

INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: A Universal Toolkit for Flowering Plant Phylogenomics

On the potential of Angiosperms353 for population genomic studies

Madeline Slimp¹ , Lindsay D. Williams¹ , Haley Hale¹ , and Matthew G. Johnson^{1,2} 

Manuscript received 16 February 2021; revision accepted 15 March 2021.

¹Department of Biological Sciences, Texas Tech University, 2901 Main Street, Lubbock, Texas 79409, USA

²Author for correspondence: matt.johnson@ttu.edu

Citation: Slimp, M., L. D. Williams, H. Hale, and M. G. Johnson. 2021. On the potential of Angiosperms353 for population genomic studies. *Applications in Plant Sciences* 9(7): e11419.

doi:10.1002/aps.11419

PREMISE: The successful application of universal targeted sequencing markers, such as those developed for the Angiosperms353 probe set, within populations could reduce or eliminate the need for specific marker development, while retaining the benefits of full-gene sequences in population-level analyses. However, whether the Angiosperms353 markers provide sufficient variation within species to calculate demographic parameters is untested.

METHODS: Using herbarium specimens from a 50-year-old floristic survey in Texas, we sequenced 95 samples from 24 species using the Angiosperms353 probe set. Our data workflow calls variants within species and prepares data for population genetic analysis using standard metrics. In our case study, gene recovery was affected by genomic library concentration only at low concentrations and displayed limited phylogenetic bias.

RESULTS: We identified over 1000 segregating variants with zero missing data for 92% of species and demonstrate that Angiosperms353 markers contain sufficient variation to estimate pairwise nucleotide diversity (π)—typically between 0.002 and 0.010, with most variation found in flanking non-coding regions. In a subset of variants that were filtered to reduce linkage, we uncovered high heterozygosity in many species, suggesting that denser sampling within species should permit estimation of gene flow and population dynamics.

DISCUSSION: Angiosperms353 should benefit conservation genetic studies by providing universal repeatable markers, low missing data, and haplotype information, while permitting inclusion of decades-old herbarium specimens.

KEY WORDS Guadalupe Mountains National Park; herbarium; Hyb-Seq; target capture.

Universal target capture

The estimation of demographic parameters in natural populations is a critical tool for species delimitation (Duminil and Di Michele, 2009), biogeography studies (Overcast et al., 2019), and monitoring of populations and species in a dynamically changing environment (Allendorf et al., 2010). The feasibility of estimating demographic parameters (including heterozygosity, effective population size, and levels of introgression) in non-model taxa relies on retrieving homologous markers that allow detection of sufficient variation across the genome, while remaining cost-effective for the analysis

of hundreds of individuals. In plants, population genomic studies could benefit from markers that enable the further unlocking of herbarium specimens for botanical research, paralleling the impact of herbarium specimens in phylogenomics (e.g., Shee et al., 2020), microbiome research (e.g., Heberling and Burke, 2019), and studies of the effects of climate change on plant populations (e.g., Miller-Rushing et al., 2009).

Traditional Sanger sequencing of PCR amplicons often employs universal primer sequences, but the genes targeted (e.g., the plastid markers *matK* and *rbcl*, or the nuclear ribosomal ITS) do not

contain sufficient variation to allow species discrimination, due to high levels of gene tree paraphyly (e.g., Fazekas et al., 2009). If species cannot be differentiated, plastid marker analysis is also unlikely to provide satisfactory results within populations (e.g., Supple and Shapiro, 2018). Whole genome methods are also not feasible for many plant species with large genome sizes (e.g., *Fritillaria* L., 86 Gbp) (Kelly et al., 2015). Genome skimming, including extraction of organellar genomes and the ribosomal cistron, is popular for phylogenetics of non-model organisms (e.g., McKain et al., 2018). However, organellar regions do not provide sufficient numbers of unlinked sites to generate unbiased estimates of genetic diversity (McMahon et al., 2014). More recent efforts have focused on optimization of reduced-representation sequencing, including restriction site-associated DNA sequencing (RAD-seq), RNA sequencing (RNA-seq), and target capture via hybrid enrichment (Hyb-Seq). Sequencing RNA of non-model species (e.g., Johnson et al., 2016; Yang et al., 2018), while generating data for tens of thousands of genes, is still prohibitively expensive at the population level and its use is therefore limited in conservation genomics. In contrast, RAD-seq (and related methods like genotyping by sequencing [GBS] and double-digest RAD sequencing) has gained in popularity for population genetics because it does not require taxon-specific primers and may be used in any organism to potentially generate thousands of markers (Andrews et al., 2016). However, using RAD-seq without a reference genome or reference sequence requires that short sequences must be clustered by sequence similarity across samples to identify loci (Eaton and Overcast, 2020), often generating large amounts of missing data. Sites must therefore be rigorously filtered to ensure that null alleles do not bias population estimates (O’Leary et al., 2018).

In contrast, target-capture sequencing of exons and flanking non-coding regions (Hyb-Seq; Weitemier et al., 2014) provides several possible advantages for estimating demographic parameters. Target-capture methods generate data for hundreds of loci at a reasonable per-sample cost (Hale et al., 2020), result in data sets that often have limited missing data compared with RAD-seq (Carter et al., 2019), and can work well with degraded DNA such as what can be retrieved from herbarium specimens (Brewer et al.,

2019). The development of universal probe sets for Hyb-Seq across large groups (Buddenhagen et al., 2016; Johnson et al., 2019) further suggests a role for Hyb-Seq at the population level, if the loci prove to be variable. In this case study, we utilize target-capture sequencing on a collection of herbarium specimens from Guadalupe Mountains National Park. We use Angiosperms353 (Johnson et al., 2019), a probe set designed to capture 353 nuclear protein-coding genes from any flowering plant. Although successful target capture has been demonstrated for phylogenetic scales (e.g., Antonelli et al., 2021; Shah et al., 2021; Zuntini et al., 2021), whether the loci are sufficiently variable within species to calculate demographic parameters is not established. We sample 24 species from the Guadalupe Mountains National Park collection at the population level in order to explore the suitability of Angiosperms353 for population-focused genomic studies.

The GUMO collection as a test case

The Guadalupe Mountains National Park (GUMO), which is located in the Trans-Pecos region of far-west Texas, has been described as the “most botanically diverse area of Texas” (Eason, 2018) due to geological variation and the intersection of three ecoregions: desert scrublands, high plains, and montane forest at higher elevations (Fig. 1). The park contains the basin-range system of the Chihuahuan Desert and abrupt elevation changes, including the Guadalupe Peak at 2667 m. For a century prior to GUMO opening in 1973, it was used for agriculture, including livestock husbandry. Goats grazed the landscape of the park, heavily disturbing natural vegetation balances within populations (Glass et al., 1974). The Guadalupe Mountains are isolated from other high-altitude peaks in the southern United States and Mexico, which may impact the future persistence of many plant species in a warming climate. Characterizing the combined impacts of land use and climate change is an important goal for the park’s staff, and ideally would include conservation genomic studies for species of concern (Allen, 2013). Northington and Burgess (1976) made the first comprehensive floristic survey of the newly created National Park, which was previously a goat ranch, by collecting an estimated 3000 specimens from over 400 species.

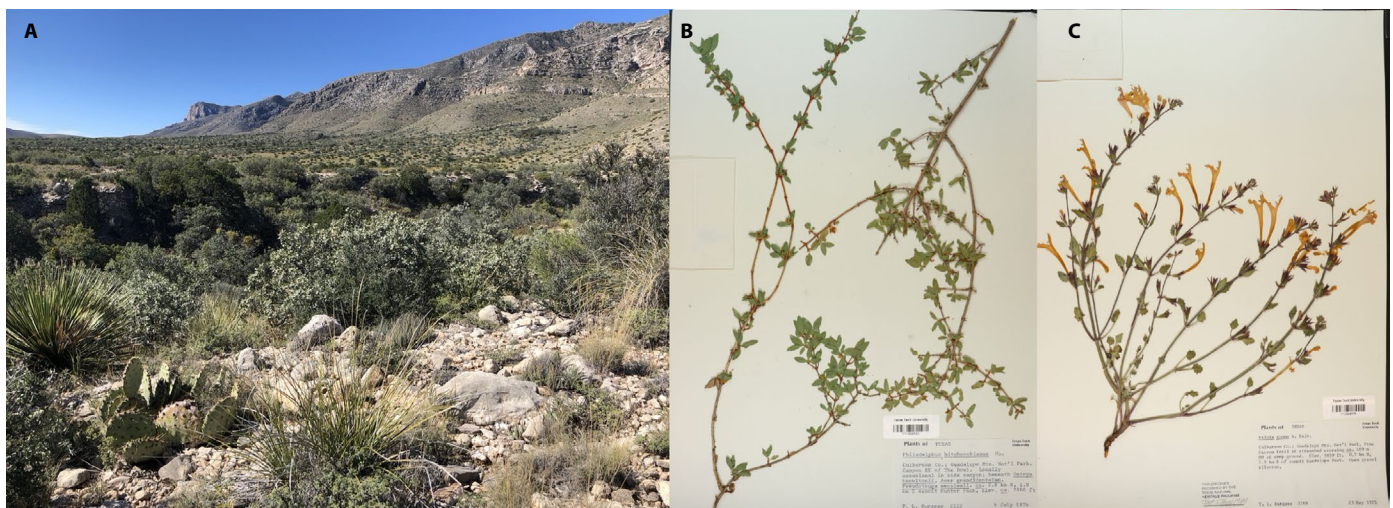


FIGURE 1. (A) Typical desert scrubland habitat in Guadalupe Mountains National Park (GUMO). (B, C) Representative herbarium specimens from the 1970s GUMO collection, *Philadelphus hitchcockianus* (B) and *Salvia summa* (C), two species with ranges restricted to the GUMO region.

They described 14 species as endemic to a limited area in and/or around GUMO, and an additional 37 species as having small ranges with their furthest extent in GUMO. Their collection, now housed at the E. L. Reed Herbarium at Texas Tech University (TTC; Thiers, 2021), includes 55 species, each with at least six unique specimens, raising the potential for estimating within-species demographic parameters based on these specimens. The extensive collection, which was not digitized until recently, provides a unique snapshot of a past botanical community, including rare species such as *Philadelphus hitchcockianus* S. Y. Hu and *Salvia summa* A. Nelson (Fig. 1B, C).

Determining whether land-use change and conservation have affected the genetic diversity of populations requires molecular markers that can work across distantly related species and can be used on herbarium specimens. We therefore aim to determine whether there is sufficient variability within Angiosperms353 genes to calculate population genetic parameters, and assess the suitability of this approach for working with herbarium-based resources like the GUMO collection, which requires high levels of gene recovery from degraded DNA typical of these specimens (Brewer et al., 2019). We focus on analyzing recovered gene data from the GUMO collection and calculating inbreeding coefficients, heterozygosity, and Tajima's D as a predictor of whether Angiosperms353 will be appropriate for larger population genomics projects. We aim to use our test-case study of the GUMO collection to demonstrate that Angiosperms353 is a universal, cost-effective, and high-output approach to reliably generate gene sequences for population-level study of any angiosperms of interest (Johnson et al., 2019).

MATERIALS AND METHODS

Sampling

We sampled species based on the availability of at least four distinct specimens per species in the GUMO collection at the E. L. Reed Herbarium, with a goal of sampling approximately 1 cm² of tissue from 24 species: eight grasses, six forbs, six shrubs, and four tree species (Appendix 1). We extracted DNA in 1.1-mL round-bottom tubes, to which we added tissue and two steel bearings. We froze the tissue in a SPEX Cryo-Station (SPEX SamplePrep, Metuchen, New Jersey, USA) and then ground the tissue using a SPEX Geno/Grinder MiniG tissue homogenizer (SPEX SamplePrep), processing all 95 specimens simultaneously. Due to the diverse species representation on the plate, grind quality differed between samples and was poor in species with extremely fibrous or glabrous tissue. We repeated freezing and grinding twice more for difficult tissues until a powder was observed for all specimens. We then proceeded with a typical cetyltrimethylammonium bromide (CTAB)/chloroform DNA extraction method (Doyle and Doyle, 1987) with three modifications: 0.4% 2-mercaptoethanol added to the CTAB buffer, incubation in CTAB for 10 h, and precipitation of DNA in isopropanol in a −20°C freezer for five days. These modifications have been shown to maximize the yield of DNA from herbarium specimens (Brewer et al., 2019). Samples for which DNA yield exceeded 5 ng/μL (estimated using a Qubit fluorometer [Thermo Fisher Scientific, Waltham, Massachusetts, USA]) and in which most DNA fragments were larger than 500 bp (via agarose gel) were sheared using the fragmentase enzyme (New England Biolabs, Ipswich, Massachusetts, USA) to a desired fragment size of 500 bp. We did not shear extracts with a DNA concentration or size distributions below either threshold, and

where necessary we concentrated DNA using a Savant SpeedVac vacuum centrifuge (Thermo Fisher Scientific).

Enriched library preparation

Using the NEBNext Ultra II DNA kit (New England Biolabs), we prepared dual-indexed Illumina sequencing libraries with all reagents in half-volume, targeting an insert size of 500–700 bp. Before library preparation, samples were either diluted or concentrated to achieve an input DNA concentration of ~200 ng in 25 μL. Following final PCR (eight cycles) and cleanup of libraries with prepared SPRI beads (Beckman Coulter, Brea, California, USA) (Rohland and Reich, 2012), we assessed concentration using a Qubit fluorometer. We measured fragment size distribution for libraries with a concentration less than 5 ng/μL using an Agilent TapeStation 2200 (Agilent Technologies, Santa Clara, California, USA) on 15 libraries, to reduce the cost of assessing all 95 libraries. The samples chosen for measurement using the TapeStation were selected based on concentration, as readings that are less than 5 ng/μL could mostly be adapter-dimer fragments instead of libraries (Hale et al., 2020). Target capture of the Angiosperms353 loci was done in four separate reactions, each of which contained 24 libraries pooled by their final concentration (pool 1: 0.139–10.8 ng/μL; pool 2: 10.9–15.1 ng/μL; pool 3: 15.4–22.9 ng/μL; pool 4: 23.1–39.3 ng/μL), regardless of species. Library pools were then concentrated by using a SpeedVac centrifuge (on low heat) until all liquid was removed, and we then re-eluted the libraries in 7 μL of nuclease-free water. The target-capture reactions followed Arbor Biosciences' myBaits hybridization capture for targeted NGS (manual version 4.01), except for a 3 : 1 dilution of the RNA probes compared with manufacturer recommendations. This modification has proven efficient for pooling 96 samples in targeted sequencing studies (Hale et al., 2020). Hybridization with the Angiosperms353 probes was performed at 65°C for 26 h, followed by PCR enrichment of the approximately 500-bp-insert libraries for 14 cycles. Each pool was analyzed on the TapeStation to combine the pools in equimolar fractions (10 nM/pool) before sequencing on the Illumina MiSeq v3 2 × 300 platform.

Data processing

We first processed sequence reads using Trimmomatic (Bolger et al., 2014) to remove adapter sequences and reads with an average Q score below 25. Angiosperms353 genes were recovered from cleaned Illumina reads using HybPiper (Johnson et al., 2016a). We used the “mega353” target file (McLay et al., 2021) to recover sequences; this file contains substantially more representative sequences per gene than the original Angiosperms353 target file. Briefly, the HybPiper workflow identifies potential matches between reads and target genes (using BLASTX [Camacho et al., 2009]) to allow for high phylogenetic distance between targeted sequences and sample reads, sorts the reads by gene, and conducts de novo assembly separately on each gene, mitigating the need for a reference sequence for non-model organisms. We used a target file containing representative amino acid sequences of the Angiosperms353 loci obtained from <https://github.com/mossmatters/Angiosperms353>. Two types of sequences were output from HybPiper: (1) coding sequences corresponding to the targeted genes and (2) “supercontigs” that contain both the exon sequences and flanking non-coding regions (i.e., introns and other untranslated regions). To test for genetic diversity in each species, we generated a set of reference

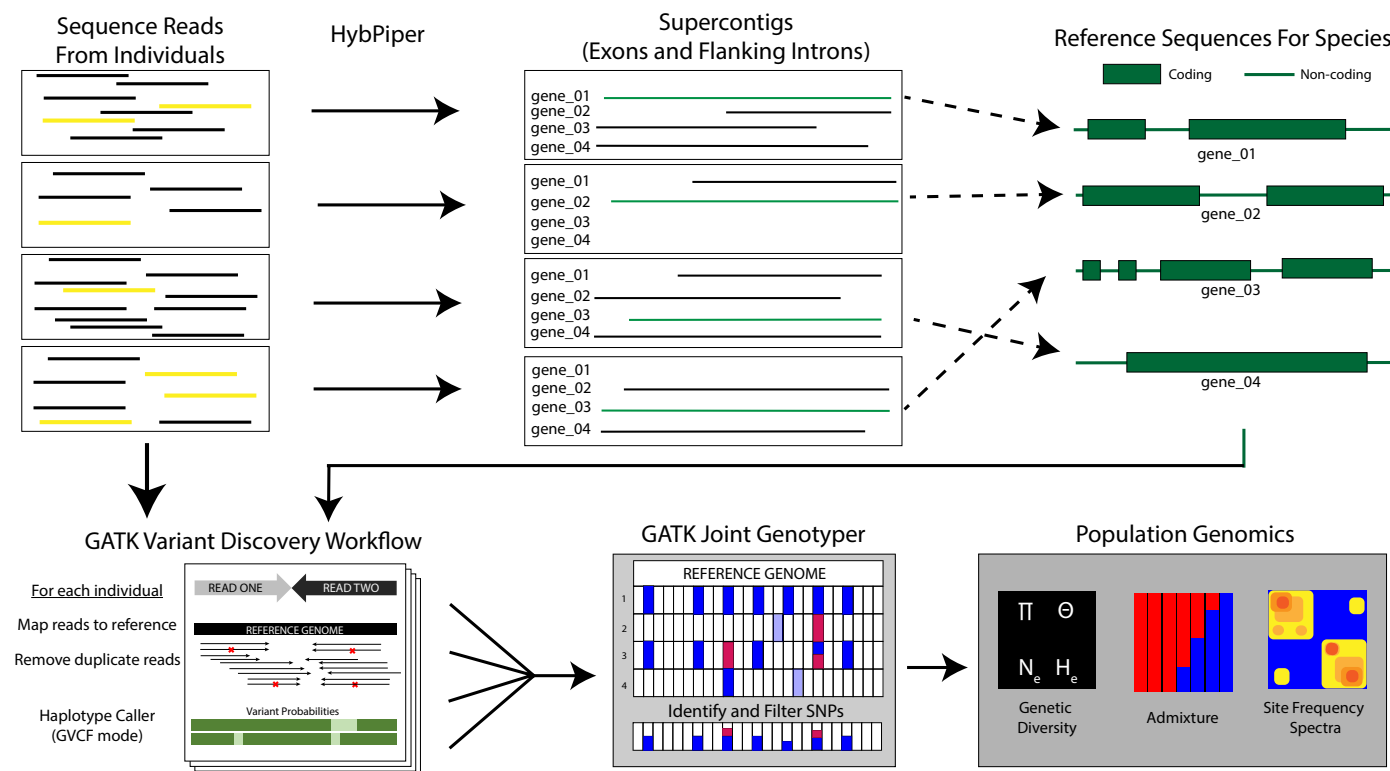


FIGURE 2. Data workflow for identifying SNPs within each species from target capture data. Reference sequences for each species were constructed by selecting the longest supercontig from each gene for each species.

sequences by selecting the longest supercontig recovered for that species for each gene (Fig. 2).

Variant detection and filtering

To detect variants within species, we followed the Germline Exome Best Practices Pipeline (DePristo et al., 2011) in GATK4 (McKenna et al., 2010). Briefly, we aligned reads to the reference sequences and merged these with unaligned reads (GATK MergeBamAlignment), followed by identifying duplicate read clusters based on read mapping rather than sequence identity (GATK MarkDuplicates). For each individual, we called provisional variants separately (GATK HaplotypeCaller in GVCF mode) and then called genotypes jointly for all individuals in a species (GATK JointGenotypeCaller) (Poplin et al., 2017). This produced an initial variant file containing both indels and single-nucleotide polymorphisms (SNPs), but only the latter were retained for genetic diversity analysis. We employed a hard filtering “ $QD < 5.0 \parallel FS > 60.0 \parallel MQ < 40.0 \parallel MQRankSum < -12.5 \parallel ReadPosRankSum < -8.0$ ” to remove SNPs only supported by low base quality or depth of coverage. Before calculating statistics on remaining variable SNP sites, we generated a reduced SNP data set to remove all SNPs with missing data within a species using PLINK (Purcell et al., 2007). Some population genetic analysis requires the use of unlinked SNPs, so a second unlinked data set was generated for each species by filtering SNPs exceeding a variance inflation factor (VIF) using a window (5 kbp), variant count (5 ct), and linkage threshold ($-indep\ 50\ 5\ 2$). The linkage threshold tests for linkage disequilibrium between pairs of SNPs within a defined window, but may be an overestimate due to the small sample size within each species.

Population genetics statistics

We calculated heterozygosity and inbreeding coefficients on the unlinked SNP data sets for each species using PLINK. We estimated Tajima’s D (which accounts for allelic variation) on the larger SNP data set (linked SNPs but no missing data) using vcfTools (Danecek et al., 2011). Critical values for detecting deviation from mutation-shift balance with four individuals were taken from Tajima (1989). All scripts used to conduct variant detection and population genetics statistics are publicly available at <https://github.com/lindsawi/HybSeq-SNP-Extraction>.

RESULTS

Gene recovery was successful for all species, although gene recovery rate (i.e., at least 25% of the targeted gene recovered) was variable—from an average of 64 genes recovered in *Bothriochloa springfieldii* (Gould) Parodi to 321 genes in *Nerisyrenia camporum* (A. Gray) Greene (Fig. 3C, Appendix S1). We found that gene recovery was typically above 200 genes when the number of mapped reads exceeded 25,000 (Fig. 3B). Target efficiency (percentage of reads mapping to target loci in HybPiper) ranged between 5% and 20%, except for *N. camporum*, which had an average target efficiency of 43%. Although much of the DNA extracted from the herbarium specimens was degraded, a third of the samples required further fragmentation before library preparation to increase the number of fragments in our desired size range (~500 bp). We found no significant difference in target efficiency between fragmented and nonfragmented samples (Welch’s two-sample *t*-test [$t(42.7) = -1.07$],

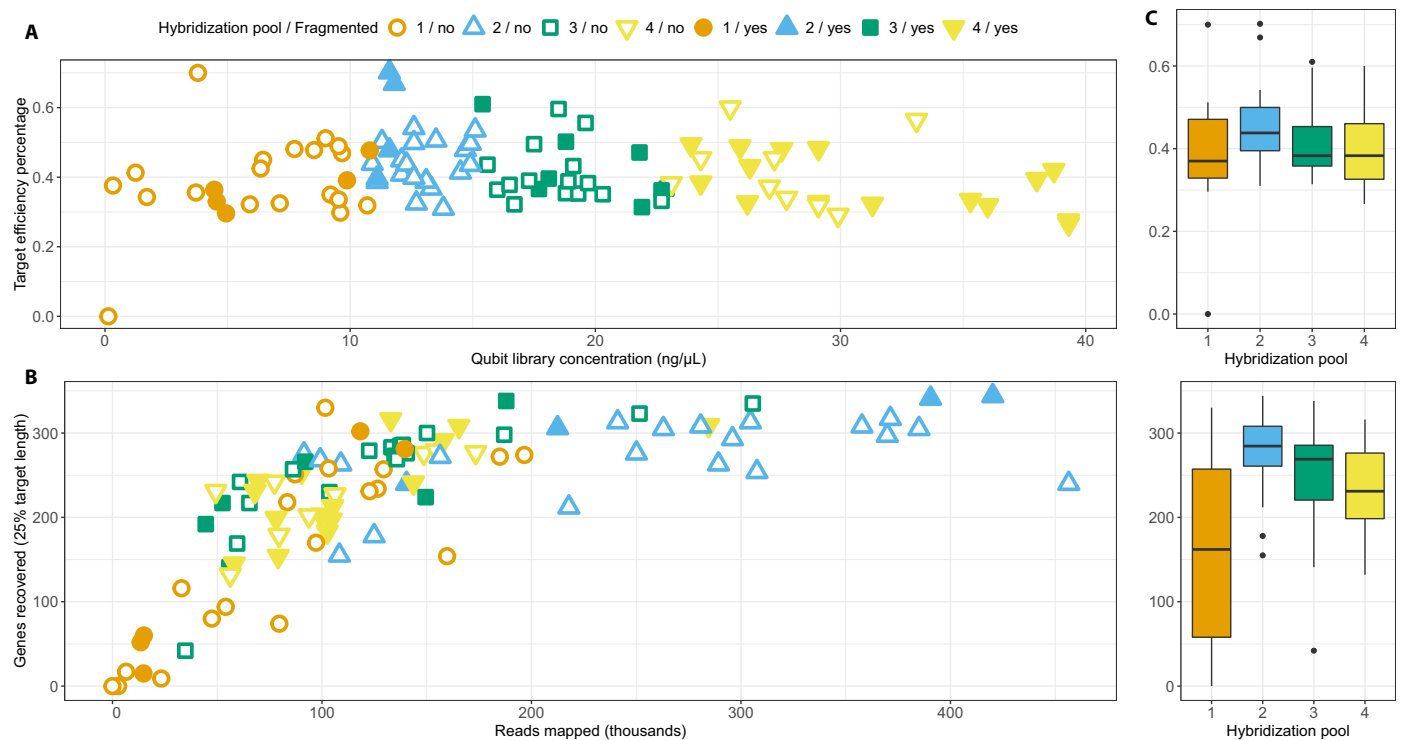


FIGURE 3. (A) Relationship between sequencing library concentration (measured on a Qubit fluorometer) and target capture efficiency (percentage of reads on target). (B) Relationship between gene recovery (at least 25% of the targeted genes recovered) and the number of reads on target. (C) Box plots showing the distribution of target efficiency (top) and gene recovery (bottom) by hybridization pool. Hybridization pools are color coded: pool 1 = orange, pool 2 = blue, pool 3 = green, pool 4 = yellow.

$P = 0.292$). The use of the “mega353” target file (McLay et al., 2021) substantially increased the gene recovery for many samples: more than 50 additional genes were recovered at 50% target length for *Morus celtidifolia* Kunth, *Sporobolus cryptandrus* (Torr.) A. Gray, and *Salvia summa* (Appendix S1).

Library concentration and pooling strategy leading into hybridization reactions were among the biggest contributors to target enrichment efficiency (Fig. 3). Pool 1, which had the lowest library concentrations (0.139–10.8 ng/μL), also had the lowest target capture efficiency and the most samples with poor gene recovery (<100 genes, Fig. 4). In each of the other pools, there was no association between target efficiency and library concentration.

We used the longest gene recovered for each species by HybPiper to serve as a reference sequence. Notably, two species (*Zuloagaea bulbosa* (Kunth) E. Bess. and *Quercus pungens* Liebm.) contained individual supercontigs (coding sequence and flanking non-coding sequence) with lengths over 10,000 bp. We retrieved a maximum supercontig length of over 5000 bp for an additional 14 species. We manually checked the longest supercontigs via manual inspection of sequences and depth, and did not find any evidence of misassembly. Total supercontig length from HybPiper was high compared to the total targeted region of Angiosperms353 (260 kbp of coding sequences). Three species (*Nerisyrenia camporum*, *Ungnadia speciosa* Endl., and *Zuloagaea bulbosa*) had a total supercontig length over 500 kbp (Fig. 4), with *N. camporum* being the most notable of these at 637,216 bp. Only one species (*Cyphomeris gypsophiloides* (M. Martens & Galeotti) Standl.) did not recover supercontigs greater than 100 kbp.

Our data workflow (Fig. 2) resulted in a high number of variants detected within species, even for species where gene recovery was poor. Fourteen species initially had over 9000 variants detected from just four individuals per species, prior to filtering with PLINK. Four species were found to have fewer than 5000 variants, including *Aristida adscensionis* L. (1913 variants) and *Asclepias asperula* (Decne.) Woodson (3213 variants). However, unlike the poor gene recovery noted above for *Bothriochloa springfieldii*, the *A. adscensionis* samples had an average of 177 genes recovered. Because many of the called variants are linked, analysis of heterozygosity within species was assessed at a subset of SNPs that contained no missing data, with one SNP chosen per linkage-disequilibrium (LD) block identified by PLINK. On average during filtering, 4.6–50.3% of variants within a species were removed by this filter. Average percent heterozygosity was high (above 20%) in most species (Fig. 5) and was highest in several widespread grass species including *Digitaria cognata* (Schult.) Pilg. and *Zuloagaea bulbosa*. Heterozygosity was also high for the rare forb *Salvia summa*, a NatureServe G3 vulnerable species endemic to Texas and New Mexico (New Mexico Rare Plant Technical Council, 1999), and was lowest for *Aristida adscensionis*.

Nerisyrenia camporum contained the most variants of any of the species studied in the GUMO collection, with over 46,000 SNPs recovered. After filtering to remove SNPs with missing data, *N. camporum* still contained 26,355 variants with 1700 unlinked variants identified across LD blocks, showing no evidence of linkage by pairwise correlation. Other species had substantial numbers of unlinked SNPs, including *Digitaria cognata* (1017) and *Sporobolus contractus* Hitchc. (1029). Pairwise nucleotide diversity (π) was lowest in

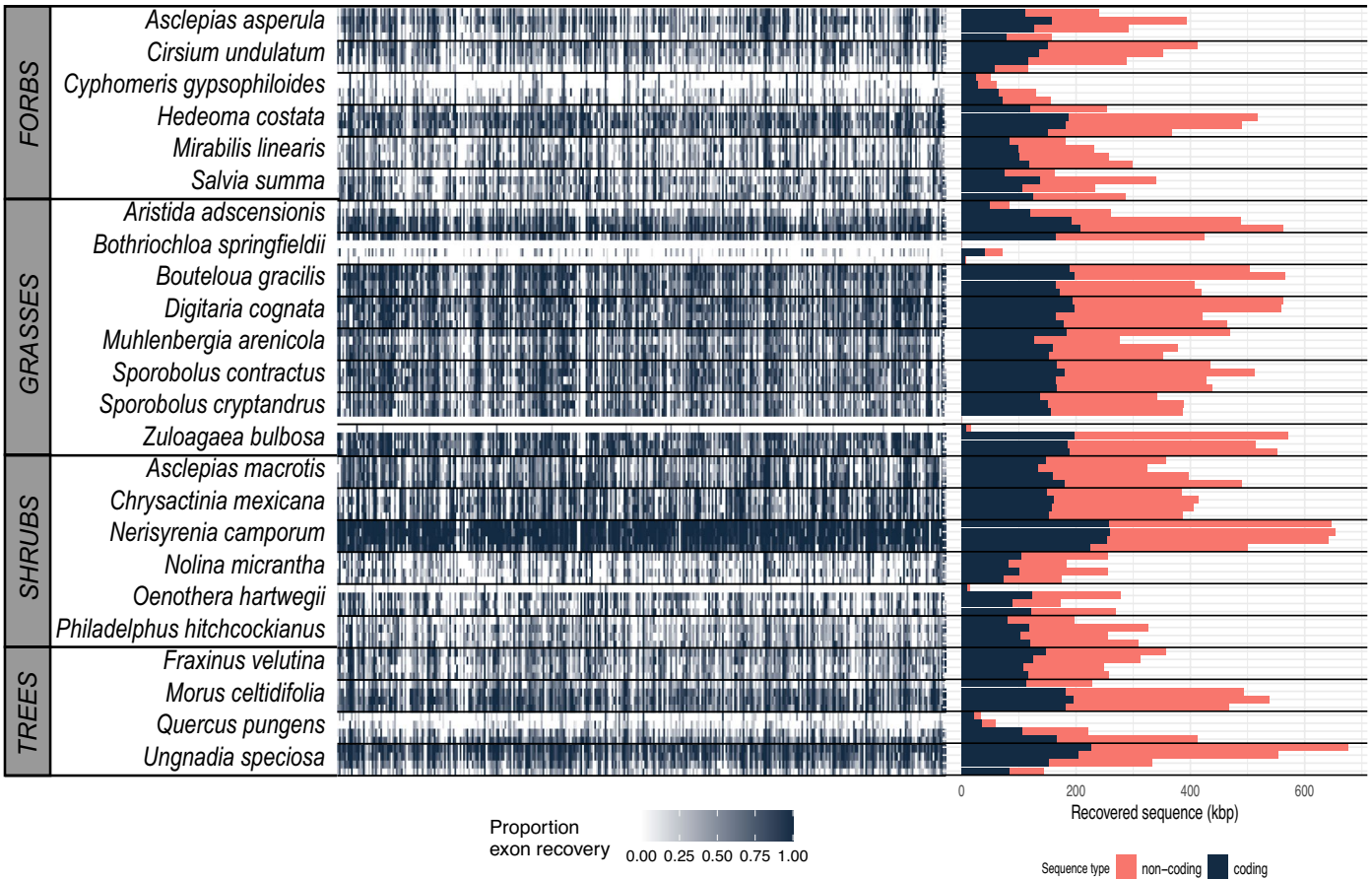


FIGURE 4. Left: Heatmap showing the recovery of each gene (column) for each sample (row), organized by species. The shading of each cell shows the percentage of the targeted gene length recovered. Right: Bar chart showing the total recovered sequence length for each specimen, categorized by targeted exon regions (dark blue) and flanking non-coding “splash zone” regions (pink).

Aristida adscensionis (0.0017 across 25,378 bp) and highest in *N. camporum* (0.028 across 385,422 bp). Comparing across loci, π was most variable in *Nolina micrantha* I. M. Johnst. and *Bothriochloa springfieldii* and lowest in *Asclepias macrotis* Torr., and 14 species had more than 100 genes where π was greater than 0.01 (Fig. 5). In *B. springfieldii*, sequence diversity was high despite poor recovery of loci from two of the four individuals sequenced. Overall, 83% of the sequence variation within species was located in the non-coding regions flanking the targeted exons (Table 1). The species with the highest percentage of SNPs identified within exonic sequences was *Quercus pungens* (34%), while *Mirabilis linearis* (Pursh) Heimerl had the lowest percentage of SNPs in exons (6%).

Tajima’s D is a measurement of sequence variability that can be an indicator of deviation from mutation-drift balance (Tajima, 1989). For most species, many Angiosperms353 loci show significantly negative Tajima’s D, which can be evidence of recent directional selection. With a sample of just four individuals, the critical values needed to show significant (95%) deviation from mutation-drift balance ($D = 0$) are -0.876 and 2.232 . In our data, a much larger number of loci had significantly negative Tajima’s D compared with significantly positive values (Appendix S2). At the species level, 16 species were found to have average Tajima’s D > 0 . Notably, four species (*Digitaria cognata*, *Sporobolus cryptandrus*, *Hedeoma costata* A. Gray, and *Salvia summa*) revealed an average Tajima’s D value > 1.0 , which is indicative of strong balancing

selection or recent population contraction. We found several species with Tajima’s D < 0 , including *Nerisyrenia camporum*.

Inbreeding coefficients (F) calculated for each individual were averaged across individuals in each species. *Morus celtidifolia* had the lowest value (-0.77), while *Aristida adscensionis* (0.765) and *Mirabilis linearis* (0.511) had the highest inbreeding coefficients of the GUMO collection. Seven species were found to have negative F values (Table 1).

DISCUSSION

Recovery of Angiosperms353 loci

Gene recovery in our sample did not display an obvious taxonomic bias; sample DNA quality and library concentration had the largest impacts on gene recovery (Fig. 3). However, target-capture efficiency was highest in *Nerisyrenia camporum*. This species, a member of the Brassicaceae, may have benefited from the probe design of Angiosperms353, which includes a representative from *Arabidopsis thaliana* for each gene (Johnson et al., 2019). Two of the species with the poorest gene recovery were grasses (*Aristida adscensionis* and *Bothriochloa springfieldii*); we believe this low recovery does not relate to poor recovery in Poaceae per se, as a large number of loci were recovered from the grasses *Bouteloua*

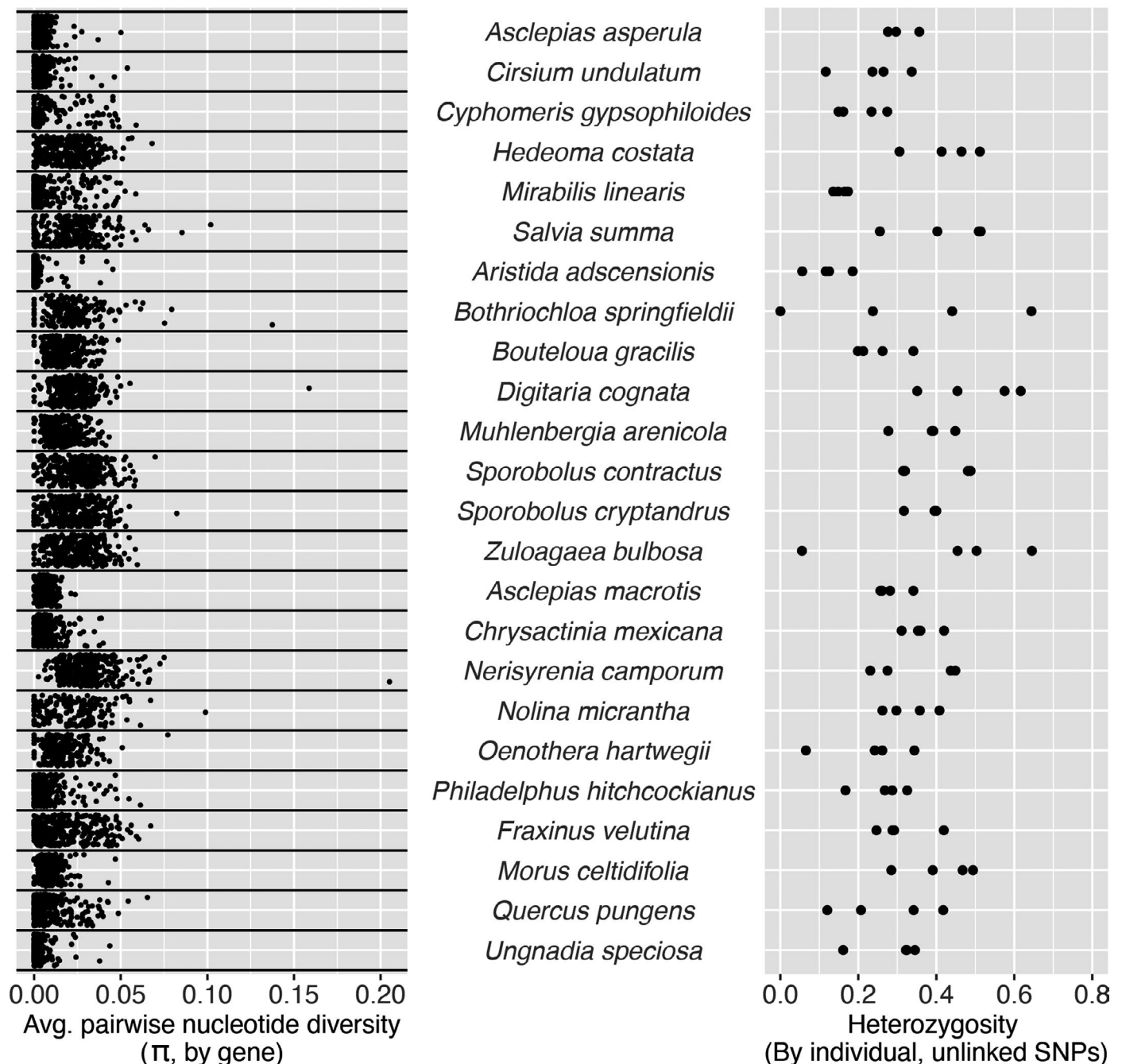


FIGURE 5. Left: Pairwise nucleotide diversity statistic (π) within each species for each gene recovered. Right: Percent heterozygosity for each individual at Angiosperms353 loci, calculated using a subset of SNPs that passed hard filtering, had no missing data, and lacked evidence of linkage disequilibrium (unlinked SNPs).

gracilis (Willd. ex Kunth) Lag. ex Griffiths and *Sporobolus contractus*. Rather, low-concentration libraries in *Aristida* and *Bothriochloa* and low-input DNA for several samples (Fig. 3) are more likely explanations for low recovery. Although low DNA and library concentrations will not lead to optimal target efficiency, it is still possible to obtain hundreds of gene sequences from these samples (Fig. 3). However, it is important for sample success that these samples have equal representation compared to higher-concentration samples, when pooling by equalizing molarity

between sample inputs (Hale et al., 2020). We pooled based on concentration alone, and while the concentrations of libraries within pools 2, 3, and 4 were fairly close in value, pool 1 contained a few samples that were clearly undersequenced compared to other samples within the pool, which resulted in poor enrichment efficiency and gene recovery for them (Fig. 3). For future studies, we suggest that low-concentration libraries (<5 ng/ μ L) should not be pooled with higher-concentration libraries for target capture. Greater effort should also be taken to maximize DNA input when

TABLE 1. SNP and population statistics for 24 species used in this study.

Species	Family	Filtered SNPs	Filtered no missing SNPs	Unlinked SNPs	Intron SNPs	Total reference length (bp)	Avg. π (all genes)	Tajima's D	Average F
<i>Aristida adscensionis</i>	Poaceae	1913	712	233	1527	253,878	0.002	0.392	0.765
<i>Asclepias asperula</i>	Asclepiadaceae	3213	1900	536	2524	154,837	0.004	−0.163	0.173
<i>Asclepias macrotis</i>	Asclepiadaceae	6814	4795	686	5687	219,807	0.006	−0.27	0.207
<i>Bothriochloa springfieldii</i>	Poaceae	12,455	197	59	8735	57,623	0.023	0.47	0.208
<i>Bouteloua gracilis</i>	Poaceae	23,420	14,027	1026	20,015	273,532	0.016	−0.52	0.275
<i>Chrysactinia mexicana</i>	Asteraceae	7111	4276	643	6401	203,568	0.008	0.06	0.003
<i>Cirsium undulatum</i>	Asteraceae	5064	2382	635	4292	143,531	0.005	−0.605	0.261
<i>Cyphomeris gypsophilioides</i>	Nyctaginaceae	3043	1509	248	2551	41,358	0.014	0.82	0.48
<i>Digitaria cognata</i>	Poaceae	24,035	14,907	1017	21,384	314,274	0.022	1.54	−0.164
<i>Fraxinus velutina</i>	Oleaceae	11,876	7151	706	9739	158,481	0.018	0.467	0.186
<i>Hedeoma costata</i>	Lamiaceae	17,661	10,165	805	14,619	249,662	0.019	1.15	−0.43
<i>Mirabilis linearis</i>	Nyctaginaceae	4712	2780	479	4386	107,288	0.010	0.54	0.511
<i>Morus celtidifolia</i>	Moraceae	9534	3680	678	8315	252,738	0.009	0.1	−0.77
<i>Muhlenbergia arenicola</i>	Poaceae	16,076	7490	780	13,581	224,046	0.016	0.11	−0.023
<i>Nerisyrenia camporum</i>	Brassicaceae	46,800	26,355	1700	41,703	385,422	0.028	−0.311	0.049
<i>Nolina micrantha</i>	Liliaceae	11,182	6424	615	10,109	67,657	0.026	0.56	0.048
<i>Oenothera hartwegii</i>	Onagraceae	10,523	2381	413	8508	120,909	0.016	−0.142	0.359
<i>Philadelphus hitchcockianus</i>	Hydrangeaceae	5280	3261	575	4700	107,138	0.008	0.049	0.274
<i>Quercus pungens</i>	Fagaceae	6877	1993	474	4474	160,055	0.010	−0.077	0.256
<i>Salvia summa</i>	Lamiaceae	12,064	6640	611	9855	122,800	0.022	1.21	−0.08
<i>Sporobolus contractus</i>	Poaceae	22,760	15,501	1029	20,185	248,960	0.024	0.956	−0.048
<i>Sporobolus cryptandrus</i>	Poaceae	15,525	11,446	650	13,245	188,922	0.023	1.028	0.125
<i>Ungnadia speciosa</i>	Sapindaceae	3323	1518	478	2608	324,434	0.002	0.149	0.237
<i>Zuloagaea bulbosa</i>	Poaceae	27,576	4678	451	24,269	288,821	0.023	0.407	−0.025

possible, and this could include employing alternative DNA extraction techniques for recalcitrant samples (Hale et al., 2020).

The poor relationship between library concentration and enrichment efficiency (for pools 2, 3, and 4) suggests that pooling via library concentration (ng/μL) is generally sufficient and using a fragment analyzer to precisely estimate molarity is not necessary. Efficient recovery of more than 200 loci from each sample in HybPiper could be achieved with as few as 25,000 reads on target (Fig. 3). Given that enrichment efficiency was typically around 20–30%, this suggests that 1000,000 reads would be necessary to ensure high gene recovery. With newer sequencing technologies including the Illumina NovaSeq SP, which has an output of 400,000,000 reads per lane, our results suggest that massive pooling of samples (at least 1500, given flexible library indices) could be achieved. We also saw a marked increase in gene recovery when using the “mega353” (McLay et al., 2021) target file in HybPiper. This target file was constructed using a larger curated set of orthologs for each Angiosperms353 probe set, compared with a target file constructed only from the representative sequences used to design the Angiosperms353 probes (Johnson et al., 2019). The “mega353” file doubled the enrichment efficiency (percentage of reads on target) on average (Appendices S1, S3) and led to dozens more recovered genes compared with the original target file.

Genetic variability within species

Estimates of heterozygosity within species should not be taken as fully representative of the GUMO populations; with only four specimens per species, the minimum minor allele frequency we could detect is 12.5%. However, the values should be roughly predictive of the amount of within-species genetic diversity that can be expected from both protein-coding and flanking non-coding regions of the Angiosperms353 genes. Even when using conservative filters—removing all missing data and retaining only SNPs with no evidence of linkage—hundreds of SNPs were variable within most species. We identified high heterozygosity for several species where low genetic variation might be expected, such as *Salvia summa* (a G3 vulnerable species) and other range-restricted species (e.g., *Philadelphus hitchcockianus*, *Nerisyrenia camporum*, and *Quercus pungens*).

Significantly negative values of Tajima's D can be found for some loci in most species. As this could be a result of directional selection, it provides a way to identify loci whose SNPs should be removed from analysis in future genetic studies. Species-level patterns are difficult to interpret with only a few individuals sampled here, but several species have an average Tajima's D that is positive, which is potentially an indicator of population contraction or balancing selection (Tajima, 1989). Notably, nearly all loci for *Digitaria cognata* showed positive Tajima's D. Because genome-wide balancing selection seems unlikely, this result suggests recent population contraction in this species. Although *D. cognata* is among the most widespread species in our study, in GUMO it is at the edge of its range, and it is not frequently observed at the altitudes found here (<1800 m) (Powell and Worthington, 2018).

Few prior estimates of genetic diversity are available for the species selected here. The species with the most prior population-level analyses is *Bouteloua gracilis*: Aguado-Santacruz et al. (2004) and Phan (2000) found high variability in *B. gracilis* populations at opposite ends of its range, but these studies used RAPD markers, and so heterozygosity estimates could not be made. More recently,

Avendaño-González et al. (2019) used RAD-seq to characterize genetic structure across this species' range and found a maximum heterozygosity of 0.32% and a Tajima's D of −1.75. These results are in stark contrast to the values estimated here (Fig. 4), but the RAD-seq study used far more individuals of *B. gracilis* (78, compared to our 4) and far more SNPs (132,000 compared to our 906). It would be useful to directly compare RAD-seq and target capture with equivalent sampling as a better comparison of the two methods. A main advantage of using Angiosperms353 loci is that studies will be more repeatable (not requiring rediscovery of loci each time new specimens are added), will be able to incorporate longer gene haplotypes (Leitwein et al., 2020), and may be more amenable to damaged DNA from herbarium specimens (but see Jordon-Thaden et al., 2020).

A study aimed at the genetic differentiation of species in *Aristida* (Thiv et al., 2019) noted little divergence of *A. adscensionis* from other species using traditional single-gene sequencing (ITS, *trnL-F*, and *rpl16*). Our results also show low genetic diversity in *A. adscensionis*, even when accounting for a low gene recovery rate. A study involving *Quercus pungens* sampling from GUMO found high heterozygosity in microsatellite markers (>0.8); this study aimed at identifying the origin of the federally threatened hybrid species *Quercus hinckleyi* C. H. Mull. (*Q. pungens* is a putative parent; Backs et al., 2016). We found slightly lower levels of heterozygosity for *Q. pungens* compared with the microsatellite markers, using the Angiosperms353 genes (Fig. 4). This may be attributable to several factors, including the difference in population parameters when comparing hypervariable markers (microsatellites) to DNA sequences (Hyb-Seq).

Potential limitations

Although our results suggest a promising potential for Angiosperms353 for population genomics, we feel it is important to identify probable limitations of the method. The Angiosperms353 loci were selected from a set of single-copy orthologs for all flowering plants (One Thousand Plant Transcriptomes Initiative et al., 2019) for the purpose of phylogenetic reconstruction. However, the genes also have an ontological bias—the genes are enriched for functions involving the chloroplast and photosynthesis (Johnson et al., 2019). As a result, their recovery may be limited in some taxa, especially those with reduced dependence on photosynthesis, including parasitic plants and mycoheterotrophs. In addition, single-copy genes may also be more likely to be under strong purifying selection (Duarte et al., 2010), which would limit the utility of the Angiosperms353 loci for genome-wide analysis (e.g., association studies or identifying selective sweeps and regions of introgression). Depending on genomic resources, RAD-seq or even whole genome sequencing would be more appropriate for genome-wide studies.

Polyploidy may have an effect on the nucleotide diversity metrics listed here. Some species, including *Nerisyrenia camporum* and *Digitaria cognata*, are likely recent polyploids and had the most genes marked with “paralog warnings” by HybPiper (Appendix S1). Although it is possible to assess ploidy in target capture data using allele ratios (Viruel et al., 2019), our data did not have enough sequencing depth to use these methods. Ideally, paralogy of individual loci should also be investigated (Gardner et al., 2020), but this would require analysis with more than four individuals per species.

Future applications

Our results illustrate that Angiosperms353 can be a cost-effective tool (Hale et al., 2020) for estimating population parameters. Although other methods may be more appropriate, depending on the goals of the study, we feel that Angiosperms353 has the highest population genomics potential in two areas: conservation genomics and environmental DNA studies. For example, Angiosperms353 could be used to conduct population-level analyses for all flowering plant species in an ecosystem, without the need to develop and optimize molecular markers for each species. This application could enrich the field of conservation genomics, which is often focused on demographic parameters within individual threatened and endangered species, due to limited funds. However, areas like GUMO, where many species are at the edge of their ranges, represent special cases for biodiversity preservation, and a new floristic survey of GUMO 50 years after the survey by Northington and Burgess (1976) could provide evidence for changes in genetic diversity and identify species of concern. Risk of local extinction for individual species may be elevated due to low genetic diversity and distance from the rest of the population (Vucetich and Waite, 2003). Local extinction, in turn, risks the ecosystem's health, creating cascading effects on biotic systems by removing species with critical ecosystem services (Pejchar and Mooney, 2009). Our results for *Digitaria cognata* provide a preliminary example, in which a widespread species has evidence of population contraction in an extreme environment. An additional advantage for target capture sequencing methods is the ability to incorporate damaged DNA from herbarium specimens, broadening the sampling opportunities in sequencing projects.

Several studies have noted that comparison among related species can enhance the understanding of conservation implications drawn from population demographics (Moran et al., 1989; Barrett and Kohn, 1991; Johnson et al., 2016b). Species within the same genus could be compared with respect to differences in species ranges (e.g., in our data *Sporobolus cryptandrus* vs. *S. contractus*) or life history traits (e.g., in our data *Asclepias asperula* vs. *A. macrotis*). In the present study, differences among related species were limited; future investigations with deeper sampling within and among species could expand comparative analysis, to identify whether genetic diversity metrics contain phylogenetic signal. By using Angiosperms353, conservation genomics projects could ensure that sufficient data are collected for rare species, while allowing for direct comparisons to the often-overlooked genetic diversity of widespread species.

Widespread application of the Angiosperms353 loci at and below the species level would enable its use as a tool for detecting plant DNA in environmental, ancient, and sedimentary samples. These metabarcoding studies frequently rely on PCR-based methods targeting very short regions of the chloroplast genome and suffer from primer universality and poor sequence differentiation among species. The results here suggest Angiosperms353 loci would provide ample sequence variation for identification of taxa in these studies, but this would likely require the development of new bioinformatics tools to handle sequence assembly and database comparison.

ACKNOWLEDGMENTS

The authors thank Sean Graham and two anonymous reviewers for comments on an earlier draft of the manuscript, and L. Pokorný (Center for Biotechnology and Genomics of Plants, Madrid, Spain)

and E. Gardner (National Tropical Botanical Garden, Singapore) for extensive advice in laboratory protocols. We thank the Guadalupe Mountains National Park for permission to destructively sample herbarium specimens, and M. Fokar (Texas Tech University Center for Bioinformatics and Genomics) for assistance with sequencing. M.S. is a Texas Tech University Honors College Undergraduate Research Scholar supported by the CH Foundation. This work was funded by contributions to M.G.J. from the Texas Tech University College of Arts and Sciences and by the National Science Foundation (NSF-DEB 1753800).

AUTHOR CONTRIBUTIONS

M.S., H.H., and M.G.J. designed the study. M.S. and H.H. sampled herbarium specimens and extracted DNA. H.H. prepared enriched sequencing libraries. All authors conducted data analysis; M.S. recovered gene sequences using HybPiper, L.D.W. built the variant calling and population genetics analysis workflow with help from M.G.J., and H.H. analyzed relationships between wet lab and bioinformatics results. All authors participated in writing the manuscript, which was approved by all authors prior to submission.

DATA AVAILABILITY

All sequencing reads are accessible from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA667845. HybPiper output, vcf files, and reference sequences recovered for all 24 species are available on Dryad (<https://doi.org/10.5061/dryad.76hdr7sv3>; Johnson et al., 2021). All herbarium specimen vouchers are described in Appendix 1. Scripts used for data analysis are available at <https://github.com/lindsawi/HybSeq-SNP-Extraction>.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Statistics on DNA concentrations, Illumina libraries, and HybPiper gene recovery statistics using the Mega353 target file.

APPENDIX S2. Tajima's D statistic within each species for each gene recovered, using SNPs that passed hard filtering and had no missing data. Vertical lines indicate 95% confidence intervals for Tajima's D calculated for four individuals.

APPENDIX S3. Statistics on DNA concentrations, Illumina libraries, and HybPiper gene recovery statistics using the default target file.

LITERATURE CITED

Aguado-Santacruz, G. A., N. E. Leyva-López, K. I. Pérez-Márquez, E. García-Moya, J. T. Arredondo-Moreno, and J. P. Martínez-Soriano. 2004. Genetic

- variability of *Bouteloua gracilis* populations differing in forage production at the southernmost part of the North American Graminetum. *Plant Ecology* 170: 287–299.
- Allen, K. K. 2013. Guadalupe Mountains National Park: Natural resource condition assessment. NPS/GUMO/NRR-2013/668. National Park Service, Fort Collins, Colorado, USA.
- Allendorf, F. W., P. A. Hohenlohe, and G. Luikart. 2010. Genomics and the future of conservation genetics. *Nature Reviews Genetics* 11: 697–709.
- Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17: 81–92.
- Antonelli, A., J. J. Clarkson, K. Kainulainen, O. Maurin, G. E. Brewer, A. P. Davis, N. Epitawalage, et al. 2021. Settling a family feud: A high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes. *American Journal of Botany* 108: in press.
- Avendaño-González, M., J. F. Morales-Domínguez, and M. E. Siqueiros-Delgado. 2019. Genetic structure, phylogeography, and migration routes of *Bouteloua gracilis* (Kunth) Lag. ex Griffiths (Poaceae: Chloridoideae). *Molecular Phylogenetics and Evolution* 134: 50–60.
- Backs, J. R., M. Terry, and M. V. Ashley. 2016. Using genetic analysis to evaluate hybridization as a conservation concern for the threatened species *Quercus hinckleyi* C.H. Muller (Fagaceae). *International Journal of Plant Sciences* 177: 122–131.
- Barrett, S. H., and J. A. Kohn. 1991. Genetic and evolutionary consequences of small population size in plants: Implications for conservation. In D. Falk and K. Holsinger [eds.], *Genetics and conservation of rare plants*. Oxford University Press, New York, New York, USA.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.
- Brewer, G. E., J. J. Clarkson, O. Maurin, A. R. Zuntini, V. Barber, S. Bellot, N. Biggs, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10: 1102.
- Buddenhagen, C., A. R. Lemmon, E. M. Lemmon, J. Bruhl, J. Cappa, W. L. Clement, M. Donoghue, et al. 2016. Anchored phylogenomics of Angiosperms I: Assessing the robustness of phylogenetic estimates. *bioRxiv* 086298 [Preprint] [published 28 November 2016]. Available at <https://doi.org/10.1101/086298> [accessed 17 March 2021].
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.
- Carter, K. A., A. Liston, N. V. Bassil, L. A. Alice, J. M. Bushakra, B. L. Sutherland, T. C. Mockler, et al. 2019. Target capture sequencing unravels *Rubus* evolution. *Frontiers in Plant Science* 10: 1615.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
- DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43: 491–498.
- Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebens-Mack, and C. W. dePamphilis. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.
- Duminil, J., and M. Di Michele. 2009. Plant species delimitation: A comparison of morphological and molecular markers. *Plant Biosystems* 143: 528–542.
- Eason, M. 2018. *Wildflowers of Texas*. Timber Press, Portland, Oregon, USA.
- Eaton, D. A. R., and I. Overcast. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36: 2592–2594.
- Fazekas, A. J., P. R. Kesanakurti, K. S. Burgess, D. M. Percy, S. W. Graham, S. C. H. Barrett, S. G. Newmaster, et al. 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* 9(s1): 130–139.
- Gardner, E. M., M. G. Johnson, J. T. Pereira, A. S. A. Puad, D. Arifiani, N. J. Wickett, and N. J. C. Zerega. 2020. Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Systematic Biology* <https://doi.org/10.1093/sysbio/syaa073>.
- Glass, M. R., R. E. Reisch, and G. M. Ahlstrand. 1974. Range condition survey and wildlife browse analysis. Carlsbad Caverns and Guadalupe Mountains National Parks. Type-written report: 45.
- Hale, H., E. M. Gardner, J. Viruel, L. Pokorny, and M. G. Johnson. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences* 8: e11337. <https://doi.org/10.1002/aps3.11337>.
- Heberling, J. M., and D. J. Burke. 2019. Utilizing herbarium specimens to quantify historical mycorrhizal communities. *Applications in Plant Sciences* 7: e01223. <https://doi.org/10.1002/aps3.1223>.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. Wickett. 2016a. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4(7): 1600016. <https://doi.org/10.3732/apps.1600016>.
- Johnson, M. G., K. Lang, P. Manos, G. H. Golet, and K. A. Schierenbeck. 2016b. Evidence for genetic erosion of a California native tree, *Platanus racemosa*, via recent, ongoing introgressive hybridization with an introduced ornamental species. *Conservation Genetics* 17: 593–602.
- Johnson, M. G., C. Malley, B. Goffinet, A. J. Shaw, and N. J. Wickett. 2016c. A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). *Molecular Phylogenetics and Evolution* 98: 29–40.
- Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68(4): 594–606.
- Johnson, M. G., M. Slimp, L. D. Williams, and H. Hale. 2021. Data from: On the potential of Angiosperms353 for population genomics studies. Dryad Dataset. <https://doi.org/10.5061/dryad.76hdr7sv3>
- Jordon-Thaden, I. E., J. B. Beck, C. A. Rushworth, M. D. Windham, N. Diaz, J. T. Cantley, C. T. Martine, and C. J. Rothfels. 2020. A basic ddRADseq two-enzyme protocol performs well with herbarium and silica-dried tissues across four genera. *Applications in Plant Sciences* 8: e11344. <https://doi.org/10.1002/aps3.11344>.
- Kelly, L. J., S. Renny-Byfield, J. Pellicer, J. Macas, P. Novák, P. Neumann, M. A. Lysak, et al. 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *The New Phytologist* 208: 596–607.
- Leitwein, M., M. Duranton, Q. Rougemont, P.-A. Gagnaire, and L. Bernatchez. 2020. Using haplotype information for conservation genomics. *Trends in Ecology & Evolution* 35: 245–258.
- McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6(3): e01038. <https://doi.org/10.1002/aps3.1038>.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernysky, K. Garimella, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.
- McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9: e11420.
- McMahon, B. J., E. C. Teeling, and J. Höglund. 2014. How and why should we implement genomics into conservation? *Evolutionary Applications* 7: 999–1007.
- Miller-Rushing, A. J., R. B. Primack, P. H. Templer, S. Rathbone, and S. Mukunda. 2009. Long-term relationships among atmospheric CO₂, stomata, and intrinsic water use efficiency in individual trees. *American Journal of Botany* 96: 1779–1786.
- Moran, G. F., O. Muona, and J. C. Bell. 1989. Breeding systems and genetic diversity in *Acacia auriculiformis* and *A. crassiparva*. *Biotropica* 21: 250–256.

- New Mexico Rare Plant Technical Council. 1999. New Mexico rare plants. Website <https://nmrareplants.unm.edu> [latest update 13 April 2021; accessed 16 April 2021].
- Northington, D. K., and T. L. Burgess. 1976. Summary of the vegetative zones of the Guadalupe Mountains National Park, Texas. In H. Genoways and R. Baker [eds.], Transactions—Symposium on the biological investigations in the Guadalupe Mountains. National Park Service, Washington, D.C., USA.
- O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy. 2018. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* 27: 3193–3206.
- One Thousand Plant Transcriptomes Initiative, J. H. Leebens-Mack, M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Overcast, I., B. C. Emerson, and M. J. Hickerson. 2019. An integrated model of population genetics and community ecology. *Journal of Biogeography* 46: 816–829.
- Pejchar, L., and H. A. Mooney. 2009. Invasive species, ecosystem services and human well-being. *Trends in Ecology & Evolution* 24: 497–504.
- Phan, A. T. 2000. Genetic diversity of blue grama, *Bouteloua gracilis*, and little bluestem, *Schizachyrium scoparium*, as affected by selection. M.S. thesis, University of Manitoba, Winnipeg, Manitoba, Canada. Available at: <http://hdl.handle.net/1993/1745> [accessed 8 March 2021].
- Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 [Preprint] [published 24 July 2018]. Available at <https://doi.org/10.1101/201178> [accessed 13 April 2021].
- Powell, M. A., and R. D. Worthington. 2018. Flowering plants of Trans-Pecos Texas and adjacent areas, 1st ed. BRIT Press, Fort Worth, Texas, USA.
- 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. *American Journal of Human Genetics* 81: 559–575.
- Rohland, N., and D. Reich. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* 22: 939–946.
- Shah, T., J. V. Schneider, G. Zizka, O. Maurin, W. Baker, F. Forest, G. E. Brewer, et al. 2021. Joining forces in Ochnaceae phylogenomics: A tale of two targeted sequencing probe kits. *American Journal of Botany* 108: in press.
- Shee, Z. Q., D. G. Frodin, R. Cámara-Leret, and L. Pokorny. 2020. Reconstructing the complex evolutionary history of the Papuanian *Schefflera* radiation through herbariomics. *Frontiers in Plant Science* 11: 258.
- Supple, M. A., and B. Shapiro. 2018. Conservation of biodiversity in the genomics era. *Genome Biology* 19: 131.
- Thiers, B. 2021 (continuously updated). Index Herbariorum. Website <http://sweetgum.nybg.org/science/ih/> [accessed 8 March 2021].
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Thiv, M., J. Alfredo Reyes-Betancort, and O. Fragman-Sapir. 2019. Lifeforms as criterion for species delimitation: Are *Aristida adscensionis* and *A. coerulescens* (Aristidoideae, Poaceae) two species? *Phytotaxa* 393: 67–74.
- Viruel, J., M. Conejero, O. Hidalgo, L. Pokorny, R. F. Powell, F. Forest, M. B. Kantar, et al. 2019. A target capture-based method to estimate ploidy from herbarium specimens. *Frontiers in Plant Science* 10: 937.
- Vucetich, J. A., and T. A. Waite. 2003. Spatial patterns of demography and genetic processes across the species' range: Null hypotheses for landscape conservation genetics. *Conservation Genetics* 4: 639–645.
- Weitemier, K., S. Straub, R. C. Cronn, and M. Fishbein. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042. <https://doi.org/10.3732/app.1400042>.
- Yang, Y., M. J. Moore, S. F. Brockington, J. Mikenas, J. Olivieri, J. F. Walker, and S. A. Smith. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *The New Phytologist* 217: 855–870.
- Zuntini, A. R., L. P. Frankel, L. Pokorny, F. Forest, and W. J. Baker. 2021. A comprehensive phylogenomic study of the monocot order Commelinales, with a new classification of Commelinaceae. *American Journal of Botany* 108: in press.

APPENDIX 1. Voucher information for 95 herbarium specimens sampled for this study. All specimens were collected within Hudspeth and Culberson counties in Texas, including Guadalupe Mountains National Park. All specimens are housed at the E. L. Reed Herbarium (TTC).

Taxon; DNA voucher information (TTC ID, collector and number)

Aristida adscensionis L.; TTC020876 T.L. Burgess 730; TTC021123 T.L. Burgess 874; TTC020822 T.L. Burgess 531; TTC020821 T.L. Burgess 638

Asclepias asperula (Decne.) Woodson; TTC021110 T.L. Burgess 2036; TTC020017 T.L. Burgess 3276; TTC020020 T.L. Burgess 1952; TTC020019 T.L. Burgess 802

Asclepias macrotis Torr.; TTC020009 L.T. Green 71; TTC020011 T.L. Burgess 1296; TTC020010 T.L. Burgess 2477; TTC020012 T.L. Burgess 2069

Bothriochloa springfieldii (Gould) Parodi; TTC020471 T.L. Burgess 3594; TTC020614 T.L. Burgess 2069 3991; TTC020890 T.L. Burgess 827; TTC020613 T.L. Burgess 3995

Bouteloua gracilis (Willd. ex Kunth) Lag. ex Griffiths; TTC021101 T.L. Burgess 632; TTC021102 T.L. Burgess 4667; TTC021100 T.L. Burgess 547; TTC021099 T.L. Burgess 838

Chrysactinia mexicana A. Gray; TTC020106 T.L. Burgess 1494; TTC020111 T.L. Burgess 342; TTC020108 T.L. Burgess 994; TTC020109 T.L. Burgess 444

Cirsium undulatum (Nutt.) Spreng.; TTC021116 T.L. Burgess 2017; TTC021114 T.L. Burgess 2082-1; TTC021115 T.L. Burgess 2031; TTC021113 T.L. Burgess 4545

Cyphomeris gypsophiloides (M. Martens & Galeotti) Standl.; TTC021112 T.L. Burgess 679; TTC020664 T.L. Burgess 3984; TTC020669 T.L. Burgess 1479; TTC020666 T.L. Burgess 2349

Digitaria cognata (Schult.) Pilg.; TTC021096 T.L. Burgess 828; TTC021097 D.K. Northington 573; TTC020956 T.L. Burgess 655; TTC021098 T.L. Burgess 977

Fraxinus velutina Torr.; TTC020639 D.K. Northington 487; TTC020636 T.L. Burgess 2028; TTC020637 T.L. Burgess 2027; TTC020717 T.L. Burgess 806

Hedeoma costata A. Gray; TTC020434 T.L. Burgess 3232; TTC020436 T.L. Burgess 2335; TTC020439 D.K. Northington 436; TTC020438 D.K. Northington 507

Mirabilis linearis (Pursh) Heimerl; TTC021109 T.L. Burgess 2548; TTC021108 T.L. Burgess 3893; TTC020774 T.L. Burgess 678; TTC020771 T.L. Burgess 4126

Morus celtidifolia Kunth; TTC020656 T.L. Burgess 3236; TTC020657 T.L. Burgess 2465; TCC021107 T.L. Burgess 1177; TTC020658 T.L. Burgess 2038

Muhlenbergia arenicola Buckley; TTC021117 T.L. Burgess 786; TTC020921 T.L. Burgess 830; TTC020918 T.L. Burgess 582; TTC020925 T.L. Burgess 721

Nerisyrenia camporum (A. Gray) Greene; TTC020355 T.L. Burgess 2140; TTC020357 T.L. Burgess 1410; TTC020770 D.K. Northington 744; TTC020359 D.K. Northington 322

Nolina micrantha I. M. Johnston; TTC020683 T.L. Burgess 927; TTC020684 D.K. Northington 367; TTC020804 T.L. Burgess 4641; TTC020681 T.L. Burgess 2381

Oenothera hartwegii Benth.; TTC020643 T.L. Burgess 2488; TTC021122 T.L. Burgess 4486; TTC020735 T.L. Burgess 4405; TTC020736 T.L. Burgess 4457

Philadelphus hitchcockianus S. Y. Hu; TTC021118 T.L. Burgess 1142; TTC021119 T.L. Burgess 4243; TTC021120 T.L. Burgess 1955; TTC021121 D.K. Northington 1031

Quercus pungens Liebm.; TTC020392 T.L. Burgess 975; TTC021093 T.L. Burgess 868; TTC020391 T.L. Burgess 1651; TTC021094 T.L. Burgess 4654

Salvia summa A. Nelson; TTC020496 T.L. Burgess 3280; TTC020502 D.K. Northington 476; TTC020503 D.K. Northington 475; TTC020497 D.K. Northington 1032

Sporobolus contractus Hitchc.; TTC020906 T.L. Burgess 656; TTC020908 T.L. Burgess 892; TTC021105 T.L. Burgess 537; TTC021106 T.L. Burgess 636

Sporobolus cryptandrus (Torr.) A. Gray; TTC021104 T.L. Burgess 536; TTC020900 T.L. Burgess 722; TTC021103 T.L. Burgess 832

Ungnadia speciosa Endl.; TTC020570 D.K. Northington 465; TTC020569 T.L. Burgess 1250; TTC021095 T.L. Burgess 1827; TTC020567 T.L. Burgess 4099

Zuloagaea bulbosa (Kunth) E. Bess; TTC020962 T.L. Burgess 864; TTC021111 T.L. Burgess 704; TTC020815 T.L. Burgess 755; TTC020461 T.L. Burgess 3595