Divergent lineages in a young species: the case of Datilillo (*Yucca valida*), a broadly distributed plant from the Baja California Peninsula

Alberto Aleman[1,2], Maria Clara Arteaga[1]*, Jaime Gasca-Pineda[3], Rafael Bello-Bedoy[4]

[1]Departamento de Biología de la Conservación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, Baja California, México

[2]Environmental and Life Sciences Graduate Program, Trent University, Peterborough, Ontario, Canada

[3]Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, México

[3]Facultad de Ciencias, Universidad Autónoma de Baja California (UABC), Ensenada, Baja California, Mexico

***Author for correspondence:** Maria Clara Arteaga: arteaga@cicese.mx

Abstract

Globally, barriers triggered by geological and climatic events have produced temporary and permanent environmental breaks, habitat fragmentation, and species divergence. Notably, diverse plant species from the Baja California Peninsula, in western North America, exhibit strong genetic structure and vicariant lineages along their range. A representative taxon of the Peninsula is the genus *Yucca*, with *Yucca valida* having the most extensive range. Although its presence is dominant, there is an extensive distribution discontinuity between 26° N and 27° N, where any populations have been identified, suggesting restricted gene flow among its populations; moreover, the potential distribution models indicate a substantial reduction in their range during the Last Interglacial. We examined the phylogeographic patterns of *Y. valida* across the species' range, expecting to find vicariant lineages. Using nextRAD sequencing, we recovered 4,411 SNPs on 147 plants from 20 locations, identifying three allopatric lineages. The simulation of historical scenarios supported the simultaneous divergence of these lineages. Finally, we retrieved whole-chloroplast-genome sequences and estimated an age below a million years old for the MRCA of *Y. valida* and its sister species, *Y. capensis*. Our results indicate that the origin of the nuclear lineages can be the consequence of habitat loss during historical climatic changes.

Subject areas: Genomics**,** Evolution, Phylogeography.

Keywords: Isolation, genetic differentiation, chloroplast-genome assembly, *Yucca* phylogenetics, nextRAD sequencing.

1. Introduction

Genetic differentiation, the formation of two or more independent lineages from an ancestral one, can be described by two main frameworks: enforced divergence, in which demographic fluctuations or populations' extrinsic barriers lead to genetic isolation [1–3], and divergence in the face of gene flow, in which selection and local adaptation drive differentiation by restricting the genetic exchange of certain loci [4–6]. Understanding the mechanisms that lead to genetic differentiation is fundamental for reconstructing the evolutionary histories of species and permits relating these histories to a given region's climatic and geological history [7].

Globally, barriers triggered by geological and climatic events have caused temporary and permanent physical and environmental breaks, leading to habitats fragmentation and, consequently, populations enforced divergence [8]. One example of this are the Quaternary climatic fluctuations, which had an important role in the North American wildlife distributions [9]; correspondingly, diverse North American species' genetic divergence can be correlated to their Quaternary demographic histories [10, e.g., 11]. Besides, since climatic cyclic oscillations also produced habitats relinkages, in many instances, divergence followed by a secondary contact has been a common process for multiple species across the Northern Hemisphere [e.g., 12,13].

Broadly distributed plant species are valuable models for studying the causes and consequences of genetic differentiation [14]. Under wide geographic ranges, plant populations are expected to be particularly subject to demographic fluctuations related to genetic drift (e.g., habitat fragmentation) and gene flow (e.g., changes in their dispersal dynamics) [15]. Notably, diverse desert plant species from the Baja California Peninsula,

located in western North America, exhibit strong genetic structure and highly spatially concordant differentiated lineages across their ranges. Populations of cacti, ocotillos, palms, trees, and agaves from Baja California were isolated during the Last Interglacial (LIG, ~120,000 years ago) and became genetically differentiated over time [16–22]; as, while during the Last Glacial Maximum (LGM, ~20,000 years ago), the deserts in the Northern Hemisphere increased their distribution when the climate was getting drier, the Interglacial was highly humid, producing a significant fragmentation on arid ecosystems [23,24].

A representative taxon of the Baja California Peninsula is the genus *Yucca* (Asparagaceae), a monocotyledonous, long-lived, semi-evergreen tree with woody stems, lanceolate succulent green leaves, and white rosette-shaped inflorescences. It is characterised by a long-life cycle and slow growth and population recruitment rates [25]. Its pollination depends on an obligate mutualism with moths belonging to the genera *Tegeticula* and *Parategeticula*, which disperse pollen over short distances [26]. Seed dispersal has been related to the extinct North American megafauna and is currently mediated by rodents, birds, and ruminants [27]. Across the Peninsula, there are three species of the genus: *Yucca schidigera*, *Yucca capensis,* and *Yucca valida*, the latter two considered sister and endemic species [28], in addition to being pollinated by a single moth species, *Tegeticula baja* [29]. Among the *Yuccas* of the Peninsula, *Y. valida* has the most extensive range, from ~29.8° N to ~25.6° N, across the Central and Vizcaino Deserts, and the Magdalena Plains [25,30]. Although its presence is dominant and continuous across its northern distribution, there is an extensive discontinuity between the 26° N and 27° N (our personal observation), where any individuals have been identified,

83 suggesting restricted gene flow rates among its populations. Furthermore, while *Y. valida*
84 is currently a dominant species in the Baja California desert, the historical distribution
85 models indicate a substantial reduction in its geographic range during the LIG [31]. As a
86 result, this species is an interesting model for studying genetic differentiation across the
87 Baja California Peninsula.

88 We examined the phylogeographic patterns of *Y. valida* throughout the species'
89 geographic range. We hypothesised that past climatic fluctuations caused populations
90 isolation, and since the species has a short-distance dispersal [32], we expected to find
91 divergent lineages across its distribution. We genotyped 160 plants from 20 locations by
92 NextRAD sequencing and aimed to i) identify the number of genetic lineages of *Y. valida*
93 across its range, ii) reconstruct its populations' demographic history and iii) estimate the
94 species' age. This information will contribute to the understanding of the evolutionary
95 processes shaping species' genetic differentiation in North America and the Baja
96 California Peninsula.

2. Materials and methods

(a) Sampling, DNA extraction and sequencing

Fresh leaf tissue samples of 160 plants were collected from 20 locations across the species' range (Figure 1a, Supplementary Table S1). All samples were dried in desiccant silicate gel. We used 100 mg of each dried sample for DNA extraction following the Qiagen DNeasy Plant Mini Kit protocol (Qiagen, Hilden, Germany). The extractions' quality and concentration were evaluated using GelRed 1.5% agarose gels and a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Each DNA sample was prepared for high-throughput sequencing libraries by enzymatic fragmentation followed by ligation of short adapter sequences with Nextera reagent (SNPsaurus, LLC) [33]. The reaction was scaled to fragment 20 ng of DNA, and 50 ng of DNA were used to compensate for the amount of degradation DNA in the samples and to increase the fragments' size. The fragmented DNA was amplified for 27 cycles at 74°C. The libraries were sequenced on a HiSeq4000 150 bp single-end line (University of Oregon).

(b) Raw data processing and SNP–calling

The quality of the demultiplexed raw sequences was evaluated in FastQC 0.11.9 [34] and MultiQC 1.14 [35]. Adapter removal was carried out with custom scripts (SNPsaurus, LLC) using the *bbduk* module of BBTools 1.0 [36]. Clean reads were forced to 110 bp in length. A *de novo* loci assembly was performed through Stacks 2.6 (*denovo_map.pl*) [37]. The parameter optimisation was based on [38] and [39], and 13 samples with less than one million raw reads were removed as a quality control filter before the assembly. We established a minimum stack depth equal to eight (*-m* 8), a

120 maximum of three mismatches between stacks (*-M* 3), and a maximum of four
121 mismatches between catalogue loci (*-n* 4). In the module *populations*, we called all the
122 SNPs present across the remaining 147 samples with a minimum genotyping rate of 0.5
123 and a minor allele frequency of 0.05 (*-r 0.5 --maf 0.05*), producing a variant call format file.
124 When a 110 bp locus showed more than one SNP, only one variant was chosen, following
125 the priorities in [40] to ensure the SNPs were mostly unlinked [41].

126 **(c) Population structure and genetic diversity**

127 We first used nuclear SNPs to assess the most likely number of genetic clusters
128 across all samples and the membership of each plant to these clusters with three
129 complementary approaches. ADMIXTURE 1.3.0 [42] was run (K = 1-20), and the optimal
130 number of clusters was chosen via the cross-validation procedure. A neighbour-joining
131 tree from the samples' pairwise genetic distance matrix (transformed on the R 4.2.2
132 package ape [43,44]) and a Principal Component Analysis (PCA) without an *a priori*
133 assumption were produced in Plink 1.90 [45]. To avoid over- or underestimating
134 population genetic structure [46], we verified and conciliated the assignment of samples
135 to genetic clusters for the three approaches. Once the samples were allocated into
136 discrete genetic clusters (corresponding to three lineages, see *Results*), the genetic
137 diversity descriptors (H_E, H.O., and F_{IS}) for each lineage and the mean pairwise Weir and
138 Cockerham's weighted values of genetic differentiation (*standardised* F_{ST}; [47]) were
139 calculated with GENODIVE 3.0 [48]. To assess the isolation by distance null hypothesis,
140 i.e., the correlation between the geographical and genetic distance between samples [49],
141 a Mantel test was carried out correlating the individuals' Euclidean genetic distance and
142 the pairwise distance between sampling locations (in kilometres, produced with [50]),

143 introducing a random jitter (± 0.05°) to samples belonging to shared sampling locations.
144 The significance of the Mantel test was evaluated with 99,999 permutations. Any required
145 genetic data conversion was completed in PGDSpider 2.1.1.5 [51]

146 **(d) Outlier SNPs and signatures of differentiation by local adaptation**

147 First, we scanned for outlier SNPs in PCAdapt 4.0.1 [52], using K = 3, and α = 0.1
148 for the false discovery rate boundary. Then, to evaluate if the genetic differentiation was
149 a consequence of local adaptation, we used the Latent Factor Mixed-effect Model [53] to
150 characterise the presence of SNPs with a strong association with environmental variables.
151 The optimal number of latent factors (K) was determined using the cross-entropy criterion
152 for K = 1-20, and LFMM was run ten times with K = 3 with the R package LEA 3.2.0 [54].
153 A total of 20 environmental variables obtained from WorldClim v1.4 [55] were tested for
154 each sampling location. The Gibbs sampler algorithm was used for 10,000 cycles (burn-
155 in of 5,000 cycles), and the median of the correlation scores (z-score) was calculated
156 across all runs, recalibrating the p-values frequency distribution with lambda. We avoided
157 any complimentary correction procedures for the false discovery rate in both analyses to
158 increase the probability of detecting outliers and SNPs correlated with local adaptation.
159 LFMM and PCAdapt results were compared to search for overlaps on outlier loci and
160 SNPs with environmental associations as a signature of selection and local adaptation.
161 The SNPs detected by any procedure were removed from the dataset to produce a
162 catalogue of neutral SNPs for the demographic history analyses.

163 **(e) Demographic history**

164 To understand the succession of events that could originate the observed nuclear
165 genetic lineages, the reconstruction of the demographic history of *Y. valida* was executed
166 through the diffusion approximation method in δaδi [56]. Simulations included the four
167 possible three-populations simple models of i) simultaneous, ii) admixed, and consecutive
168 divergence from iii) North to South and iv) South to North, without gene flow or changes
169 in populations' sizes, as developed by [57]. The site frequency spectrum was computed
170 as folded, and the best model was selected using the maximum likelihood values and the
171 Akaike information criterion (AIC).

172 **(f) Putative whole chloroplast genome sequences recovery**

173 Following [58], we implemented a reference-guided workflow to reconstruct whole-
174 chloroplast-genome sequences (as a by-product of the enzymatic fragmentation for the
175 high-throughput sequencing libraries preparation without isolating cpDNA). The clean
176 reads of *Y. valida* and 40 clean reads of *Y. capensis* from [31] were mapped to the
177 chloroplast reference genome of *Y. schidigera* (GenBank: NC_032714.1) in bowtie 2.2.6
178 under the very-sensitive-local parameter set [59]. Mapping statistics were evaluated with
179 the samtools 1.15.1 flagstat and coverage modules [60]. Nucleotide calling was performed
180 individually for each of the 200 samples in ANGSD [61] using the reads that mapped to
181 the chloroplast genome reference, requiring minimum mapping and base qualities of 20,
182 and using Ns as missing data *(-dofasta 2 -minMapQ 20 -minQ 20 -doCounts 1*).

183 **(g) Phylogenetic relationships and molecular clock analyses**

184 The intraspecific phylogenetic relationships of *Y. valida* were reconstructed using
185 the sequences with a breadth of coverage >0.5 (130 samples) through RAxML 8.2.12 [62],
186 with *Y. schidigera* as the outgroup. A general time reversible model was established,
187 assuming gamma-distributed rates with four substitution categories and estimating the
188 proportion of invariant sites (GTR + G + I), letting RaxML halt automatic bootstrapping.
189 The best-scoring maximum likelihood tree was plotted with iTOL 6 [63].

190 As the intraspecific phylogenetic analysis revealed no signals of divergence among
191 the chloroplast genomes of *Y. valida* (see *Results*), the divergence time of *Y. valida* and
192 *Y. capensis* within the genus was estimated as an approximation to the age of *Y. valida*.
193 The phylogenetic relationships of the Agavoideae subfamily were reconstructed based on
194 a relaxed molecular clock replicating [64] with BEAST 1.8.4 [65]. Based on [66] and [67],
195 we produced consensus sequences for *Y. valida* and *Y. capensis*, using all the available
196 sequences for each species. The consensus sequences were aligned to the chloroplast
197 reference genomes of 16 species of the Agavoideae (*Agave attenuata*, *Beschorneria*
198 *septentrionalis*, *Camassia scilloides*, *Chlorogalum pomeridianum*, *Hesperaloe*
199 *campanulata*, *Hesperaloe parviflora*, *Hesperocallis undulata*, *Hesperoyucca whipplei*,
200 *Hosta ventricosa*, *Manfreda virginica*, *Schoenolirion croceum*, *Yucca brevifolia*, *Yucca*
201 *filamentosa*, *Yucca queretaroensis*, *Yucca treculeana* and *Yucca schidigera*)
202 (Supplementary Table S2) following [64] by applying MAFFT 7 default settings [68], with
203 the flag *--nzero*. A general time reversible model estimating the proportion of invariant
204 sites (GTR + I) was employed. Gamma-distributed priors for the five substitution types
205 and uniform priors (0-1) for the base frequencies and the proportion of invariant sites were

206 established. An uncorrelated relaxed clock model with log-normal distributed rates was
207 set up (mean = 1, standard deviation = 0.33), assuming a Yule process for the tree prior,
208 starting from a random tree, establishing *Hosta ventricosa* as the outgroup, and forcing
209 the ingroup's monophyly [65]. A log-normal prior distribution on the age of the ingroup was
210 specified to 14.5 million years (offset, mean = 1 standard deviation = 1.4), based on the
211 minimum age of the fossil *Protoyucca shadesii* [69,70], completing two independent runs
212 of 60 million steps long with a sample every 6,000,000 generations. The runs were joined
213 in LogCombiner 1.8.4, considering a 10% burn-in for each input. A maximum credibility
214 tree was produced in TreeAnnotator 1.8.4, and each node's means, medians, and 95%
215 confidence intervals were outputted.

Results

Population structure, genetic diversity and demographic history

Using 4,411 SNPs, the nuclear genetic structure analyses established the most likely number of genetic lineages as three ($K = 3$) (Figure 2). The lineages were named after their geographic location as Northern, Central, and Southern. The three lineages were significantly different ($p = 0.001$), with the F_{ST} between North and South being the greatest (North-South: 0.11; North-Central: 0.05; Central-South: 0.09). The overall observed heterozygosity was lower than the expected (H.O. = 0.26, $H_E = 0.30$), and a mean coefficient of inbreeding (F_{IS}) of 0.17 (Table 1) was observed. The Mantel test detected a positive and significant association between the genetic distance and the geographic differentiation; however, the correlation and slope values were neglectable (p = >0.0000, r = 0.25, y = 0.01x + 33, $r^2 = 0.062$); thus the isolation by distance hypothesis was rejected.

LFMM identified 63 (1%) unique SNPs potentially associated with adaptation to multiple environmental variables, and PCAdapt detected 20 (0.5%) putative outlier SNPs. Most SNPs associated with environmental adaptation were related to annual precipitation. PCAdapt detected none of the SNPs detected by LFMM.

The maximum likelihood results and the AIC indicated that the optimal demographic model was the simultaneous divergence of the three nuclear genetic lineages from a common ancestral population.

236 **Phylogenetic relationships and molecular clock analyses**

237 The intraspecific phylogenetic analysis grouped the *Y. valida* chloroplast
238 sequences (mean depth = 30×) into a single monophyletic clade without intraspecific
239 lineages. Low support nodes were observed, revealing no differentiation within the
240 chloroplast genomes.

241 The interspecific phylogenetic relationships and age estimates for the Agavoideae
242 subfamily results mirrored [64], and all interspecific nodes were represented with high
243 support values (100%). The molecular clock calculated an age below a million years old
244 for the MRCA of *Y. valida* and *Y. capensis* (mean ≈ 731,000; median ≈ 650,000; 95% C.I.
245 ≈ 310,000 – 1,445,000).

246 **Discussion**

247 The genetic differentiation patterns of taxa from the Baja California Peninsula
248 desert suggest that the distribution of the genetic diversity of the species in this region has
249 been influenced by the geological processes that gave rise to the Peninsula, as well as
250 by the climatic oscillations that occurred throughout the last glacial cycles [71]. Here, we
251 identified three monophyletic nuclear lineages throughout the range of *Y. valida*, with a
252 single chloroplast lineage, whose MRCA with its sister species' estimated age (below one
253 million years) suggests that the nuclear differentiation in *Y. valida* is a consequence of a
254 recent event. While the time of origin of the three nuclear lineages is uncertain, ecological
255 niche model projections predicted that during the Last Interglacial, the area with suitable
256 conditions for the species suffered a substantial reduction and started growing back since
257 the LGM [31]. These climatic oscillations could act as a vicariant event, which, coupled
258 with a limited dispersal capacity and an extensive geographical range gap impeding the
259 secondary contact of populations, led to populations' differentiation [72].

260 Subtle values of genetic differentiation were detected among the three nuclear
261 lineages, the highest being between bordering lineages (paired F_{ST} North-South = 0.11).
262 This value is greater than what was reported in microsatellites for *Y. valida* (F_{ST} = 0.059;
263 [32]), *Y. brevifolia* (F_{ST} = 0.061; [73]), *Y. schidigera* (F_{ST} = 0.067; [74]), and its sister
264 species, *Y. capensis* (F_{ST} = 0.022; [75]). The levels of genetic differentiation between the
265 Central and the Southern lineages (F_{ST} = 0.09) support our idea that the absence of
266 populations between the 26° N and 27° N has influenced the differentiation of these
267 lineages. The overall observed subtle differentiation can be explained by potentially large
268 historical effective population sizes [31] and reflects limited gene flow between the

269 neighbouring lineages. Previous estimates of the dispersal rates indicate that pollen
270 movement in *Y. valida* is approximately 50 m [32], and seed-mediated dispersal seems to
271 also occur at short distances [76]. In this sense, the genetic differentiation patterns
272 suggest that the historical fragmentation and the accumulation of restricted dispersal over
273 long distances have been the major drivers of genetic isolation in this species.

274 The estimated genetic diversity seems to be equally distributed among the three
275 lineages since we found similar levels of observed heterozygosity. The genetic diversity
276 estimates (H.O. = 0.26, H_E = 0.31) were higher than those previously described for this
277 species (H.O. = 0.11, H_E = 0.17; [31]) with the same kind of molecular markers. The
278 disparity among these values reiterates the need for additional discussions to ascertain
279 the accuracy of SNP discovery in population genomics [77]. *De novo* assemblies seek to
280 maximise the number of SNPs shared by a representative number of individuals [37]. [31]
281 assembled loci from samples of three taxa (*Y. valida*, *Y. capensis*, and their hybrid).
282 Creating an SNP catalogue with samples of a single species modifies the matrix of
283 recovered sites, thus, the measurements of genetic diversity [78]. Performing SNP-calling
284 among differentiated taxa or structured samples implies retrieving loci fixed for a species
285 or a population that are polymorphic for all the samples as a whole (e.g., species-specific
286 SNPs, that are loci with alternative SNPs fixed between two taxa) or that are variable only
287 within a structured lineage. This untrivial catalogue-size bias highlights that diversity
288 estimations based on SNPs should be carefully exposed. Correspondingly, the inbreeding
289 coefficient had lower levels (F_{IS} = 0.17) than those previously registered ([31] F_{IS} = 0.35).
290 The estimates of relatively high genetic diversity and low inbreeding detected in this work

291 are congruent with the hypothetical high historical population sizes and the *Yuccas*
292 reproductive system, which supports interbreeding and limits self-pollination [79].

293 The historical demographic models' simulations indicate that the nuclear lineages'
294 divergence is associated with a single vicariant. The divergence time estimate from
295 chloroplast genome sequences between *Y. valida* and *Y. capensis*, under one million
296 years ago, and the lack of intraspecific differentiation within the chloroplast genome
297 indicate that the drivers of genetic differentiation in *Y. valida* happened long enough to be
298 recorded in the nucleus, and so recently to not be registered in the chloroplast; supporting
299 the role of a young event, as the Quaternary climatic changes, which have shaped the
300 distribution of the genetic variation of other plant species across the Peninsula [16–22],
301 by reducing population sizes during the last Interglacial. Additionally, the overall
302 comparable high genetic diversity suggests that the species was already distributed
303 before this divergent phenomenon occurred; thus, the history of the nuclear lineages does
304 not correspond to a colonisation, but a divergent process.

305 While the contribution of selection in the demographic history of *Y. valida* was
306 unrevealed in this study, the consequences of this force on the genetic diversity of this
307 species should not be ruled out [80]. The geographic extent where each of the genetic
308 lineages is distributed coincides with three different ecoregions: the Central Desert
309 (Northern lineage), the Vizcaino Desert (Central lineage) and the Magdalena Plains
310 (Southern lineage) [81]. Further studies should address the contribution of these
311 environments' ecological differences on *Y. valida*. Furthermore, generating a high-quality
312 nuclear reference genome could provide an essential resource for future genomic
313 analyses to deepen our understanding of this evolutionarily dynamic genus.

Acknowledgements

We respectfully acknowledge the territory where the samples were collected as the ancestral homelands of the Cochimí peoples. Data analyses were conducted at Centro de Investigación Científica y Educación Superior de Ensenada, which is on the traditional territory of the Ku'ahles, Cochimíes, Pa ipais and Kiliwas. We are grateful to Cynthia Rocío Álamo-Herrera, Leonardo de la Rosa-Conroy, Astrid Luna-Ortiz, Mario Salazar, José Delgadillo, and Lita Castañeda for their contributions in the laboratory and during the field. We thank Camille Kessler and Flavia Termignoni-García for their bioinformatic advice.

Funding

This work was financially supported by Consejo Nacional de Ciencia y Tecnología (CONACyT) (CB-2014-01-238843, infra-2014-1-226339), from which Alberto Aleman also received a master's degree scholarship (Grant 748805). The Rufford Foundation (RSG 13704-1) and the JiJi Foundation provided financial support for part of this study. The Life Science Postgraduate Program at CICESE provided the facilities to complete this study.

References

1. Barton NH. 2008 The effect of a barrier to gene flow on patterns of geographic variation. *Genetics Research* **90**, 139–149. (doi:10.1017/S0016672307009081)

2. Mayr E, Mayr AAP of Z (Emeritus) M of CZE, Mayr EW. 1963 *Animal Species and Evolution*. Belknap Press of Harvard University Press.

3. Scheiner SM, Mindell DP, editors. 2020 *The Theory of Evolution: Principles, Concepts, and Assumptions*. Chicago, IL: University of Chicago Press. See https://press.uchicago.edu/ucp/books/book/chicago/T/bo45713136.html.

4. Darwin C. 1859 *On the Origin of Species by Means of Natural Selection, Or The Preservation of Favoured Races in the Struggle for Life*. John Murray, Albemarle Street.

5. Stankowski S, Ravinet M. 2021 Defining the speciation continuum. *Evolution* **75**, 1256–1273. (doi:10.1111/evo.14215)

6. Tigano A, Friesen VL. 2016 Genomics of local adaptation with gene flow. *Molecular Ecology* **25**, 2144–2164. (doi:10.1111/mec.13606)

7. Avise JC. 2000 *Phylogeography: The History and Formation of Species*. Harvard University Press. (doi:10.2307/j.ctv1nzfgj7)

8. Kumar R, Kumar V. 2018 A review of phylogeography: biotic and abiotic factors. *Geology, Ecology, and Landscapes* **2**, 268–274. (doi:10.1080/24749508.2018.1452486)

9. Shafer AB, Cullingham CI, Cote SD, Coltman DW. 2010 Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular ecology* **19**, 4589–4621.

10. Hewitt G. 2000 The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913. (doi:10.1038/35016000)

11. de Oliveira FFR, Gehara M, Solé M, Lyra M, Haddad CFB, Silva DP, de Magalhães RF, Leite FSF, Burbrink FT. 2021 Quaternary climatic fluctuations influence the demographic history of two species of sky-island endemic amphibians in the Neotropics. *Molecular Phylogenetics and Evolution* **160**, 107113. (doi:10.1016/j.ympev.2021.107113)

12. Harrington SM, Hollingsworth BD, Higham TE, Reeder TW. 2018 Pleistocene climatic fluctuations drive isolation and secondary contact in the red diamond rattlesnake (Crotalus ruber) in Baja California. *Journal of Biogeography* **45**, 64–75. (doi:10.1111/jbi.13114)

13. Rödin-Mörch P, Luquet E, Meyer-Lucht Y, Richter-Boix A, Höglund J, Laurila A. 2019 Latitudinal divergence in a widespread amphibian: Contrasting patterns of neutral and adaptive genomic variation. *Molecular Ecology* **28**, 2996–3011. (doi:10.1111/mec.15132)

14. Petit RJ, Hampe A. 2006 Some Evolutionary Consequences of Being a Tree. *Annual Review of Ecology, Evolution, and Systematics* **37**, 187–214. (doi:10.1146/annurev.ecolsys.37.091305.110215)

15. Sork VL, Gugger PF, Chen J-M, Werth S. 2016 Evolutionary lessons from California plant phylogeography. *Proceedings of the National Academy of Sciences* **113**, 8064–8071. (doi:10.1073/pnas.1602675113)

16. Clark-Tapia R, Molina-Freaner F. 2003 The genetic structure of a columnar cactus with a disjunct distribution: Stenocereus gummosus in the Sonoran desert. *Heredity* **90**, 443–450. (doi:10.1038/sj.hdy.6800252)

17. Fehlberg SD, Ranker TA. 2009 Evolutionary history and phylogeography of Encelia farinosa (Asteraceae) from the Sonoran, Mojave, and Peninsular Deserts. *Molecular Phylogenetics and Evolution* **50**, 326–335. (doi:10.1016/j.ympev.2008.11.011)

18. Garrick RC, Nason JD, Meadows CA, Dyer RJ. 2009 Not just vicariance: phylogeography of a Sonoran Desert euphorb indicates a major role of range expansion along the Baja peninsula. *Molecular Ecology* **18**, 1916–1931. (doi:10.1111/j.1365-294X.2009.04148.x)

19. Klimova A, Hoffman JI, Gutierrez-Rivera JN, Leon de la Luz J, Ortega-Rubio A. 2017 Molecular genetic analysis of two native desert palm genera, Washingtonia and Brahea, from the Baja California Peninsula and Guadalupe Island. *Ecology and Evolution* **7**, 4919–4935. (doi:10.1002/ece3.3036)

20. Klimova A, Ortega-Rubio A, Vendrami DLJ, Hoffman JI. 2018 Genotyping by sequencing reveals contrasting patterns of population structure, ecologically mediated divergence, and long-distance dispersal in North American palms. *Ecology and Evolution* **8**, 5873–5890. (doi:10.1002/ece3.4125)

21. Martínez-Noguez JJ, León de la Luz JL, Delgadillo Rodríguez J, García-De León FJ. 2020 Phylogeography and genetic structure of an iconic tree of the Sonoran Desert, the Cirio (Fouquieria columnaris), based on chloroplast DNA. *Biological Journal of the Linnean Society* **130**, 433–446. (doi:10.1093/biolinnean/blaa065)

22. Nason JD, Hamrick JL, Fleming TH. 2002 Historical Vicariance and Postglacial Colonization Effects on the Evolution of Genetic Structure in Lophocereus, a Sonoran Desert Columnar Cactus. *Evolution* **56**, 2214–2226. (doi:10.1111/j.0014-3820.2002.tb00146.x)

398 23. Meng H-H, Gao X-Y, Huang J-F, Zhang M-L. 2015 Plant phylogeography in arid
399 Northwest China: Retrospectives and perspectives. *Journal of Systematics and*
400 *Evolution* **53**, 33–46. (doi:10.1111/jse.12088)

401 24. Qiu Y-X, Fu C-X, Comes H.P. 2011 Plant molecular phylogeography in China and
402 adjacent regions: Tracing the genetic imprints of Quaternary climate and
403 environmental change in the world's most diverse temperate flora. *Molecular*
404 *Phylogenetics and Evolution* **59**, 225–244. (doi:10.1016/j.ympev.2011.01.012)

405 25. Turner RM, Bowers JE, Burgess TL. 2005 *Sonoran Desert Plants: An Ecological*
406 *Atlas*. University of Arizona Press.

407 26. Pellmyr O. 2003 Yuccas, Yucca Moths, and Coevolution: A Review. *Annals of the*
408 *Missouri Botanical Garden* **90**, 35–55. (doi:10.2307/3298524)

409 27. Lenz L. 2001 Seed Dispersal in Yucca Brevifolia (Agavaceae)-Present and Past,
410 With Consideration of the Future of the Species. *aliso* **20**, 61–74.
411 (doi:10.5642/aliso.20012002.03)

412 28. Pellmyr O, Segraves KA, Althoff DM, Balcázar-Lara M, Leebens-Mack J. 2007 The
413 phylogeny of yuccas. *Molecular Phylogenetics and Evolution* **43**, 493–501.
414 (doi:10.1016/j.ympev.2006.12.015)

415 29. Pellmyr O, Balcázar-Lara M, Segraves KA, Althoff DM, Littlefield RJ. 2008
416 Phylogeny of the pollinating yucca moths, with revision of Mexican species
417 (Tegeticula and Parategeticula; Lepidoptera, Prodoxidae). *Zool J Linn Soc* **152**, 297–
418 314. (doi:10.1111/j.1096-3642.2007.00361.x)

419 30. Thiede J. 2020 Yucca AGAVACEAE. In *Monocotyledons* (eds U Eggli, R Nyffeler),
420 pp. 363–421. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-662-56486-8_109)

421 31. Arteaga MC, Bello-Bedoy R, Gasca-Pineda J. 2020 Hybridisation Between Yuccas
422 From Baja California: Genomic and Environmental Patterns. *Frontiers in Plant*
423 *Science* **11**.

424 32. Alamo-Herrera CR, Arteaga MC, Bello-Bedoy R, Rosas-Pacheco F. 2022 Pollen
425 dispersal and genetic diversity of Yucca valida (Asparagaceae), a plant involved in
426 an obligate pollination mutualism. *Biological Journal of the Linnean Society* **136**,
427 364–374. (doi:10.1093/biolinnean/blac031)

428 33. Russello MA, Waterhouse MD, Etter PD, Johnson EA. 2015 From promise to
429 practice: pairing non-invasive sampling with genomics in conservation. *PeerJ* **3**,
430 e1106. (doi:10.7717/peerj.1106)

431 34. Andrews S. 2017 FastQC: a quality control tool for high throughput sequence data.
432 2010.

35. Ewels P, Magnusson M, Lundin S, Käller M. 2016 MultiQC: summarise analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048. (doi:10.1093/bioinformatics/btw354)

36. In press. BBMap. *SourceForge*. See https://sourceforge.net/projects/bbmap/ (accessed on 11 January 2023).

37. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013 Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124–3140. (doi:10.1111/mec.12354)

38. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. 2015 Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimisation for population genetic inference. *Molecular Ecology Resources* **15**, 28–41. (doi:10.1111/1755-0998.12291)

39. Paris JR, Stevens JR, Catchen JM. 2017 Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360–1373. (doi:10.1111/2041-210X.12775)

40. Alexander A. 2020 GBS_SNP_filter v1.17.

41. Pearman WS, Urban L, Alexander A. 2022 Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular Ecology Resources* **22**, 2599–2613. (doi:10.1111/1755-0998.13646)

42. Alexander DH, Lange K. 2011 Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246. (doi:10.1186/1471-2105-12-246)

43. Paradis E, Schliep K. 2019 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. (doi:10.1093/bioinformatics/bty633)

44. R Core Team. 2022 R: The R Project for Statistical Computing. See https://www.r-project.org/ (accessed on 4 October 2022).

45. Purcell S *et al.* 2007 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575. (doi:10.1086/519795)

46. Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017 The K = 2 conundrum. *Molecular Ecology* **26**, 3594–3602. (doi:10.1111/mec.14187)

47. Weir BS, Cockerham CC. 1984 Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370. (doi:10.2307/2408641)

48. Meirmans PG. 2020 genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Molecular Ecology Resources* **20**, 1126–1131. (doi:10.1111/1755-0998.13145)

49. Meirmans PG. 2012 The trouble with isolation by distance. *Mol Ecol* **21**, 2839–2846. (doi:10.1111/j.1365-294X.2012.05578.x)

50. In press. Geographic Distance Matrix Generator. See https://biodiversityinformatics.amnh.org/open_source/gdmg/ (accessed on 11 April 2023).

51. Lischer HEL, Excoffier L. 2012 PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299. (doi:10.1093/bioinformatics/btr642)

52. Luu K, Bazin E, Blum MGB. 2017 pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17**, 67–77. (doi:10.1111/1755-0998.12592)

53. Frichot E, Schoville SD, Bouchard G, François O. 2013 Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Mol Biol Evol* **30**, 1687–1699. (doi:10.1093/molbev/mst063)

54. Frichot E, François O. 2015 LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**, 925–929. (doi:10.1111/2041-210X.12382)

55. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965–1978. (doi:10.1002/joc.1276)

56. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* **5**, e1000695. (doi:10.1371/journal.pgen.1000695)

57. Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, Rödel M-O, Blackburn DC, Fujita MK. 2017 Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Molecular Ecology* **26**, 5245–5263. (doi:10.1111/mec.14266)

58. Aleman A, Dorken ME, Shafer A, Patel T, Volkova PA, Freeland JR. 2023 Development of genomic resources for cattails (Typha), a globally important macrophyte genus. , 2023.04.21.537876. (doi:10.1101/2023.04.21.537876)

59. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359. (doi:10.1038/nmeth.1923)

60. Li H *et al.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

61. Korneliussen TS, Albrechtsen A, Nielsen R. 2014 ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356. (doi:10.1186/s12859-014-0356-4)

62. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)

63. Letunic I, Bork P. 2021 Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296. (doi:10.1093/nar/gkab301)

64. Smith CI, McKain MR, Guimond A, Flatz R. 2021 Genome-scale data resolves the timing of divergence in Joshua trees. *American Journal of Botany* **108**, 647–663. (doi:10.1002/ajb2.1633)

65. Drummond AJ, Bouckaert RR. 2015 *Bayesian Evolutionary Analysis with BEAST*. Cambridge: Cambridge University Press. (doi:10.1017/CBO9781139095112)

66. Scheunert A, Dorfner M, Lingl T, Oberprieler C. 2020 Can we use it? On the utility of de novo and reference-based assembly of Nanopore data for plant plastome sequencing. *PLOS ONE* **15**, e0226234. (doi:10.1371/journal.pone.0226234)

67. Roy NS, Jeong U, Na M, Choi I-Y, Cheong EJ. 2020 Genomic analysis and a consensus chloroplast genome sequence of Prunus yedoensis for DNA marker development. *Hortic. Environ. Biotechnol.* **61**, 859–867. (doi:10.1007/s13580-020-00265-3)

68. Katoh K, Rozewicki J, Yamada KD. 2019 MAFFT online service: multiple sequence alignment, interactive sequence choice and visualisation. *Briefings in Bioinformatics* **20**, 1160–1166. (doi:10.1093/bib/bbx108)

69. Iles WJD, Smith SY, Gandolfo MA, Graham SW. 2015 Monocot fossils suitable for molecular dating analyses. *Botanical Journal of the Linnean Society* **178**, 346–374. (doi:10.1111/boj.12233)

70. Tidwell WD, Parker LR. 1990 Protoyucca shadishii gen. et sp. nov., An arborescent monocotyledon with secondary growth from the Middle Miocene of Northwestern Nevada, U.S.A. *Review of Palaeobotany and Palynology* **62**, 79–95. (doi:10.1016/0034-6667(90)90018-E)

71. Dolby GA, Bennett SEK, Lira-Noriega A, Wilder BT, Munguía-Vega A. 2015 Assessing the Geological and Climatic Forcing of Biodiversity and Evolution

Surrounding the Gulf of California. *Journal of the Southwest* **57**, 391–455. (doi:10.1353/jsw.2015.0005)

72. Irwin DE. 2002 Phylogeographic Breaks Without Geographic Barriers to Gene Flow. *Evolution* **56**, 2383–2394. (doi:10.1111/j.0014-3820.2002.tb00164.x)

73. Starr TN, Gadek KE, Yoder JB, Flatz R, Smith CI. 2013 Asymmetric hybridisation and gene flow between Joshua trees (Agavaceae: Yucca) reflect differences in pollinator host specificity. *Molecular Ecology* **22**, 437–449. (doi:10.1111/mec.12124)

74. De la Rosa-Conroy L, Gasca-Pineda J, Bello-Bedoy R, Eguiarte LE, Arteaga MC. 2020 Genetic patterns and changes in availability of suitable habitat support a colonisation history of a North American perennial plant. *Plant Biology* **22**, 233–242. (doi:10.1111/plb.13053)

75. Luna-Ortiz A, Arteaga MC, Bello-Bedoy R, Gasca-Pineda J, León de la Luz JL, Domínguez-Cadena R, Medel-Narváez A. 2022 High genetic diversity and low structure in an endemic long-lived tree, Yucca capensis (Asparagaceae). *Plant Biology* **24**, 185–191. (doi:10.1111/plb.13346)

76. Waitman BA, Vander Wall S.B., Esque T.C. 2012 Seed dispersal and seed fate in Joshua tree (Yucca brevifolia). *Journal of Arid Environments* **81**, 1–8. (doi:10.1016/j.jaridenv.2011.12.012)

77. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. 2018 These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* **27**, 3193–3206. (doi:10.1111/mec.14792)

78. Shafer AB, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JB. 2017 Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution* **8**, 907–917.

79. Huth CJ, Pellmyr O. 2000 Pollen-Mediated Selective Abortion in Yuccas and Its Consequences for the Plant-Pollinator Mutualism. *Ecology* **81**, 1100–1107. (doi:10.2307/177181)

80. Rivas MJ, Domínguez-García S, Carvajal-Rodríguez A. 2015 Detecting the Genomic Signature of Divergent Selection in Presence of Gene Flow. *Curr Genomics* **16**, 203–212. (doi:10.2174/1389202916666150313230943)

81. González-Abraham CE, Garcillán PP, Ezcurra E. 2010 Ecorregiones de la península de Baja California: Una síntesis. *Boletín de la Sociedad Botánica de México* , 69–82.

82. Wickham H. 2016 *ggplot2*. Cham: Springer International Publishing. (doi:10.1007/978-3-319-24277-4)

83. In press. FigTree. See http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 21 May 2023).

84. McKain MR, McNeal JR, Kellar PR, Eguiarte LE, Pires JC, Leebens-Mack J. 2016 Timing of rapid diversification and convergent origins of active pollination within Agavoideae (Asparagaceae). *Am J Bot* **103**, 1717–1729. (doi:10.3732/ajb.1600198)

578 **Data Accessibility Statement**

579 The raw reads, SNPs, chloroplast sequences and the related metadata supporting
580 this study's findings will be deposited into online databases upon publication. In addition,
581 all the scripts used will be available, providing unique sample identifier tags that can be
582 matched to the deposited data.

Tables and Figures

Table 1. Global genetic diversity of *Y. valida* and its lineages using 4,411 nuclear SNPs. Number of individuals per lineage (n) and descriptors of genetic diversity (H.O., observed heterozygosity; H_E, expected heterozygosity; F_{IS}, inbreeding coefficient).

Lineage	N	H_O	H_E	F_{IS}
Northern	55	0.25	0.30	0.17
Central	72	0.26	0.30	0.17
Southern	21	0.26	0.32	0.17
Total	148	0.26	0.31	0.17

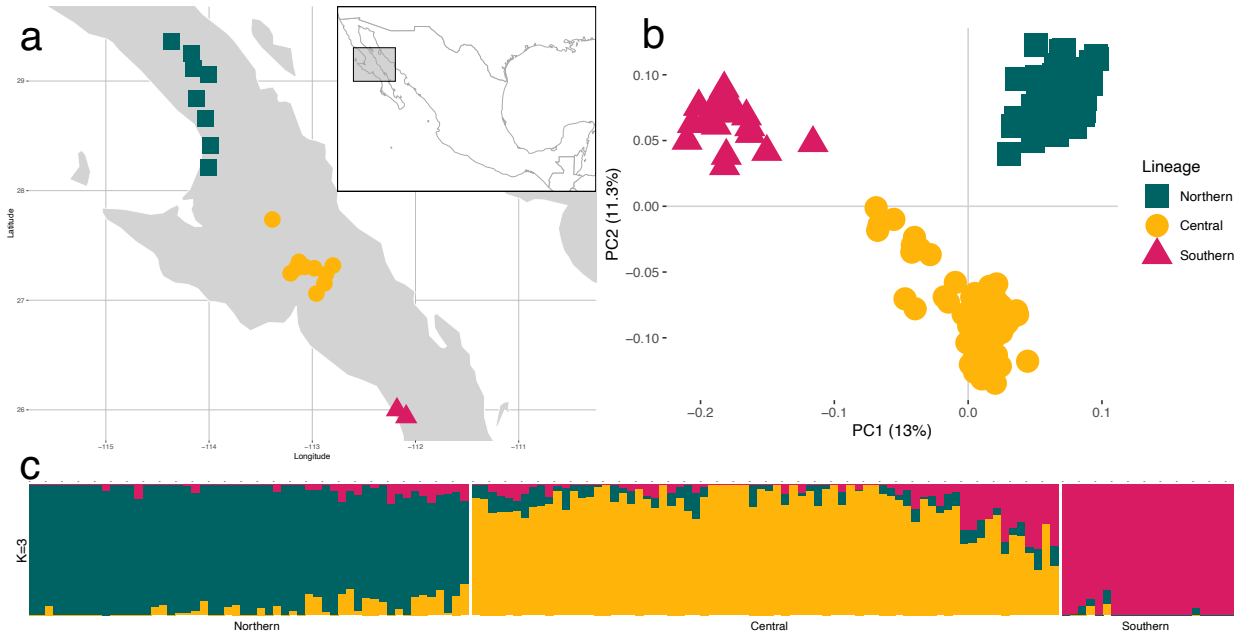

Figure 1. Genetic structure results of 4,411 nuclear SNPs obtained for 147 *Y. valida* samples (a) Sampling locations (20) in this study, colours and shapes represent genetic lineages (b) Principal Component Analysis for PC1 and 2. Shapes represent individuals, and colours represent genetic lineages, (c) ADMIXTURE analysis for K = 3. Horizontal bars represent individuals, and the genetic components are coded with different colours. Colour and shape code as box in b: Northern (green squares), Central (yellow circles), Southern (pink triangles). a-b plotted in ggplot2 [82] and d plotted in R.

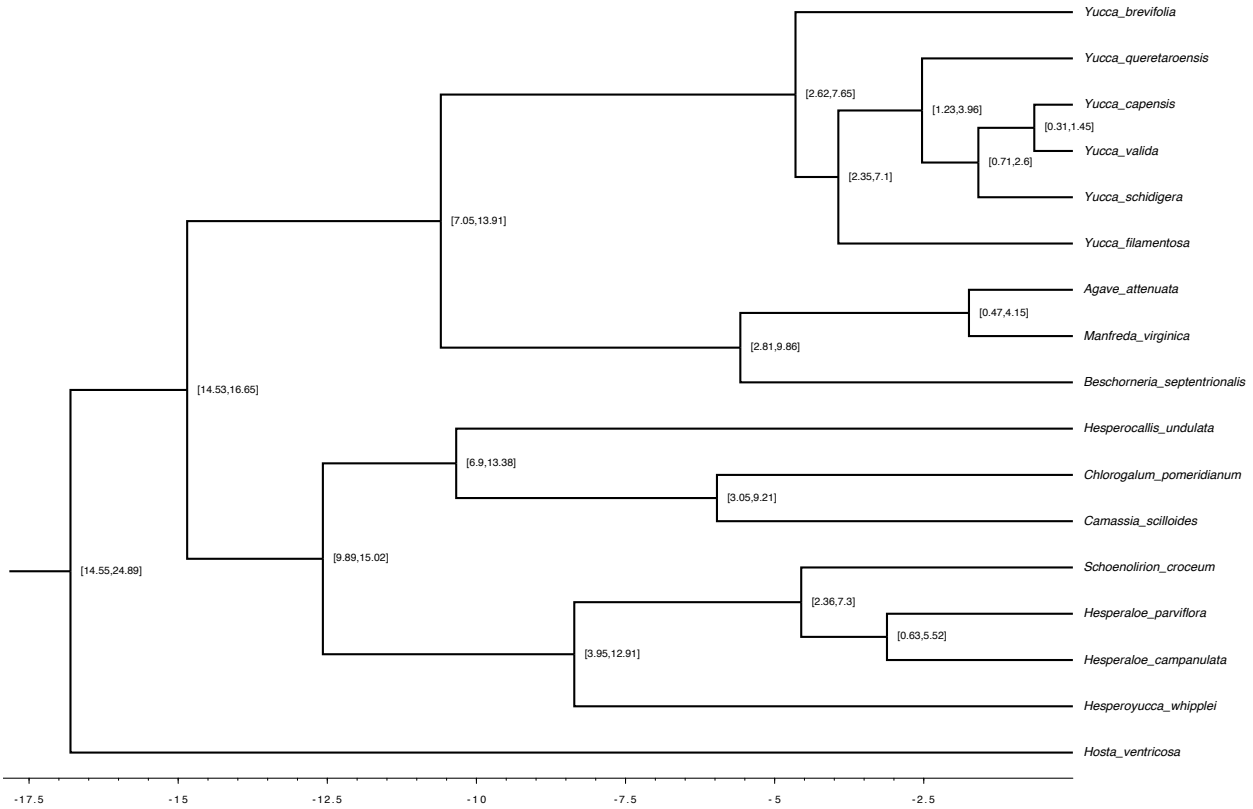

Figure 3. Chloroplast phylogeny and divergence times estimated by BEAST 1.8.4 from wide-chloroplast-genome sequences of 18 Agavoideae species. The tree was produced with *Hosta ventricosa* at the root. The numbers on the scale represent millions of years from the present. Numbers in brackets are 95% confidence intervals. Plotted in FigTree [83].

603 **Supplementary Table S1.** Coordinates of the sampling locations and sample sizes (n)
604 collected throughout the distribution of *Y. valida* for this study.
605

Location ID	n	Latitude	Longitude
100	3	29.36467	-114.36412
16	6	29.25202	-114.16822
15	6	29.11758	-114.14969
26	5	29.05869	-114.00852
105	10	28.83702	-114.12524
104	4	28.65867	-114.03167
103	10	28.40771	-113.99003
14	12	28.20819	-114.00066
1	12	27.73669	-113.38708
13	7	27.31694	-112.79994
7	9	27.31363	-113.12936
8	7	27.30480	-113.07413
9	5	27.2925	-112.98147
4	12	27.29052	-113.15716
6	8	27.24591	-113.21036
10	2	27.23780	-112.86741
102	8	27.15399	-112.88222
101	12	27.06119	-112.96024
90	7	26.000366	-112.17968
91	15	25.936442	-112.09022

606

607 **Supplementary Table S2**. Outgroup Taxa and GenBank accessions used to analyse
608 the phylogenetic relationships and time of origin of *Y. valida* for this study.

Taxon	Accession No.	Reference
Agave attenuata	NC_032696.1	[84]
Beschorneria septentrionalis	NC_032699.1	
Camassia scilloides	NC_032700.1	
Chlorogalum pomeridianum	NC_032701.1	
Hesperaloe campanulata	NC_032702.1	
Hesperaloe parviflora	NC_032703.1	
Hesperocallis undulata	NC_032704.1	
Hesperoyucca whipplei	NC_032705.1	
Hosta ventricosa	NC_032706.1	
Manfreda virginica	NC_032707.1	
Schoenolirion croceum	NC_032710.1	
Y. brevifolia	NC_032711.1	
Y. filamentosa	NC_032712.1	
Y. queretaroensis	NC_032713.1	
Y. schidigera	NC_032714.1	
Y. treculeana	NC_061319.1	Wang, F., 2021

609