Intragenomic polymorphisms among high-copy loci: A genus-wide study of nuclear ribosomal DNA in *Asclepias* (Apocynaceae)

Despite knowledge that concerted evolution of high-copy loci is often imperfect, few studies investigate the extent of intragenomic polymorphisms and comparisons across a large number of species are rarely made. We present a bioinformatic pipeline for characterizing polymorphisms within an individual among copies of a high-copy locus. Results are presented for nuclear ribosomal DNA (nrDNA) across the milkweed genus, *Asclepias*. The 18S-26S portion of the nrDNA cistron of *Asclepias syriaca* served as a reference for assembly of the region from 124 samples representing 90 species of *Asclepias*. Reads were mapped back to each individual's consensus and at each position reads differing from the consensus were tallied using a custom perl script. Low frequency polymorphisms existed in all individuals (mean = 5.8%). Most nrDNA positions (91%) were polymorphic in at least one individual, with polymorphic sites being less frequent in subunit regions and loops. Highly polymorphic sites existed in each individual, with highest abundance in the "noncoding" ITS regions. Phylogenetic signal was present in the distribution of intragenomic polymorphisms across the genus. Intragenomic polymorphisms in nrDNA are common in *Asclepias*, being found at higher frequency than any other study to date. The high and variable frequency of polymorphisms across species highlights concerns that phylogenetic applications of nrDNA may be error-prone. The new analytical approach provided here is applicable to other taxa and other high-copy regions characterized by low coverage genome sequencing (genome skimming).

1 Intragenomic polymorphisms among high-copy loci: A genus-wide study of nuclear
2   ribosomal DNA in *Asclepias* (Apocynaceae)

3   Kevin Weitemier[1,2], Shannon C. K. Straub[3], Mark Fishbein[4], and Aaron Liston[2]

4   [2]Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall,
5   Corvallis, OR 97331 USA; [3]Department of Biology, Hobart and William Smith Colleges, 300
6   Pulteney Street, Geneva, NY 14456 USA; and [4]Department of Botany, Oklahoma State
7   University, 301 Physical Sciences, Stillwater, OK 74078 USA

8   Email addresses:

9   KW: kevin.weitemier@science.oregonstate.edu

10   SCKS: straub@hws.edu

11   MF: mark.fishbein@okstate.edu

12   AL: listona@science.oregonstate.edu

13 **INTRODUCTION**

1   1: Author for correspondence: kevin.weitemier@science.oregonstate.edu, 541-737-5275

14   With the advent of DNA sequencing technology to infer phylogenetic relationships,

15   investigators began searching for genetic loci that were both phylogenetically informative and

16   readily sequenced in most organisms. The use of nuclear ribosomal DNA (nrDNA) soon became

17   a popular choice for phylogenetic inference (Hamby & Zimmer, 1988; Hillis & Dixon, 1991;

18   Baldwin, 1992; Baldwin et al., 1995; Álvarez & Wendel, 2003). Nuclear ribosomal DNA offered

19   several advantages over other loci: the combination of highly conserved and variable regions

20   allowed phylogenetic inference across a broad range of evolutionary time scales, conserved

21   regions allowed the use of "universal" PCR primers applicable to a wide range of taxa, the high

22   copy number of nrDNA repeats allowed reliable amplification from lower quality DNA

23   extractions, and the process of concerted evolution ensured that these copies were similar within

24   individuals (Baldwin et al., 1995). The use of nrDNA, particularly the variable internal

25   transcribed spacer (ITS) regions, became widespread, to the extent that many studies were based

26   exclusively on ITS data (Álvarez & Wendel, 2003).

27   However, nrDNA loci have been shown to harbor limitations in their phylogenetic utility.

28   Nuclear ribosomal DNA copies are assembled as tandem repeats at one or more loci in the

29   genome, with each locus being known as an array. The number of repeats present within an array

30   is labile, as is the number and location of arrays (Álvarez & Wendel, 2003). The process of

31   nrDNA copy homogenization is thought to occur much more frequently within than among arrays

32   (Schlötterer & Tautz, 1994). Thus, differing nrDNA alleles may become fixed in different arrays

33   within a genome, creating paralogy that, if unrecognized, may confound phylogenetic inference

34   (Álvarez & Wendel, 2003). Moreover, these events can create pseudogenes which, freed from

35   selective pressures, may evolve through processes quite different from the functional loci and

36   provide misleading evidence for between-individual genetic divergences if compared to

37   functional copies. These events may occur at a greater rate than inter-array homogenization via

38   concerted evolution (Gernandt & Liston, 1999).

39   Due to the technical difficulty of systematically sequencing individual nrDNA loci

40   because of their high copy number, studies characterizing the abundance and patterns of

41   intragenomic polymorphisms have been rare. Recently, studies utilizing whole-genome shotgun

42   sequencing have begun to reveal levels of intragenomic polymorphism in *Drosophila* (Stage &

43   Eickbush, 2007)*,* nematodes (Bik et al., 2013), and fungi (Ganley & Kobayashi, 2007). However,

44   these studies included a small number of species (12, 6, and 5, respectively) and did not attempt

45   to place patterns of polymorphism in a phylogenetic context. Other studies have used

46   intragenomic polymorphisms to identify populations of *Arabidopsis* (Simon et al., 2012) and

47   infer intraspecific phylogenies of *Saccharomyces* (West et al., 2014). Studies of intragenomic

48   polymorphism patterns across many species have not been performed in plants (but see Straub et

49   al., 2012).

50   The current study utilizes high throughput technology to survey a large sample of species

51   and individuals in the angiosperm genus *Asclepias* (Apocynaceae) in order to characterize levels

52   of intragenomic nrDNA polymorphism and place these within a phylogenetic context. The

53   methods presented here are expanded from those we have previously developed as part of the

54   Milkweed Genome Project (Straub et al., 2011, 2012), and generalized for use with a large

55   number of taxa and any high-copy locus, such as those that may be obtained from a genome-

56   skimming or Hyb-Seq study (Straub et al., 2012; Weitemier et al., 2014).

57   **METHODS**

58   ***Sampling and sequencing***

59   One hundred twenty-five individuals representing 90 *Asclepias* species and subspecies

60   were sampled (Table 1) and sequencing libraries were produced as described in Straub et al.

61   (2012). Two individuals of putatively hybrid origin were included: *A. albicans* × *subulata* and *A.*

62   *speciosa* × *syriaca*. These individuals were collected from wild populations and identified as

63   hybrids through expression of intermediate morphological characteristics (M. Fishbein, unpubl.

64 data; see also Fishbein et al., 2011). Samples were multiplexed in approximately equimolar

65 ratios, with up to 21 individuals per lane, and sequenced with 80 bp single-end reads on an

66 Illumina GAIIx instrument (Illumina, San Diego California, USA). *Asclepias subverticillata* was

67 multiplexed in a lane with 32 samples and sequenced with 101 bp paired-end reads on an

68 Illumina HiSeq 2000 instrument, with reads analyzed as though they were single-end. One

69 individual of *A. syriaca* was sequenced at higher coverage: this individual was sequenced in a

70 single lane on an Illumina GAIIx with 40 bp single-end reads (Straub et al., 2011). To allow more

71 efficient assembly downstream, read pools were filtered to remove plastid reads (using the

72 custom script sort_fastq_v1.pl modified to retain Ns; Knaus, 2010).

73 An *A. syriaca* haploid genome size estimate of 420 Mbp and a nrDNA copy number

74 estimate of 960 were used for estimates of sequencing depth and for comparisons with other

75 organisms. These estimates are modified from Straub et al. (2011), where an incorrect estimate of

76 the average *A. syriaca* 2C value led to a haploid genome size estimate of 820 Mbp and a nrDNA

77 copy number estimate of 1,845. The current values are based on 2C estimates from Bai et al.

78 (2012) and Bainard et al. (2012).

79 ***Polymorphism quantification***

80 The method for determining polymorphisms present among nrDNA copies within an

81 individual while retaining information about position homology across a group of distantly

82 related individuals included four general steps: 1) A sequence was selected to serve as a reference

83 for the whole group. 2) A consensus sequence was obtained for each individual taxon and aligned

84 against the group reference, allowing tailored read-mapping for each individual while associating

85 positions along the individual consensus with their homologous positions in the group reference.

86 3) Reads for each individual were mapped onto that individual's consensus sequence. 4) At each

87 position the reads differing from the individual consensus were tallied.

88 *1) Group reference*

89   The nrDNA cistron of the high-coverage *A. syriaca* individual was previously assembled

90   (Straub et al., 2011; GenBank JF312046). The nontranscribed spacer and external transcribed

91   spacers from each end were removed due to the presence of internal repeats, and the more

92   conserved 18S and 26S subunits used as the boundaries of the alignment. The resulting reference

93   sequence contained 5839 bp.

94   *2) Individual consensus sequences*

95   Read pools were examined prior to consensus assembly and reads that were exact

96   duplicates were reduced to a single representative, retaining the highest average quality score

97   (using the custom script fastq_collapse.py, available at https://github.com/listonlab). Sequences

98   for each individual were constructed via reference-guided assembly, with *A. syriaca* as the

99   reference, using Alignreads ver. 2.25 (Straub et al., 2011). Alignreads is a pipeline that includes

100   the short-read assembler YASRA (Ratan, 2009), utilities from the MUMmer ver. 3.0 suite

101   (Delcher et al., 2002; Kurtz et al., 2004), and custom scripts. Parameters were selected to ensure

102   high identity of reads mapping to the nrDNA reference (95%), but allow the reconstructed

103   sequence to differ from the reference. In addition to assembling the individual consensus

104   sequence, Alignreads outputs a file associating each position in the group reference with those in

105   the individual consensus.

106   *3) Read mapping*

107   Prior to mapping reads from an individual onto its consensus, reads with an average

108   quality (Phred) score below 20 were removed, and bases in the remaining reads with a score

109   below 20 were converted to Ns using FASTX-Toolkit ver. 0.0.13 (Gordon, 2008). Read mapping

110   was performed with the program BWA ver. 0.5.7 (Li & Durbin, 2009), and output files processed

111   with the SAMtools ver. 0.1.13 utilities (Li et al., 2009).

112   *4) Polymorphism counting*

113   A perl script was developed to tally the number of reads differing from the consensus at

114   each position (polymorphic_read_counter_bwaPileup.pl ver. 3.03b, available at

115   https://github.com/listonlab). For example, a base covered by 10 reads might have 7 reads with a

116   G in that position and 3 with a C. In this case the consensus would have called a G at that

117   position, with 30% of the reads differing. See Supplemental File 1 for exact parameters and a

118   pipeline of commands used.

119   Positions with 2% or more of reads differing from the consensus base were considered

120   polymorphic. This cutoff is the same used by Straub et al. (2011) and comparable to that used by

121   Nguyen et al. (2011) under a similar quality-filtering scheme. The control *PhiX* lane of the higher

122   coverage *A. syriaca* individual was examined by Straub et al. (2011, Additional file 1 from that

123   study) and found to have an error rate much less than 2%, indicating that the cutoff used here

124   may be somewhat conservative. In addition to counting positions that were polymorphic,

125   positions were recorded as "highly polymorphic" if 10% or more of the reads differed from the

126   consensus.

127   In order to keep homologous bases aligned across individuals, only those positions that

128   were present in the *A. syriaca* (group) reference were kept in the analysis (i.e., insertions relative

129   to the reference were discarded). Note that deletions (relative to the reference) fixed within an

130   individual are also not considered because zero reads called a base at that position.

131   ***RNA structure determination***

132   The secondary structure of each subunit and spacer region was predicted for the *A.*

133   *syriaca* reference from the minimum free energy structure found by the program RNAfold for the

134   18S, ITS1, and ITS2 regions and RNAcofold for the 5.8S+26S regions (Lorenz et al., 2011).

135   Program default model parameters were used (37° C, unpaired bases can participate in up to one

136   dangling end). Positions along the cistron were then categorized as either paired (stems) or

137   unpaired (loops).

138   *Polymorphisms within the cistron*

139   The effects of cistron position (subunit or spacer) and secondary structure position (stem

140   or loop) on the likelihood of a position being polymorphic or highly polymorphic were assessed

141   in two ways. Differences in the likelihood that at least one individual was polymorphic at a

142   position (e.g. positions coded as either polymorphic or invariant) were assessed via a two factor

143   multiple logistic regression, as implemented in R ver. 3.1.0 using the MASS ver. 7.3.33 package

144   (Venables & Ripley, 2002; R Core Team, 2014). Differences in the abundance of polymorphic

145   individuals at a position were assessed using square-root transformed data with a two-way

146   ANOVA and type III sum of squares for unbalanced design, as implemented using the car ver.

147   2.0.20 package (Fox & Weisberg, 2011). The ANOVA analysis is not reported for the highly

148   polymorphic individuals because the data diverge substantially from assumptions of a normal

149   distribution.

150   *Phylogenetic context*

151   A maximum likelihood estimate of phylogenetic relationships within *Asclepias* was

152   produced by Fishbein et al. (2011, Figure 2, TreeBase #27576). This tree was pruned to match the

153   sampling in this study, and counts of polymorphic positions were recorded for each taxon. Taxa

154   sampled in this study, but not present in Fishbein et al. (2011), were omitted from further

155   analyses. Counts were averaged for taxa with multiple individuals in this study, but sampled only

156   once in Fishbein et al. (2011). Ancestral states for the number of polymorphic positions in the

157   rDNA cistron were reconstructed using squared-change parsimony in the Mesquite phylogenetic

158   suite ver. 2.75+ build 573 (Maddison & Maddison, 2011; Maddison, Maddison & Midford,

159   2011).

160   Phylogenetic signal in the distribution of the number polymorphic positions was tested in

161   two ways. In the first method, the total length of the tree (parsimony steps) in terms of changes in

162   number of polymorphic positions was compared to a distribution of tree lengths created by

163 10,000 permutations of polymorphic positions across tips. In the second method, a likelihood

164 ratio test (LRT) was performed between models where character evolution followed a Brownian

165 motion model across the tree. In the first model, the parameter lambda (describing how well the

166 phylogeny correctly predicts the covariance among taxa for a trait) was found that maximized the

167 model's likelihood (Pagel, 1999). This was compared to the likelihood found when lambda was

168 held at zero, representing phylogenetic independence among species for that trait. Parsimony

169 permutations were performed in Mesquite (Maddison & Maddison, 2011; Maddison, Maddison

170 & Midford, 2011), and likelihood ratio tests were performed in R with the phytools ver. 0.4.05

171 and ape ver. 3.1.2 packages (Paradis, Claude & Strimmer, 2004; Revell, 2012; R Core Team,

172 2014).

173   To determine if any clades held intragenomic polymorphism frequencies that were

174 significantly high or low relative to the rest of the phylogeny, polymorphism rates were simulated

175 along the tree and the true polymorphism counts compared to the distribution of simulated counts

176 (Garland et al., 1993). Two tips in the tree separated by zero branch length (*A. asperula* ssp.

177 *asperula* and ssp. *capricornu*) were collapsed, and the average of the polymorphism counts for

178 the four sampled individuals of the species used for the new tip. A model of trait evolution under

179 Brownian motion, using the lambda parameter estimated from the LRT above, was found from

180 10,000 random starting points using the fitContinuous function from the geiger ver. 2.0.3 R

181 package (Harmon et al., 2008). This model was used to simulate polymorphism counts across the

182 phylogeny 10,000 times, with lower and upper bounds of 0 and infinity, respectively, using the

183 fastBM function in phytools ver. 0.4.31 (Revell, 2012). For each node, the ancestral state was

184 estimated for the true data and the simulated data using squared change parsimony as

185 implemented in Mesquite (Maddison & Maddison, 2011; Maddison, Maddison & Midford,

186 2011).

187 **RESULTS**

188   Average coverage of the nrDNA region was ~97x for all individuals (median = 88x;

189   Sequence Read Archive ###-###). Despite high overall coverage of the nrDNA region, not all

190   positions of the cistron were assembled for all individuals; therefore results for polymorphic

191   positions are presented both as counts and as percentages of sequenced bases. The *A. syriaca*

192   reference had total genome coverage of ~0.8x (Sequence Read Archive SRX040889).

193   ***Intragenomic polymorphism***

194   All individuals were polymorphic at several positions homologous with the *A. syriaca*

195   reference (Table 1). Number of polymorphic positions ranged from 23 (*A. verticillata*, 0.41% of

196   sequenced bases) to 882 (*A. solanoana*, 15.13%), with a mean of 333 (5.77%, Fig. 1). A very

197   high percentage (91%) of positions in the *A. syriaca* reference were polymorphic in at least one

198   individual (Fig. 2A).

199   Positions were significantly more likely to be polymorphic if they were in a spacer region

200   (ITS1, ITS2; Fig. 2A) or stem (Fig. 3A). This is true both when considering the number of

201   individuals polymorphic at a position (Table 2), and whether any sample was polymorphic at that

202   position (Table 3A).

203   The number of highly polymorphic positions ranged from 2 (*A. arenaria*, 0.03%) to 111

204   (*A. boliviensis*, 1.91%) with a mean of 28 (0.50%). Positions highly polymorphic in more than 10

205   individuals were found in the 18S, ITS1, ITS2, and 26S regions (Fig. 2B). The most polymorphic

206   position was 4172 (using the *A. syriaca* reference), in the 26S region, which was highly

207   polymorphic in 29 individuals. Highly polymorphic positions were dramatically less frequent in

208   the subunit regions (18S=19%, 5.8S=26%, 26S=31%) than in the spacer regions (ITS1=61%,

209   ITS2=57%; Table 3B, Fig. 2B). Positions in secondary structure stems were moderately more

210   likely to be highly polymorphic in at least one individual than loop positions (Table 3B, Fig. 3B).

211   ***Phylogenetic Signal***

212   The number of polymorphic base pair positions exhibited strong phylogenetic signal

213   across *Asclepias* under both the permutaion test ($P<0.0012$) and the likelihood ratio test

214   (estimated lambda=0.51, P=0.0067; Table 4; Fig. 4). This signal remained even after the ITS

215   regions were removed from the dataset (permutation test $P<0.0200$, lambda=0.45, LRT

216   P=0.0112). Highly polymorphic base pair abundance, however, was not significantly influenced

217   by phylogenetic history (Table 4; Fig. 5), even when considering only the subunit regions.

218   Of the 53 resolved clades in the phylogeny, none showed polymorphism values more

219   extreme than expected under a Brownian motion model after correcting for multiple comparisons.

220   The most extreme clade was that formed by *A. hypoleuca* and *A. otarioides*, which had an

221   ancestral number of polymorphic positions exceeded by 79 of the 10,000 simulations (when

222   excluding the spacer regions this node was exceeded by 60 simulations). The following most

223   extreme nodes were those ancestral to *A. rosea* and *A. lemmonii* with more polymorphic positions

224   than all but 130 (79) simulations, and ancestral to *A. boliviensis* and *A. mellodora* with more

225   polymorphic positions than all but 246 (150) simulations. Under a Bonferroni correction a clade

226   would require 9 or fewer simulations more extreme than the observed value (α=0.05) to reject a

227   hypothesis of no divergence from Brownian motion.

228   ***Identification of Mixed Ancestry***

229   The number of polymorphic sites for hybrid individuals, 299 for *A. albicans* × *subulata*

230   and 161 for *A. speciosa* × *syriaca*, are less than the mean number of 333 polymorphic sites; and

231   the number of highly polymorphic sites, 12 and 37, are less than or greater than the mean of 28.

232   Of those positions that are highly polymorphic, 4 of 37 in *A. speciosa* × *syriaca* have a minor

233   allele frequency of 0.3 or higher, while none of the positions in *A. albicans* × *subulata* have a

234   minor allele frequency above 0.2.

235   **DISCUSSION**

236 Absolute levels of intragenomic polymorphism among the copies of nrDNA in *Asclepias*

237 (mean=333 positions) were found to be much higher than levels reported for nematodes (<250;

238 Bik et al., 2013), fungi (3-37; Ganley & Kobayashi, 2007), and *Drosophila* (3-18; Stage &

239 Eickbush, 2007) when including all polymorphic positions. When considering only positions that

240 are highly polymorphic, *Asclepias* exhibits slightly higher rates (mean=28.4 positions) than fungi

241 and *Drosophila*, but much lower rates than nematodes. However, these comparisons may be

242 misleading: First, the number of nrDNA copies varies greatly between these taxa, estimated to

243 range from about 50-180 in the fungal species, 200-250 in *Drosophila melanogaster*, 56-323 in

244 the nematodes, and about 960 in *Asclepias* (Ganley & Kobayashi, 2007; Stage & Eickbush, 2007;

245 Straub et al., 2011; Bik et al., 2013). Second, polymorphic base pair counts are confounded by

246 differing criteria for scoring polymorphism (i.e., methods for excluding sequencing errors). The

247 levels listed for the fungal species include both "high-confidence" and "low-confidence"

248 polymorphisms, based primarily on sequence quality (Ganley & Kobayashi, 2007). The levels

249 listed for *Drosophila* are polymorphisms present in $\geq$ 3% of loci (Stage & Eickbush, 2007). Bik

250 et al. (2013) tallied read counts using a method similar to the method presented here, but called

251 positions polymorphic when the count of differing reads exceeded what would be expected for a

252 single copy locus. Nevertheless, given that *Asclepias* has absolute counts of polymorphic

253 positions at least 33% higher than the other organisms studied, and that the sequencing depth was

254 nearly two orders of magnitude greater in the nematodes than in *Asclepias* (6.3-10x per nrDNA

255 copy in nematodes, vs. ~0.1x in *Asclepias*; Bik et al., 2013), it is likely that *Asclepias* harbors

256 greater rates of intragenomic polymorphism within the nrDNA cistron than the organisms studied

257 to date.

258 ***Polymorphism patterns across the nrDNA cistron***

259 Spacer regions (ITS1, ITS2) had higher frequencies of polymorphic positions than subunit

260 regions (18S, 5.8S, 26S; Fig. 2). However, positions with low polymorphism frequencies are

261 distributed much more evenly across the nrDNA cistron (Fig. 2A) than highly polymorphic

262 positions, which show strong differentiation between the subunit and spacer regions (Fig. 2B).

263 The lower frequency of highly polymorphic positions within the subunit regions suggests that

264 these regions are under selection to remain homogenous within individual genomes. The lower

265 difference in low polymorphism frequency between subunit and spacer regions suggests that this

266 selection pressure is positively correlated with the proportion of nrDNA copies that differ from

267 the majority. These findings are in contrast to those reported for nematodes (Bik et al., 2013),

268 where the subunit regions had much higher levels of polymorphism abundance than the spacer

269 regions.

270 Positions in stem regions were more likely to be polymorphic than loop positions (Fig. 3).

271 This was strongly significant for all polymorphic positions (Tables 2, 3A) and moderately

272 significant for highly polymorphic positions (Table 3B). This would seem to contradict the

273 hypothesis that stem sites in general should be more highly conserved in order to maintain a

274 functional RNA secondary structure. Indeed, this finding agrees with those from Rhetsky (1995),

275 who not only found that trees estimated from stem regions contained longer branch lengths than

276 those from loop regions, but that those stem sites least likely to affect secondary structure tended

277 to be less conserved. Loop sites, on the other hand, contain a large proportion of the sites critical

278 to ribosomal function (Rzhetsky, 1995), and may be under stronger stabilizing selection than stem

279 sites.

280 Among the subunit regions, the 26S region harbored the highest frequency of highly

281 polymorphic positions. This agrees with Stage & Eickbush (2007) who showed the *Drosophila*

282 28S region to have a higher mutation rate than the other subunit regions. However, in that study

283 the 28S region also had a lower frequency of polymorphic base pairs. Stage & Eickbush (2007)

284 hypothesize that this is due to the action of two retrotransposable elements found in many

285 *Drosophila* 28S copies, with instances of aborted insertions causing the cell to repair the region

286   using a nearby template and thereby homogenize the copies. They predict that levels of

287   polymorphism will be higher in the 28S region of organisms lacking the retrotransposable

288   elements, as seen here in the homologous angiosperm 26S region.

289   Within the subunit regions, highly polymorphic positions may be more common within

290   expansion regions of the RNA gene (less conserved regions that tend to incorporate sequence

291   insertions without affecting functionality; Clark et al., 1984). This is strongly implied by the

292   recovery of clusters of highly polymorphic sites near *A. syriaca* positions 4440 and 5020, which

293   are directly within the 26S expansion regions seven and eight, respectively (Kolosha & Fodor,

294   1990; Fig. 1). This agrees with results found in *Drosophila* (Stage & Eickbush, 2007). However,

295   this may not be true for all highly polymorphic sites, as the high peak at position 4172 is not

296   within an expansion region.

297   ***Phylogenetic context***

298   Mapping the number of polymorphic positions onto the phylogeny of *Asclepias*

299   demonstrates strong and significant phylogenetic signal when counting all polymorphisms (Fig.

300   4), but this signal is not significant when only counting highly polymorphic positions (Fig. 5).

301   While this study demonstrates phylogenetic signal in polymorphism abundance, it remains

302   unknown whether abundance within a lineage is influenced by selection or purely neutral causes.

303   Different demographic histories across clades could create this effect via neutral causes, while

304   selection against organisms that retain too many variant nrDNA copies could be variable across

305   the genus. Polymorphisms present in a high proportion of nrDNA copies may be uniformly

306   selected against, while low polymorphism sites are tolerated at varying levels. This could explain

307   the lack of phylogenetic signal at the highly polymorphic sites, and its presence when including

308   all polymorphic sites.

309   These results have implications for phylogenetic inference based on nrDNA data. As

310   previously cautioned, it cannot be assumed that all copies of nrDNA are identical within a

311   genome. This is especially true for the spacer regions, but also for the subunit regions (Álvarez &

312   Wendel, 2003). The discovery of phylogenetic signal in intragenomic polymorphism abundance

313   demonstrates that those positions likely to lead to ambiguities are not distributed evenly across

314   the phylogeny (Fig. 4). In addition to topology, variable polymorphism rates among lineages may

315   affect the estimation of branch lengths on a tree. This may become especially problematic when

316   nrDNA is used to date phylogenies. Recently developed methods of phylogenetic inference

317   incorporating information about polymorphic positions may be able to alleviate difficulties in tree

318   building caused by uneven levels of polymorphism abundance (Potts, Hedderson & Grimm,

319   2014).

320   ***Identification of mixed ancestry***

321   Characterization of intragenomic nrDNA polymorphisms may allow for the identification

322   of hybrid offspring between parents with differing nrDNA sequences (Zimmer, Jupe & Walbot,

323   1988), and may be able to provide an estimate of the number of generations since hybridization.

324   This is especially true for early generation hybrids, as nrDNA homogenization can occur in a

325   small number of generations (Kovarik et al., 2005), and mosaic *Saccharomyces* genomes have

326   been identified based on intragenomic polymorphism abundance (West et al., 2014). Neither of

327   the polymorphism profiles for the two wild-collected putative hybrid individuals in this study

328   strongly indicate that they are early generation hybrids. This is in contrast to evidence from

329   nuclear gene sequences that show heterozygosity consistent with inheritance of divergent alleles

330   from the putative parents in the *A. albicans* × *A. subulata* hybrid (B. Haack and M. Fishbein,

331   unpupl. data). Detection of mixed ancestry may be hampered in this case by a lack of fixed

332   differences between parental haplotypes. Consensus sequences for the *A. albicans* and *A.*

333   *subulata* individuals in this study show no fixed differences, while one difference is found

334   between the *A. syriaca* and *A. speciosa* consensus sequences. Unexpectedly, this position is

335   monomorphic in the hybrid, while the positions with minor allele frequencies >0.3 are in

336   positions not shown to differ between *A. syriaca* and *A. speciosa*. This likely indicates ancestry

337   from an unsampled nrDNA haplotype.

338   ***Conclusions***

339   Nuclear ribosomal DNA copies within individuals of *Asclepias* are not identical, with

340   intragenomic polymorphisms present at a higher rate than reported for other organisms.

341   Polymorphism frequencies across the genus vary by more than an order of magnitude and

342   demonstrate strong phylogenetic signal. Stem positions of ribosomal subunits are more likely to

343   be polymorphic than loop positions. Distribution of polymorphic sites across the nrDNA cistron

344   are consistent with strong selection on nrDNA subunits, with polymorphic sites being more

345   frequent in the spacer regions, and this difference being amplified for sites that are highly

346   polymorphic. These results reinforce the need for caution when using nrDNA for phylogenetic

347   inference, especially when using the spacer regions or for applications requiring the precise

348   estimates of branch lengths or divergence times.

349   **ACKNOWLEDGMENTS**

350   The authors thank Richard Cronn for helpful comments and suggestions during the

351   execution of this work. Expert laboratory assistance from Shakuntala Fathepure, Ben Haack,

352   LaRinda Holland, Angela McDonnell, Laura Mealy, Nicole Nasholm, Matthew Parks, and

353   Lauren Ziemian was a crucial contribution to this research. Kind thanks go to James Riser, who

354   generously shared DNA extractions. Computer infrastructure was provided by the Center for

355   Genome Research and Biocomputing at Oregon State University. Processing of sequences for the

356   Sequence Read Archive was performed by Sanjuro Jogdeo.

357   **LITERATURE CITED**

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Plant Molecular Evolution* 29:417–434.

Bai C, Alverson WS, Follansbee A, Waller DM. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany* 110:1623–1629.

Bainard JD, Bainard LD, Henry TA, Fazekas AJ, Newmaster SG. 2012. A multivariate analysis of variation in genome size and endoreduplication in angiosperms reveals strong phylogenetic signal and association with phenotypic traits. *New Phytologist* 196:1240–1250.

Baldwin BG. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae. *Molecular Phylogenetics and Evolution* 1:3–16.

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82:247–277.

Bik HM, Fournier D, Sung W, Bergeron RD, Thomas WK. 2013. Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS ONE* 8:e78230.

Clark CG, Tague BW, Ware VC, Gerbi SA. 1984. *Xenopus laevis* 28S ribosomal RNA: A secondary structure model and its evolutionary and functional implications. *Nucleic Acids Research* 12:6197–6220.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* 30:2478 –2483.

Fishbein M, Chuba D, Ellison C, Mason-Gamer RJ, Lynch SP. 2011. Phylogenetic relationships of *Asclepias* (Apocynaceae) inferred from non-coding chloroplast DNA sequences. *Systematic Botany* 36:1008–1023.

Fox J, Weisberg S. 2011. *An R Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Ganley ARD, Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* 17:184–191.

Garland T, Dickerman AW, Janis CM, Jones JA. 1993. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* 42:265–292.

Gernandt DS, Liston A. 1999. Internal transcribed spacer region evolution in *Larix* and *Pseudotsuga* (Pinaceae). *American Journal of Botany* 86:711–723.

Gordon A. 2008. *FASTX toolkit*. Accessed June 6, 2011. Available at
http://hannonlab.cshl.edu/fastx_toolkit/index.html

Hamby RK, Zimmer EA. 1988. Ribosomal RNA sequences for inferring phylogeny within the grass
family (Poaceae). *Plant Systematics and Evolution* 160:29–37.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: Investigating evolutionary
radiations. *Bioinformatics* 24:129–131.

Hillis DM, Dixon MT. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference.
Quarterly Review of Biology 66:411–453.

Knaus B. 2010. Short read toolbox. Accessed January 28, 2011. Available at http://brianknaus.com

Kolosha VO, Fodor I. 1990. Nucleotide sequence of *Citrus limon* 26S ribosomal-RNA gene and secondary
structure model of its RNA. *Plant Molecular Biology* 14:147–161.

Kovarik A, Pires JC, Leitch AR, Lim KY, Sherwood AM, Matyasek R, Rocca J, Soltis DE, Soltis PS.
2005. Rapid concerted evolution of nuclear ribosomal DNA in two *Tragopogon* allopolyploids of
recent and recurrent origin. *Genetics* 169:931–944.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and
open software for comparing large genomes. *Genome Biol* 5:R12.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000
Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and
SAMtools. *Bioinformatics* 25:2078 –2079.

Lorenz R, Bernhart S, Honer zu Siederdissen C, Tafer H, Flamm C, Stadler P, Hofacker I. 2011.
ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 6:26.

Maddison WP, Maddison DR. 2011. *Mesquite: a modular system for evolutionary analysis*. Available at
http://mesquiteproject.org.

Maddison WP, Maddison DR, Midford PE. 2011. *Tree Farm package for Mesquite*. Available at
http://mesquiteproject.org

Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger T. 2011. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 12:106.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Potts AJ, Hedderson TA, Grimm GW. 2014. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Systematic Biology* 63:1–16.

Ratan A. 2009. Assembly algorithms for next-generation sequence data. PhD Dissertation Thesis. University Park, Pennsylvania, USA: Pennsylvania State University.

R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3:217–223.

Rzhetsky A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics* 141:771–783.

Schlötterer C, Tautz D. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Current Biology* 4:777–783.

Simon UK, Trajanoski S, Kroneis T, Sedlmayr P, Guelly C, Guttenberger H. 2012. Accession-specific haplotypes of the internal transcribed spacer region in *Arabidopsis thaliana*—A means for barcoding populations. *Molecular Biology and Evolution* 29:2231–2239.

Stage DE, Eickbush TH. 2007. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Research* 17:1888–1897.

Straub S, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn R, Liston A. 2011. Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12:211.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99:349–364.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. New York: Springer.

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*:1400042.

West C, James SA, Davey RP, Dicks J, Roberts IN. 2014. Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and predicts genome structure in two contrasting yeast species. *Systematic Biology* 63:543–554.

Zimmer EA, Jupe ER, Walbot V. 1988. Ribosomal gene structure, variation and inheritance in maize and its ancestors. *Genetics* 120:1125–1136.

Polymorphic site abundance in *Asclepias*

Polymorphic site abundance in *Asclepias* taxa. **Voucher**: Collector, collection #, [herbarium]. **Poly #**: Number of polymorphic positions. **Poly %**: Percentage of the assembled positions that are polymorphic. **High #**, **High %**: Number and percentage of highly polymorphic positions. **SRA**: NCBI Sequence Read Archive accession number.

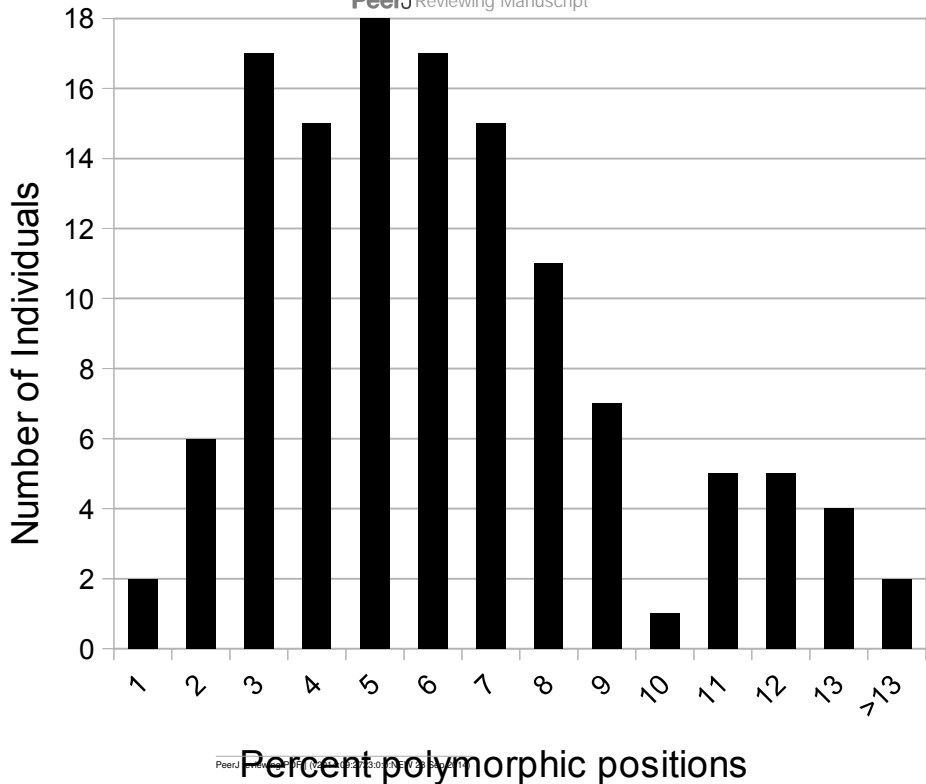
Asclepias taxon	Voucher	Poly #	Poly %	High #	High %	SRA
A. albicans S. Watson	Fishbein 3146 [WS]	174	3.12	16	0.29	SRX000000
A. albicans	Fishbein 6463 [OKLA]	273	4.70	11	0.19	
A. alticola E. Fourn.	Steinmann 5243 [IEB]	668	11.45	31	0.53	
A. amplexicaulis Sm.	Lynch 12652 [OKLA]	248	4.25	25	0.43	
A. angustifolia Schweigg.	Reina 2004-1315 [ARIZ]	55	1.04	36	0.68	
A. angustifolia	Reina 2008-203 [OKLA]	169	2.99	26	0.46	
A. arenaria Torr.	Lynch 11495 [OKLA]	174	2.98	2	0.03	
A. asperula (Decne.) Woodson ssp. *asperula*	Lynch 12037 [OKLA]	342	5.87	7	0.12	
A. asperula ssp. *asperula*	Fishbein 6536 [OKLA]	391	6.73	24	0.41	
A. asperula ssp. *capricornu* (Woodson) Woodson	Lynch 13314 [OKLA]	474	8.15	38	0.65	
A. asperula ssp. *capricornu*	Fishbein 6486 [OKLA]	347	5.95	6	0.10	
A. atroviolacea Woodson	Fishbein 3612 [ARIZ]	451	7.73	26	0.45	
A. auriculata Kunth	Lynch 1694 [OKLA]	513	8.85	33	0.57	
A. auriculata	Fishbein 5833 [OKLA]	378	6.54	39	0.68	
A. boliviensis E. Fourn.	Fishbein 6072 [OKLA]	875	15.08	111	1.91	
A. brachystephana Engelm. ex Torr.	Lynch 10642 [OKLA]	309	5.32	22	0.38	
A. californica Greene	Lynch 10779 [OKLA]	472	8.09	24	0.41	
A. aff. *candida* Vell.	Fishbein 6347 [OKLA]	245	4.20	14	0.24	
A. cinerea Walter	Fishbein 4793 [OKLA]	297	5.10	27	0.46	
A. circinalis (Decne.) Woodson	Webster 17186 [OKLA]	464	7.95	59	1.01	
A. connivens Baldwin ex Elliott	Lynch 12336 [OKLA]	394	6.75	16	0.27	
A. cordifolia (Benth.) Jeps.	Lynch 10942 [OKLA]	344	5.96	30	0.52	
A. cordifolia	Fishbein 5772 [OKLA]	308	5.35	13	0.23	
A. coulteri A. Gray	Ventura & Lopez 7986 [TEX]	471	8.09	27	0.46	
A. cryptoceras S. Watson ssp. *cryptoceras*	Fishbein 6504 [OKLA]	230	4.00	35	0.61	
A. cryptoceras ssp. *davisii* (Woodson) Woodson	Fishbein 5723 [OKLA]	350	6.00	13	0.22	
A. curassavica L.	Zuloaga & Morrone 7087 [OKLA]	258	4.42	24	0.41	
A. cutleri Woodson	Fishbein 6511 [OKLA]	167	2.90	18	0.31	
A. cutleri	Fishbein 6500 [OKLA]	157	2.69	8	0.14	
A. emoryi (Greene) Vail ex Small	Carr 12032 [TEX]	124	2.29	44	0.81	
A. engelmanniana Woodson	Lynch 11224 [OKLA]	141	2.54	33	0.59	

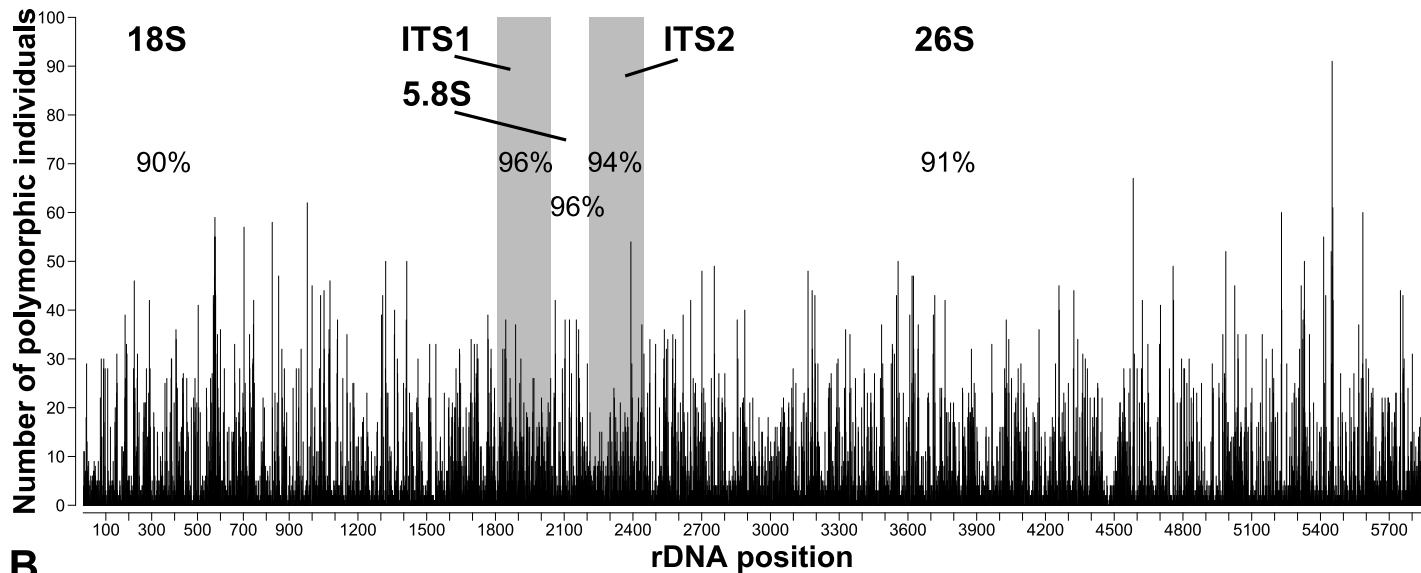
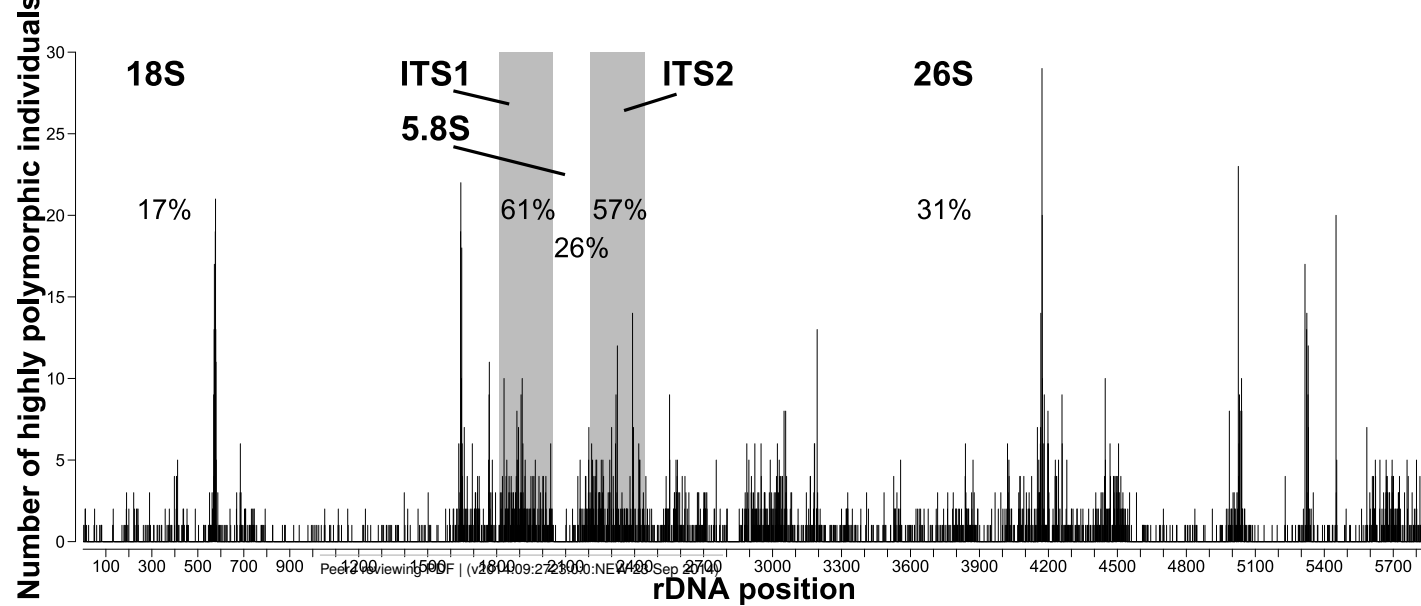
A. engelmanniana	Lynch 11029 [OKLA]	192	3.30	7	0.12
A. eriocarpa Benth.	Lynch 10923 [OKLA]	610	10.49	32	0.55
A. eriocarpa	Lynch 10799 [OKLA]	492	8.43	17	0.29
A. feayi Chapm. ex A. Gray	Fishbein 5586 [OKLA]	329	5.71	42	0.73
A. fournieri Woodson	Fishbein 3660 [ARIZ]	556	9.58	47	0.81
A. fournieri	Lynch 1655 [OKLA]	711	12.25	49	0.84
A. glaucescens Kunth	Lynch 14142 [OKLA]	190	3.33	19	0.33
A. glaucescens	Lynch 1623 [OKLA]	90	1.69	85	1.60
A. glaucescens	Fishbein 5097 [OKLA]	215	3.72	18	0.31
A. sp. nov. aff. *glaucescens*	Fishbein 3671 [ARIZ]	225	3.87	25	0.43
A. hallii A. Gray	Lynch 11299 [OKLA]	600	10.36	54	0.93
A. humistrata Walter	Fishbein 5596 [OKLA]	331	5.67	13	0.22
A. hypoleuca (A. Gray) Woodson	Lynch 11374 [OKLA]	670	11.51	44	0.76
A. incarnata L.	Lynch 12567 [OKLA]	434	7.45	28	0.48
A. involucrata Engelm. ex Torr.	Lynch 12050 [OKLA]	326	5.65	35	0.61
A. involucrata	Fishbein 6531 [OKLA]	217	3.72	13	0.22
A. jaliscana B.L. Rob.	Fishbein 2493 [ARIZ]	140	2.97	74	1.57
A. jaliscana	Fishbein 3657 [WS]	165	3.00	65	1.18
A. jorgeana Fishbein & S.P. Lynch	Vásquez & Alvarez 4905 [IEB]	160	2.76	12	0.21
A. lanceolata Walter	Fishbein 5605 [MISSA]	249	4.27	8	0.14
A. lanuginosa Nutt.	Lynch 12661 [OKLA]	343	5.92	11	0.19
A. latifolia (Torr.) Raf.	Lynch 11018 [OKLA]	236	4.05	11	0.19
A. lemmonii A. Gray	Lynch 11453 [OKLA]	620	10.70	23	0.40
A. leptopus I.M. Johnst.	Fishbein 6263 [OKLA]	282	4.84	9	0.15
A. longifolia Michx.	Lynch 12447 [OKLA]	427	7.34	24	0.41
A. lynchiana Fishbein	Venable & Becerra s.n. [ARIZ]	129	2.39	31	0.57
A. macrosperma Eastw.	Gierisch 4191 [ARIZ]	76	1.42	30	0.56
A. macrosperma	Fishbein 6518 [OKLA]	391	6.72	21	0.36
A. macrotis Torr.	Lynch 11260 [OKLA]	364	6.27	11	0.19
A. macrotis	Lynch 11263 [OKLA]	260	4.46	5	0.09
A. masonii Woodson	Fishbein 3101 [OKLA]	151	2.59	7	0.12
A. meadii Torr. ex A. Gray	Freeman 9106 [KANU]	208	3.62	20	0.35
A. mellodora A. St.-Hil.	Zuloaga & Morrone 7168 [OKLA]	377	6.47	18	0.31
A. mexicana Cav.	Fishbein 3009 [ARIZ]	186	3.22	18	0.31
A. michauxii Decne.	Lynch 12316 [OKLA]	458	7.87	39	0.67
A. notha W.D. Stevens	Lynch 14113 [OKLA]	375	6.45	44	0.76
A. notha	Fishbein 5389 [OKLA]	249	4.31	41	0.71
A. notha	Nee 32966 [NY]	432	7.47	17	0.29
A. sp. nov. cf. *notha*	Fishbein 5816 [OKLA]	376	6.45	16	0.27

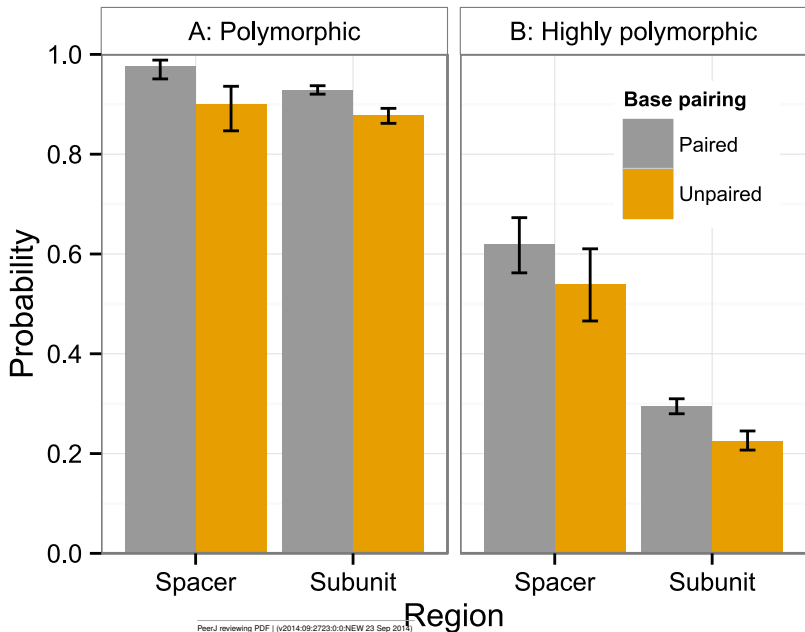
A. nyctaginifolia A. Gray	Fishbein 6268 [OKLA]	109	1.92	23	0.40
A. obovata Elliott	Lynch 11543 [OKLA]	708	12.75	87	1.57
A. oenotheriodes Schltdl. & Cham.	Fishbein 5819 [OKLA]	230	3.95	15	0.26
A. oenotheroides	Lynch 13339 [OKLA]	100	1.74	16	0.28
A. oenotheroides	Lynch 11477 [OKLA]	134	2.30	7	0.12
A. otarioides E. Fourn.	Bellsey 97-5 [ARIZ]	745	12.80	35	0.60
A. otarioides	Lynch 1533 [OKLA]	659	11.32	20	0.34
A. otarioides	Fishbein 5857 [OKLA]	697	11.95	48	0.82
A. ovalifolia Decne.	Lynch 13546 [OKLA]	333	5.91	42	0.75
A. ovata M. Martens & Galeotti	Laferrière 1478 [MO]	324	5.59	39	0.67
A. pellucida E. Fourn.	Fishbein 5136 [OKLA]	439	7.53	9	0.15
A. perennis Walter	Lynch 12408 [OKLA]	356	6.11	15	0.26
A. pilgeriana Schltr. ("*flava*" in Fishbein et al., 2011)	Zuloaga & Morrone 7069 [OKLA]	458	7.85	23	0.39
A. pratensis Benth.	Fishbein 5143 [OKLA]	410	7.11	29	0.50
A. pratensis	Pérez 1850 [MO]	609	10.49	34	0.59
A. prostrata W.H. Blackw.	Fishbein 2432 [ARIZ]	465	7.99	41	0.70
A. purpurascens L.	Lynch 12847 [OKLA]	233	4.05	22	0.38
A. purpurascens	Fishbein 5654 [MISSA]	216	3.73	7	0.12
A. quadrifolia Jacq.	Webb s.n. [ARIZ]	116	2.33	41	0.82
A. quadrifolia	Fishbein 6545 [OKLA]	204	3.55	31	0.54
A. rosea Kunth	Lynch 1656 [OKLA]	747	12.86	33	0.57
A. ruthiae Maguire	Riser 329 [WS]	31	0.61	23	0.45
A. sanjuanensis K.D. Heil, J.M. Porter & S.L. Welsh	Ellison s.n. [HPSU]	158	2.96	33	0.62
A. sanjuanensis	Fishbein 6525 [OKLA]	237	4.11	38	0.66
A. sanjuanensis	Riser 335 [WS]	68	1.18	26	0.45
A. scaposa Vail	Fishbein 2951 [ARIZ]	500	8.61	35	0.60
A. schaffneri A. Gray	Fishbein 5846 [OKLA]	327	5.65	36	0.62
A. scheryi Woodson	Fishbein 5137 [OKLA]	586	10.05	15	0.26
A. scheryi	Zamudio 5234 [MEXU]	272	4.88	35	0.63
A. similis Hemsl.	Fishbein 3000 [ARIZ]	294	5.05	8	0.14
A. similis	Fishbein 5148 [MISSA]	386	6.71	45	0.78
A. solanoana Woodson	Lynch 10884 [OKLA]	882	15.13	35	0.60
A. speciosa Torr.	Lynch 10981 [OKLA]	242	4.19	23	0.40
A. aff. *standleyi* Woodson	Reina 98-579 [WS]	150	2.57	19	0.33
A. subaphylla Woodson	Fishbein 3518 [WS]	182	3.31	23	0.42
A. subaphylla	Lynch 1008 [OKLA]	383	6.86	77	1.38
A. subulata Decne.	Fishbein 6434 [OKLA]	269	4.64	7	0.12
A. subulata	Fishbein 6446 [OKLA]	370	6.36	18	0.31
A. subulata × *albicans*	Fishbein 3142 [WS]	299	5.14	12	0.21

A. subverticillata (A. Gray) Vail	Fishbein 2948 [ARIZ]	160	3.09	92	1.77	
A. syriaca L.	Lynch 11138 [OKLA]	204	3.51	11	0.19	
A. syriaca	Fishbein 4885 [OKLA]	298	5.10	7	0.12	SRX040889
A. syriaca × speciosa	Fishbein 2810 [ARIZ]	161	2.89	37	0.66	
A. tomentosa Elliott	Fishbein 5608 [MISSA]	198	3.39	11	0.19	
A. tuberosa L. ssp. *interior* Woodson	Fishbein 2816 [ARIZ]	685	11.84	29	0.50	
A. tuberosa ssp. *interior*	Fishbein 4825 [MISSA]	297	5.10	31	0.53	
A. tuberosa ssp. *rolfsii* (Britton ex Vail) Woodson	Lynch 12526 [OKLA]	251	4.33	11	0.19	
A. uncialis Greene	Fishbein 6494 [OKLA]	282	4.86	19	0.33	
A. variegata L.	Lynch 12787 [OKLA]	375	6.50	29	0.50	
A. verticillata L.	Lynch 11102 [OKLA]	23	0.41	21	0.37	
A. vestita Hook. & Arn. ssp. *parishii* (Jeps.) Woodson	Lynch 10735 [OKLA]	506	8.68	14	0.24	
A. viridis Walter	Lynch 12955 [OKLA]	261	4.55	40	0.70	
A. viridula Chapm.	Fishbein 4806 [MISSA]	425	7.37	46	0.80	
A. welshii N.H. Holmgren & P.K. Holmgren	Lynch 11369 [OKLA]	364	6.53	73	1.31	
A. woodsoniana Standl. & Steyerm.	D. A. Neil 242 [MO]	122	2.28	30	0.56	

Polymorphic site frequency among species of *Asclepias.*

Histogram of polymorphic site frequency among species of *Asclepias*. Individuals contained from 0.4% to 15.1% polymorphic sites.

Polymorphic sites across the nrDNA cistron of *Asclepias.*

Number of individuals that are A) polymorphic and B) highly polymorphic at each position. Polymorphic positions are those with ≥2% of reads differing from the consensus; highly polymorphic positions are those with ≥10% differing reads. Subunit regions, white background; spacer regions, shaded background. Numbers in each region are the percentage of sites polymorphic or highly polymorphic in at least one individual.

Polymorphism probability by region and structure.

Probability that at least one individual is A) polymorphic or B) highly polymorphic at a position that is either within a spacer (ITS1, ITS2) or subunit region (18S, 5.8S, 26S), and either paired (stems) or unpaired (loops). Error bars indicate 95% confidence intervals. Values derived from two-factor multiple logistic regressions (Table 3).

ANOVA of the number of polymorphic individuals at nrDNA positions, by position type.

Two-way ANOVA of the number of polymorphic individuals at nrDNA positions categorized as either subunit (18S, 5.8S, 26S) or spacer regions (ITS1, ITS2), and as either paired (stems) or unpaired (loops). More individuals are likely to be polymorphic at sites that are in spacer regions over subunit regions, and that are paired over unpaired. **Bold** values indicate categories that significantly affect polymorphism abundance ($P<0.05$). *df* = degrees of freedom.

Source of variation	Sum of Squares	df	F value	P
(Intercept)	8310.6	1		
Paired	109.1	1	**53.5887**	**<0.0001**
Subunit	16.4	1	**8.0664**	**0.0045**
Paired*Subunit	0.4	1	0.1859	0.6663
Residuals	11878.9	5835		

Multiple logistic regression of the likelihood of nrDNA position polymorphism, by position type.

Two-factor multiple logistic regression of the likelihood of nrDNA positions being A) polymorphic or B) highly polymorphic in at least one individual. Positions are categorized as either within a subunit (18S, 5.8S, 26S) or spacer region (ITS1, ITS2), and as either paired (stem) or unpaired (loop). Odds ratios indicate whether a category decreases (<1) or increases (>1) the likelihood a position is polymorphic or highly polymorphic. The intercept represents paired, spacer positions. Categories that significantly affect polymorphism likelihood are indicated by *italics* ($P<0.1$) or **boldface** ($P<0.05$). CI = confidence interval, Std. Error = standard error. Polymorphism probabilities for each category are presented in Fig. 3.

A: Polymorphic

Source	Odds Ratio	95% CI	Coefficient estimate	Std. Error	Z-value	P
(Intercept)			3.7136	0.3825		
Subunit	**0.320**	**0.150 - 0.686**	**-1.1382**	**0.3881**	**-2.933**	**0.0034**
Unpaired	**0.220**	**0.090 - 0.537**	**-1.5163**	**0.4561**	**-3.324**	**0.0009**
*Subunit*Unpaired*	*2.488*	*0.998 - 6.206*	*0.9115*	*0.4663*	*1.955*	*0.0506*

B: Highly polymorphic

Source	Odds Ratio	95% CI	Coefficient estimate	Std. Error	Z-value	P
(Intercept)			0.48551	0.1201		
Subunit	**0.257**	**0.201 - 0.329**	**-1.35881**	**0.12564**	**-10.815**	**<0.0001**
Unpaired	*0.719*	*0.494 - 1.047*	*-0.32964*	*0.19178*	*-1.719*	*0.0856*
Subunit*Unpaired	0.970	0.651 - 1.444	-0.03061	0.20315	-0.151	0.8802

Phylogenetic signal tests across *Asclepias* for the number of polymorphic or highly polymorphic sites.

Tests for phylogenetic signal across *Asclepias* for the number of polymorphic or highly polymorphic positions across the entire nrDNA cistron (Subunits + spacers) or just the subunits (Subunits only). **Parsimony permutations**: The proportion of permutations with a shorter tree length than the true data. **Lambda**: The maximum likelihood estimate of lambda. **logL**: The log-likelihood ratio of the unconstrained model including the estimated lambda over the constrained model with lambda=0. **P**: The probability of obtaining a likelihood ratio this small or smaller by chance alone.

		Parsimony permutations	Lambda	$\log L$	P
Polymorphic sites	Subunits + spacers	<0.0012	0.51	$\dfrac{-594.192}{-597.863}$	0.0067
	Subunits only	<0.0200	0.45	$\dfrac{-587.547}{-590.765}$	0.0112
Highly polymorphic sites	Subunits + spacers	>0.6050	<0.0001	$\dfrac{-387.226}{-387.225}$	1
	Subunits only	>0.5300	<0.0001	$\dfrac{-376.150}{-376.149}$	1

Ancestral state reconstruction of polymorphic site abundance.

Ancestral state reconstruction of the number of polymorphic positions in nrDNA in *Asclepias* obtained with squared-change parsimony. The tree topology is that pruned from Fig. 2 of Fishbein et al. (2011) with clades indicated by letters, following that study.

Asclepias pratensis
Asclepias cordifolia
Asclepias macrotis
Asclepias coulteri
Asclepias cutleri
Asclepias leptopus
Asclepias albicans
Asclepias subaphylla
Asclepias masonii
Asclepias subulata
Asclepias subulata x Asclepias albicans
Asclepias californica ssp. californica
Asclepias vestita ssp. parishii
Asclepias curassavica
Asclepias pilgeriana
Asclepias candida
Asclepias boliviensis
Asclepias mellodora
Asclepias woodsoniana
Asclepias mexicana
Asclepias angustifolia
Asclepias perennis
Asclepias incarnata ssp. incarnata
Asclepias verticillata
Asclepias subverticillata NL
Asclepias scaposa
Asclepias humistrata
Asclepias cryptoceras ssp. davisii
Asclepias cryptoceras ssp. cryptoceras
Asclepias atroviolacea
Asclepias schaffneri
Asclepias hypoleuca
Asclepias otarioides
Asclepias sp. nov. cf. notha
Asclepias circinalis
Asclepias fournieri
Asclepias brachystephana
Asclepias rosea
Asclepias lemmonii
Asclepias ovata
Asclepias pellucida
Asclepias auriculata
Asclepias jorgeana
Asclepias jaliscana
Asclepias notha
Asclepias similis NL
Asclepias similis QRO
Asclepias scheryi JAL
Asclepias scheryi MICH
Asclepias feayi
Asclepias cinerea
Asclepias meadii
Asclepias tomentosa
Asclepias connivens
Asclepias viridula
Asclepias prostrata
Asclepias arenaria
Asclepias emoryi
Asclepias oenotheroides
Asclepias nyctaginifolia
Asclepias solanoana
Asclepias latifolia
Asclepias eriocarpa
Asclepias engelmanniana
Asclepias lynchiana
Asclepias glaucescens
Asclepias sp. nov. aff. glaucescens
Asclepias ruthiae
Asclepias uncialis
Asclepias sanjuanensis
Asclepias viridis
Asclepias asperula ssp. capricornu
Asclepias asperula ssp. asperula
Asclepias macrosperma
Asclepias welshii
Asclepias involucrata
Asclepias longifolia
Asclepias amplexicaulis
Asclepias lanceolata
Asclepias ovalifolia
Asclepias variegata
Asclepias purpurascens
Asclepias speciosa CA
Asclepias hallii
Asclepias quadrifolia
Asclepias syriaca
Asclepias tuberosa ssp. rolfsii
Asclepias michauxii
Asclepias tuberosa ssp. interior KS
Asclepias tuberosa ssp. interior MS
Asclepias obovata

	23.0 to 109.0
	109.0 to 195.0
	195.0 to 280.0
	280.0 to 367.0
	367.0 to 453.0
	453.0 to 538.0
	538.0 to 624.0
	624.0 to 710.0
	710.0 to 796.0
	796.0 to 850.0
	850.0 to 882.0

Number of
polymorphic
sites

Ancestral state reconstruction of highly polymorphic site abundance.

Ancestral state reconstruction of the number of highly polymorphic positions in nrDNA in *Asclepias* obtained with squared-change parsimony. The tree topology is that pruned from Fig. 2 of Fishbein et al. (2011) with clades indicated by letters, following that study.

Number of highly
polymorphic
sites

2.0 to 10.0
10.0 to 20.0
20.0 to 30.0
30.0 to 40.0
40.0 to 50.0
50.0 to 60.0
60.0 to 70.0
70.0 to 80.0
80.0 to 90.0
90.0 to 100.0
100.0 to 111.0