequence identity of 90\% or better (Hamming distance of 3 as used previously by Saski et al. 2005). To further investigate dispersed repeat sequences, we extracted the sequence of each repeat separated by more than 1000 bp. Dispersed repeat sequences were aligned using SeqMan (DNASTAR, Madison WI, Version 2.2.0.56). In many cases, repeats listed separately in the REPuter output were from the same location, having the same sequence, and the only difference was the length; for example, a 36 bp repeat could also be listed for the same position as a 30 bp repeat. Similarly, repeats were listed as pairs, and if the repeat was found in a third or fourth location, the original
location was listed again with the third location, and again with the fourth location hence, the overall number of repeats was inflated. Repeat analysis was also performed using RepeatScout and Repeat Masker (Price et al. 2005; Smit et al. 1996) using default parameters in the MAKER 2.10 (Cantarel et al. 2008) genome annotation software.

A Bayesian approach was used to generate a phylogenetic tree for the chloroplast genome sequences. The phylogenetic tree was generated with BEAST (Drummond et al. 2012), using a prior assumption that Glycine perennials are a monophyletic group compared with the annual Glycine species (G. max), and that Phaseolus vulgaris is an outgroup to the Glycine genus. The node date of 19.2 mya for P. vulgaris and Glycine in the plastome tree was used for calibration, and is based on the mean age estimated from matK sequence divergence in a comprehensive legume phylogeny (Lavin et al. 2005). The substitution model was General Time Reversible (GTR). The analysis was run for a MCMC length of 10,000,000 iterations, sampling every 1000 iterations. The molecular clock was an uncorrelated relaxed clock with a log normal distribution model (Drummond et al. 2006). FigTree was used to plot the tree (http://tree.bio.ed.ac.uk/software/figtree/).

Percent identity across the chloroplast genomes, excluding the second IR, was visualized using the VISTA tools server (Brudno et al. 2003). The LAGAN shuffle option was used to facilitate inclusion of chloroplast sequences from GenBank for species with inversions not present in Glycine (Phaseolus vulgaris var. Negro Jamapa (DQ886273), Millettia pinnata (JN673818) and Vigna radiata (GQ893027). Pairwise differences in single nucleotides were recorded for each species in an alignment of Glycine species
and *Phaseolus vulgaris* plastomes. Pairwise differences were used to calculate the nucleotide substitution rate by dividing by the length of the alignment and the divergence dates calculated with histone H3D alignments (González-Orozco et al. 2012; Sherman-Broyles et al. In Press).

**RESULTS and DISCUSSION**

**Sequencing, size, gene content and extent of inverted repeat regions:** The two *G. falcata* BAC sequences used to create a complete chloroplast sequence were polymorphic. No other species had polymorphisms between the two BAC sequences. BAC libraries for all taxa were created by collecting tissue from up to five plants in each accession. The region of overlap between the two BACs was 86,523 bp. There were 72 segregating sites in this region, resulting in pairwise nucleotide diversity or $\pi = 0.00083$. This very low level of diversity rules out contamination with chloroplast DNAs from other *Glycine* species studied here. To determine levels of chloroplast genome polymorphism in *G. falcata*, we sequenced a non-coding region of the plastome known to be variable in *Glycine* species, the intergenic spacer between *trnL* and *trnF*, in 11 additional accessions of *G. falcata* (Sherman-Broyles, J.A. Doyle and J.J. Doyle, unpublished data). We found that this sample was an order of magnitude more diverse ($\pi = 0.00225$) than the estimate from the two BACs, consistent with the two BACs representing variation among individuals in accession G1718. This could be due to polymorphisms in the chloroplast genome, or to one BAC representing the chloroplast genome and the other being derived from a recent, large nuclear integrant of plastid DNA (NuPT).

Similar types of variation were found in the cultivated plant *Pelargonium X hortorum* (Chumley et al. 2006). CHUMLEY et al. (2006) pointed out that *Pelargonium* has
biparental inheritance of chloroplasts that may contribute to the detection of variation between individuals, but also commented that they found no evidence of heteroplasmy. In *Glycine*, maternal inheritance of chloroplasts was confirmed in *G. max* (Hatfield *et al.* 1985) and in the perennials, determined by Southern blot analysis of two synthetic allopolyplloid taxa (Doyle *et al.* 1990b), so it seems more likely that there are differences between individuals rather than within a single plant. Further analysis of nuclear genome sequences is required to determine if a NuPT was sequenced (Yoshida *et al.* 2013; Michalovova *et al.* 2013).

The *G. syndetika* genome map (Figure 2) is representative of the seven perennial *Glycine* species’ chloroplast sequences discussed here. No structural rearrangements were detected among the perennial *Glycine* plastomes. The *G. syndetika* plastome sequence is 152,794 base pairs (bp) in length, just slightly larger than *G. max* (152,218 bp: SASKI *et al.* 2005; Table 1). The chloroplast genome length in *Glycine* ranges from *G. max*, with the smallest genome, to *G. falcata* at 153,023 bp with the largest genome (Table 1).

*Glycine* plastomes contain 111 unique genes, including 77 protein coding genes, 30 tRNA genes, and 4 rRNA genes. Six protein coding genes, seven tRNA genes and all of the rRNA genes are completely duplicated in the IRs. The perennial *Glycine* plastomes are similar to the annual soybean plastome, with nineteen genes containing introns, of which six are tRNA genes. Two genes have alternative start codons (*psbL* and *ndhD*) in all of the *Glycine* chloroplast sequences as found in other angiosperm plastomes (Sugiura 2013). The GC content in *G. max* is 34% (Saski *et al.* 2005) whereas in the perennial taxa the GC content of all species is 35% (Table 1). Codon
bias is consistent with the A/T-rich aspect of angiosperm chloroplast genomes, codons with A/T(U) in the third position being more common.

The total number of genes reported above does not include *ycf4*, which in legumes is divergent from other angiosperm *ycf4* sequences (Stefanovic et al. 2009) and was not annotated by Dogma (Wyman et al. 2004) or GenBank for our submissions. Stefanovic et al. (2009) were the first to draw attention to the fact that the gene was present and highly divergent in legumes (Stefanovic et al. 2009) rather than absent as suggested by DNA hybridization surveys (Doyle et al. 1995) and the *G. max* annotation (Saski et al. 2005). This gene is thought to be a nonessential Photosystem I assembly factor in higher plants (Krech et al. 2012), and this level of sequence divergence may indicate that if its function is retained in *Glycine*, it is replaced by a nuclear copy of a plastid gene (NuPT). This region flanks one of several NuPTs found in the soybean genome sequence (Bombarely, A., D. Robinson, S. Sherman-Broyles and J.J. Doyle, unpublished data). Although *ycf4* is divergent from other legume *ycf4* sequences, there is no indication that it is a mutational hotspot in *Glycine*, as it is known to be in the legume genus *Lathyrus* (Magee et al. 2010). There are high levels of sequence similarity in genes between *rps16* and *cemA*, the hypervariable region in Magee et al. (Figure 3). Pairwise Ka/Ks values for all *Glycine* species (data not shown) are misleading because there are so few differences that ratios were based on too few polymorphic sites.

The genes that mark the beginning and end of the IR are only partially duplicated: 1) 61 bp of *rps19* for all species except *G. falcata* (59 bp); 2) hypothetical chloroplast RF1 (*ycf1*), with 478 bp duplicated in *G. max*, 490 bp in *G. dolichocarpa*,

Sherman-Broyles *et al.* 13
and 463 bp in the remaining perennial *Glycine* taxa have (Table 1). The full length *ycf1* sequence in *G. cyrtoloba*, spanning the IRb and SSC region, has a premature stop codon in the single copy region; further analyses are necessary to determine if this is a sequencing error, an edited base, or if *ycf1* is a pseudogene in *G. cyrtoloba*. The function of *ycf1* was recently determined to be a translocon protein of the inner chloroplast membrane and in *Arabidiposis* was renamed Tic214 (Kikuchi et al. 2013). It has been found to have relaxed substitution rates except at its 5’ end, which may coincide with an origin of replication but might also be explained by its duplication in the IR regions, which have lower substitution rates than other chloroplast regions (Wicke et al. 2011).

**Repetitive Sequences:** Repeat sequences are of interest as a possible mechanism for inversions or as remnants of the inversion process and thus are frequently reported in descriptions of legume plastomes and other taxa that are known to be characterized by one or more inversions.

RePuter detected 104 tandem and dispersed repeats of 30 bp in length or longer in *G. max* (as reported in SASKI et al. 2005); similarly in *G. syndetika* there are 103 repeats. RePuter detected the fewest (85) repeats from the *G. falcata* plastome (Table 2). The number of repeats detected in *Glycine* plastomes is much higher than the number of repeats detected previously in *Arabidopsis* (57 repeats, Table 2 and SASKI et al. 2005). To gain a more thorough understanding of the RePuter output we investigated the sequence of each repeat identified. If repeat sequences were within 1000 bp of each other, we considered them tandem repeats. If dispersed repeats occurred only in the IR regions, we did not consider them further, but if they were
repeated both within and outside the IR regions, the repeat sequences were
investigated further. An alignment of all the repeats identified by REPuter and separated
by more than 1000 bp showed that *Glycine* dispersed repeats with high levels of
sequence similarity are found in multiple locations in the chloroplast genome. This
reduced the number of unique sequences to 10 or fewer in each species and allowed us
to plot their locations (Figure 4). For example, Repeat sequence 1,
TATATATCTATMTATMTATAGATAGATATATAGATAT, is the most common repeat
sequence and it was found in up to eleven positions in a single plastome (represented
by light green squares in Figure 4). The nine remaining repeat sequences are found
less frequently (four or fewer positions, depicted by dark green squares in Figure 4.)
Repeats present within a coding region are depicted by an asterisk.

Unlike the Geraniaceae (Guisinger et al. 2011) and *Oenothera* (Onagraceae)
(Greiner et al. 2008), where inversions are often flanked by tandem or palindromic
repeats, in soybean (Saski et al. 2005) or in the plastomes of any of the other *Glycine*
species reported here, there is not a higher percentage of repeats in the area flanking
the 51 kb inversion (Figure 4). The 51 kb inversion shared by most papilionoid species
dates back about 56 million years ago (mya) (Lavin et al. 2005). The lack of repeats
flanking the inversion in modern genomes sequences is not surprising considering the
age of the inversions.

The size of repeats detected in the *Glycine* plastomes are mostly less than 50
bp, but within the range necessary for illegitimate recombination. Microhomologies as
little as 16 bp in length have been shown to be sufficient for illegitimate recombination
while regions of sequence similarity longer than 50 bp are necessary for homologous
recombination (Maréchal and Brisson 2010). The higher incidence of repeated sequences in legume plastomes and several other taxa with inversions suggests that there may have been mutations in chloroplast DNA repair genes; this could be tested by examining candidate genes such as genes recently described as responsible for organelle stability through recombination surveillance (Maréchal and Brisson 2010).

RepeatMasker and RepeatScout (Smit et al. 1996; Price et al. 2005) detected 101 repetitive sequences in *G. syndetika*, (Table 2) including short sequences with similarity to transposable elements, low complexity repeats and simple sequence repeats (SSRs). The locations of low complexity and SSRs are also shown in Figure 4. Comparison of *G. syndetika* to *G. dolichocarpa* indicates that the repetitive sequence profiles are identical except for a 76 bp region in *G. syndetika* with sequence similarity to hAT DNA transposon (Figure 4). All sequences with similarity to transposable elements are shorter than 506 bp and probably do not represent true transposable element insertions but rather sequences that are coincidentally similar to transposable elements and therefore detected by the software.

**Evolution:** The alignment of the *Glycine* sequences was deposited in GenBank as Popset 514252878. Distances estimated from this alignment reveal that all species of *Glycine* outside of the B-, C-, and G-genome groups have highly similar chloroplast genome sequences, consistent with their previous recognition as the A-plastome group (Doyle et al. 1990b, Table 1). A Bayesian inference tree (Figure S1), calibrated with *Phaseolus vulgaris* (Guo et al. 2007) as an outgroup, has a topology that is congruent with previous *Glycine* chloroplast gene trees (Doyle et al. 1990b; Sakai et al. 2003). All chloroplast-based trees are incongruent with topologies based on nuclear DNA
sequences (Doyle et al. 2004b), including histone H3D (Figure 5; Doyle et al. 1996b, Gonzalez-Orozco et al. 2012), nrDNA ITS (Rauscher et al. 2004), and genome-wide single nucleotide polymorphisms (ILUT, D., P. Cregan and J.J. Doyle, unpublished). The most conspicuous incongruence involves the placement of G. falcata as sister to A-genome and D-genome taxa in the chloroplast phylogeny rather than sister to all other perennial taxa as indicated by nuclear data (Figure 5).

The histone tree lacks a P. vulgaris sequence because although this locus is useful for nuclear phylogenies, it is mostly intron sequence and the P. vulgaris sequence is too diverged to align properly. The histone tree is instead calibrated using a mean divergence date between annual and perennial Glycine subgenera of 5.25 mya, an average estimated from the divergence dates of several low copy nuclear genes from G. max and G. tomentella D3 (Egan and Doyle 2010; Innes et al. 2008). That date is similar to the 6.17 my estimated from the BEAST plastome phylogeny, and is well within the credible interval for that node (Figure 5A). These estimates of relative divergence times rather than absolute node ages, are dependent on calibration using estimates from other studies and not on dates from fossils.

The plastome and nuclear gene trees differ in the depiction of the relationship between B and C genome groups. The plastome tree shows G. stenophita and G. cyrtoloba as sister to each other, sharing a common ancestor 2.61 my, while the histone tree (Figure 5B) is similar to previously published trees that vary in the placement of the B and C genomes relative to each other depending on taxon sampling and the locus that was used. Regardless of the data set used, the B and C genome groups are always derived after the F genome group and before the remaining genome groups.
The presence of a B-C clade is consistent with genomewide SNP data (ILUT, D., P. Cregan and J.J. Doyle, unpublished).  

The strong similarities among A-plastomes lead to very recent divergence time estimates: less than 500,000 years to the common ancestor of the *G. canescens* and *G. syndetika* plastomes, whereas their histone H3D genes are estimated to have diverged 1.28 mya. Similarly, the date for the divergence of the *G. tomentella* D3 plastome from the plastomes of these A-genome species is under 1 mya, whereas the histone divergence date is 2.84 mya. As expected, the most dramatic difference involves *G. falcata*. As sister to the remainder of the perennial subgenus in the histone H3D tree it is estimated to last shared a common ancestor with the other species approximately 4.6 mya, yet its chloroplast genome is estimated to have diverged from those of the other A-plastome species only 1.64 mya (Figure 5).

DOYLE et al. (1996, 2004b) hypothesized that the plastome from the common ancestor of the entire A-plastome group was introgressed into *G. falcata*, through a hybridization event and subsequent backcrossing to *G. falcata*. The plastome phylogeny divergence dates suggest that the introgression occurred as recently as 1.6 mya. The transfer of a maternally inherited chloroplast replaced the *G. falcata* plastome, in what is called a chloroplast capture event (Tsitrone et al. 2003; Baack and Rieseberg 2007). The current range of *G. falcata* overlaps with both A and D genome species however little is known about the historical ranges of the taxa (Hwang, E.Y., and P. Cregan, in prep.). Based on the sample of 11 *G. falcata* accessions surveyed here for the *trnL-trnF* spacer region, there is no evidence of polymorphism involving a second, deeply coalescing chloroplast genome in this species.
In the subsequent generations, with back crossing to *G. falcata*, the nuclear sequences in *G. falcata* have eliminated all but a few genes that demarcate the introgression event. Analysis of genome-wide SNP data from transcriptome libraries of the same taxa detected SNPs from only 80 genes out of over 27,000 genes (< 0.3%), consistent with introgression from the A-genome into *G. falcata* (ILUT, D., P. Cregan and J.J. Doyle, unpublished). Even relatively recent introgression events have been reported to result in only a very low number of retained genes from the non-backcross parent, perhaps because only genes that are selectively advantageous are retained (McKinnon *et al.* 2010). To further test the hypothesis of introgression from the common ancestor of A/D-genome ancestors into *G. falcata*, we are interested in employing genotyping by sequencing, an inexpensive method for generating genome-wide markers from many individuals (Twyford and Ennos 2012).

The plastome tree has low posterior probabilities for the placement of *G. canescens*, *G. syndetika* and *G. dolichocarpa*. The depicted tree shows the plastomes of *G. syndetika* and *G. canescens* diverging 0.43 mya and that of *G. dolichocarpa* as slightly older at 0.50 mya. The allopolyploid, *G. dolichocarpa*, is polymorphic for chloroplast sequences from its two diploid progenitors, and this result supports previous work (Rauscher, J.T., A.H.D. Brown and J.J. Doyle, unpublished) in showing that the accession sampled here has the *G. syndetika* plastome. A divergence date of approximately 0.5 my is consistent with results from nuclear genes sampled from transcriptome data (Bombarely *et al.* 2014). Hybridization events, resulting in sterile diploid offspring and fertile allopolyploids following whole genome duplication, between taxa that had diverged as many as 3 mya are common in *Glycine*, a prime example.
being *G. dolichocarpa*. The allopolyploid taxa lend credence to the hypothesis that a hybridization event, 1 to 3 mya, led to the capture of an A-plastome within what is probably the most earliest diverging perennial *Glycine* species, *G. falcata*.

Most differences between *Glycine* plastomes involve a handful of insertion/deletion polymorphisms, and A-plastome sequences are virtually identical to one another, with low levels of genetic diversity among taxa (Figure 3). The number of pairwise nucleotide substitutions (Table 3) can be used to estimate the plastome substitution rates in the *Glycine* taxa, if it is assumed that the annuals and perennials diverged about 5.25 mya (Egan and Doyle 2010; Innes et al. 2008) and we use the other divergence dates estimated from the histone H3D phylogeny (Figure 5). In comparisons between *G. max* and the other taxa (including not only other *Glycine* species but also *Phaseolus vulgaris*), the average substitution rate is $2.68 \times 10^{-9}$ substitutions per site per year (Table 3). These rates are slow relative to other reported chloroplast nucleotide substitution rates (Zhong et al. 2009; Guo et al. 2007). The similarity of the substitution rates between *G. max* and the other species suggests a clock-like rate of evolution. However, when divergence dates from the nuclear gene are used, substitution rates appear much slower among the A-plastome species, averaging $7.95 \times 10^{-10}$ substitutions per site per year. Because comparisons of these same taxa with *G. max* are clock-like, this suggests that the divergence date estimates from the nuclear genome are inappropriately high for estimating substitution rates of chloroplast genomes within this group of taxa. Recent divergence of plastomes within the A-plastome group supports a hypothesis of introgression.
CONCLUSIONS

Our understanding of the tempo and mode of plastome evolution is enhanced as sequencing technologies have led to the increase in complete plastome sequences, including multiple species within key genera. Rearrangements observed in legumes, when only a few sequences were available, are now being shown to be specific to individual lineages at key times in legume evolution and not indicative of continuous processes leading to changes in gene order. Within Glycine, the chloroplast genome is very stable. Repetitive sequences do not flank the 51 kb inversion shared by the papilionoid legumes and are not as numerous as they might seem without further investigation of their location and sequence. Despite being distributed across the Glycine chloroplast genomes, repetitive sequences have not changed the order of genes among species. Phylogenetic analyses of complete plastome sequences corroborate previous restriction mapping studies (Doyle et al. 1990b) in suggesting an introgression of the A-plastome into G. falcata. Bayesian analysis and dating divergence allows us to hypothesize the introgression occurred 1-3 mya.

ACKNOWLEDGEMENTS NSF Grant 0822258, SoyMap2, Jane Doyle, Dan Ilut and Ashley Egan for technical support and access to unpublished data.
Figure 1. Partial representation of legume phylogeny. Fourteen genera with sequenced plastomes are named. Asterisks indicate genera having multiple species with plastome sequences.

Figure 2. Gene map of *Glycine syndetika* chloroplast genome. Gray bars on inner circle represent the extent of the inverted repeats (IRa and IRb). Yellow bar on middle circle represents 51 kb inversion found in papilionoid legumes. Genes on the outside of the circle are transcribed in a clockwise direction, genes on the inside of the circle are transcribed in a counter-clockwise direction.

Figure 3. Sequence similarity plot generated by VISTA tools. Base sequence is *G. syndetika*. Intergenic regions are pink. Shuffle LAGAN option used to align *Phaseolus* and *Vigna* that have 78 kb inversions. Hypervariable region identified by MAGEE et al. (2010) is demarcated by red line. Gene with low sequence similarity to other angiosperms, *ycf4* is indicated, along with five other genes for orientation. Only one inverted repeat is shown.

Figure 4. Locations of repetitive sequences in *Glycine* plastomes. Black horizontal lines represent coding regions in each plastome. Light green squares mark the location of dispersed repeat sequence 1 from REPuter. Dark green squares mark the location of ten additional dispersed sequences. The locations of low complexity repeats are shown by blue lines, SSRs by red lines as determined by RepeatScout. Asterisks represent

Sherman-Broyles et al. 22
repeats within coding sequence. Black diamonds represent repeats with sequence similarity to hAT repeat transposons.

Figure 5. Comparison of Bayesian inference trees based on whole plastome alignment (A) and histone H3D alignment (B). Divergence dates calculated using BEAST. Bars represent the 95% HPD (Highest Posterior Density) interval.
Sherman-Broyles et al. 24

LITERATURE CITED


Sherman-Broyles et al. 24


Table 1. Summary statistics from complete *Glycine* chloroplast sequences. Length in base pairs of the chloroplast and the three major divisions. Extent of border genes within the IR. GC content in each chloroplast genome.

<table>
<thead>
<tr>
<th>Plastome Group</th>
<th>Genome Group</th>
<th><em>Glycine</em> species</th>
<th>Total</th>
<th>LSC</th>
<th>SSC</th>
<th>IR</th>
<th>duplicated ycf1 in IR</th>
<th>rps19 in IR</th>
<th>%GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G</td>
<td><em>G. max</em>(^a)</td>
<td>152218</td>
<td>83174</td>
<td>17896</td>
<td>25574</td>
<td>478</td>
<td></td>
<td>0.34*</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td><em>G. syndetika</em></td>
<td>152794</td>
<td>83844</td>
<td>17840</td>
<td>25555</td>
<td>463</td>
<td>61</td>
<td>0.353</td>
</tr>
<tr>
<td>A</td>
<td>A / D</td>
<td><em>G. dolichocarpa</em></td>
<td>152804</td>
<td>83815</td>
<td>17807</td>
<td>25591</td>
<td>490</td>
<td>61</td>
<td>0.353</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td><em>G. canescens</em></td>
<td>152533</td>
<td>83559</td>
<td>17844</td>
<td>25565</td>
<td>463</td>
<td>61</td>
<td>0.353</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td><em>G. tomentella D3</em></td>
<td>152728</td>
<td>83773</td>
<td>17829</td>
<td>25563</td>
<td>463</td>
<td>61</td>
<td>0.353</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
<td><em>G. falcata</em></td>
<td>153023</td>
<td>84027</td>
<td>17846</td>
<td>25575</td>
<td>463</td>
<td>59</td>
<td>0.353</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td><em>G. stenophita</em></td>
<td>152618</td>
<td>83937</td>
<td>17817</td>
<td>25432</td>
<td>463</td>
<td>61</td>
<td>0.353</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td><em>G. cyrtoloba</em></td>
<td>152381</td>
<td>89368</td>
<td>17801</td>
<td>25505</td>
<td>463</td>
<td>61</td>
<td>0.353</td>
</tr>
</tbody>
</table>

\(^a\)Saski et al. 2005
Table 2. Number and types of repeated sequences. Repeated sequences were detected by RePuter, RepeatMasker and Repeat Finder. Further analysis of dispersed repeats showed that many of the sequences shared similarity, reducing the number of unique repeat sequences and locations.

<table>
<thead>
<tr>
<th>Species</th>
<th>RePuter Repeat Sequences</th>
<th>Dispersed Repeats Unique Locations</th>
<th>Dispersed Repeats Unique Sequences</th>
<th>Tandem Repeats</th>
<th>RepeatMasker/Repeat Finder Repeat Sequences</th>
<th>RepeatFinder Low complexity</th>
<th>Low Complexity</th>
<th>SSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>57&lt;sup&gt;a&lt;/sup&gt;</td>
<td>19</td>
<td>9</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. max&lt;sup&gt;+&lt;/sup&gt;</td>
<td>104&lt;sup&gt;a&lt;/sup&gt;</td>
<td>22</td>
<td>10</td>
<td>28</td>
<td>83</td>
<td>56</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>G. synegetika</td>
<td>103</td>
<td>20</td>
<td>5</td>
<td>32</td>
<td>95</td>
<td>69</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>G. dolichocarpa</td>
<td>104</td>
<td>20</td>
<td>5</td>
<td>29</td>
<td>94</td>
<td>69</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>G. canescens</td>
<td>86</td>
<td>16</td>
<td>5</td>
<td>29</td>
<td>95</td>
<td>71</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>G. tomentella D3</td>
<td>108</td>
<td>28</td>
<td>8</td>
<td>30</td>
<td>92</td>
<td>69</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>G. falcata</td>
<td>85</td>
<td>28</td>
<td>7</td>
<td>27</td>
<td>89</td>
<td>64</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>G. stenophita</td>
<td>96</td>
<td>19</td>
<td>6</td>
<td>31</td>
<td>82</td>
<td>58</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>G. cyrtoloba</td>
<td>101</td>
<td>20</td>
<td>6</td>
<td>34</td>
<td>78</td>
<td>54</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Phaseolus</td>
<td>62</td>
<td>9</td>
<td>3</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna</td>
<td>68</td>
<td>10</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milletia</td>
<td>46</td>
<td>10</td>
<td>3</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Saski et al. 2005
Table 3. Pairwise SNPs and substitution rates. The number of SNPs between species appears above the diagonal, the substitution rate was calculated using histone H3D divergence dates. Substitution rates in bold represent A plastome taxa with substitution rates that appear slower because of inferred chloroplast capture.

<table>
<thead>
<tr>
<th>Rate/SNPs</th>
<th>G. max</th>
<th>G. syndetika</th>
<th>G. dolichocarpa</th>
<th>G. canescens</th>
<th>G. tomentella D3</th>
<th>G. falcata</th>
<th>G. stenophita</th>
<th>G. cyrtoloba</th>
<th>P. vulgaris</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. max</td>
<td>2066</td>
<td>2044</td>
<td>2074</td>
<td>2104</td>
<td>2094</td>
<td>2022</td>
<td>2384</td>
<td>8234</td>
<td></td>
</tr>
<tr>
<td>G. syndetika</td>
<td>2.49E-09</td>
<td></td>
<td>175</td>
<td>210</td>
<td>307</td>
<td>638</td>
<td>1608</td>
<td>2026</td>
<td>8353</td>
</tr>
<tr>
<td>G. dolichocarpa</td>
<td>2.47E-09</td>
<td>2.22E-09</td>
<td></td>
<td>303</td>
<td>635</td>
<td>1596</td>
<td>2032</td>
<td>8341</td>
<td></td>
</tr>
<tr>
<td>G. canescens</td>
<td>2.50E-09</td>
<td>1.04E-09</td>
<td>2.70E-09</td>
<td>322</td>
<td>660</td>
<td>1627</td>
<td>2054</td>
<td>8331</td>
<td></td>
</tr>
<tr>
<td>G. tomentella</td>
<td>2.54E-09</td>
<td>5.60E-10</td>
<td>5.53E-10</td>
<td>5.87E-10</td>
<td>687</td>
<td>1640</td>
<td>2066</td>
<td>8368</td>
<td></td>
</tr>
<tr>
<td>G. falcata</td>
<td>2.53E-09</td>
<td>8.80E-10</td>
<td>8.76E-10</td>
<td>9.11E-10</td>
<td>9.48E-10</td>
<td>1640</td>
<td></td>
<td>8365</td>
<td></td>
</tr>
<tr>
<td>G. stenophita</td>
<td>2.44E-09</td>
<td>2.74E-09</td>
<td>2.72E-09</td>
<td>2.77E-09</td>
<td>2.79E-09</td>
<td>2.79E-09</td>
<td></td>
<td>1284</td>
<td>8351</td>
</tr>
<tr>
<td>G. cyrtoloba</td>
<td>2.88E-09</td>
<td>3.70E-09</td>
<td>3.71E-09</td>
<td>3.75E-09</td>
<td>3.77E-09</td>
<td>3.78E-09</td>
<td>2.34E-09</td>
<td></td>
<td>8616</td>
</tr>
<tr>
<td>P. vulgaris</td>
<td>2.72E-09</td>
<td>2.76E-09</td>
<td>2.76E-09</td>
<td>2.75E-09</td>
<td>2.77E-09</td>
<td>2.77E-09</td>
<td>2.76E-09</td>
<td>2.85E-09</td>
<td></td>
</tr>
</tbody>
</table>