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

Aleman, Alberto [1,2]; Arteaga, Maria Clara[1]*; Gasca-Pineda, Jaime [3], Bello-Bedoy, Rafael[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 (UNAM), 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

Manuscript received _____; revision accepted _____.

Running head: Divergent lineages in a young species: the case of *Yucca valida*

ABSTRACT

Premise: Globally, barriers triggered by climatic changes have caused habitat fragmentation and population allopatric divergence. One example of this are the oscillations of the Quaternary, which had an important role in the historical distribution of wildlife across North America. Notably, diverse plant species from the Baja California Peninsula, in western North America, exhibit strong genetic structure and highly concordant differentiated lineages across their ranges, as they were isolated during the Pleistocene glacial-interglacial cycles and thus accumulated genetic differences in their genomes. A representative plant genus of this Peninsula is *Yucca*, with *Yucca valida* having the widest range. Whereas *Y. valida* is a dominant species, there is an extensive distribution discontinuity between 26° N and 27° N, where no individuals have been identified, suggesting restricted gene flow. Moreover, the historical distribution models indicate a substantial reduction in its range during the Last Interglacial, making it an interesting model for studying genetic divergence. **Methods**: We examined the phylogeography of *Y. valida* throughout its range to identify the number of genetic lineages, quantify their genetic differentiation, reconstruct their demographic history and estimate the species' age. **Key results**: We assembled 4,411 SNPs from 147 plants, identifying three allopatric lineages. Our analyses supported that genetic drift is the driver of genetic differentiation among these lineages. We estimated an age under one million years for the common ancestor of *Y. valida* and its sister species. **Conclusions**: Habitat fragmentation caused by climatic changes, low dispersal, and an extensive geographical range gap acted as cumulative mechanisms, leading to allopatric divergence in *Y. valida*.

Keywords: Allopatric divergence; chloroplast-genome assembly; climate-driven barriers; evolutionary history; genetic diversity and differentiation; genetic drift; isolation; natural selection; nextRAD sequencing; *Yucca* phylogenetics.

INTRODUCTION

Genetic divergence, the formation of two or more independent lineages from an ancestral one, can be understood by two broad non-mutually exclusive frameworks: allopatric divergence, in which low dispersal, demographic fluctuations, or population extrinsic barriers lead to genetic differentiation (Mayr et al., 1963; Barton, 2008; Scheiner and Mindell, 2020) and ecological divergence, in which selection drives differentiation by cumulatively restricting the exchange of certain loci (Darwin, 1859; Tigano and Friesen, 2016; Stankowski and Ravinet, 2021). Identifying how genetic differentiation is accumulated between and within species is fundamental for understanding their evolutionary histories (Han et al., 2017; Irwin et al., 2018; Shang et al., 2022) and permits linking these histories to geographical regions' geologic and climatic histories (Avise, 2000).

Globally, barriers triggered by climatic changes have caused habitat fragmentation and population allopatric divergence (Kumar and Kumar, 2018). One example of this are the climatic oscillations of the Quaternary, which had an important role in the historical distribution of wildlife across North America (Shafer et al., 2010), causing genetic differentiation in numerous taxa (Hewitt, 2000; e.g., de Oliveira et al., 2021). Likewise, since these climatic oscillations have been cyclic, divergence followed by a secondary contact due to habitat relinkages has been exposed as a common process in the Northern Hemisphere (e.g., Harrington et al., 2018; Rödin-Mörch et al., 2019).

Broadly distributed plant species are valuable models for studying genetic divergence (Petit and Hampe, 2006). Under wide geographic ranges, plant populations

68 are expected to be particularly subject to demographic changes related to genetic drift
69 (e.g., habitat fragmentation) and gene flow (e.g., changes in their dispersers dynamics)
70 (Sork et al., 2016). Notably, diverse plant species from the Baja California Peninsula
71 (BCP), located in western North America, exhibit strong genetic structure and highly
72 concordant differentiated lineages across their ranges. Populations of cacti, ocotillos,
73 palms, trees, and agaves from the BCP were isolated during the Pleistocene glacial-
74 interglacial cycles and accumulated genetic differences in their genomes (Clark-Tapia
75 and Molina-Freaner, 2003; Fehlberg and Ranker, 2009; Rebernig et al., 2010; Lira-
76 Noriega et al., 2015; Gutiérrez-Flores et al., 2016; Klimova et al., 2017; Martínez-
77 Noguez et al., 2020; Pérez-Alquicira et al., 2023). This is because, while during the
78 glacial periods, the deserts in the Northern Hemisphere increased their distribution as
79 the climate was drier, the interglacial periods were highly humid, producing a significant
80 fragmentation of these ecosystems (Qiu et al., 2011; Meng et al., 2015). Collectively,
81 these patterns of genetic variation throughout the BCP indicate that divergence in this
82 region has been strongly shaped by allopatry (Dolby et al., 2015, 2022).
83 A representative plant genus of the BCP is *Yucca* (Asparagaceae), a
84 monocotyledonous, long-lived, semi-evergreen tree with woody stems, lanceolate
85 succulent green leaves, and white rosette-shaped inflorescences. It is characterized by
86 a long life cycle and slow population growth and recruitment rates (Turner et al., 2005).
87 Its pollination depends on an obligate mutualism with moths of the genera *Tegeticula*
88 and *Parategeticula*, which disperse pollen over short distances (Pellmyr, 2003), while
89 seed dispersal has been related to extinct megafauna and is currently mediated by
90 rodents, birds, and ruminants (Lenz, 2001). Three *Yucca* species are distributed across

91 the BCP: *Yucca schidigera*, *Yucca capensis,* and *Yucca valida*, the latter two being
92 endemic and sister taxa that are pollinated by a single moth species, *Tegeticula baja*
93 (Pellmyr et al., 2007, 2008). Among the *Yuccas* of the BCP, *Y. valida* has the widest
94 range, from ~29.8° N to ~25.6° N, across the Central and Vizcaino deserts, and the
95 Magdalena Plains (Turner et al., 2005; Thiede, 2020). Whereas populations of *Y. valida*
96 are continuous across their northern range, there is an extensive gap between 26° N
97 and 27° N (our personal observation), where no individuals have been identified,
98 suggesting restricted gene flow. Furthermore, while *Y. valida* is a widespread species,
99 the historical distribution models indicate a substantial reduction in its range during the
100 Last Interglacial (Arteaga et al., 2020). As a result, this species is an interesting model
101 for studying genetic divergence.
102 We examined the phylogeography of *Y. valida* throughout its range. We
103 hypothesized that its populations have experienced allopatric divergence, and since this
104 species has a short-distance dispersal that limits genetic interchange between distant
105 populations (Alamo-Herrera et al., 2022), we expected to find differentiated genetic
106 lineages across its distribution. We aimed to i) identify the number of genetic lineages of
107 *Y. valida* throughout its range, ii) quantify the levels of genetic differentiation between
108 these lineages, iii) reconstruct their demographic history and iv) estimate the species'
109 age. This information will contribute to the understanding of the processes shaping
110 genetic divergence in the BCP and Western North America.

111 **MATERIALS AND METHODS**

112 ***Sampling, DNA extraction and sequencing*** — Fresh leaf tissue samples of 120 plants
113 were collected from 20 locations across the species range (Figure 1, Supplementary
114 Table S1). All samples were dried in desiccant silicate gel. We used 100 mg of each
115 dried sample for DNA extraction following the Qiagen DNeasy Plant Mini Kit protocol
116 (Qiagen, Hilden, Germany). The extractions' quality and concentration were evaluated
117 using GelRed 1.5% agarose gels and a NanoDrop spectrophotometer (Thermo Fisher
118 Scientific, Waltham, MA, USA).

119 Genomic DNA was converted into nextRAD sequencing libraries (SNPsaurus,
120 LLC) as in Russello et al. (2015). Briefly, DNA was fragmented with Nextera reagent
121 (Illumina, Inc), which also ligates short adapter sequences to the ends of the fragments.
122 The reaction was scaled to fragment 20 ng of DNA, and 50 ng of DNA were added to
123 compensate for DNA degradation and to increase the number of fragments. The DNA
124 was amplified for 27 cycles at 74°C. The libraries were sequenced on a HiSeq4000 150
125 bp single-end line (University of Oregon). The dataset was enriched with 40 raw
126 sequences of *Y. valida* (NCBI SRA: SRR11514779) and 40 of *Y. capensis* (NCBI SRA:
127 SRR11514778) from Arteaga et al. (2020) that were obtained for previous population
128 genetic analyses following the same extraction and sequencing protocols described in
129 this study; thus, representing a consistent dataset across analyses.

130 ***Raw data processing and SNP–calling*** — The quality of the 160 demultiplexed raw
131 sequences of *Y. valida* was evaluated through FastQC 0.11.9 (Andrews, 2017) and
132 MultiQC 1.14 (Ewels et al., 2016). Adapter removal was done with custom scripts
133 (SNPsaurus, LLC) using the *bbduk* module of BBTools 1.0 (Bushnell, 2023). Clean

134 reads were forced to 110 bp in length. A *de novo* assembly was performed in Stacks 2.6
135 (Catchen et al., 2013). Parameter optimization followed Mastretta-Yanes et al. (2015)
136 and Paris et al. (2017), and 13 samples with less than 1M reads were removed as a
137 quality control step before the assembly. We established a minimum stack depth equal
138 to eight (*-m* 8), a maximum of three mismatches between stacks (*-M* 3), and a maximum
139 of four mismatches between catalogue loci (*-n* 4). We called the SNPs present across
140 the remaining 147 samples in the module populations, with a minimum genotyping rate
141 of 0.5 and a minor allele frequency of 0.05. Hodel et al. (2017) and Shafer et al. (2017)
142 exposed accurate resolutions on summary statistics, phylogenetic inferences, and
143 demographic parameters using low completion cut-offs in reduced representation
144 sequencing data; hence, we did not apply more astringent filtering to our SNP
145 identification. When a 110 bp locus showed more than one SNP, only one variant was
146 chosen, following the priorities in Alexander (2020) to ensure SNPs were unlinked
147 (Pearman et al., 2022).
148 ***Population structure*** — We used nuclear SNPs to assess the most likely number of
149 genetic clusters across all samples and the membership of each sample to these
150 clusters using two approaches: i) ADMIXTURE 1.3.0 (Alexander & Lange, 2011) was
151 run with $K = 1 – 20$, and the optimal number of clusters was chosen using the cross–
152 validation procedure, and ii) a Principal Components Analysis (PCA) was produced with
153 Plink 1.90 (Purcell et al., 2007). To avoid over- or underestimating population structure
154 (Janes et al., 2017), we verified that the assignment of samples to genetic clusters
155 (corresponding to three lineages, see *Results*) was consistent for both approaches.

156 ***Genetic diversity and divergence*** — Once the samples were allocated into discrete
157 genetic clusters (corresponding to three lineages, see *Results*), the genetic diversity
158 descriptors (H_E, H_O, and F_{IS}) for each lineage and the relative genetic divergence (paired
159 F_{ST}) between lineages were calculated with GENODIVE 3.0 (Meirmans, 2020), using the
160 SNP dataset. Data conversion was completed using PGDSpider 2.1.1.5 (Lischer and
161 Excoffier, 2012).
162 To examine the isolation by distance null hypothesis, i.e., the correlation between
163 the geographic and genetic distance between samples (Meirmans, 2012), a Mantel test
164 was carried out using the individuals' Euclidean genetic distance and the pairwise
165 distance between sampling locations in meters, introducing a random jitter (±0.05°) to
166 samples from a shared location. To evaluate the significance of the Mantel test, 99,999
167 permutations were used. However, as the correlation and slope values were found
168 negligible ($p = >0.0000$, $r = 0.25$, $y = 0.0001x + 32$), hence rejecting the isolation by
169 distance hypothesis, we do not present further results from this test.
170 ***Role of selection vs drift in genetic divergence*** — To evaluate if population structure
171 is a consequence of ecological selection or genetic drift, we classified the levels of
172 genetic diversity and differentiation within and between lineages according to the
173 evolutionary forces shaping their divergence. All the nucleotides (variant and invariant)
174 from the sequences (110 bp long) containing the 4,411 SNPs (above) used to assess
175 population structure were called in Stacks. Then, the mean values of relative genetic
176 divergence (F_{ST}; Weir and Cockerham, 1984), absolute genetic divergence (d_{XY}; Nei and
177 Miller, 1990), and nucleotide diversity (π; Nei and Li, 1979) for every sequence of 110 bp
178 were computed in population pairs using pixy 1.2.7 (Korunes and Samuk, 2021). The

179 correlations between pairs of estimates were compared. Following Kessler et al. (2023)
180 and Shang et al. (2022), we categorized each correlation according to the relation
181 among paired F_{ST}, d_{XY}, and π, following four models of divergence as follows: 1) a
182 positive correlation between F_{ST} and d_{XY} and a negative correlation between both F_{ST}
183 and d_{XY} to π (local adaptation with gene flow); 2) uniform levels of d_{XY} across different
184 values of F_{ST} and π, with a negative correlation between F_{ST} and π (selection under
185 allopatry); a negative correlation between both d_{XY} and π to F_{ST}, and a positive
186 correlation between F_{ST} and π, characterized by 3) overall high F_{ST}, low π and d_{XY}
187 (background selection) or 4) overall low F_{ST} with high π and d_{XY} (balancing selection). As
188 our interest were the sequences with highly increased or reduced levels of diversity and
189 differentiation, the estimates were classified based on their percentiles' distribution as
190 high (> 95%) and low (< 5%) –or uniform (> 45%, < 55%) for d_{XY}, and then each
191 sequence was assigned to one of the models described. As neutrality was our null
192 hypothesis, any sequence that could not be classified according to these models of
193 divergence was identified as consistent with genetic drift.
194 ***Outlier SNPs and SNPs associated with environmental variables*** — We tested the
195 role of current ecological adaptation in promoting genetic divergence. First, we scanned
196 for outlier SNPs using PCAdapt 4.0.1 (Luu et al., 2017), with $K = 3$ and $\alpha = 0.1$ for the
197 false discovery rate boundary. Then, we used the Latent Factor Mixed-effect Model
198 (Frichot et al., 2013) to characterize the presence of SNPs with a strong association with
199 environmental variables. LFMM was run ten times ($K = 3$) on the R package LEA 3.2.0
200 (Frichot and François, 2015). A total of 20 environmental variables obtained from
201 WorldClim v1.4 (Hijmans et al., 2005) were tested for each sampling location. The Gibbs

202 sampler algorithm was used for 10,000 cycles (burn-in = 5,000), and the median of the
203 correlation scores (z-score) was calculated across all runs, recalibrating the p-values
204 frequency distribution with lambda. We avoided any complimentary correction
205 procedures for the false discovery rate in both PCAdapt and LFMM to increase the
206 probability of detecting outliers and SNPs correlated with local adaptation. The SNPs
207 detected by any procedure were removed from the dataset to produce a neutral
208 catalogue for the demographic history analyses.
209 ***Demographic history*** — To understand the succession of events that could originate
210 the observed nuclear genetic lineages, the demographic history of *Y. valida* was
211 reconstructed using the diffusion approximation method δaδi (Gutenkunst et al., 2009).
212 Simulations included four possible three-populations simple models of i) simultaneous,
213 ii) admixed, and sequential divergence from iii) North to South and iv) South to North,
214 without gene flow or changes in population sizes, as developed by Portik et al. (2017).
215 The site frequency spectrum was computed as folded, and the best model was selected
216 using the maximum likelihood and the Akaike information criterion (AIC).
217 ***Putative whole chloroplast genome sequences recovery*** — We implemented a
218 reference-guided workflow to reconstruct whole-chloroplast-genome sequences
219 following Aleman et al. (2024). The 200 raw sequences of *Y. capensis* and *Y. valida*
220 were cleaned as above, without forcing a maximum length and establishing a minimum
221 length = 100 bp. The clean reads were mapped to the plastome of *Y. schidigera*
222 (GenBank: NC_032714.1) in bowtie 2.2.6 under the --*very-sensitive-local* option
223 (Langmead and Salzberg, 2012). Mapping statistics were evaluated with the flagstat and
224 coverage modules of SAMtools 1.15.1 (Li et al., 2009). Nucleotide calling was performed

225 individually for each sample in ANGSD (Korneliussen et al., 2014), producing linear
226 sequences by requiring minimum mapping and base qualities of 20, and outputting
227 missing data as Ns *(-dofasta 2 -minMapQ 20 -minQ 20 -doCounts 1*).
228 ***Phylogenetic relationships and divergence time*** — We reconstructed the intraspecific
229 phylogenetic relationships of the samples in this study to understand their evolutionary
230 relationships. For this, we used the plastome sequences and chose the individuals with
231 higher data completeness (130 samples, maximum missing data < 0.5) to run RAxML
232 8.2.12 (Stamatakis, 2014), establishing *Y. schidigera* as the outgroup. A GTR + G + I
233 model was used, letting RaxML halt automatic bootstrapping.
234 As the intraspecific phylogenetic analysis revealed a lack of differentiation among
235 the plastomes of *Y. valida* (see *Results*), the divergence time of *Y. valida* and *Y. capensis*
236 was estimated as an approximation to the age of *Y. valida*. The phylogenetic
237 relationships of the Agavoideae subfamily were reconstructed based on the molecular
238 clock, replicating Smith et al. (2021) in BEAST 1.8.4 (Drummond and Bouckaert, 2015).
239 Based on Scheunert et al. (2020) and Roy et al. (2020), we produced consensus
240 sequences for each *Y. valida* and *Y. capensis*, using all the sequences available on each
241 species. The consensus sequences were aligned to the plastomes of 15 Agavoideae
242 species (Supplementary Table S2, (McKain et al., 2016)) with MAFFT 7 default settings
243 (Katoh et al., 2019). A GTR + I model was employed for the molecular clock analysis.
244 Gamma-distributed priors for the five substitution types and uniform priors (0-1) for the
245 base frequencies and the proportion of invariant sites were established. An uncorrelated
246 relaxed clock model with log-normal distributed rates was set up (mean = 1, standard
247 deviation = 0.33), assuming a Yule process for the tree prior, starting from a random

248 tree, establishing *Hosta ventricosa* as the outgroup, and forcing the ingroup's monophyly
249 (Drummond and Bouckaert, 2015). A log-normal prior distribution on the age of *Yucca +*
250 *Agave* was specified to 14.5 million years (offset, mean = 1 standard deviation = 1.4),
251 based on the minimum age of the fossil *Protoyucca shadesii* (Tidwell and Parker, 1990;
252 Iles et al., 2015). We completed two independent runs of 60 million steps long with a
253 sample every 6,000,000 steps, and the runs were joined in LogCombiner 1.8.4,
254 considering a 10% burn-in for each input. A maximum credibility tree was produced in
255 TreeAnnotator 1.8.4. Nodes' medians and 95% confidence intervals were outputted.

256 **RESULTS**

257 ***Raw data processing and SNP–calling*** — A total of 400M reads of 150 bp from 160
258 samples were processed (~2.5M reads per sample). After quality trimming and adapter
259 removal, ~300M reads of 110 bp distributed in 147 samples were kept (~2M reads per
260 sample). The *de novo* assembly in Stacks recovered 10,511 SNPs in 4,411 loci
261 (minimum depth = 15x, mean missing data = 0.27, minimum sequencing quality = 39).

262 ***Population structure, genetic diversity and divergence*** — Using 4,411 unlinked
263 SNPs, the nuclear genetic structure analyses established the most likely number of
264 lineages as three ($K = 3$) (Figure 1). The lineages were named after their geographic
265 location as Northern, Central, and Southern, and had 54, 72, and 21 individuals,
266 respectively. The overall observed heterozygosity was lower than the expected (H_O =
267 0.26, H_E = 0.30), and a mean coefficient of inbreeding (F_{IS}) of 0.17 was observed
268 (Supplementary Table S3). The relative divergence (F_{ST}) between lineages was subtle,
269 with the values between the Northern and Southern lineages being the greatest (North-
270 South: 0.11; North-Central: 0.05; Central-South: 0.09).

271 ***Role of selection vs drift in genetic divergence*** — The correlations between
272 measures of genetic diversity and divergence showed similar patterns between pairs of
273 lineages (Figure 2): low d_{XY} with an insignificant positive correlation to F_{ST}, low π with an
274 insignificant inverse correlation to F_{ST}, and a significant direct correlation between d_{XY} to
275 π (p-value < .05). A total of 41 sequences of 110 bp were consistent with one of the four
276 divergent scenarios. Of those, 3 were consistent with background selection and 38 with
277 balancing selection. A total of 4,370 sequences were identified as consistent with
278 genetic drift.

Outlier SNPs and SNPs associated with environmental variables — PCAdapt detected 20 outlier SNPs. LFMM classified 63 SNPs associated with adaptation to multiple environmental variables, most of which were related to annual precipitation (BIO12 to BIO19). None of the SNPs detected by PCAdapt were also detected by LFMM, and the 83 SNPs detected were removed to produce the neutral dataset.

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

Phylogenetic relationships and divergence time — The intraspecific phylogenetic analysis grouped the *Y. valida* plastomes into a single monophyletic clade without intraspecific lineages. Low support nodes were observed, exposing an absence of differentiation across these genomes. The interspecific phylogenetic relationships and age estimates for the Agavoideae subfamily results were consistent with Smith et al. (2021), and all interspecific nodes were represented with high support values (100%). The molecular clock calculated an age below a million years old for the last common ancestor of *Y. valida* and *Y. capensis* (Figure 3; mean ≈ 819,900; median ≈ 792,600; 95% CI ≈ 33,500 – 1,650,800).

296 **DISCUSSION**

297 We examined the phylogeography of *Y. valida* throughout its range and identified
298 three monophyletic nuclear lineages accompanied by a single chloroplast lineage. The
299 age estimated for the last common ancestor of *Y. valida* and its sister species –below 1
300 MYA, suggests that the nuclear differentiation in *Y. valida* is a consequence of recent
301 events. While the time of origin of the three nuclear lineages is uncertain, the ecological
302 niche model projections predicted that during the Last Interglacial, the area with suitable
303 conditions for *Y. valida* had a substantial reduction (Arteaga et al., 2020). However,
304 climatic oscillations are not the only possible explanation for the observed patterns of
305 divergence (Dolby et al., 2022). Low dispersal without a climatic barrier can also result in
306 genetic differentiation (Araya-Donoso et al., 2022); thus, the role of restricted gene flow
307 should be considered. Likewise, the absence of an isolation by distance pattern
308 indicates that divergence cannot be attributed solely to the distance between
309 populations. Collectively, this suggests that habitat fragmentation caused by climatic
310 changes, low dispersal, and an extensive geographical range gap have acted as
311 cumulative mechanisms, leading to allopatric divergence, as indicated by the patterns of
312 genetic differentiation consistent with drift.

313 The values of genetic differentiation between the three nuclear lineages were
314 subtle (mean paired F_{ST} = 0.08), the highest being between bordering lineages. These
315 values are in agreement with those observed with microsatellite loci for *Y. valida* (F_{ST} =
316 0.059; (Alamo-Herrera et al., 2022)), *Y. brevifolia* (F_{ST} = 0.061; (Starr et al., 2013)), and
317 *Y. schidigera* (F_{ST} = 0.067; (De la Rosa-Conroy et al., 2020)), and greater than those in
318 its sister species, *Y. capensis* (F_{ST} = 0.022; (Luna-Ortiz et al., 2022)). The low levels of

319 differentiation can result from potentially large historical effective population sizes
320 (Arteaga et al., 2020) and reflect some level of gene flow between lineages. Estimates
321 of the dispersal rates indicate that pollen movement in *Y. valida* is approximately 50 m
322 (Alamo-Herrera et al., 2022), and seed-mediated dispersal also occurs at short
323 distances (Waitman et al., 2012), which could maintain connectivity between populations
324 through a stepping-stone model. In this sense, the presence of genetic structure, the
325 patterns of genetic diversity and divergence, and the low differentiation coincide with
326 contemporary genetic drift to be the leading cause of genetic divergence in this species.
327 Genetic diversity was similar across the three lineages. The diversity estimates
328 (H_O = 0.26, H_E = 0.31) were higher than those described for this species with the same
329 type of genetic markers (H_O = 0.11, H_E = 0.17; (Arteaga et al., 2020)). The disparity
330 among these values reiterates the effect of different methodological variables on SNP
331 calling in population genomics (O'Leary et al., 2018). Arteaga et al. (2020) assembled
332 loci from samples of two species —*Y. valida*, *Y. capensis*, and their hybrid. Here, we
333 generated a SNP dataset with samples of a single species, thus modifying the number
334 of polymorphisms detected. SNP-calling among differentiated taxa implies retrieving
335 alleles fixed for a species or a population that are polymorphic between species or
336 populations, affecting diversity estimates (Shafer et al., 2017). This catalogue-size bias
337 highlights that conclusions based on genetic diversity estimates should be made on a
338 case-by-case basis. Similarly, the inbreeding coefficient had lower levels (F_{IS} = 0.17)
339 than those in Arteaga et al. (2020) (F_{IS} = 0.35). The high genetic diversity and low
340 inbreeding detected in this work are congruent with hypothetical high historical

population sizes and the *Yuccas* reproductive system, which favours interbreeding and limits self-pollination (Huth and Pellmyr, 2000).

The historical demographic models' simulations indicated that the nuclear lineages' divergence is associated with a simultaneous event. The lack of intraspecific differentiation in the plastome of *Y. valida* indicates that the drivers of genetic divergence in *Y. valida* have been accumulating long enough to be recorded in the nucleus and so recently not to be registered in the chloroplast. The lack of congruence between outlier SNPs and SNPs associated with environmental variables, plus the absence of signals of divergent selection, rule out the role of ecological divergence in promoting differentiation in *Y. valida*. Our results support the role of drift under allopatry in shaping the genetic divergence of plant species across the BCP (Clark-Tapia and Molina-Freaner, 2003; Fehlberg and Ranker, 2009; Rebernig et al., 2010; Lira-Noriega et al., 2015; Gutiérrez-Flores et al., 2016; Klimova et al., 2017; Martínez-Noguez et al., 2020; Pérez-Alquicira et al., 2023).

CONCLUSIONS

Determining the impact of the drivers of species' genetic diversity and differentiation is a crucial aspect of evolutionary biology. Here, we tested the role of drift vs selection in promoting divergence on a representative taxon of the BCP. The history of *Y. valida* is one of allopatric differentiation. Addressing the origin and colonization of *Yucca* in the Peninsula remains an essential research gap in the evolutionary history of this genus.

ACKNOWLEDGEMENTS

We respectfully acknowledge that the territory where the samples were collected are the ancestral homelands of the Cochimí people. This study was completed 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, to whom we show our respect. The authors thank 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 the field. This work was supported by Consejo Nacional de Humanidades, Ciencias y Tecnologías (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. Finally, we thank Camille Kessler and Flavia Termignoni-García for their bioinformatic advice and Enrique Ruiz for his work in Fig. 3.

AUTHOR CONTRIBUTIONS

Conceptualisation: MCA. Developing methods, conducting the research, data interpretation and writing: AA, JGP, MCA, and RBB. Data analysis and preparation of figures and tables: AA. All authors contributed to the manuscript and approved its final version.

DATA ACCESSIBILITY STATEMENT

Data and scripts for this study can be found at https://github.com/al-aleman/datillilo. High-throughput sequencing raw data is available at https://datadryad.org/stash/share/Ujo8Mi1d2g50_VAAsdXzMWKgsYcffCGD59_VLFGyc3U and https://doi.org/10.5061/dryad.jh9w0vtgv.

REFERENCES

Alamo-Herrera, C. R., M. C. Arteaga, R. Bello-Bedoy, and F. Rosas-Pacheco. 2022. Pollen dispersal and genetic diversity of Yucca valida (Asparagaceae), a plant involved in an obligate pollination mutualism. *Biological Journal of the Linnean Society* 136: 364–374.

Aleman, A., M. E. Dorken, A. B. A. Shafer, T. Patel, P. A. Volkova, and J. R. Freeland. 2024. Development of genomic resources for cattails (Typha), a globally important macrophyte genus. *Freshwater Biology* 69: 74–83.

Alexander, A. 2020. GBS_SNP_filter v1.17.

Alexander, D. H., and K. Lange. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12: 246.

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

Araya-Donoso, R., S. M. Baty, P. Alonso-Alonso, M. J. Sanín, B. T. Wilder, A. Munguia-Vega, and G. A. Dolby. 2022. Implications of barrier ephemerality in geogenomic research. *Journal of Biogeography* 49: 2050–2063.

Arteaga, M. C., R. Bello-Bedoy, and J. Gasca-Pineda. 2020. Hybridization Between Yuccas From Baja California: Genomic and Environmental Patterns. *Frontiers in Plant Science* 11.

Avise, J. C. 2000. Phylogeography: The History and Formation of Species. Harvard University Press.

Barton, N. H. 2008. The effect of a barrier to gene flow on patterns of geographic variation. *Genetics Research* 90: 139–149.

Bushnell. 2023. BBTools.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22: 3124–3140.

Clark-Tapia, R., and F. Molina-Freaner. 2003. The genetic structure of a columnar cactus with a disjunct distribution: Stenocereus gummosus in the Sonoran desert. *Heredity* 90: 443–450.

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.

De la Rosa-Conroy, L., J. Gasca-Pineda, R. Bello-Bedoy, L. E. Eguiarte, and M. C. Arteaga. 2020. Genetic patterns and changes in availability of suitable habitat

419 support a colonisation history of a North American perennial plant. *Plant Biology*
420 22: 233–242.

421 Dolby, G. A., S. E. K. Bennett, R. J. Dorsey, M. F. Stokes, B. R. Riddle, A. Lira-Noriega,
422 A. Munguia-Vega, and B. T. Wilder. 2022. Integrating Earth–life systems: a
423 geogenomic approach. *Trends in Ecology & Evolution* 37: 371–384.

424 Dolby, G. A., S. E. K. Bennett, A. Lira-Noriega, B. T. Wilder, and A. Munguía-Vega.
425 2015. Assessing the Geological and Climatic Forcing of Biodiversity and
426 Evolution Surrounding the Gulf of California. *Journal of the Southwest* 57: 391–
427 455.

428 Drummond, A. J., and R. R. Bouckaert. 2015. Bayesian Evolutionary Analysis with
429 BEAST. Cambridge University Press, Cambridge.

430 Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: summarize analysis
431 results for multiple tools and samples in a single report. *Bioinformatics* 32: 3047–
432 3048.

433 Fehlberg, S. D., and T. A. Ranker. 2009. Evolutionary history and phylogeography of
434 Encelia farinosa (Asteraceae) from the Sonoran, Mojave, and Peninsular Deserts.
435 *Molecular Phylogenetics and Evolution* 50: 326–335.

436 Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological
437 association studies. *Methods in Ecology and Evolution* 6: 925–929.

438 Frichot, E., S. D. Schoville, G. Bouchard, and O. François. 2013. Testing for
439 Associations between Loci and Environmental Gradients Using Latent Factor
440 Mixed Models. *Molecular Biology and Evolution* 30: 1687–1699.

441 Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009.
442 Inferring the Joint Demographic History of Multiple Populations from
443 Multidimensional SNP Frequency Data. *PLOS Genetics* 5: e1000695.

444 Gutiérrez-Flores, C., F. J. García-De León, J. L. León-de la luz, and J. H. Cota-Sánchez.
445 2016. Microsatellite genetic diversity and mating systems in the columnar cactus
446 Pachycereus pringlei (Cactaceae). *Perspectives in Plant Ecology, Evolution and*
447 *Systematics* 22: 1–10.

448 Han, F., S. Lamichhaney, B. R. Grant, P. R. Grant, L. Andersson, and M. T. Webster.
449 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the
450 genomic landscape of divergence among Darwin's finches. *Genome Research*
451 27: 1004–1015.

452 Harrington, S. M., B. D. Hollingsworth, T. E. Higham, and T. W. Reeder. 2018.
453 Pleistocene climatic fluctuations drive isolation and secondary contact in the red

diamond rattlesnake (Crotalus ruber) in Baja California. *Journal of Biogeography* 45: 64–75.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* 405: 907–913.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.

Hodel, R. G. J., S. Chen, A. C. Payton, S. F. McDaniel, P. Soltis, and D. E. Soltis. 2017. Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. *Scientific Reports* 7: 17598.

Huth, C. J., and O. Pellmyr. 2000. Pollen-Mediated Selective Abortion in Yuccas and Its Consequences for the Plant-Pollinator Mutualism. *Ecology* 81: 1100–1107.

Iles, W. J. D., S. Y. Smith, M. A. Gandolfo, and S. W. Graham. 2015. Monocot fossils suitable for molecular dating analyses. *Botanical Journal of the Linnean Society* 178: 346–374.

Irwin, D. E., B. Milá, D. P. L. Toews, A. Brelsford, H. L. Kenyon, A. N. Porter, C. Grossen, et al. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. *Molecular Ecology* 27: 4839–4855.

Janes, J. K., J. M. Miller, J. R. Dupuis, R. M. Malenfant, J. C. Gorrell, C. I. Cullingham, and R. L. Andrew. 2017. The K = 2 conundrum. *Molecular Ecology* 26: 3594–3602.

Katoh, K., J. Rozewicki, and K. D. Yamada. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* 20: 1160–1166.

Kessler, C., E. Wootton, and A. B. A. Shafer. 2023. Speciation without gene-flow in hybridizing deer. *Molecular Ecology* 32: 1117–1132.

Klimova, A., J. I. Hoffman, J. N. Gutierrez-Rivera, J. Leon de la Luz, and A. Ortega-Rubio. 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.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15: 356.

Korunes, K. L., and K. Samuk. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* 21: 1359–1368.

Kumar, R., and V. Kumar. 2018. A review of phylogeography: biotic and abiotic factors. *Geology, Ecology, and Landscapes* 2: 268–274.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9: 357–359.

Lenz, L. 2001. Seed Dispersal in Yucca Brevifolia (Agavaceae)-Present and Past, With Consideration of the Future of the Species. *Aliso* 20: 61–74.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Lira-Noriega, A., O. Toro-Núñez, J. R. Oaks, and M. E. Mort. 2015. The roles of history and ecology in chloroplast phylogeographic patterns of the bird-dispersed plant parasite Phoradendron californicum (Viscaceae) in the Sonoran Desert. *American Journal of Botany* 102: 149–164.

Lischer, H. E. L., and L. Excoffier. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28: 298–299.

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

Luu, K., E. Bazin, and M. G. B. Blum. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* 17: 67–77.

Martínez-Noguez, J. J., J. L. León de la Luz, J. Delgadillo Rodríguez, and F. J. García-De León. 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.

Mastretta-Yanes, A., N. Arrigo, N. Alvarez, T. H. Jorgensen, D. Piñero, and B. C. Emerson. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* 15: 28–41.

Mayr, E., A. A. P. of Z. (Emeritus) M. of C. Z. E. Mayr, and E. W. Mayr. 1963. Animal Species and Evolution. Belknap Press of Harvard University Press.

McKain, M. R., J. R. McNeal, P. R. Kellar, L. E. Eguiarte, J. C. Pires, and J. Leebens-Mack. 2016. Timing of rapid diversification and convergent origins of active

524 pollination within Agavoideae (Asparagaceae). *American Journal of Botany* 103:
525 1717–1729.

526 Meirmans, P. G. 2020. genodive version 3.0: Easy-to-use software for the analysis of
527 genetic data of diploids and polyploids. *Molecular Ecology Resources* 20: 1126–
528 1131.

529 Meirmans, P. G. 2012. The trouble with isolation by distance. *Molecular Ecology* 21:
530 2839–2846.

531 Meng, H.-H., X.-Y. Gao, J.-F. Huang, and M.-L. Zhang. 2015. Plant phylogeography in
532 arid Northwest China: Retrospectives and perspectives. *Journal of Systematics*
533 *and Evolution* 53: 33–46.

534 Nei, M., and W. H. Li. 1979. Mathematical model for studying genetic variation in terms
535 of restriction endonucleases. *Proceedings of the National Academy of Sciences*
536 *of the United States of America* 76: 5269–5273.

537 Nei, M., and J. C. Miller. 1990. A simple method for estimating average number of
538 nucleotide substitutions within and between populations from restriction data.
539 *Genetics* 125: 873–879.

540 O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy. 2018.
541 These aren't the loci you'e looking for: Principles of effective SNP filtering for
542 molecular ecologists. *Molecular Ecology* 27: 3193–3206.

543 de Oliveira, F. F. R., M. Gehara, M. Solé, M. Lyra, C. F. B. Haddad, D. P. Silva, R. F. de
544 Magalhães, et al. 2021. Quaternary climatic fluctuations influence the
545 demographic history of two species of sky-island endemic amphibians in the
546 Neotropics. *Molecular Phylogenetics and Evolution* 160: 107113.

547 Paris, J. R., J. R. Stevens, and J. M. Catchen. 2017. Lost in parameter space: a road
548 map for stacks. *Methods in Ecology and Evolution* 8: 1360–1373.

549 Pearman, W. S., L. Urban, and A. Alexander. 2022. Commonly used Hardy–Weinberg
550 equilibrium filtering schemes impact population structure inferences using
551 RADseq data. *Molecular Ecology Resources* 22: 2599–2613.

552 Pellmyr, O. 2003. Yuccas, Yucca Moths, and Coevolution: A Review. *Annals of the*
553 *Missouri Botanical Garden* 90: 35–55.

554 Pellmyr, O., M. Balcázar-Lara, K. A. Segraves, D. M. Althoff, and R. J. Littlefield. 2008.
555 Phylogeny of the pollinating yucca moths, with revision of Mexican species
556 (Tegeticula and Parategeticula; Lepidoptera, Prodoxidae). *Zoological Journal of*
557 *the Linnean Society* 152: 297–314.

Pellmyr, O., K. A. Segraves, D. M. Althoff, M. Balcázar-Lara, and J. Leebens-Mack. 2007. The phylogeny of yuccas. *Molecular Phylogenetics and Evolution* 43: 493–501.

Petit, R. J., and A. Hampe. 2006. Some Evolutionary Consequences of Being a Tree. *Annual Review of Ecology, Evolution, and Systematics* 37: 187–214.

Portik, D. M., A. D. Leaché, D. Rivera, M. F. Barej, M. Burger, M. Hirschfeld, M.-O. Rödel, et al. 2017. Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Molecular Ecology* 26: 5245–5263.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, 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.

Qiu, Y.-X., C.-X. Fu, and H. P. Comes. 2011. Plant molecular phylogeography in China and adjacent regions: Tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics and Evolution* 59: 225–244.

Rebernig, C. A., G. M. Schneeweiss, K. E. Bardy, P. Schönswetter, J. L. Villaseñor, R. Obermayer, T. F. Stuessy, and H. Weiss-Schneeweiss. 2010. Multiple Pleistocene refugia and Holocene range expansion of an abundant southwestern American desert plant species (Melampodium leucanthum, Asteraceae). *Molecular Ecology* 19: 3421–3443.

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

Roy, N. S., U. Jeong, M. Na, I.-Y. Choi, and E. J. Cheong. 2020. Genomic analysis and a consensus chloroplast genome sequence of Prunus yedoensis for DNA marker development. *Horticulture, Environment, and Biotechnology* 61: 859–867.

Russello, M. A., M. D. Waterhouse, P. D. Etter, and E. A. Johnson. 2015. From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ* 3: e1106.

Scheiner, S. M., and D. P. Mindell eds. . 2020. The Theory of Evolution: Principles, Concepts, and Assumptions. University of Chicago Press, Chicago, IL.

Scheunert, A., M. Dorfner, T. Lingl, and C. Oberprieler. 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.

Shafer, A. B., C. I. Cullingham, S. D. Cote, and D. W. Coltman. 2010. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular ecology* 19: 4589–4621.

Shafer, A. B., C. R. Peart, S. Tusso, I. Maayan, A. Brelsford, C. W. Wheat, and J. B. Wolf. 2017. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution* 8: 907–917.

Shang, H., M. Rendón-Anaya, O. Paun, D. L. Field, J. Hess, C. Vogl, J. Liu, et al. 2022. Drivers of genomic landscapes of differentiation across Populus divergence gradient. 2021.08.26.457771.

Smith, C. I., M. R. McKain, A. Guimond, and R. Flatz. 2021. Genome-scale data resolves the timing of divergence in Joshua trees. *American Journal of Botany* 108: 647–663.

Sork, V. L., P. F. Gugger, J.-M. Chen, and S. Werth. 2016. Evolutionary lessons from California plant phylogeography. *Proceedings of the National Academy of Sciences* 113: 8064–8071.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stankowski, S., and M. Ravinet. 2021. Defining the speciation continuum. *Evolution* 75: 1256–1273.

Starr, T. N., K. E. Gadek, J. B. Yoder, R. Flatz, and C. I. Smith. 2013. Asymmetric hybridization and gene flow between Joshua trees (Agavaceae: Yucca) reflect differences in pollinator host specificity. *Molecular Ecology* 22: 437–449.

Thiede, J. 2020. Yucca AGAVACEAE. *In* U. Eggli, and R. Nyffeler [eds.], Monocotyledons, Illustrated Handbook of Succulent Plants, 363–421. Springer, Berlin, Heidelberg.

Tidwell, W. D., and L. R. Parker. 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.

Tigano, A., and V. L. Friesen. 2016. Genomics of local adaptation with gene flow. *Molecular Ecology* 25: 2144–2164.

Turner, R. M., J. E. Bowers, and T. L. Burgess. 2005. Sonoran Desert Plants: An Ecological Atlas. University of Arizona Press.

Waitman, B. A., S. B. Vander Wall, and T. C. Esque. 2012. Seed dispersal and seed fate in Joshua tree (Yucca brevifolia). *Journal of Arid Environments* 81: 1–8.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38: 1358–1370.

FIGURES AND SUPPLEMENTARY MATERIAL

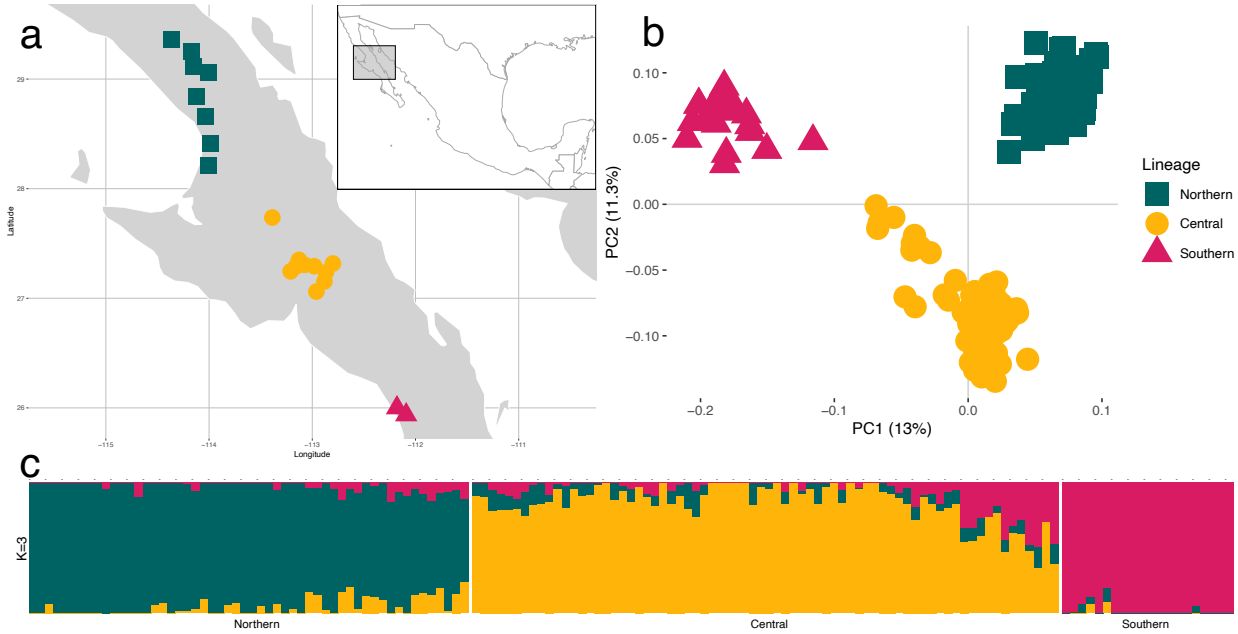

Figure 1. Genetic structure results of 4,411 nuclear SNPs obtained from 147 *Y. valida* samples. (a) Sampling locations (20) for this study. Shapes represent sampling locations, and colours represent genetic lineages. (b) Principal Component Analysis for PC1 and PC2. Shapes represent individuals, and colours represent genetic lineages, (c) ADMIXTURE ($K = 3$). Vertical bars represent individuals, and the genetic components are coded with different colours. Colour and shape code as the box in b: Northern (green, squares), Central (yellow, circles), Southern (pink, triangles).

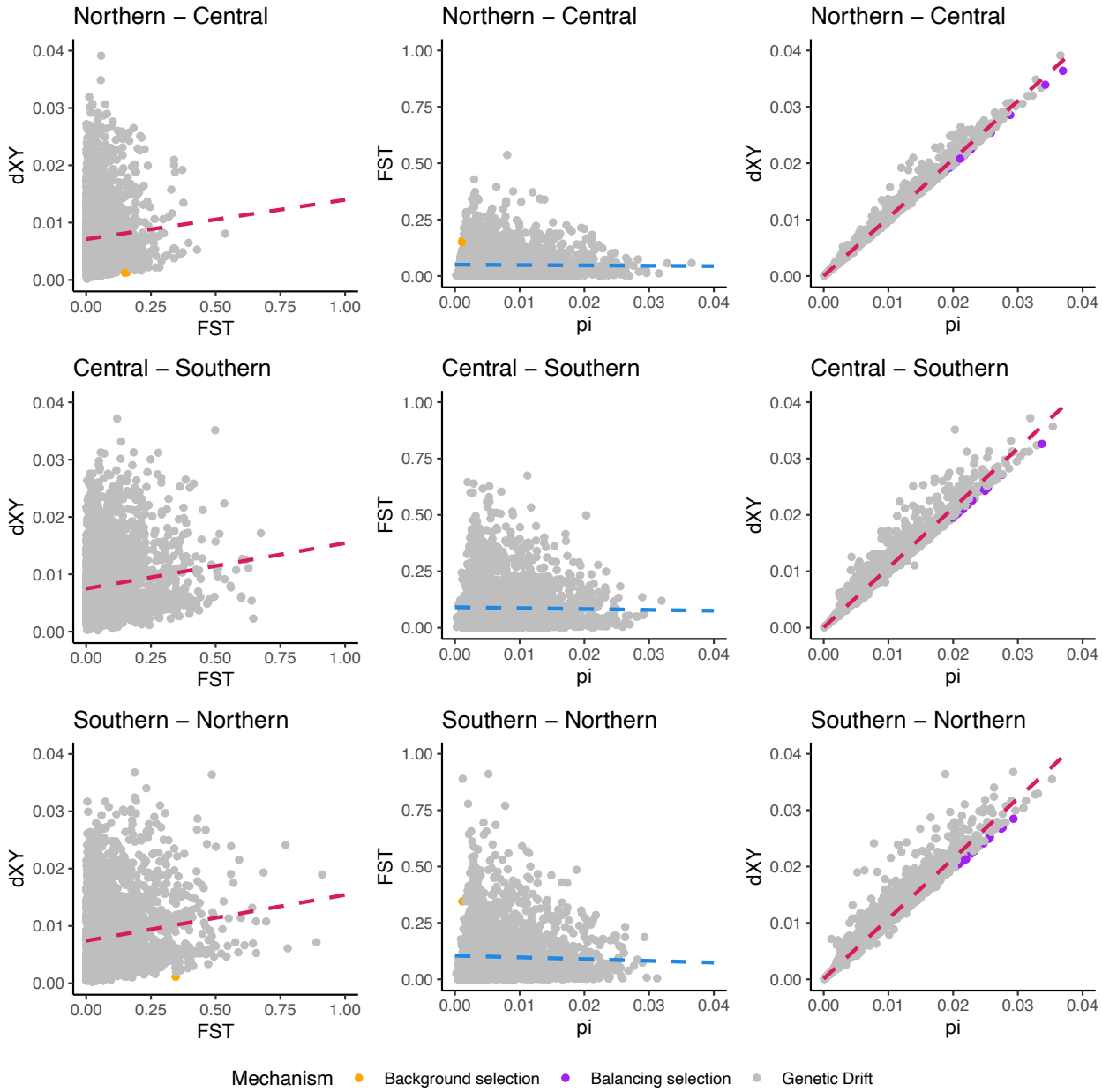

Figure 2. Correlation analysis for F_{ST}, d_{XY} and π across 4,411 sequences (110 bp each) between pairs of nuclear genetic lineages of *Y. valida* found in this study. Each point represents a sequence, and it is coloured according to the potential evolutionary mechanism driving genetic differentiation according to the labels at the bottom. Trend lines are coloured as positive (red) or negative (blue).

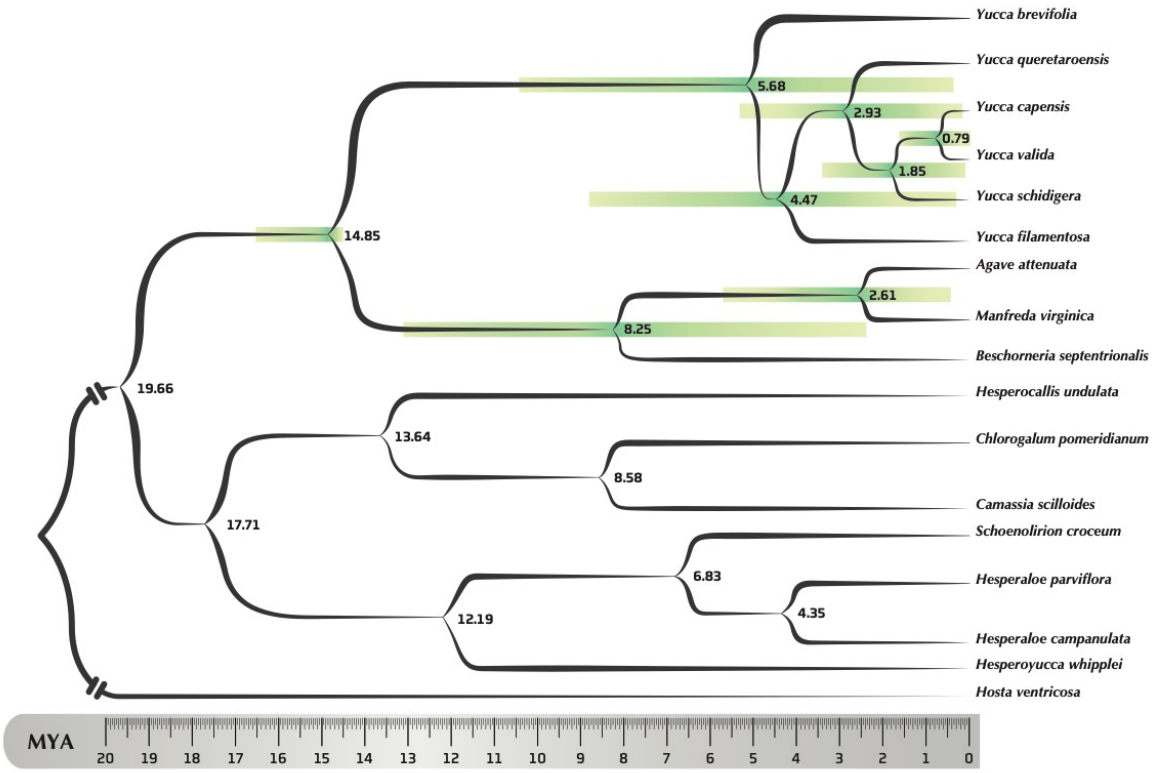

Figure 3. Chloroplast phylogeny of 17 Agavoideae species, estimated in BEAST 1.8.4. The contribution of this study is solely the position and divergence time of *Y. capensis* and *Y. valida*. The phylogenetic relationships among the chloroplast genome sequences within the Agavoideae are discussed in Smith et al. (2021). The tree was produced with *Hosta ventricosa* at the root and calibrated with a log-normal distribution on the age of *Yucca* and *Agave* at 14.5 million years. Numbers represent the estimated medians of the nodes' divergence times as millions of years from now. Horizontal bars represent the 95% confidence intervals. Branch support on every node = 1.

Supplementary Table S1. Site IDs, approximate coordinates, and number of samples (n) collected across the distribution of *Y. valida* for this study. The number of samples includes 40 individuals collected by Arteaga et al. (2020) and used to enrich the final dataset.

Site ID	Latitude	Longitude	n
100	29.36467	-114.36412	3
16	29.25202	-114.16822	6
15	29.11758	-114.14969	6
26	29.05869	-114.00852	5
105	28.83702	-114.12524	10
104	28.65867	-114.03167	4
103	28.40771	-113.99003	9
14	28.20819	-114.00066	11
1	27.73669	-113.38708	12
13	27.31694	-112.79994	4
7	27.31363	-113.12936	8
8	27.30480	-113.07413	5
9	27.2925	-112.98147	5
4	27.29052	-113.15716	11
6	27.24591	-113.21036	7
10	27.23780	-112.86741	2
102	27.15399	-112.88222	6
101	27.06119	-112.96024	12
90	26.000366	-112.17968	6
91	25.936442	-112.09022	15

662 **Supplementary Table S2**. Taxa and GenBank accessions used to investigate the
663 phylogenetic relationships and time of origin of *Y. valida* for this study.

Taxon	Accession No.
Agave attenuata	NC_032696.1
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
Yucca brevifolia	NC_032711.1
Yucca filamentosa	NC_032712.1
Yucca queretaroensis	NC_032713.1
Yucca schidigera	NC_032714.1

664

665 **Supplementary Table S3.** Global genetic diversity of *Y. valida* and by lineage in 4,411
666 nuclear SNPs. Number of individuals per lineage (n) and genetic diversity descriptors
667 (H_O, observed heterozygosity; H_E expected heterozygosity; F_{IS}, inbreeding coefficient).
668

Lineage	**n**	H_O	H_E	F_{IS}
Northern	54	0.25	0.30	0.17
Central	72	0.26	0.30	0.17
Southern	21	0.26	0.32	0.17
Global	**147**	**0.26**	**0.31**	**0.17**

669