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 events have caused habitats fragmentation and populations divergence. One example 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, located in western North America, exhibit strong genetic structure and highly concordant differentiated lineages across their ranges, as they were isolated during the Last Interglacial and became genetically differentiated over time. A representative taxon of this Peninsula is the genus *Yucca*, with *Yucca valida* having the widest range. Although Y. valida is a dominant species, there is an extensive distribution discontinuity between 26° N and 27° N, where any individuals have been identified, suggesting restricted gene flow rates; moreover, the historical distribution models indicate a substantial reduction in its range during the Last Interglacial, making it an interesting model for studying genetic differentiation. **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 the accumulation of genetic differentiation was driven by genetic drift. We estimated an age under one million years old for the last common ancestor of *Y. valida* and its sister species. **Conclusions**: The drivers of genetic divergence in Y. valida have been historical climate changes, limited dispersal capacity, and an extensive geographical range gap.

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 demographic fluctuations or populations' extrinsic barriers lead to differentiation (Mayr et al., 1963; Barton, 2008; Scheiner and Mindell, 2020), and divergence in the face of gene flow, in which selection and local adaptation drive this differentiation by cumulatively restricting the exchange of barrier loci (Darwin, 1859; Tigano and Friesen, 2016; Stankowski and Ravinet, 2021). Identifying how genetic differentiation is accumulated between and within taxa is fundamental for understanding the evolution of species (Han et al., 2017; Irwin et al., 2018; Shang et al., 2022) and permits linking species' demographic histories to geographical regions' climatic and geological histories (Avise, 2000).

Globally, barriers triggered by climatic events have caused habitats fragmentation and, consequently, populations divergence (Kumar and Kumar, 2018). One example of this are the 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 oscillations have been cyclic, divergence followed by a secondary contact due to habitat's 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 the causes and consequences of genetic divergence (Petit and Hampe, 2006). Under wide geographic ranges, plant populations are expected to be particularly subject to demographic changes related to genetic drift (e.g., habitat fragmentation) and gene flow (e.g.,

changes in their dispersers dynamics) (Sork et al., 2016). Notably, diverse plant species from the Baja California Peninsula (BCP), located in western North America, exhibit strong genetic structure and highly concordant differentiated lineages across their ranges. Populations of cacti, ocotillos, palms, trees, and agaves across the BCP were isolated during the Last Interglacial (LIG, 120 Kya) and have accumulated genetic differences in their genomes (Nason et al., 2002; Clark-Tapia and Molina-Freaner, 2003; Fehlberg and Ranker, 2009; Garrick et al., 2009; Klimova et al., 2017, 2018; Martínez-Noguez et al., 2020). This is because, while during the Last Glacial Maximum (LGM, 20 Kya), the deserts in the Northern Hemisphere increased their distribution as the climate was drier, the Interglacial was highly humid, producing a significant fragmentation of these ecosystems (Qiu et al., 2011; Meng et al., 2015).

A representative taxon of the BCP 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 characterized by a long-life cycle and slow population growth and recruitment rates (Turner et al., 2005). Its pollination depends on an obligate mutualism with moths of the genera *Tegeticula* and *Parategeticula*, which disperse pollen over short distances (Pellmyr, 2003), while seed dispersal has been related to extinct megafauna and is currently mediated by rodents, birds, and ruminants (Lenz, 2001). Three *Yucca* species are distributed across the BCP: *Yucca schidigera*, *Yucca capensis,* and *Yucca valida*, the latter two being endemic and sister taxa that are pollinated by a single moth species, *Tegeticula baja* (Pellmyr et al., 2007, 2008). Among the *Yuccas* of the BCP, *Y. valida* has the widest range, from ~29.8° N to ~25.6° N, across the Central and Vizcaino deserts, and the Magdalena Plains (Turner et al., 2005; Thiede, 2020). Whereas populations of *Y. valida*

are continuous across their northern range, there is an extensive gap between 26° N and 27° N (our personal observation), where any individuals have been identified, suggesting restricted gene flow. Furthermore, while *Y. valida* is a widespread species, the historical distribution models indicate a substantial reduction in its range during the LIG (Arteaga et al., 2020). As a result, this species is an interesting model for studying genetic differentiation.

We examined the phylogeography of *Y. valida* throughout its range. We hypothesized that past climatic fluctuations caused populations isolation, and since this species has a short-distance dispersal that limits genetic interchange between distant populations (Alamo-Herrera et al., 2022), we expected to find divergent lineages across its distribution. We aimed to i) identify the number of genetic lineages of *Y. valida* across its range, ii) quantify the levels of genetic differentiation between these lineages, iii) reconstruct their demographic history and iv) estimate the species' age. This information will contribute to the understanding of the processes shaping genetic differentiation in the BCP and Western North America.

MATERIALS AND METHODS

Sampling, DNA extraction and sequencing — Fresh leaf tissue samples of 120 plants were collected from 20 locations across the species range (Figure 1, 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).

Genomic DNA was converted into nextRAD sequencing libraries (SNPsaurus, LLC) as in Russello et al. (2015). Briefly, DNA was fragmented with Nextera reagent (Illumina, Inc), which also ligates short adapter sequences to the ends of the fragments. The reaction was scaled to fragment 20 ng of DNA, and 50 ng of DNA were added to compensate for DNA degradation and to increase the number of fragments. The DNA was amplified for 27 cycles at 74°C. The libraries were sequenced on a HiSeq4000 150 bp single-end line (University of Oregon). The dataset was enriched with 40 raw sequences of *Y. valida* (NCBI SRA: SRR11514779) and 40 of *Y. capensis* (NCBI SRA: SRR11514778) from Arteaga et al. (2020).

Raw data processing and SNP–calling — The quality of the 160 demultiplexed raw sequences of *Y. valida* was evaluated through FastQC 0.11.9 (Andrews, 2017) and MultiQC 1.14 (Ewels et al., 2016). Adapter removal was done with custom scripts (SNPsaurus, LLC) using the *bbduk* module of BBTools 1.0 (Bushnell, 2023). Clean reads were forced to 110 bp in length. A *de novo* assembly was performed in Stacks 2.6 (Catchen et al., 2013). Parameter optimization followed Mastretta-Yanes et al. (2015) and Paris et al. (2017), and 13 samples with less than 1M reads were removed as a

quality control filter before the assembly. We established a minimum stack depth equal to eight (*-m* 8), a maximum of three mismatches between stacks (*-M* 3), and a maximum of four mismatches between catalogue loci (*-n* 4). In the module *populations*, we called the SNPs present across the remaining 147 samples, with a minimum genotyping rate of 0.5 and a minor allele frequency of 0.05. When a 110 bp locus showed more than one SNP, only one variant was chosen, following the priorities in Alexander (2020) to ensure SNPs were mostly unlinked (Pearman et al., 2022).

Populations' structure — We used nuclear SNPs to assess the most likely number of genetic clusters across all samples and the membership of each sample to these clusters using three approaches: i) ADMIXTURE 1.3.0 (Alexander & Lange, 2011) was run with K = 1 – 20, and the optimal number of clusters was chosen using the cross–validation procedure, ii) a neighbour–joining tree from the samples' pairwise genetic distance matrix (as allele counts, converted on the R package ape (Paradis et al., 2004; R Core Team, 2022)), and iii) a Principal Components Analysis (PCA) were produced in Plink 1.90 (Purcell et al., 2007). To avoid over- or underestimating populations' structure (Janes et al., 2017), we verified and conciliated the assignment of samples to genetic clusters for the three approaches.

Genetic diversity and divergence — Once the samples were allocated into discrete genetic clusters (corresponding to three lineages, see *Results*), the genetic diversity descriptors (H_E, H_O, and F_{IS}) for each lineage were calculated in GENODIVE 3.0 (Meirmans, 2020). Genetic data conversion was completed using PGDSpider 2.1.1.5 (Lischer and Excoffier, 2012).

To associate the levels of genetic diversity and differentiation within and between lineages to the evolutionary forces shaping their divergence, all the nucleotides (variant

and invariant) of the scaffolds containing the 4,411 SNPs (see *Results*) used to assess populations' structure were called with Stacks, then, the mean values of Weir and Cockerham's genetic differentiation (F_{ST}; Weir and Cockerham, 1984), genetic divergence (d_{XY}; Nei and Miller, 1990), and nucleotide diversity (π; Nei and Li, 1979) for every scaffold of 110 bp were computed in pixy 1.2.7 (Korunes and Samuk, 2021), and the correlations between estimates by pairs of populations were compared. Following Kessler et al. (2023) and Shang et al. (2022), we characterized each correlation according to the potential evolutionary mechanisms driving it as 1) direct proportion between F_{ST} to d_{XY} and inverse proportion between both F_{ST} and d_{XY} to π (Divergence under gene flow); 2) uniform d_{XY} across different values of F_{ST} and π, with an inverse proportion between F_{ST} to π (Selection under allopatry); inverse proportion between both d_{XY} and π to F_{ST}, and direct proportion between F_{ST} to π, characterized by 3) correlations to be subtle, with mostly high F_{ST}, low π and d_{XY} (Background selection) or 4) strong correlations, with mostly low F_{ST} with high π and d_{XY} (Balancing selection). As our interest were the loci shaping divergence (with highly increased or reduced differentiation), the estimates were classified based on their percentiles' distribution as: high (> 95%) and low (< 5%) -or uniform (> 45%, < 55%) for d_{XY}, and then assigned to each of the mechanisms described.

Outlier SNPs and SNPs associated with environmental variables — First, we scanned for outlier SNPs using PCAdapt 4.0.1 (Luu et al., 2017), with K = 3 and $\alpha = 0.1$ for the false discovery rate boundary. Then, we used the Latent Factor Mixed-effect Model (Frichot et al., 2013) to characterize the presence of SNPs with a strong association with environmental variables. LFMM was run ten times (K = 3) on the R package LEA 3.2.0 (Frichot and François, 2015). A total of 20 environmental variables

obtained from WorldClim v1.4 (Hijmans et al., 2005) were tested for each sampling location. The Gibbs sampler algorithm was used for 10,000 cycles (burn-in = 5,000), and the median of the correlation scores (z-score) was calculated across all runs, recalibrating the p-values frequency distribution with lambda. We avoided any complimentary correction procedures for the false discovery rate in both PCAdapt and LFMM to increase the probability of detecting outliers and SNPs correlated with local adaptation. The SNPs detected by any procedure were removed from the dataset to produce a neutral catalogue for the demographic history analyses.

Demographic history — To understand the succession of events that could originate the observed nuclear genetic lineages, the demographic history of *Y. valida* was reconstructed using the diffusion approximation method δaδi (Gutenkunst et al., 2009). Simulations included all four possible three-populations simple models of i) simultaneous, ii) admixed, and sequential divergence from iii) North to South and iv) South to North, without gene flow or changes in populations' sizes, as developed by Portik et al. (2017). The site frequency spectrum was computed as folded, and the best model was selected using the maximum likelihood and the Akaike information criterion (AIC) values.

Putative whole chloroplast genome sequences recovery — We implemented a reference-guided workflow to reconstruct whole-chloroplast-genome sequences following Aleman et al. (2023). The 200 raw sequences of *Y. capensis* and *Y. valida* were cleaned as above, without forcing a maximum length and establishing a minimum length = 100 bp. The clean reads were mapped to the plastome of *Y. schidigera* (GenBank: NC_032714.1) in bowtie 2.2.6 under the *--very-sensitive-local* option (Langmead and Salzberg, 2012). Mapping statistics were evaluated with the flagstat and

coverage modules of SAMtools 1.15.1 (Li et al., 2009). Nucleotide calling was performed individually for each sample in ANGSD (Korneliussen et al., 2014), producing linear sequences by requiring minimum mapping and base qualities of 20, and outputting missing data as Ns *(-dofasta 2 -minMapQ 20 -minQ 20 -doCounts 1*).

Phylogenetic relationships and molecular clock analyses — The intraspecific phylogenetic relationships of *Y. valida* were reconstructed using the plastome sequences with a breadth of coverage >0.5 (130 samples) in RAxML 8.2.12 (Stamatakis, 2014), establishing *Y. schidigera* as the outgroup. A GTR + G + I model was used, letting RaxML halt automatic bootstrapping.

As the intraspecific phylogenetic analysis revealed a lack of differentiation among the plastomes of *Y. valida* (see *Results*), the divergence time of *Y. valida* and *Y. capensis* was estimated as an approximation to the age of *Y. valida*. The phylogenetic relationships of the Agavoideae subfamily were reconstructed based on the molecular clock, replicating Smith et al. (2021) in BEAST 1.8.4 (Drummond and Bouckaert, 2015). Based on Scheunert et al. (2020) and Roy et al. (2020), we produced consensus sequences for each *Y. valida* and *Y. capensis*, using all the sequences available on each species. The consensus sequences were aligned to the plastomes of 15 Agavoideae species (Supplementary Table S2, (McKain et al., 2016)) with MAFFT 7 default settings (Katoh et al., 2019) and the flag *--nzero*. A GTR + I model was employed. Gamma-distributed priors for the five substitution types and uniform priors (0-1) for the base frequencies and the proportion of invariant sites were established. An uncorrelated relaxed clock model with log-normal distributed rates was set up (mean = 1, standard deviation = 0.33), assuming a Yule process for the tree prior, starting from a random tree, establishing *Hosta ventricosa* as the outgroup, and forcing the ingroup's monophyly

(Drummond and Bouckaert, 2015). A log-normal prior distribution on the age of the ingroup was specified to 14.5 million years (offset, mean = 1 standard deviation = 1.4), based on the minimum age of the fossil *Protoyucca shadesii* (Tidwell and Parker, 1990; Iles et al., 2015). We completed two independent runs of 60 million steps long with a sample every 6,000,000 steps, and the runs were joined in LogCombiner 1.8.4, considering a 10% burn-in for each input. A maximum credibility tree was produced in TreeAnnotator 1.8.4, and each node's means, medians, and 95% confidence intervals were outputted.

RESULTS

Population structure, genetic diversity and divergence, demographic history — Using 4,411 SNPs, the nuclear genetic structure analyses established the most likely number of lineages as three (*K* = 3) (Figure 1). The lineages were named after their geographic location as Northern, Central, and Southern. 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 was observed (Supplementary Table S3).

The estimates of genetic differentiation (F_{ST}) between populations were subtle, with the values between the Northern and Southern lineages being the greatest (North-South: 0.06; North-Central: 0.03; Central-South: 0.05). The correlations between measures of genetic diversity and divergence showed similar patterns between pairs of lineages (Figure 2): low d_{XY} with a minimal positive correlation to F_{ST}, low π with a minimal inverse correlation to F_{ST}, and a direct proportion between d_{XY} to π. A total of 41 scaffolds were consistent with one of the four divergent scenarios, 3 to Background selection and 38 to Balancing selection, and 4370 scaffolds were identified as consistent with genetic drift.

PCAdapt detected 20 outlier SNPs, and 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 were removed to produce the neutral dataset. 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 molecular clock analyses — 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 for this genome. 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 ≈ 730,900; median ≈ 646,900; 95% CI ≈ 310,000 – 1,450,000).

DISCUSSION

The patterns of genetic divergence of taxa within the BCP reflect the influence of geological processes shaping its formation and climatic oscillations during past glacial cycles. (Dolby et al., 2015; e.g., Pérez-Alquicira et al., 2023). Here, we identified three monophyletic nuclear lineages across the range of *Y. valida*, accompanied by a single chloroplast lineage. The age estimated for the last common ancestor of *Y. capensis* and *Y. valida* –below 1 MYA, suggests that the nuclear differentiation in *Y. valida* is a consequence of a recent event. While the time of origin of the three nuclear lineages is uncertain, the ecological niche model projections predicted that during the LIG, the area with suitable conditions for *Y. valida* had a substantial reduction (Arteaga et al., 2020). Climate fragmentation, limited dispersal capacity, and an extensive geographical range gap have inhibited secondary contact between populations. This has led to the accumulation of genetic differentiation, as indicated by patterns of genetic divergence driven by drift.

The values of genetic differentiation between the three nuclear lineages were subtle, the highest being between bordering lineages (paired F_{ST} North-South = 0.06). This value is in agreement with those observed on microsatellite loci for *Y. valida* (F_{ST} = 0.059; (Alamo-Herrera et al., 2022)), *Y. brevifolia* (F_{ST} = 0.061; (Starr et al., 2013)), and *Y. schidigera* (F_{ST} = 0.067; (De la Rosa-Conroy et al., 2020)), and greater than its sister species, *Y. capensis* (F_{ST} = 0.022; (Luna-Ortiz et al., 2022)). The low levels of differentiation can be explained by potentially large historical effective population sizes (Arteaga et al., 2020) and reflect restricted gene flow between neighbouring lineages. Previous estimates of the dispersal rates indicate that pollen movement in *Y. valida* is approximately 50 m (Alamo-Herrera et al., 2022), and seed-mediated dispersal seems

to also occur at short distances (Waitman et al., 2012). In this sense, the patterns of genetic differentiation coincide with climate-driven fragmentation and restricted dispersal over long distances to be the major causes of genetic divergence in this genus.

The estimated genetic diversity has a similar value among the three lineages. The genetic diversity estimates (H_O = 0.26, H_E = 0.31) were higher than those previously described for this species with the same type of genetic markers (H_O = 0.11, H_E = 0.17; (Arteaga et al., 2020)). The disparity among these values reiterates the need for additional research to ascertain the accuracy of SNP calling for population genomics (O'Leary et al., 2018). Arteaga et al. (2020) assembled loci from samples of two species —*Y. valida*, *Y. capensis*, and their hybrid. Here, we generated a SNP dataset with samples of a single species, thus modifying the number of sites assembled. SNP-calling among differentiated taxa implies retrieving alleles fixed for a species or a population that are polymorphic between species or populations, affecting the estimates of genetic diversity (Shafer et al., 2017). This catalogue-size bias highlights that diversity estimations based on SNPs should be exposed carefully. Similarly, the inbreeding coefficient had lower levels (F_{IS} = 0.17) than those in Arteaga et al. (2020) (F_{IS} = 0.35). The high genetic diversity and low inbreeding levels 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 single vicariant 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. These findings support

the role of the Quaternary climate oscillations, which reduced population sizes during the LIG, in shaping the genetic diversity distribution of plant species across the BCP (Nason et al., 2002; Clark-Tapia and Molina-Freaner, 2003; Fehlberg and Ranker, 2009; Garrick et al., 2009; Klimova et al., 2017, 2018; Martínez-Noguez et al., 2020).

CONCLUSIONS

Determining the impact of the drivers of species' genetic diversity and differentiation is a crucial aspect of evolutionary biology. Our results highlight the consequences of climate-driven fragmentation during the Quaternary on populations divergence across the BCP. The overall comparable genetic diversity among the three lineages suggests that the species was already distributed before this divergent mechanism occurred. The history of *Y. valida* is one of allopatric differentiation; therefore, addressing the origin and colonization of *Yucca* in the Peninsula is an essential research gap in the evolutionary history of this genus.

ACKNOWLEDGEMENTS

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. We thank Camille Kessler and Flavia Termignoni-García for their bioinformatic advice and Enrique Ruiz for his work in Fig. 3. This work was financially 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. We respectfully acknowledge that the territory where the samples were collected are 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.

AUTHOR CONTRIBUTIONS

Conceptualisation: MCA. Developing methods, conducting the research, data interpretation and writing: AA, JGP, MCA, and RBB. Data analysis and preparation of figures & 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. Shafer, T. Patel, P. A. Volkova, and J. R. Freeland. 2023. Development of genomic resources for cattails (Typha), a globally important macrophyte genus. 2023.04.21.537876.

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.

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

Dolby, G. A., S. E. K. Bennett, A. Lira-Noriega, B. T. Wilder, and A. Munguía-Vega. 2015. Assessing the Geological and Climatic Forcing of Biodiversity and

Evolution Surrounding the Gulf of California. *Journal of the Southwest* 57: 391–455.

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

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

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

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

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

Garrick, R. C., J. D. Nason, C. A. Meadows, and R. J. Dyer. 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.

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

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

Harrington, S. M., B. D. Hollingsworth, T. E. Higham, and T. W. Reeder. 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.

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.

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.

Klimova, A., A. Ortega-Rubio, D. L. J. Vendrami, and J. I. Hoffman. 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.

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.

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

Meirmans, P. G. 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.

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

Nason, J. D., J. L. Hamrick, and T. H. Fleming. 2002. Historical Vicariance and Postglacial Colonization Effects on the Evolution of Genetic Structure in Lophocereus, a Sonoran Desert Columnar Cactus. *Evolution* 56: 2214–2226.

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

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

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

de Oliveira, F. F. R., M. Gehara, M. Solé, M. Lyra, C. F. B. Haddad, D. P. Silva, R. F. de Magalhães, et al. 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.

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

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

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

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

Pellmyr, O., M. Balcázar-Lara, K. A. Segraves, D. M. Althoff, and R. J. Littlefield. 2008. Phylogeny of the pollinating yucca moths, with revision of Mexican species (Tegeticula and Parategeticula; Lepidoptera, Prodoxidae). *Zoological Journal of 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.

Pérez-Alquicira, J., E. V. Wehncke, G. A. García-Loza, B. C. Carstens, C. A. Domínguez, R. Pérez-Ishiwara, F. E. Molina-Freaner, et al. 2023. Geographic isolation and long-distance gene flow influence the genetic structure of the blue fan palm Brahea armata (Arecaceae). *Journal of Plant Research* 136: 277–290.

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.

R Core Team. 2022. R: The R Project for Statistical Computing. Website https://www.r-project.org/ [accessed 4 October 2022].

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

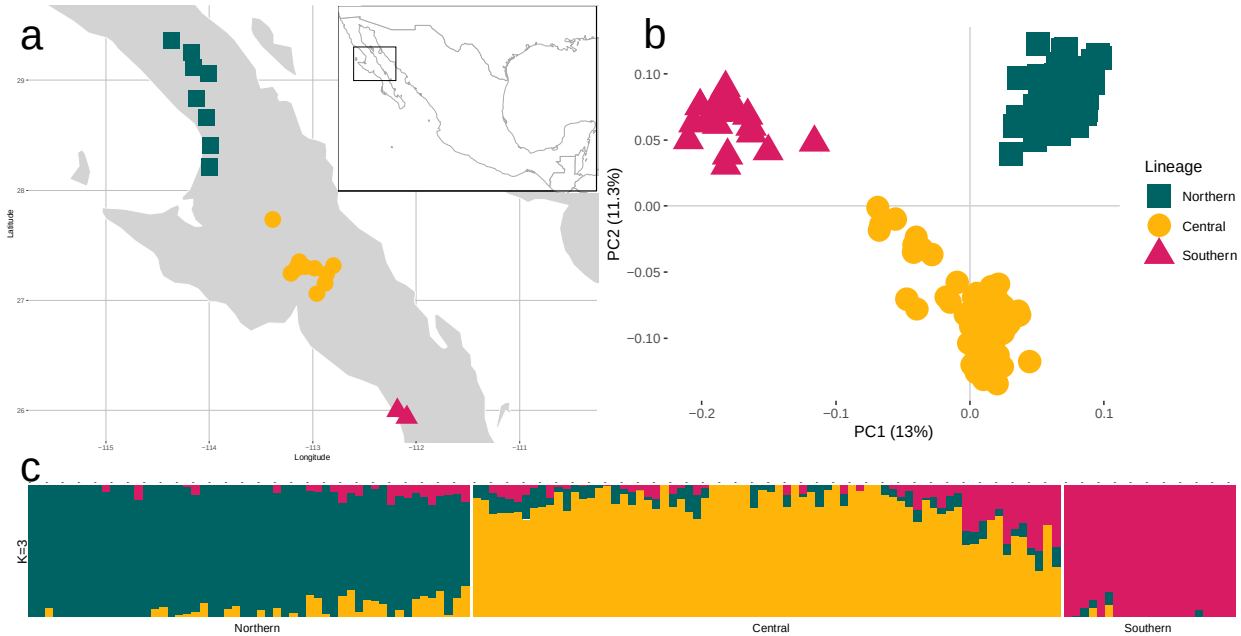

Figure 1. Genetic structure results of 4,411 nuclear SNPs obtained from 147 *Y. valida* samples. (a) Sampling locations (20) for this study. Shapes represent sampling locations, and colours represent genetic lineages. (b) Principal Component Analysis for PC1 and PC2. Shapes represent individuals, and 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).

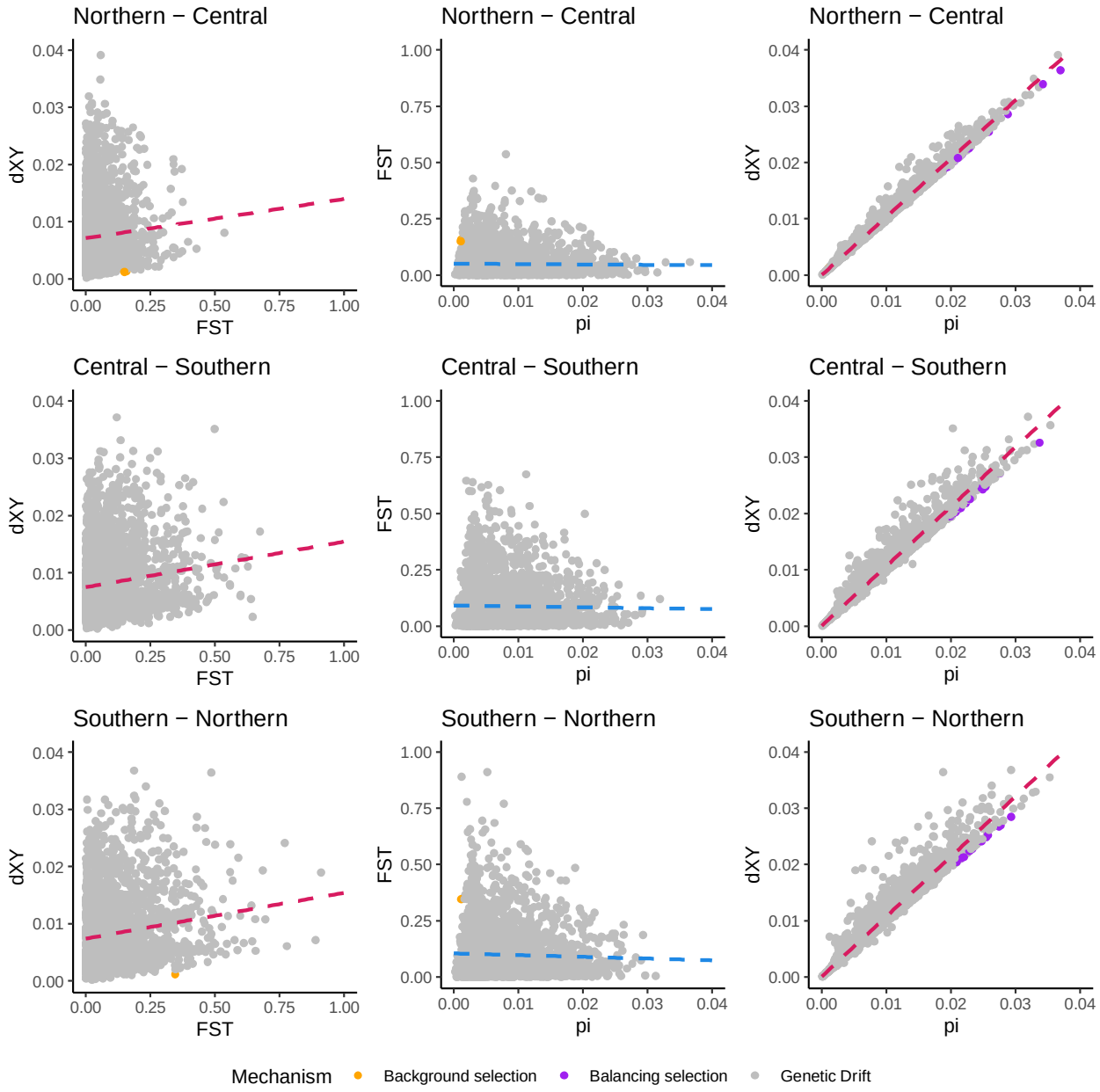

Figure 2. Correlation analysis for F_{ST}, d_{XY} and π across 4,411 scaffolds (110bp) between pairs of nuclear genetic lineages of *Y. valida* found in this study. Each point represents a scaffold, and it is coloured according to the potential evolutionary mechanism driving genetic differentiation following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

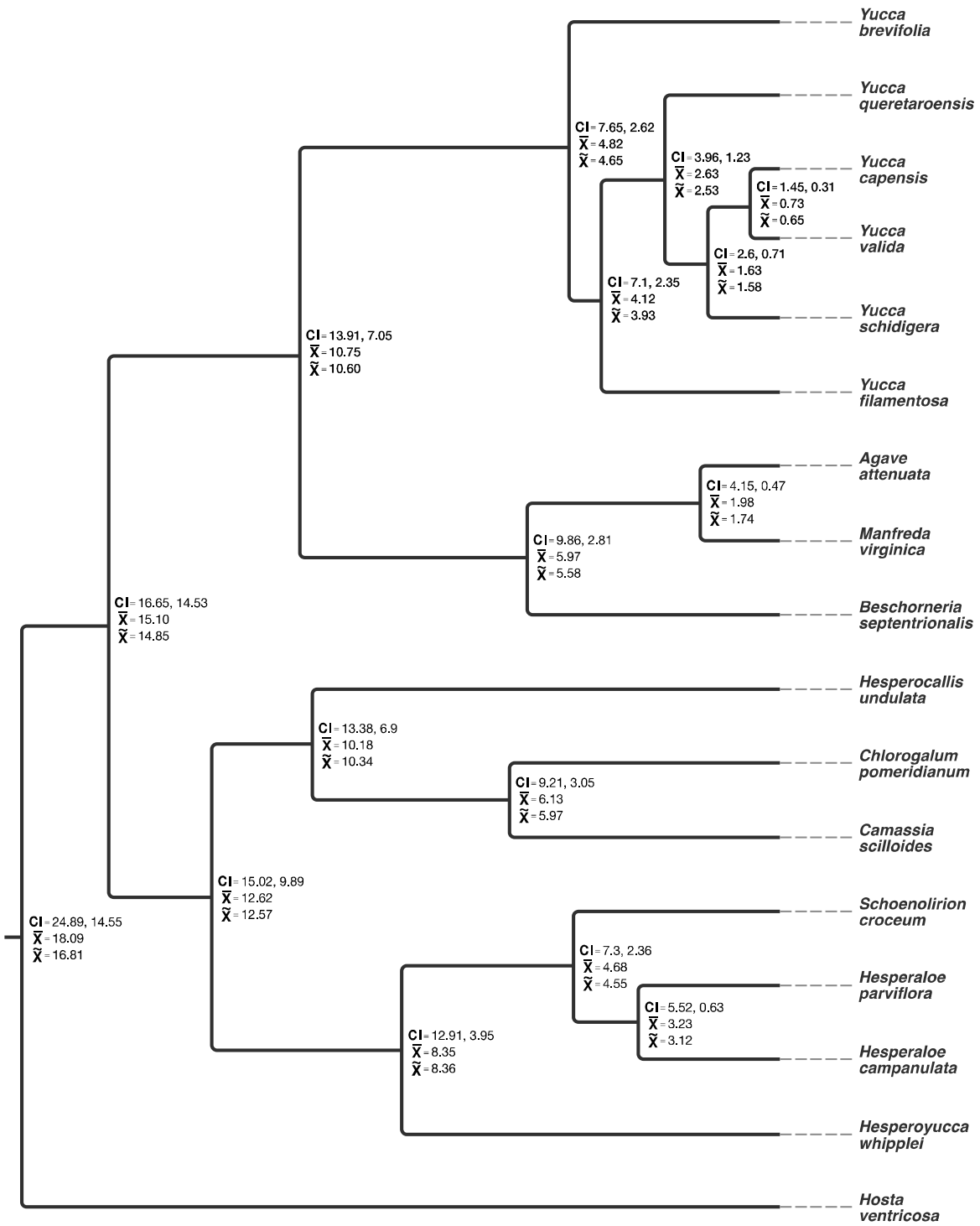

Figure 3. Chloroplast phylogeny of 17 Agavoideae species, estimated in BEAST 1.8.4. The tree was produced with *Hosta ventricosa* as the outgroup. Mean ($\bar{x}$), median ($\tilde{x}$), and 95% confidence intervals of nodes' divergence times are shown as millions of years from the present. Branch support on every node = 1.

Supplementary Table S1. Site IDs and coordinates of the sampling locations across the distribution of *Y. valida* for this study.

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

Supplementary Table S2. Taxa and GenBank accessions used to investigate the 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

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

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**