Received: 2 August 2021 | Revised: 8 December 2021 | Accepted: 10 December 2021

DOI: 10.1111/1755-0998.13577

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

A multispecies amplicon sequencing approach for genetic diversity assessments in grassland plant species

Miguel Loera-Sánchez | Bruno Studer | Roland Kölliker

Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland

Correspondence
Roland Kölliker, Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland.
Email: roland.koelliker@usys.ethz.ch

Funding information
This work was funded by the Swiss Federal Office of Agriculture (FOAG) grant FH627000554

Abstract

Grasslands are widespread and economically relevant ecosystems at the basis of sustainable roughage production. Plant genetic diversity (PGD; i.e., within-species diversity) is related to many beneficial effects on the ecosystem functioning of grasslands. The monitoring of PGD in temperate grasslands is complicated by the multiplicity of species present and by a shortage of methods for large-scale assessments. However, the continuous advancement of high-throughput DNA sequencing approaches has improved the prospects of broad, multispecies PGD monitoring. Among them, amplicon sequencing stands out as a robust and cost-effective method. Here, we report a set of 12 multispecies primer pairs that can be used for high-throughput PGD assessments in multiple grassland plant species. The target loci were selected and tested in two phases: a "discovery phase" based on a sequence capture assay (611 nuclear loci assessed in 16 grassland plant species), which resulted in the selection of 11 loci; and a "validation phase", in which the selected loci were targeted and sequenced using multispecies primers in test populations of *Dactylis glomerata* L., *Lolium perenne* L., *Festuca pratensis* Huds., *Trifolium pratense* L. and *T. repens* L. The multispecies amplicons had nucleotide diversities per species from 5.19×10^{-3} to 1.29×10^{-2}, which is in the range of flowering-related genes but slightly lower than pathogen resistance genes. We conclude that the methodology, the DNA sequence resources, and the primer pairs reported in this study provide the basis for large-scale, multispecies PGD monitoring in grassland plants.

KEYWORDS
amplicon sequencing, genetic diversity, grasses, grassland, legumes, sequence capture

1 | INTRODUCTION

Grasslands cover an estimated 40% of Earth's land, fulfilling many ecosystem services including the preservation of soil integrity and regulation of water, carbon and nitrogen flows (Bengtsson et al., 2019; Reynolds, 2005; Zhao et al., 2020). Multispecies grasslands are important sources of roughage for ruminant livestock and provide the basis for sustainable meat and dairy production (Zhao et al., 2020). They are also of unique cultural relevance and serve as places of recreation (Huber & Finger, 2020; Zhao et al., 2020). Plant biodiversity, including genetic diversity (i.e., within-species diversity), plays an important role in the ecosystem functioning of grasslands.

Recent work has shown that high levels of plant genetic diversity in grasslands confer resistance against invasive plants (Hadincová et al., 2020) and increase yield stability under environmental stress conditions, in part due to interactions between plant species richness and genetic diversity (Malyshev et al., 2016; Meilhac et al., 2019; Prieto et al., 2015). Furthermore, valuable genetic resources for forage breeding are to be found along the wide geographical range for cultivated and wild grasses (Poaceae) and legumes (Fabaceae), the two most economically relevant families of forage crop species found in grasslands. Nevertheless, genetic diversity is still underexplored for most taxonomic groups in all domains of life, including many grass and legume species from temperate grasslands. This is mainly because common genetic diversity assessment methods are time- and resource-demanding, particularly for constant and broad monitoring. Enabling large-scale, multispecies plant genetic diversity assessments will benefit the study of its effect on grassland ecosystem services, including the provisioning of roughage and genetic diversity resources. Large-scale, multispecies genetic diversity assessments will also be key to reach worldwide biodiversity protection targets, such as those of the Convention of Biological Diversity of the United Nations (Hoban et al., 2020; Laikre et al., 2020; Pärli et al., 2021).

Advances in high-throughput sequencing technologies may soon enable large-scale, multispecies genetic diversity monitoring in grassland plants. Such technologies have facilitated genetic diversity assessments in non-model organisms, including the forage crop species from temperate regions (Loera-Sánchez et al., 2019). Compared to traditional methods for genetic diversity assessments (e.g., using simple sequence repeats or SSR), hundreds of samples and tens of thousands of markers may be processed simultaneously with high-throughput sequencing approaches. High-throughput approaches suitable for genetic diversity assessments in non-model organism include complexity reduction methods (e.g., genotyping-by-sequencing [GBS], restriction site-associated DNA sequencing [RADseq]) and target enrichment methods (e.g., sequence capture and amplicon sequencing). Those methods differ in the amount of DNA they require, the marker density they produce (i.e., the number of single nucleotide polymorphisms, or SNPs) and the costs they imply (Carroll et al., 2018; Harvey et al., 2016).

The choice of a method to assess genetic diversity in grassland plant species at a large-scale would require making a balance of its technical features along with the biological features of grassland plants. Most grass and legume species of temperate grasslands are outcrossing species exhibiting a highly effective self-incompatibility system (Annicchiarico et al., 2015; Cropano et al., 2021). As a result, this kind of species display high levels of intrapopulation genetic diversity and low levels of interpopulation differentiation (Hamrick & Godt, 1996). Outcrossing species also show lower levels of linkage disequilibrium compared to autogamous plants. This is the case for many important forage crops, including ryegrasses (*Lolium* spp.), fescues (*Festuca* spp.), orchardgrass (*Dactylis glomerata* L.) and clovers (*Trifolium* spp.; Collins et al., 2012; Cuyeu et al., 2013; Last et al., 2013; Liu et al., 2018). Because of the high levels of intrapopulation genetic diversity, large sample numbers per population are usually necessary to estimate genetic diversity in these species (Kölliker et al., 2009). In contrast, a high marker density is not necessary to detect genetic differentiation in populations of such taxa. This has been observed in *Lolium perenne* L. (Liu et al., 2018), as well as in other highly diverse species, such as open-pollinated maize landraces (*Zea mays* L.; Caldu-Primo et al., 2017) and *Arabidopsis halleri*, (L.) O'Kane & Al-Shehbaz, an outbreeding model species (Fischer et al., 2017).

Common sequence-based genetic diversity metrics, like nucleotide diversity (π), rely on SNP calling. SNPs are mostly biallelic DNA markers that are widely spread across the genome. However, although high-quality SNPs can produce accurate estimations of genetic diversity, they do not consider all the information that is present in high-throughput sequencing reads (Voichek & Weigel, 2020). SNPs do not take into account insertions or deletions (i.e., indels), which can be informative for genetic diversity estimations. Furthermore, SNP calling depends on aligning reads to a reference genome. Any genomic variant that is not present in the reference genome assembly will not produce SNP calls. Alignment-independent metrics, like *k*-mer counting, can compensate for these issues. *K*-mer analysis is commonly used to compare large sequences (e.g., entire genomes) in an efficient manner, all the while accounting for indels and genomic rearrangements (Zielezinski et al., 2017). As a tool to detect intraspecific variation, *k*-mer richness (or *k*-mer count) analysis has been applied in barley (*Hordeum vulgare* L.) to detect apparent heterozygous mappings, that is, clusters of divergent sequencing reads that map to genomic loci that are closely related (Pérez-Cantalapiedra et al., 2018). *K*-mer richness analysis has also been applied to infer haplotype diversity in viral populations (Malhotra et al., 2013). Having both kinds of diversity metrics (i.e., SNP- and *k*-mer-based) can therefore provide a better estimate of the genetic diversity from sequencing data, combining stringent SNP calling with the indel-sensitive *k*-mer analysis.

In this study, we present a novel genetic diversity assessment approach that is based on amplicon sequencing and that can be applied in multiple grass and legume species. Our aim is to lay the groundwork for cost-effective, multispecies genetic diversity assessments of grassland plant species. Taking advantage of the naturally high levels of genetic diversity and low levels of linkage disequilibrium of such outcrossing species, we hypothesized that a reduced set of nuclear loci would contain enough sequence-level polymorphisms for genetic diversity analyses. We followed a sequence capture approach to produce a shortlist of loci (the "discovery" phase of this study), designed multispecies primers to target them, and confirmed that the resulting amplicons were genetically diverse in test populations (the "validation" phase). To reduce analysis costs, we used pooled-plant samples in the validation phase. Furthermore, to fully characterize the sequence diversity of the target loci, we used SNP-based metrics along with

normalized *k*-mer richness (NKR), which we define as the ratio between the observed count of unique *k*-mers from sequencing reads and the expected maximum number of unique *k*-mers for each locus.

2 | MATERIALS AND METHODS

The development of multispecies amplicons for genetic diversity assessments in grassland plant species followed a two-phase approach. In the "discovery phase", the genetic diversity of 611 conserved loci was assessed in 16 forage species with targeted sequencing. Then, in the "validation phase", selected multispecies amplicons were sequenced in test populations of five species (Figure 1).

2.1 | Discovery phase

2.1.1 | Bait design and synthesis

A set of 734 putatively single-copy, orthologous genes (SCOGs) was identified with OrthoMCL (Li et al., 2003) in a selection of reference genomes including model and agriculturally relevant flowering plant species. Such reference genomes were selected to increase the chances of finding loci that are conserved in Poaceae and Fabaceae species using, as far as possible, high-quality genome assemblies. The reference genomes included two species of the Poaceae family: *Brachypodium distachyon* (L.) P. Beauv. (GCA_000005505.4; Vogel et al., 2010) and *L. perenne* (GCA_001735685.1); two Fabaceae species: *Glycine max* (L.) Merr., 1917 (GCA_000004515.4; Schmutz et al., 2010) and

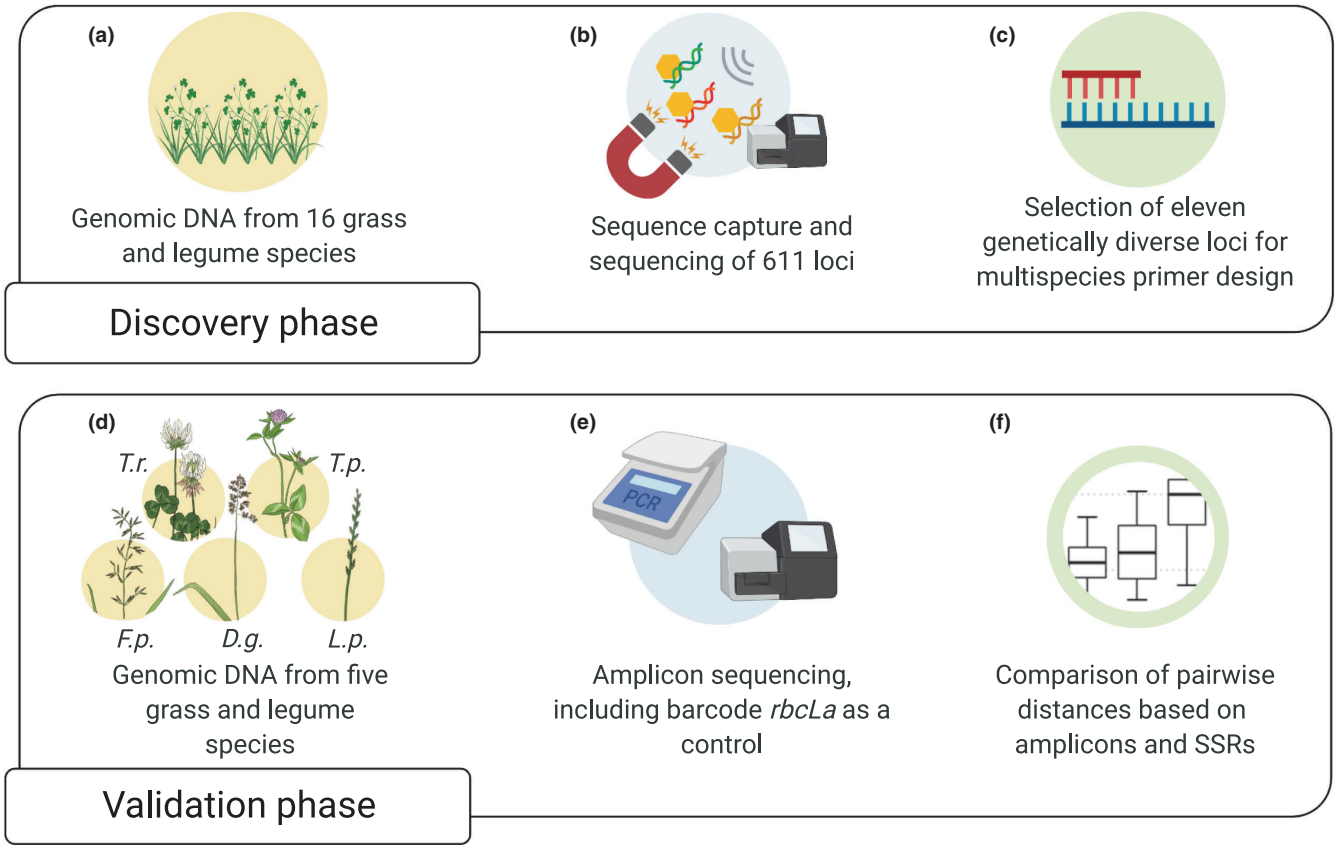

FIGURE 1 Graphical summary of the steps leading to the identification of multispecies amplicons that were used for genetic diversity assessments. (a) Genomic DNA was extracted from 16 grass and legume species (five plants from different cultivars per species). Each species was represented by five single-plant samples and one pooled-plant sample. The pooled-plant sample of each species contained the same five plants used for single-plant samples. (b) Illumina dual-indexed libraries were prepared (~550 bp fragment size) and sequenced. Libraries were then used as targets for sequence capture. The target loci were 611 single-copy, orthologous genes and ultra-conserved like elements. (c) Eleven loci were selected for multispecies primer design, aiming for amplicons of ~500 bp. (d) Genomic DNA was extracted from 16 plants from five grass and legume species (16 plants from three cultivars per species). Each species was represented by 16 single-plant DNA samples and three pooled-plant DNA samples. In each species, pooled-plant samples included the same plants used for the single-plant samples. (e) Multispecies amplicons were amplified in all samples, according to the plant family specificity of the primer pairs, including the DNA barcode *rbcLa* as a reference for a low-diversity locus. Amplicons were pooled equimolarly and then used to prepare dual-indexed libraries. Dual-indexed libraries were sequenced on an Illumina MiSeq. (f) The 16 single-plant DNA samples were genotyped using eight SSR. The resulting SSR multilocus genotypes were compared to amplicon sequencing-based multilocus genotypes, in terms of their multilocus genotype count and pairwise distances. Figure created with BioRender.com. Art in (d) created by Danira León https://www.behance.net/leondanira1c28

Trifolium pratense L. (GCA_900079335.1; De Vega et al., 2015); one Solanaceae: *Solanum lycopersicum* L. (GCA_000188115.3; The Tomato Genome Consortium, 2012); one Malvaceae: *Theobroma cacao* L. (GCA_000403535.1; Motamayor et al., 2013); one Vitaceae: *Vitis vinifera* L. (GCA_000003745.2; Jaillon et al., 2007); and one Brassicaceae: *Arabidopsis thaliana* (L.), Heynh. (GCA_000001735.1; Lamesch et al., 2012). All reference genomes, except for *L. perenne*, were downloaded from EnsemblPlants (https://plants.ensembl.org/index.html). The reference genome of *L. perenne* (GCA_001735685.1) was obtained from NCBI (https://www.ncbi.nlm.nih.gov). Bait design was performed using BaitFisher v1.2.7 (Mayer et al., 2016), with noncontinuous multiple sequence alignments of the 734 SCOGs as input. The annotation of *A. thaliana* (GCA_000001735.1) was used to account for intron-exon boundaries.

Additionally, 1277 ultra conserved-like elements (ULEs) were identified with Phyluce v1.4 (Faircloth, 2016). For this purpose, the same set of genomes mentioned above was complemented with three Poaceae species: *Aegilops tauschii* Coss. (GCA_002575655.1; Luo et al., 2017), *Leersia perrieri* (A. Camus) Launer (GCA_000325765.3), and *Oryza sativa* L. subsp. *japonica* (GCA_001433935.1; Kawahara et al., 2013); and three Fabaceae species: *Lotus japonicus* (Regel) K. Larsen, 1955 (GCA_000181115.2; Sato et al., 2008), *Medicago truncatula* Gaertn. (GCA_000219495.2; Tang et al., 2014), and *Phaseolus vulgaris* L., 1753 (GCA_000499845.1; Schmutz et al., 2014). The Phyluce UCE-identification pipeline was performed thrice using one of the following guiding genomes on each iteration: *A. thaliana* (GCA_000001735.1), *B. distachyon* (GCA_000005505.4) and *M. truncatula* (GCA_000219495.2). Phyluce v1.4 was then used to find more candidate bait sequences targeting the found ULEs, in addition to the candidate bait sequences found with BaitFisher.

To control which genomic regions were targeted, the bait sequences were mapped to three annotated reference genomes: *A. thaliana* (GCA_000001735.1), *B. distachyon* (GCA_000005505.4) and *M. truncatula* (GCA_000219495.2). The bait sequences were used to synthesize a custom myBaits kit (Arbor Biosciences) from now on referred to as the "FORAGE-611" baits.

2.1.2 | Plant DNA extraction

Seeds of 16 forage species were germinated on filter paper, and their seedlings were transferred into pot trays (77 wells, 50 × 32 cm, with compost as substrate). The 16 species were: *Alopecurus pratensis* L., *Arrhenaterum elatius* L., *Cynosurus cristatus* L., *D. glomerata*, *F. pratensis*, *F. rubra* L., *L. perenne*, *L. multiflorum* Lam., *L. corniculatus* L., *M. sativa* L., *Onobrychis viciifolia* Scop., *Phleum pratense* L., *Poa pratensis* L., *T. pratense*, *T. repens* L. and *Trisetum flavescens* L. After 3–5 weeks, five single plants from different cultivars per species were sampled (Table S1). For grasses, samples consisted of three leaf fragments of ~1 cm; for legumes, samples consisted of three young leaflets. The plant material was freeze-dried for 48 h and pulverized in a Qiagen TissueLyser II (Qiagen). DNA was extracted using the NucleoSpin II kit (Macherey-Nagel) and its integrity visually inspected by agarose gel electrophoresis (1% w/v). DNA purity and concentration were determined with a NanoDrop spectrophotometer (ThermoFisher Scientific).

2.1.3 | Sequence capture and DNA sequencing

Dual-indexed libraries were constructed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs). Each species was represented by six libraries: five single-plant libraries and one pooled DNA library, for a total of 96 libraries. The pooled-plant libraries consisted of equimolarly pooled DNA from the five single plants. The average library insert size was ~550 base pairs (bp).

After indexing, the libraries were divided in four pools of 6× libraries, four pools of 8× libraries and four pools of 10× libraries. Each pool was hybridized to the FORAGE-611 baits. The target DNA fragments were enriched following manufacturer instructions. Fragments were sequenced using the Illumina MiSeq v3 Reagent Kit (2 × 300 bp, 600 cycles). The raw paired-end reads were merged using BBMerge v37.36 (Bushnell et al., 2017) with default parameters. Adapter removal and quality filtering was done using fastp v0.20.0 (Chen et al., 2018) with default parameters.

2.1.4 | Sequencing quality control

The pooled DNA libraries were used to construct pseudo-reference assemblies (pseudo-RAs) using SPAdes v3.10.0 (Bankevich et al., 2012). One pseudo-RA was assembled for each species. The sequences of the FORAGE-611 baits were mapped to the pseudo-RAs using BBMap v37.36 (Bushnell, 2014) with default parameters. Bait-mapping coordinates were determined with SAMtools v1.2 (Li, 2011; Li et al., 2009) and BEDtools v2.28.0 (Quinlan & Hall, 2010). Target locus centers were defined as the middle-point between the 3′-most and 5′-most coordinate of the mapped bait sequences that belong to the same SCOG or ULE. A target locus region was defined as the sequences spanning 500 bp up- and downstream from each locus centre. In the cases where baits mapped to more than one pseudo-RA contig, only the largest contig was kept for further analysis.

The 80 single-plant libraries were mapped to their corresponding pseudo-RAs with BBmap. Reads mapping within target locus regions were considered as "on-target reads". Duplicate reads were handled with Picard v2.23.8 MarkDuplicates (Broad Institute, 2019), marking them for variant calling and removing them for *k*-mer richness calculations.

A locus with >5 mapped reads was labelled as a quality-controlled locus (QCL). Libraries with >100 QCL and >1000 total on-target reads were labelled as quality-controlled libraries (QC-libs). Only QC-libs and on-target reads were considered for further analysis. Read-mapping statistics were determined using BEDtools.

2.1.5 | Diversity metrics

Four diversity metrics were used to rank the 611 targeted loci within each species: nucleotide diversity (π), SNP-based nucleotide diversity, SNP density and normalized *k*-mer richness. Each metric was calculated per locus within each species.

To calculate SNP-based within-species gene diversity, variant calling was done species-by-species using BCFtools v1.12 mpileup (Li, 2011; Li et al., 2009) on the BAM files from species with >3 QC-libs. The resulting variant call files (VCFs) were filtered using BCFtools for biallelic SNPs with a minimum quality of 20, a minimum allele frequency of 0.1 and a minimum read depth per sample of 5. Estimations of π were calculated per site using VCFtools v0.1.14, treating all samples as diploid (Danecek et al., 2011). Locus-level estimations were then summed and divided by locus length, which in turn were based on the scaffold lengths in pseudo-RAs and had a maximum value of 1 kbp per locus. SNP count and SNP densities were calculated using the same custom R script based on genotype tables produced with VCFtools.

$$\mathrm{NKR} = \frac{K_c}{L - k + 1} \quad (1)$$

To calculate NKR, reads were extracted from de-replicated BAM files using SAMtools bam2fq, creating separate FASTQ files for each locus within each library. *K*-mers of $k = 25$ in each FASTQ file were determined using Jellyfish v2.2.10 (Marçais & Kingsford, 2011). *K*-mer dumps were filtered by discarding *k*-mers with <5× sequencing depth. Locus-specific NKR was calculated at the species level by concatenating the filtered *k*-mer dumps, counting unique *k*-mers within each concatenated dump and then dividing the value of such unique *k*-mer counts by the maximum expected *k*-mer count per locus (*L*–*k*+1). This is summarized in equation 1, where *Kc* indicates the count of unique *k*-mers, *k* indicates *k*-mer length, and *L* indicates locus length.

Total diversity metrics were calculated for each species using Rstudio v1.3.1717 (RStudio Team, 2020) running with R4.1.0 (R Core Team, 2021). Total NKR was calculated by dividing the sum of all unique *k*-mers in all loci by the total expected *k*-mers in all loci ($(L_{\mathrm{total}}-k+1)$, where L_{total} is the sum of the lengths of all loci in bp units). Total SNP count is the sum of all SNPs found in all loci and the SNP density per kbp was calculated following the formula: $SNP_{\mathrm{total}} \times 1000/L_{\mathrm{total}}$,, where SNP_{total} is the total SNP count. Total π was calculated using formula 2, where π_i is the nucleotide diversity of the *i*-th locus, L_i is the length of the *i*-th locus and *a* is the total number of loci.

$$n_{total} = \frac{\sum_i^a \pi_i \times L_i}{L_{\mathrm{total}}} \quad (2)$$

2.1.6 | Primer design

Eleven loci out of the 611 initially captured were selected for primer design. Selection criteria included having high median gene diversity and normalized *k*-mer richness, as well as being present in as many species as possible. Primers were designed using PriMux v20_july_2014 (Hysom et al., 2012) with a target amplicon size of 500 bp and primer size of 25 nt. The input for primer design was a set of FASTA files, one per locus, containing the contigs from an assembly of each QC-lib. The QC-lib assemblies were performed using SPAdes.py and they are available in Dryad (https://doi.org/10.5061/dryad.gb5mkkwqw) Primers with successful in silico and in vitro multispecies PCRs were selected for further testing. Thirteen primer pairs were initially selected for synthesis (Microsynth AG). Six primer pairs were grass-specific, six were legume-specific and one pair was found in both plant families. However, as AMP3437 targeted the same locus as AMP3425 (Table 1), all sequencing reads from the former were accounted for the latter. Therefore, the diversity metrics results consider only 12 different amplicons.

2.2 | Validation phase

2.2.1 | Plant DNA extraction

For single-plant DNA extractions, seeds of five forage species (*D. glomerata*, *F. pratensis*, *L. perenne*, *T. pratense* and *T. repens*) were germinated and used for DNA extraction after 3–4 weeks of growth as described above. Each species was represented by 16 individual plants from three different cultivars (Table S4). For pooled-plant DNA extractions, ground leaf material from the same 16 plants per species was pooled at equal proportions (50 mg per plant), mixed and 20 mg of that mixture were used for DNA extraction. DNA was diluted to 5 ng/µl.

2.2.2 | Amplicon sequencing

Selected amplicons (Table 1) were amplified in 25 µl PCR reactions containing 15 ng of template DNA, 1× flexi buffer (Promega), 2 mM $MgCl_2$, 200 µM dNTPs, each primer at 0.4 µM and 0.75 units of GoTaq G2 Flexi DNA polymerase (Promega). PCR conditions were 5 min at 94°C followed by 30 cycles of 1 min at 94°C, 1 min at a primer-specific melting temperature (Tm; Table 1) and 1 min at 72°C, followed by a final extension cycle of 10 min at 72°C. The amplicons for each sample were then equimolarly pooled. Dual-indexed libraries were constructed using the NEBNext UltraII DNA Library Prep Kit for Illumina (New England Biolabs). In total, 80 single-plant libraries (16 per species) and 15 pooled-plant libraries (three per species) were prepared. The composition of each Illumina library is detailed in Table S5. Dual-indexed libraries were equimolarly pooled and sequenced using the Illumina MiSeq v3 Reagent Kit (2 × 300 bp reads; 600 cycles). Merging of raw pair-ended reads, adapter removal and quality filtering were done as described above. Additionally, to remove possible remnants of primer sequences, reads were trimmed 25 bp at each end using Cutadapt v3.4 (Martin, 2011). In order to test the further utility of the multispecies primer, they were tested in eleven additional species, which were not used in the

TABLE 1 Loci selected for amplicon sequencing and their corresponding multispecies primer pair sequences

Locus	Corresponding *Arabidopsis thaliana* gene	Primer sequences (5′–3′)	Tm (°C)	Plant family[a]	Amplicon name
O-1262	AT3G07080	F: AAAGATTTGGATAGTAAAGCATGGA R: TCCARCGYCCTTTYKCATCC	60	F	AMP469
U-11004576	AT5G47010	F: AAGCTTGARGCTGAYTATGA R: TGCATACCACATGACCMACT	58	F	AMP615
O-1520	AT2G42900	F: ACACAAGCTCCTTTGTKGAA R: ATACGATCATGCCACGTGTC	58	F	AMP735
O-165	AT1G50480	F: GCTGATATGCGAGAGAGGCTAG R: TAACATTCGCACCATAAGCT	58	F	AMP3411
O-1390	AT1G08460	F: GGCACWTTCYTGAACCCTGG R: GAATAACTTACTGCACTYGAGTC	58	F	AMP3711
U-11001396	AT5G67170	F: GGGAGCATTATAAYAGTGTGCG R: CTTCTGYWCCTTGTTCWGCT	58	F	AMP3794
O-1211	AT5G61540	F: GTRGCACCATTGGTTGATGTG R: GAARTHGGAGCTGTKGSTGC	58	FP	AMP3941
O-1347	AT2G44020	F: GCGTGACATTGGTCCCATGG R: GATCCTBGACTCCAAGCTGTA	58	P	AMP3376
U-11004158/ O-1519	AT1G70570	F: GCTGGCATGACAATAAGAGC R: WGTRCTCCTRAARAAGCGTGT	58	P	AMP3425
U-11004158/ O-1519	AT1G70570	F: GCTGGCATGACAATAAGAGC R: AAAAGCGTGTATTCCCATCATA	60	P	AMP3437
O-1288	AT5G38460	F: GGAAGAGATTGTTTGCAATAAAGCC R: CATTATCAGTGCATCAAAGAGGA	58	P	AMP3532
O-165	AT1G50480	F: GGAGGTGGCTACAGTCAAGT R: GTGGGATGAATAGCATCTTTCATT	58	P	AMP3625
O-1318	AT5G04050	F: GGGAGGAGGTGCACAAGAAG R: ACAACCGRTGCCARTAACGG	58	P	AMP3799
rbcLa	*rbcLa*	F: ATGTCACCACAAACAGAGACT AAAGC R: GTAAAATCAAGTCCACCRCG	55	FP	*rbcLa*
psbK-psbI	*psbK-psbI*	F: TCTMTTAGCYTTTGTTTGGCAAGCT R:ACAAAWAGTTTKAGAGTAAGCAT	55	FP	psbK-psbI

[a]Plant family abbreviations: F, Fabaceae (legumes, which included: *Trifolium pratense* and *T. repens*); P, Poaceae (grasses, which included *Dactylis glomerata*, *Festuca pratensis* and *Lolium perenne*).

discovery phase, using a touch-down PCR approach: the legumes *G. max* and *M. lupulina* L., and the grasses *F. arundinacea* Schreb., *Agrostis stolonifera* L., *B. distachyon*, *Holcus lanatus* L., *Bromus erectus* Huds., *B. inermis* Leyss., *B. catharticus* Vahl., *B. arvensis* L., and *Z. mays*. The products of those additional PCRs and the touch-down PCR protocol are shown in Figures S2 and S3. The amplicons from the additional species were not used for Illumina library preparation.

2.2.3 | Diversity metrics

To calculate SNP-based diversity metrics, the quality-controlled reads were mapped to pseudo-RAs, which consisted only of the sequences of the selected loci taken from the sequence capture data of the discovery phase. The pseudo-RAs also contained the sequences of DNA barcode *rbcLa* corresponding to each species, which were obtained from the Barcoding Of Life Datasystems (Ratnasingham & Hebert, 2007), project SWFRG (Loera-Sánchez et al., 2020). For single-plant libraries, variant calling and SNP-based diversity metrics were calculated as described above. Variant calling was performed simultaneously on the BAM files of the 16 single-plant libraries of each species. In the case of pooled-plant libraries, amplicon-wise π and SNP counts were calculated using the "Variance-at-position.pl" script from PoPoolation v1.2.2 (Kofler et al., 2011) with a minimum covered fraction of 0.5, a pool size of 32 (i.e., the equivalent of 16 diploid plants), a minimum site quality of 20, a minimum coverage of

10, and a minimum allele count of 5. The PoPoolation analysis was run individually for each BAM file of the pooled-plant libraries.

SNP haplotypes were reconstructed by concatenating each SNP position from each sequencing read for both single- and pooled-plant samples. For this, BAM files containing the read mappings to the amplicon pseudo-RAs were parsed to extract reads and SNP positions using a custom Python script ("haplotyper.2.py" available at https://github.com/mloera/gendiv/commit/9f4d4a3606c8d192d8d7dd2680e5a4728994b7d6). The SNP positions came from the mappings of the single-plant libraries. Determining the final haplotype-based genotype required filtering out spurious haplotypes with low read counts, which likely are due to PCR or sequencing errors. Briefly, for each amplicon, the haplotype with the maximum read count was determined. Such maximum read count was used to normalize read counts for the rest of the haplotypes. Normalized haplotype allele counts thus ranged from zero to one. Only haplotypes with normalized counts higher than a specific cutoff value (0.9 for diploids, 0.3 for tetraploids and 0.2 for pools) were retained. For single-species samples, haplotypes were further visually inspected, and the following was corrected: (a) haplotypes with missing values were removed, (b) haplotypes with tri- or tetra-allelic SNPs were removed and (c) co-occurring haplotypes with only one different SNP allele call were consolidated into one haplotype. No visual inspection was conducted for pooled-plant samples, but only haplotypes that were present in single-species samples were retained.

Haplotype-based multilocus genotypes, pairwise Prevosti's distances and haplotype allele count (H.Ac) were calculated using the "poppr" v2.8.6 R package (Kamvar et al., 2014).

To calculate normalized *k*-mer richness, reads were extracted from de-replicated BAM files using SAMtools bam2fq, creating separate FASTQ files for each amplicon within each library. Each FASTQ file was then subsampled to 100 reads using Seqtk v1.3 (Seqtk-1.3, 2018). *K*-mers of $k = 25$ in each subsampled FASTQ file were determined with Jellyfish (Marçais & Kingsford, 2011). The resulting *k*-mer dumps were filtered by discarding *k*-mers with <25× sequencing depth (i.e., 25% of the maximum depth). Figure S4 shows the effect of the depth cutoff on NKR for each amplicon and species. Amplicon-specific NKR was calculated at the species level as described above.

Total diversity per species was calculated using exclusively single-plant libraries. Calculations were conducted as described in the discovery phase. Genetic diversity metrics for each sample were compared to the diversity of the DNA barcode *rbcLa*, which is known to have a very low within-species diversity.

2.2.4 | SSR and multilocus analysis

Single-plant DNA samples were also used as templates for SSR analyses. In total, eight species-specific SSR primer pairs were used. SSR primers and PCR conditions are described in Table S6. Genotype tables were analysed using the R package poppr (Kamvar et al., 2014). Summary statistics per locus are shown in Table S7.

The five amplicons with the highest coverage per species were selected for multilocus analysis (AMP469 was excluded for *T. repens* as some genotypes that did not match the species' ploidy indicating that it is potentially duplicated). Only single-plant libraries were analysed. The five most diverse SSR per species were selected for multilocus analysis (four for *T. repens*).

Multilocus genotypes were produced for SSR (from now on "SSR-MLGs") and for amplicon-based haplotypes (from now on "HAP-MLGs") using the R package poppr (Kamvar et al., 2014). Pairwise Prevosti's distance was calculated for SSR- and HAP-MLGs using the R package poppr (Kamvar et al., 2014). Plants with >5% missing loci, either amplicon-based haplotypes or SSR, were removed from this analysis. Only the amplicons from single-plant libraries were considered for the multilocus distance analysis.

In addition, multilocus *k*-mer distances were calculated using sourmash (Brown & Irber, 2016). For this, the normalized FASTQ files described above were merged into multiamplicon FASTQ files according to library name. Only FASTQ files corresponding to single-plant libraries and to the amplicons selected for HAP-MLGs were merged. Multiamplicon FASTQ files were used to calculate *k*-mer signatures using sourmash v4.0.0 sketch (Brown & Irber, 2016). Finally, pairwise similarities among samples of the same species were calculated using sourmash compare (Brown & Irber, 2016). Distances were calculated from similarities by applying the formula: $d = 1-s$, were d is distance and s is simalirity.

3 | RESULTS

3.1 | Discovery phase

3.1.1 | Target loci selection and bait design

A total of 611 loci (265 ULEs and 346 SCOGs) were selected, for which 12,253, 100-nucleotide long baits were designed (5526 targeting ULEs and 6727 SCOGs; FASTA files in Dryad: https://doi.org/10.5061/dryad.gb5mkkwqw). In total, 602 loci mapped to a gene model in any of the three reference genomes used as controls (*A. thaliana*, *B. distachyon* and *M. truncatula*). Of those loci, 365 were present in all three reference genomes, 142 were present in only in *A. thaliana* and *B. distachyon*, 31 in *A. thaliana* and *M. truncatula*, 31 in *B. distachyon* and *M. truncatula*, 10 only in *A. thaliana*, 10 in *B. distachyon* and 13 in *M. truncatula* (Table S2 and Figure S1).

In any of the three reference genomes, a total of 85 genes contained two or more target loci and, conversely, 23 loci mapped to two or more genes (Table S3).

3.1.2 | Sequence capture and sequencing

Around 14.5 million raw read pairs were obtained after sequence capture, 10.3 million reads for grasses and 4.2 million for legumes (Table 2; raw FASTQ files in Dryad: https://doi.org/10.5061/

dryad.gb5mkkwqw). At the species level, total raw output ranged from 512,233 read pairs for *T. flavescens* to ~1.5 million read pairs for *L. corniculatus*. The N50 of pseudo-RAs generated from the pooled-plant samples ranged from 590 bp (*T. pratense*) to 4905 bp (*A. pratensis*).

Approximately two million read pairs were successfully quality-controlled, merged and mapped to a target locus. This constitutes a total capture efficiency of 13.45% of the total raw reads. Within grasses and legumes, capture efficiencies were 4.78% (495,858 quality-controlled, merged reads) and 35% (~1.5 million quality-controlled, merged reads), respectively. At the species level, capture efficiency varied from 1.58% for *A. elatius* (18,735 quality-controlled, merged reads) to 55.82% for *T. pratense* (537,289 quality-controlled, merged reads). Read duplication in quality-controlled, merged reads ranged from 12.75% in *F. rubra* to 38.59% in *T. pratense*.

For further analysis, only quality-controlled loci (QCL: loci with five or more quality-controlled, merged reads within a library) and quality-controlled libraries (QC-libs: libraries with >100 QCL and >1000 quality-controlled, merged reads) were considered. Read counts were always calculated as quality-controlled, merged reads. In total, 60 QC-libs with a median of 438 QCL per library (interquartile range, IQR = 355–532; Table 2) were obtained. Twenty-three loci were not captured in any library. Each QCL had a median of 37 reads (IQR = 17–84). For grasses, there were 37 QC-libs with a median of 377 QCL per library (IQR = 313–410); median reads per QCL in grasses was 21, (IQR = 12–36). For legumes, there were 23 QC-libs with a median of 548 QCL (IQR = 520–552); median reads per QCL in legumes was 84 (IQR = 48–136). At the species level, QC-lib varied from one for *T. flavecens* to a maximum of five obtained for *C. cristatus*, *D. glomerata*, *L. corniculatus*, *O. viciifolia* and *T. pratense*. Median read count per QCL varied from 13 in *L. perenne* (IQR = 9–20) to 149 in *T. pratense* (IQR = 97–227). Median QCL count varied from 213 for *A. elatius* (IQR = 190–233) to 579 for *T. repens* (IQR = 574–584).

3.1.3 | Genetic diversity estimations

Diversity statistics were calculated for each locus within each species. Total diversity metrics for each species are presented in Table 3. The genetic diversity estimates for each locus and species are presented

TABLE 2 Sequence capture output summary

Family	Species name	Raw read pairs	% capture efficiency[a]	% duplicates in captured reads	Reads per locus per library[b]	Loci per library[b]	QC-libs.
Fabaceae	*Lotus corniculatus*	1,471,339	21.05	34.34	88 (49, 143)	528 (520, 531)	5
	Medicago sativa	674,442	34.34	31.83	82 (51, 118)	552 (551, 554)	4
	Onobrychis viciifolia	539,336	29.17	22.19	53 (33, 79)	465 (462, 468)	5
	Trifolium pratense	962,538	55.82	38.59	149 (97, 227)	548 (548, 548)	5
	Trifolium repens	523,205	42.86	30.95	72 (43, 111)	579 (574, 584)	4
Poaceae	*Arrenatherum elatius*	1,185,777	1.58	26.62	15 (10, 27)	213 (190, 233)	3
	Alopecurus pratensis	1,391,349	3.55	25.31	20 (12, 33)	415 (378, 429)	4
	Cynosurus cristatus	839,504	5.78	29.16	18 (11, 30)	356 (354, 378)	5
	Dactylis glomerata	770,598	7.15	32.18	20 (12, 33)	377 (365, 387)	5
	Festuca pratensis	1,116,236	1.73	30.88	21 (12, 31)	233 (228, 238)	2
	Festuca rubra	975,051	4.36	12.75	20 (12, 32)	369 (355, 385)	4
	Lolium multiflorum	889,813	2.24	18.12	14 (10, 20)	308 (300.5, 312)	3
	Lolium perenne	624,934	2.31	13.46	13 (9, 20)	309 (307, 310)	2
	Phleum pratense	956,719	6.08	15.52	23 (14, 36)	449 (437, 455)	4
	Poa pratensis	1,111,382	11.81	33.02	41 (21, 70)	531 (514, 536)	4
	Trisetum flavescens	512,233	7.60	34.52	30 (16, 55)	407	1
Fabaceae		4,170,860	35.00	33.68	84 (48, 136)	548 (520, 552)	23
Poaceae		10,373,596	4.78	26.62	21 (12, 36)	377 (313, 410)	37
	Total	14,544,456	13.45	31.89	37 (17, 84)	438 (355, 532)	60

[a]Capture efficiency expressed as the percentage of merged, on-target reads, that is, reads that were quality-controlled and successfully merged and mapped to a pseudoreference assembly; percentage is in relation to raw read pairs.

[b]Median and interquartile range calculated considering only quality-controlled loci (QCL, loci with five or more quality-controlled, merged reads) and quality-controlled libraries (QC-libs, libraries with >100 QCL and >1000 quality-controlled, merged reads). Each QC-lib represents a single plant.

in Table S9. At the amplicon level, NKR values ranged from zero (locus ortho-1479 in *T. flavescens*) to 13.04 (locus ortho-147 in *T. repens*) with a median of 1.06 (IQR = 0.82–1.35). SNP count ranged from one (326 locus-species pairs) to 86 (loci ortho-1288/uce-11005138 in *Phleum pratense*) with a median of nine (IQR = 4–15). SNP densities ranged from one (36 locus-species pairs) to 99.83 (ortho-1178 in *T. pratense*) with a median of 10.79 (IQR = 5–19.27). Nucleotide diversity (π) ranged from 2.5×10^{-4} (ortho-1106 and uce-12001183 in *M. sativa*) to 5.17×10^{-2} (ortho-1186/uce-11003333 in *F. pratensis*) with a median of 4.99×10^{-3} (IQR = 2.33×10^{-3}–9.08×10^{-3}).

In general, NKR and π showed a very low correlation (r^2 = .08, Figure 2a). At the species-level, r^2 values ranged from .07 for *P. pratensis* to .47 in *L. multiflorum*, while for most species $r^2 \geq .1$ (Figure 2b).

Locus selection was based on median diversity values across species, as well as on species counts (i.e., the number of species where a locus was found) and the suitability for multispecies primer design. The loci assemblies used for multispecies primer design can be found in Dryad: https://doi.org/10.5061/dryad.gb5mkkwqw. Across species, median NKR values in the selected loci ranged from 0.45 (ortho-350) to 8.78 (ortho-106). Median SNP count ranged from 1 SNP (ortho-1474 and ortho-1498) to 43 SNPs (uce-11005074). Median SNP densities ranged from 1.01 (ortho-1474) to 49.7 SNPs/kbp (ortho-1198). Median π ranged from 5.3×10^{-4} (ortho-1498) to 2.1×10^{-2} (uce-11005074). Median diversity metrics across species for the selected loci are shown in Table 4. A Wilcoxon test showed significant differences (p-value <.001) between selected and nonselected loci based on the median of medians of NKR, SNP count, SNP density per kbp and π (Figure 2c).

3.2 | Validation phase

3.2.1 | Multispecies amplicon sequencing and genetic diversity analysis

Amplicon sequencing in test populations (n = 16) of *D. glomerata*, *F. pratensis*, *L. perenne*, *T. pratense* and *T. repens* produced 12.4 million raw read pairs, of which 4.3 million reads (34.81%) were successfully quality-controlled and merged (Table 5). The raw sequencing output was 6.5 million reads for grasses, with 1.8 million quality-controlled, merged reads (26.77% of raw output for this plant family). For legumes, raw output was ~5.9 million reads, with 2.5 million quality controlled, merged reads (43.72%). At the species level, raw output was 1.99, 2.25, 2.30, 3.03 and 2.86 million reads for *D. glomerata*, *F. pratensis*, *L. perenne*, *T. pratense* and *T. repens*, respectively. The raw FASTQ files can be found in Dryad: https://doi.org/10.5061/dryad.gb5mkkwqw. In turn, quality-controlled, merged read counts were 0.57 (28.79% of raw output for this plant species), 0.64 (28.34%), 0.54 (23.49%), 1.51 (49.67%) and 1.07 (37.42%) million reads for *D. glomerata*, *F. pratensis*, *L. perenne*, *T. pratense* and *T. repens*, respectively. Duplicate rate was always >90% in all groups (total, plant families, and species).

At the locus level, quality-controlled amplicons (i.e., amplicons with ≥5 mapped reads) reported a median of >1000 reads per library. Median count of quality-controlled amplicons was eight in *L. perenne* and *T. pratense* libraries, and seven in the rest of the species. The DNA barcode *psbK-psbI* could not be recovered from legume samples, so it was excluded from analysis. Furthermore, amplicon AMP3794 had zero coverage in *T. repens* libraries. In addition, amplicon AMP3437 had zero coverage in *D. glomerata* and *F. pratense* libraries. The latter case is because AMP3437 targets the same locus as AMP3425 so, from here on, AMP3437 reads were treated as AMP3425 reads.

In single-plant libraries and excluding DNA barcode *rbcLa*, total NKR values ranged from 1 (*F. pratensis*) to 2.05 (*T. repens*). For most amplicons and species, NKR showed a steady decline as *k*-mer depth cutoff values increased, except for *rbcLa*, AMP3625, AMP3794, and AMP3941, which had the lowest total NKR and whose NKR values start declining only at depth cutoff values >50% (Figure S4). Total SNP counts ranged from 27 (*F. pratensis*) to 62 (*D. glomerata*). Total π ranged from 5.19×10^{-3} to 1.29×10^{-2}. Total SNP-based haplotype allele counts (H.Ac) ranged from 13 (*T. repens*) to 43 (*L. perenne*). Total diversity metrics per species are shown in Table 6.

In single-plant libraries and excluding DNA barcode *rbcLa*, amplicon-level NKR ranged from zero (AMP3625 in *F. pratensis*) to 2.94 (AMP3799 in *D. glomerata*) with a median of 1.53 (IQR = 1.02–1.97). SNP count ranged from zero (nine species-locus pairs) to 28 (AMP3799 in *D. glomerata*) with a median of 3 (IQR = 0–11). SNP densities ranged from zero (nine species-locus pairs) to 50.36 SNPs/kbp (AMP3799 in *D. glomerata*) with a median of 6.91 (IQR = 0–20.14). π ranged from zero (nine species-locus pairs) to 3.83×10^{-2} (AMP3799 in *D. glomerata*) with a median of 4.62×10^{-3} (IQR = 0–1.52×10^{-2}). SNP-based haplotype allele count ranged from one (11 cases) to 25 (AMP3799 in *D. glomerata*) with a median of 3 (IQR = 1–7).

In pooled-plant libraries, amplicon-level NKR values ranged from zero (AMP3625 in *F. pratensis* pool 2) to 3.62 (AMP3799 in *D. glomerata* pool 1) with a median of 1.62 (IQR = 1.01–2.21). SNP counts ranged from zero (38 species-locus pairs) to 54 (AMP3799 in *D. glomerata* pool 1) with a median of 2 (IQR = 0–19). SNP density ranged from zero (38 species-locus pairs) to 101.82 SNPs/kbp (AMP3799 in *D. glomerata* pool 3) with a median of 3.77 (IQR = 0–35.83). π ranged from 0 (38 species-locus pairs) to 6.68×10^{-2} (AMP3425 in *L. perenne* pool 1) with a median of 3.13×10^{-3} (IQR = 0–1.51×10^{-2}). H.Ac ranged from one (41 species-locus pairs) to 22 (AMP3799 in *D. glomerata* pool 1) with a median of 2 (IQR = 1–5). Median genetic diversity statistics per amplicon considering both single- and pooled-plant libraries are shown in Table 7.

The concordance between diversity metrics calculated for each species-amplicon pair from single-plant libraries and pooled-plant libraries was good. Overall, the coefficients of variation of all diversity metrics for each amplicon and species had a median of 13.38% (IQR = 3.35%–27.58%). The coefficient of variation (CV) of NKR estimates ranged from 0.11% (AMP3532 in *F. pratensis* and *rbcLa* in *T. pratense*) to 86.67% (AMP3625 in *F. pratensis*) with a median of 8.1% (IQR = 2.43%–16.68%). The range of the CV of

TABLE 3 Total diversity statistics per species for captured loci

Family	Species name	NKR[a]	SNPs	SNPs/kbp	π[b]	N[c]
Fabaceae	*Lotus corniculatus*	1.29	4339	9.68	4.25×10^{-3}	477
	Medicago sativa	1.20	5615	12.47	5.38×10^{-3}	507
	Onobrychis viciifolia	1.17	3259	8.84	4.03×10^{-3}	416
	Trifolium pratense	1.31	2766	6.36	2.76×10^{-3}	469
	Trifolium repens	1.55	8463	20.17	9.41×10^{-3}	536
Poaceae	*Alopecurus pratensis*	1.18	3819	19.66	9.50×10^{-3}	234
	Arrhenatherum elatius	0.78	287	5.29	2.66×10^{-3}	72
	Cynosurus cristatus	1.03	1040	7.3	3.35×10^{-3}	173
	Dactylis glomerata	1.36	3022	15.91	7.84×10^{-3}	229
	Festuca pratensis	1.11	3525	26.81	1.73×10^{-2}	188
	Festuca rubra	0.76	3208	14.77	7.15×10^{-3}	281
	Lolium multiflorum	0.85	3009	17.08	8.38×10^{-3}	218
	Lolium perenne	0.65	1479	8.55	5.05×10^{-3}	211
	Phleum pratense	1.13	6094	20.12	8.67×10^{-3}	354
	Poa pratensis	1.24	5723	17.4	9.28×10^{-3}	415
	Trisetum flavescens	0.53	696	4.88	4.92×10^{-3}	179

[a]NKR, normalized *k*-mer richness.

[b]π, nucleotide diversity.

[c]*N*, total quality-controlled loci captured per species.

SNP counts ranged from 0% (16 species-amplicon pairs) to 88.79% (AMP3425 in *L. perenne*) with a median of 12.87% (IQR = 0%–27.75%). In the case of SNP densities, CV ranged from 0% (12 species-amplicon pairs) to 90.60% (AMP3425 in *L. perenne*) with a median of 10.10% (IQR = 0%–28.65%). The CV of π estimates ranged from 0% (12 species-amplicon pairs) to 114.47% (AMP3425 in *L. perenne*) with a median of 15.43% (IQR = 0%–26.06%). Finally, the CV of SNP-based haplotype counts ranged from 0% (22 species-amplicon pairs) to 119.02% (AMP3425 in *L. perenne*) with a median of 0% (IQR = 0%–23.14%).

At the amplicon level and across species, median CV of NKR estimates ranged from 1.36% (AMP3532) to 56.70% (AMP3625). For SNP counts, median CV ranged from 0% (AMP615, AMP3411, AMP3941 and AMP3794) to 88.79% (AMP3425). For SNP densities, median CV ranged from 0% (*rbcLa* and AMP3941) to 90.60% (AMP3425). For π estimates, median CV ranged from 0% (*rbcLa* and AMP3941) to 114.47% (AMP3425). For SNP-based haplotype counts, median CV ranged from 0% (*rbcLa*, AMP3941, AMP615, AMP3411, AMP3625, AMP3794 and AMP3425) to 28.57% (AMP3376).

Across species, the amplicons were significantly more diverse than the DNA barcode *rbcLa* for at least one diversity metric (Figure 3a). However, amplicon AMP469 in *T. repens* produced SNP-based haplotypes that did not match the ploidy of this species (i.e., tetraploid). Therefore, this amplicon was not considered for overall diversity calculations nor for comparisons with SSR-based diversity metrics.

Sequence-based diversity metrics showed a weak positive correlation to each other ($.25 \leq r^2 \leq .32$, Figure 3b, black). Excluding amplicons without SNP calls marginally improved correlations ($0.3 \leq r^2 \leq .45$, Figure 3b, blue). Excluding amplicons without SNP calls and considering only single-plant libraries resulted in considerably higher correlations ($.51 \leq r^2 \leq .69$, Figure 3b, red).

3.2.2 | Amplification beyond the 16 species used for targeted sequencing

Five out of six multispecies amplicons were also successfully amplified in at least three of the additional grass species not included in the discovery phase (Figure S2). However, the primers for locus ortho-1347 showed only faint bands in *F. arundinacea*, *A. stolonifera* and *B. distachyon*, and locus ortho-1211 was not amplified in any of the grass species. In addition, none of the amplicons was amplified in *Z. mays*. The multispecies primers were less effective in the two additional legume species *M. lupulina* and *G. max*. The loci uce-11004576, ortho-1520, ortho-165, and ortho-1390 were amplified in *M. lupulina*, while only the locus uce-11004576 was amplified in *G. max* (Figure S3).

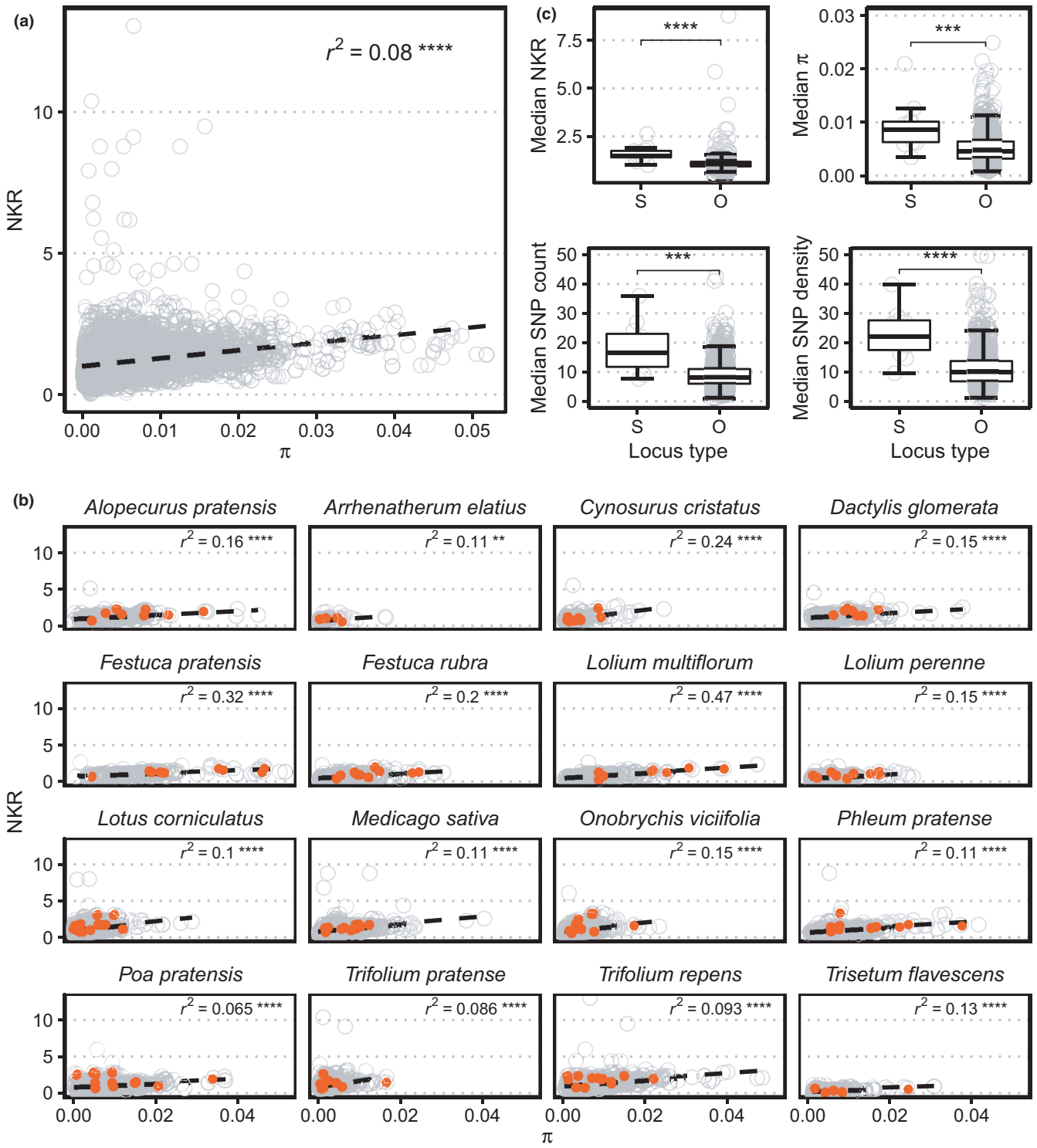

FIGURE 2 Diversity statistics of the captured loci. (a) Nucleotide diversity per kbp (π) versus normalized *k*-mer richness (NKR) for all captured loci in the 60 quality-controlled libraries. Linear regression shown in black dotted line. Regression r^2 value and *p*-value significance level shown in the upper right corner. (b) π per kbp versus NKR for all captured loci divided by species. Red dots indicate selected loci. Linear regression shown in black dotted line. Regression r^2 value and *p*-value significance level shown in the upper right corner. (c) A Wilcoxon test shows that median diversity metrics across the 16 species are higher in selected loci ("S" label, $n = 10$) than in other captured loci ("O" label, $n = 578$). Significance levels for *p*-values: .0001 = ****; .001 = ***; .01 = **; .05 = *; ns = nonsignificant. The upper and bottom ends of boxplots indicate the third and first quartiles, respectively. The black line inside of boxplots indicates the median value. The upper and lower whiskers in boxplots indicate the maximum (third quartile +1.5 times the interquartile range) and minimum (first quartile - 1.5 times the interquartile range) values, respectively

TABLE 4 Median genetic diversity statistics of the selected loci

Locus	NKR[a]	SNPs	SNPs/kbp	π[b]	N[c]
ortho-1390	1.00 (0.84–1.36)	9.50 (5.50–18.00)	16.00 (6.66–27.32)	8.09×10^{-3} (2.45×10^{-3}–1.21×10^{-2})	15
ortho-1347	1.20 (0.79–1.35)	22.00 (6.00–31.00)	22.00 (6.83–34.66)	8.65×10^{-3} (4.15×10^{-3}–1.09×10^{-2})	15
ortho-1318	1.40 (1.06–1.60)	16.50 (8.50–35.00)	26.46 (9.80–39.14)	9.68×10^{-3} (4.29×10^{-3}–1.95×10^{-2})	14
ortho-1262	1.42 (1.13–1.66)	16.50 (13.62–22.25)	28.13 (19.12–31.58)	9.94×10^{-3} (7.27×10^{-3}–1.50×10^{-2})	16
ortho-1211	1.43 (0.93–2.13)	14.00 (9.50–21.00)	19.71 (10.70–29.00)	6.36×10^{-3} (5.59×10^{-3}–8.63×10^{-3})	16
uce-11004576	1.49 (1.12–1.82)	9.00 (2.50–14.50)	9.55 (3.38–26.15)	3.42×10^{-3} (1.33×10^{-3}–8.68×10^{-3})	14
uce-11004158	1.68 (1.20–2.10)	24.00 (11.50–31.00)	27.06 (18.69–38.19)	1.03×10^{-2} (8.50×10^{-3}–1.75×10^{-2})	15
ortho-1288	1.74 (1.28–1.97)	36.00 (14.25–62.75)	39.82 (15.76–68.32)	2.10×10^{-2} (9.09×10^{-3}–3.49×10^{-2})	12
ortho-165	1.75 (1.52–2.24)	15.00 (6.00–24.75)	18.97 (7.89–28.09)	6.23×10^{-3} (2.88×10^{-3}–1.08×10^{-2})	14
ortho-1520	1.91 (1.68–2.07)	27.00 (18.00–33.00)	29.22 (19.72–39.14)	1.26×10^{-2} (8.34×10^{-3}–1.73×10^{-2})	3
uce-11001396	2.62 (1.33–2.77)	7.67 (6.00–13.00)	14.59 (11.00–35.00)	5.85×10^{-3} (3.64×10^{-3}–8.53×10^{-3})	13

[a]NKR, normalized *k*-mer richness.

[b]π, nucleotide diversity.

[c]*N*, number of species where a locus was found.

TABLE 5 Amplicon sequencing output summary

Family	Species name	Raw read pairs	% merged reads[a]	% duplicates in merged[b]	Median reads per amplicon[c]	Median amplicons per library[d]	Library count
Fabaceae	*Trifolium pratense*	3,034,018	49.67	96.32	6684 (4119–15,068)	8	19
	Trifolium repens	2,864,276	37.42	95.99	5947 (2935–11,129)	7	19
Poaceae	*Dactylis glomerata*	1,987,368	28.79	91.34	2652 (1254–5476)	7	19
	Festuca pratensis	2,250,146	28.34	92.63	2208 (945–6793)	7	19
	Lolium perenne	2,306,113	23.49	91.69	1273 (577–4876)	8 (7–8)	19
Fabaceae	All	5,898,294	43.72	96.18	6446 (3336–13,948)	7 (7–8)	38
Poaceae	All	6,543,627	26.77	91.92	1901 (868–5870)	7	57
	Total	12,441,921	34.81	94.46	4149 (1174–8411)	7 (7–8)	95

[a]Merged, on-target reads are reads that were quality-controlled and successfully merged and mapped to a reference amplicon; percentage is in relation to raw read pairs.

[b]Percentage of duplicate reads within the merged reads.

[c]Median and interquartile range calculated considering only quality-controlled amplicons (e.g., amplicons with five or more mapped reads).

[d]Interquartile range shown only for species with varying counts of amplicons per library.

3.2.3 | Multilocus analysis

K-mer based distances had an overall median of 0.29 (IQR = 0.21–0.38). At the species level, median *k*-mer-based distances were 0.34 (IQR = 0.28–0.39) for *D. glomerata*, 0.20 (IQR = 0.15–0.43) for *F. pratensis*, 0.31 (IQR = 0.22–0.40) for *L. perenne*, 0.28 (IQR = 0.19–0.39) for *T. pratense*, and 0.24 (IQR = 0.21–0.30).

Prevosti's pairwise distances based on HAP-MLGs had an overall median of 0.25 (IQR = 0.12–0.32). At the species level, median distances based on HAP-MLGs were 0.28 (IQR = 0.25–0.32) for *D. glomerata*, 0.25 (IQR = 0.16–0.30) for *F. pratensis*, 0.28 (IQR = 0.22–0.32) for *L. perenne*, 0.45 (IQR = 0.40–0.50) for *T. pratense* and 0.09 (0.06–0.12) for *T. repens*.

Prevosti's pairwise distances based on SSR-MLGs had an overall median of 0.60 (IQR = 0.50–0.70). At the species level, median distances based on SSR-MLGs were 0.55 (IQR = 0.50–0.65) for *D. glomerata*, 0.50 (IQR = 0.40–0.60) for *F. pratensis*, 0.50 (IQR = 0.40–0.55) for *L. perenne*, 0.90 (IQR = 0.83–1.00) for *T. pratense* and 0.62 (IQR = 0.56–0.69) for *T. repens*.

Distances based on SSR-MLGs were larger than both *k*-mer- and HAP-MLGs distances in all species (Figure 4). In turn, *k*-mer-based distances were significantly larger than HAP-MLGs distances in all species except *F. pratensis* (no significant differences between *k*-mer- and haplotype-based distances, Figure 4) and *T. pratense* (haplotype-based distances were larger than *k*-mer based, Figure 4).

In grasses, *k*-mer- and haplotype-based distances are very similar to each other (i.e., their ratio is close to 1; Table 8), whereas in legumes, they are more dissimilar (i.e., their ratio deviates from 1; Table 8).

4 | DISCUSSION

The sequence capture approach of the discovery phase of this work was followed as a steppingstone to find suitable loci for

TABLE 6 Genetic diversity per species based on the multispecies amplicons in single-plant libraries

Family	Species name	NKR[a]	SNPs	SNPs/kbp	π[b]	Haplotype allele count	Total length (bp)[c]	N[d]
Fabaceae	*Trifolium pratense*	1.42	36	10.28	7.63×10^{-3}	36	3503	7
	Tritfolium repens	2.01	23	11.56	1.05×10^{-2}	13	1989	5
Poaceae	*Dactylis glomerata*	1.89	62	18.01	1.29×10^{-2}	40	3442	6
	Festuca pratensis	1.00	27	7.75	5.19×10^{-3}	19	3483	6
	Lolium perenne	1.41	44	12.61	8.49×10^{-3}	43	3490	6
	Mean	1.55	38.40	12.04	8.94×10^{-3}	30.20	3181.40	6

[a]NKR, normalized *k*-mer richness.

[b]π, nucleotide diversity.

[c]Sum of the lengths of all amplicons in each species expressed in base pairs (bp).

[d]N, number of amplicons per species. All metrics exclude the DNA barcode *rbcLa* and amplicon AMP469 for *Trifolium repens*.

multispecies amplicon sequencing; however, it can per se be useful for other applications that justify its relatively high cost. For example, the FORAGE-611 bait set could help elucidating biogeographic patterns in grasses and legume populations. It could also help to resolve phylogenetic relationships in closely related clades, such as the *Festuca-Lolium* complex, which comprises some of the most widely cultivated forage grasses in temperate climates (Cheng et al., 2016). It can also be a helpful tool to study intergeneric hybridization and allopolyploidization, a prevalent feature in grasses (Tkach et al., 2020).

We observed capture efficiencies (i.e., the proportion of the total sequencing read output that maps on the target loci; Table 2) ranging from 1.58% to 55.82% per species, a range that is in line with previous reports of multispecies sequence capture assays. For example, a study of 25 legume species reported capture efficiencies ranging from 17% to 48% per species (Vatanparast et al., 2018). Another study that analysed 42 angiosperm species reported a wider range of capture efficiencies per species: 5% to 68.1% (Johnson et al., 2019). Grass and legume species had different capture efficiencies, which is likely due to grasses having larger average genome sizes (4838.5 Mbp) than legumes (958.4 Mbp; Table S8). As sequence capture libraries with similar DNA concentrations were pooled together in each sequence capture reaction, disregarding their species and genome size, this could have resulted in a lower effective concentration of target loci for grass samples in the discovery phase of this study. This is supported by a negative linear correlation observed between genome size and capture efficiency ($r^2 = .58$, p-value <.001). On average and depending on the species, 34.9% to 94.8% of the 611 target loci per species were recovered (Table 2). A similarly high, species-dependent variability of captured loci has been observed in other multispecies studies, from 54% to 98% (Vatanparast et al., 2018) or even 2% to 98% (Johnson et al., 2019).

We expected our estimates of genetic diversity based on the sequence capture assay to be lower than genome-wide genetic diversity estimates of forage crop species. This is because of the large proportion of coding sequences and ULEs in our target loci, which were included to increase the chances of finding multispecies priming sites around polymorphic sequences. SNP densities per species (Table 3) were lower than a genome-wide estimate in *L. multiflorum* (31 SNPs/kbp; Knorst et al., 2019). They were also lower than an estimate based on genic regions in *L. perenne* (35.6 SNPs/kbp; Ruttink et al., 2015). The SNP densities we found in our sequence capture data are closer to the 17 SNPs/kbp reported for a set of plastomes from a comprehensive sampling of *D. glomerata* from western Europe (Hodkinson et al., 2019).

Three main criteria to choose candidate loci for the validation phase were followed: (a) a high NKR value, (b) high SNP-based diversity (i.e., high π and SNP density) and (c) suitability for multispecies primers with predicted amplicon sizes of 450 ± 50 bp. Such an amplicon length was chosen so the longest possible haplotypes could be generated after merging Illumina MiSeq reads.

We observed a low to moderate positive correlation between NKR and π when analysing each species individually (Figure 2b), but no significant correlation was present in ungrouped data (Figure 2a). This is likely because the pseudo-RAs underlying our SNP and *k*-mer analyses were generated on a species-by-species basis. In addition, indel size variability was high (Figure S5), causing poor correlations between SNP-based metrics and NKR. By comparing the SNP- and *k*-mer-based diversity metrics, we were able to identify some loci that could be problematic for sequencing and/or genetic diversity analysis. Such loci had a low π and a high NKR (Figure 2a), indicating large indels or possible paralogous sequences. Thus, they were not considered for further analysis.

In the validation phase of this study, we focused on 11 polymorphic loci and the DNA barcode *rbcLa* (Table 1), which were amplified and sequenced from single- and pooled-plant samples of *D. glomerata*, *F. pratensis*, *L. perenne*, *T. pratense* and *T. repens*. We initially designed a set of >30 candidate multispecies primer pairs targeting >20 loci, but we only picked those with high PCR efficiency (results not shown). As expected from amplicon sequencing data, sequencing coverage per locus was very high (>1000×). Most of the selected amplicons were polymorphic, which made them useful for genetic diversity assessment (Figure 3a and Table 6). Since the primers were designed using data from multiple grass and legume genera, it is likely that they work in other genera of such plant families. Evidence in this regard is presented in Figures S2 and S3, which show the results of PCR reactions in eleven

TABLE 7 Median genetic diversity statistics per amplicon including single- and pooled-plant samples for the multispecies amplicons and DNA barcode *rbcLa*

Amplicon	Locus	NKR[a]	SNPs	SNPs/kbp	π[b]	Haplotype allele count	N[c]
rbcLa	*rbcLa*	0.98 (0.95–1.03)	2.00 (2.00–2.00)	3.71 (3.58–3.77)	2.37×10^{-3} (2.01×10^{-3}–2.56×10^{-3})	1.00 (1.00–1.00)	20
AMP3625	O-165	1.04 (0.15–1.17)	4.00 (2.50–6.50)	7.54 (4.71–11.55)	5.31×10^{-3} (3.49×10^{-3}–7.15×10^{-3})	3.00 (1.00–3.00)	12
AMP3794	U-11001396	1.05 (1.00–1.10)	2.00 (2.00–2.00)	3.77 (3.77–3.86)	6.34×10^{-3} (5.56×10^{-3}–6.50×10^{-3})	3.00 (3.00–3.00)	4
AMP3425	U-11004158/O-1519	1.06 (1.00–1.63)	4.00 (1.00–10.00)	6.91 (1.71–17.12)	8.34×10^{-3} (3.69×10^{-3}–3.77×10^{-2})	2.50 (2.00–4.00)	12
AMP3941	O-1211	1.07 (0.86–1.25)	17.00 (12.75–18.25)	32.05 (24.79–34.41)	4.71×10^{-2} (3.47×10^{-2}–5.23×10^{-2})	1.00 (1.00–1.00)	20
AMP3411	O-165	1.30 (0.94–1.70)	2.00 (2.00–2.00)	3.77 (3.77–3.79)	3.62×10^{-3} (3.41×10^{-3}–3.98×10^{-3})	2.00 (1.00–3.00)	8
AMP615	U-11004576	1.53 (0.93–2.25)	1.00 (1.00–1.00)	1.89 (1.83–1.89)	1.33×10^{-3} (1.27×10^{-3}–1.35×10^{-3})	1.50 (1.00–2.00)	8
AMP3711	O-1390	1.63 (1.52–2.10)	4.00 (3.75–7.50)	8.34 (7.54–14.14)	1.10×10^{-2} (7.20×10^{-3}–1.36×10^{-2})	2.50 (1.00–4.00)	8
AMP3376	O-1347	1.76 (1.54–2.71)	14.00 (11.50–16.50)	26.40 (20.65–31.11)	1.87×10^{-2} (1.06×10^{-2}–1.96×10^{-2})	1.00 (1.00–3.00)	12
AMP3532	O-1288	1.81 (0.87–2.03)	17.50 (9.00–29.00)	30.97 (16.97–54.68)	1.88×10^{-2} (8.23×10^{-3}–3.88×10^{-2})	1.00 (1.00–5.00)	12
AMP469	O-1262	2.09 (1.94–2.28)	21.00 (16.00–21.75)	39.60 (33.20–41.01)	3.78×10^{-2} (2.44×10^{-2}–4.38×10^{-2})	9.00 (7.00–10.50)	8
AMP3799	O-1318	2.18 (1.64–3.02)	32.00 (19.50–39.50)	59.12 (36.38–74.48)	4.25×10^{-2} (2.97×10^{-2}–5.27×10^{-2})	12.50 (8.50–17.25)	12
AMP735	O-1520	2.50 (2.48–2.61)	24.50 (20.00–28.25)	46.20 (37.92–53.27)	2.68×10^{-2} (2.55×10^{-2}–4.73×10^{-2})	10.50 (5.00–13.00)	8

[a]NKR, normalized *k*-mer richness.

[b]π, nucleotide diversity.

[c]N, the number of values of each diversity metric considered for the median and interquartile range calculations. For each amplicon, there are four values per species: one from single-plant samples and three from pooled-plant samples.

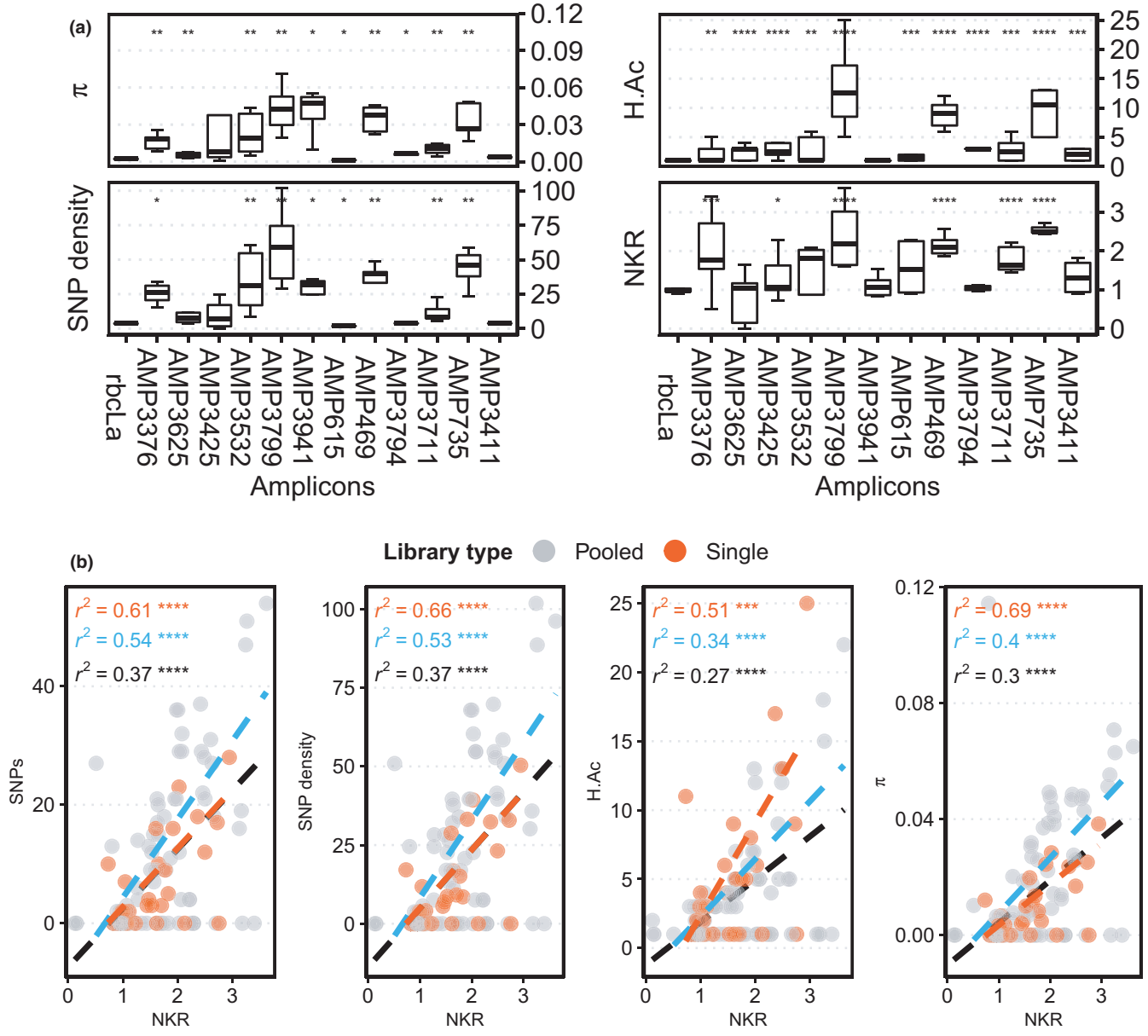

FIGURE 3 Sequence-based genetic diversity in groups of 16 plants per species. (a) SNP-based haplotype allele count (H.Ac), normalized *k*-mer richness (NKR), nucleotide diversity (π) and SNP density (SNPs/kbp) at each locus in groups of 16 plants per species. Amplicons without SNP calls were assumed to have only one haplotype allele in all 16 plants. Wilcoxon test's significance levels are in relation to *rbcLa*. (b) Regression of diversity metrics per amplicon. Red dots show metrics from libraries generated from single-plant sequencing libraries; grey dots show metrics for libraries generated from pooled-plant sequencing libraries. Each dot represents one amplicon in a single- or pooled-plant library. The r^2 value and *p*-value for all amplicons are shown in black; in blue, the same is shown only for amplicons with SNP calls; in red, only for amplicons from single-plant libraries and with SNP calls. Significance levels for *p*-values: .0001 = ****; .001 = ***; .01 = **; 0.05 = *; ns, nonsignificant. The upper and bottom ends of boxplots indicate the third and first quartiles, respectively. The black line inside of boxplots indicates the median value. The upper and lower whiskers in boxplots indicate the maximum (third quartile +1.5 times the interquartile range) and minimum (first quartile - 1.5 times the interquartile range) values, respectively

additional species (nine Poaceae and two Fabaceae species). Most multispecies primers were transferable to the additional Poaceae species, which included genera not used for primer development (e.g., *Agrostis*, *Brachypodium*, *Bromus*, and *Holcus*) and to a lower extent in the two additional legume species (*M. lupulina* and *G. max*). Although this demonstrates a more general applicability of the method for a broad range of grassland species, primer design and amplification conditions may need to be optimized for specific applications involving further related species.

Pooled-plant samples can be an efficient way to analyse large numbers of plants. Such samples can be generated by pooling leaf fragments of similar sizes or genomic DNA at equal concentration. In this study, we pooled ground leaf material in an effort to resemble the ideal scenario in which all plants in a pool are represented by

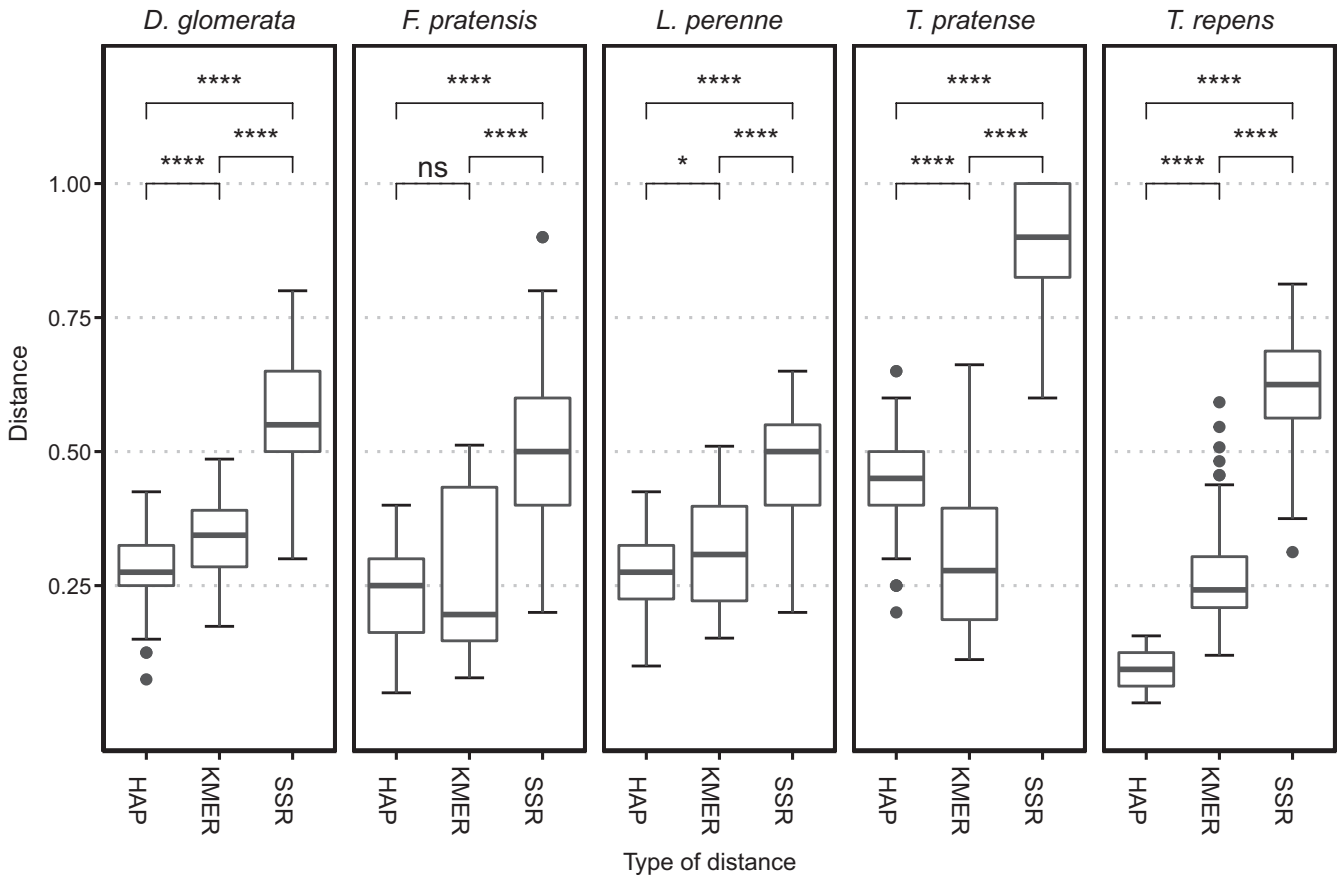

FIGURE 4 Comparison of pairwise distances derived from single-plant samples using sequence- and SSR-based genetic diversity statistics. Prevosti's distance was used for amplicon haplotype-based multilocus genotypes (HAP-MLGs) and SSR-based MLGs (SSR-MLGs). Sourmash-derived distance was used for *k*-mers. Five amplicons per species (four, in case of *T. repens*) were used to calculate distance matrices for HAP-MLGs and *k*-mers. The five most polymorphic SSR per species were used to calculate the SSR-MLGs distance matrix. Samples with >5% missing loci were removed. Wilcoxon test's significance levels are shown for all possible comparisons. Significance levels for *p*-values: .0001 = ****.001 = ***.01 = **, *.05: ns, nonsignificant. The upper and bottom ends of boxplots indicate the third and first quartiles, respectively. The black line inside of boxplots indicates the median value. The upper and lower whiskers in boxplots indicate the maximum (third quartile +1.5 times the interquartile range) and minimum (first quartile - 1.5 times the interquartile range) values, respectively. Dots on top and at the bottom of boxplots indicate outliers

equal amounts of dry biomass. In general, genetic diversity in single- and pooled-plant libraries were mostly similar, with some variation due to the diversity metric and the amplicon (median CV = 13.38%). These results can serve as a reference for further studies or monitoring efforts for grassland conservation that use other kinds of pooled-plant samples (e.g., leaf fragment pools) for genetic diversity assessments.

Among diversity metrics, NKR had the lowest variation due to sample pooling. In pooled-plant libraries, SNP-based metrics (i.e., SNP count, SNP density and π) had higher variations, probably due to the differences between the SNP calling software. PoPoolation produced more SNP calls in the pooled-plant libraries compared to the SNP calls of BCFtools mpileup in single-plant libraries. This is to be expected because PoPoolation false SNP discovery rates are known to increase with depth (Raineri et al., 2012). In the case of SNP-based haplotypes, the low variation haplotype counts stems from the way such haplotypes were generated in pooled-samples (i.e., only the haplotypes that were present in the single-plant libraries were considered).

Among amplicons, those with low variability due to sample pooling (median CV <10%) were *rbcLa*, AMP3411, AMP615, AMP3941 and AMP3794, which are also among the least diverse amplicons (Table 7). In the rest of the amplicons, the moderate levels of variability of diversity metrics (median CV between 12.54% to 33.19%) is also likely a result of differences in SNP calling software. However, correlations among amplicon diversity metrics were better than those observed in the discovery phase of this study (Figure 3b). This is probably because indels in our multispecies amplicons had less size variability than the 611 loci captured in the discovery phase (Figure S5), which in turn reduced the discrepancy between NKR and SNP-based metrics.

The effect of *k*-mer depth filtering on total NKR (assessed in single-plant libraries) reflected the sequence diversity of each amplicon (Figure S4). For the highly polymorphic amplicons, NKR

TABLE 8 Median ratios between SSR- and amplicon sequencing-based multilocus pairwise distances for single-plant samples

Family	Species name	Amplicons	SSR	SSR:*k*-mer[a]	SSR:HAP[b]	*k*-mer:HAP[c]	*N*[d]
Fabaceae	*Trifolium pratense*	AMP3411	TP45	3.32 (2.12–4.98)	2.00 (1.78–2.25)	0.68 (0.39–0.93)	66
		AMP3711	TP46				
		AMP615	TP50				
		AMP735	TP34				
		AMP469	TP10				
	Trifolium repens	AMP3411	TR04	2.42 (1.88–3.12)	7.33 (5.50–10.00)	3.31 (2.16–4.62)	115
		AMP3711	TR09				
		AMP615	TR10				
		AMP735	TR13				
Poaceae	*Dactylis glomerata*	AMP3376	DG004	1.68 (1.33–2.10)	2.16 (1.67–2.62)	1.23 (1.04–1.51)	120
		AMP3425	DG006				
		AMP3532	DG010				
		AMP3625	DG011				
		AMP3799	DG012				
	Festuca pratensis	AMP3376	FA49	2.62 (1.17–3.44)	2.00 (1.62–2.67)	0.93 (0.63–1.49)	66
		AMP3425	LM26				
		AMP3532	LM29				
		AMP3625	LP27				
		AMP3799	LP47				
	Lolium perenne	AMP3376	LP07	1.68 (1.22–2.31)	1.85 (1.48–2.20)	1.21 (0.80–1.52)	66
		AMP3425	LP20				
		AMP3532	LP23				
		AMP3625	LP13				
		AMP3799	LP16				
	Total			2.07 (1.45–2.93)	2.36 (1.80–4.50)	1.29 (0.88–2.20)	433

[a]SSR:*k*-mer, median of the ratios between SSR- and *k*-mer-based multi-locus pairwise distances (MLPD).
[b]SSR:HAP, median of the ratios between SSR- and haplotype-based MLPD.
[c]*k*-mer:HAP, median of the ratios between *k*-mer- and haplotype-based MLPD.
[d]*N*, number of comparisons.

values decreased steadily with increasing *k*-mer depth cutoff values. In the less polymorphic amplicons (i.e., *rbcLa*, AMP3625, AMP3794, and AMP3941), the sequencing read output is spread across a smaller set of unique *k*-mers, which results in a more even sequencing depth. That the NKR values of *rbcLa* were close to 1 across a long range of *k*-mer depth cutoff values indicates that PCR and sequencing errors did not greatly affect NKR estimates. This highlights the robustness of alignment-free methods for genetic diversity studies.

The average genetic diversity per species (mean $\pi = 8.94 \times 10^{-3}$; mean SNP density = 12.04 SNPs/kbp; Table 6) was higher than the diversity reported in another targeted sequencing study, which analysed nine genes putatively involved in flowering (overall $\pi = 7.90 \times 10^{-3}$; overall SNP density = 7.87 SNPs/kbp; Fiil et al., 2011). That study was carried out using 20 *L. perenne* genotypes from different European sources (i.e., the "*Lolium* Test Set" or LTS). In contrast, the overall genetic diversity per species was lower than those reported for the LTS in 11 pathogen resistance genes ($\pi = 3.14 \times 10^{-2}$; total SNP density ≈ 100 SNPs/kbp; Xing et al., 2007), which are subject to constant selection pressure, co-evolve alongside pathogens and, therefore, are highly variable. The moderate levels of genetic diversity of our multispecies amplicons compared to resistance genes is likely because our target loci are not subject to similar co-evolutionary dynamics.

To benchmark the diversity estimations from multispecies amplicons, pairwise genetic distances based on *k*-mers and SNPs were compared to pairwise distances produced with SSR, a highly polymorphic and multi-allelic marker system. Multispecies amplicons underestimated SSR-based genetic distances by ~50% (Figure 4 and Table 8). This indicates that more multispecies amplicons are needed to improve their discriminatory power. Theoretically, the discriminatory power of a few tens of SSR is equivalent to roughly a hundred neutral SNPs (Kalinowski, 2002). However, empirical evidence from an outbreeding plant species shows that the number of SNPs needed to approximate genome-wide diversity estimations is higher than theoretical predictions (Fischer et al.,

2017). Only low correlation was observed between multispecies amplicon sequencing- and SSR-based pairwise distances (data not shown), indicating that multispecies amplicons and SSR interrogate different parts of the genome of each species, which results in different genetic distance estimations between any given pair of genotypes. This highlights the large diversity present in the genomes of outbreeding forage species.

Although multispecies amplicon sequencing underestimated genetic diversity when compared to SSR, some of its features can make it an attractive method for large-scale assessments of genetic diversity in outcrossing grass and legume species. In contrast to SSR analysis, multispecies amplicon sequencing does not require expensive primer modifications and the primers are transferable across many species. The costs of high-throughput sequencing for amplicon sequencing are balanced by the possibility of processing multiple plants in a single run. This is particularly important for outcrossing species, like forage grasses and legumes, which require large sample numbers to accurately assess their genetic diversity (Kölliker et al., 2009). Compared to other sequencing approaches like GBS and RAD-seq, the output of amplicon sequencing is highly enriched in informative data, making it more resource efficient. Furthermore, amplicon sequencing data greatly simplify bioinformatic analyses (Sato et al., 2019).

In conclusion, we showed that multispecies amplicon sequencing is useful for genetic diversity assessments in outbreeding forage crop species. The approach uses inexpensive, unmodified primers and other conventional PCR reagents. The method is transferable across economically relevant grassland plant species of the Poaceae and Fabaceae families, which can further diminish the costs and labor needed for large-scale, multispecies assessments. We also showed that genetic diversity estimates based on multispecies amplicon sequencing of pooled-plant material are similar to those of single-plant material, which further reduces costs per sample. The shortcomings of multispecies amplicon sequencing in contrast to SSR analysis are compensated by its throughput and its resource- and time-saving features. This is particularly relevant for large-scale and constant monitoring efforts. Further improvements to the method can increase its resolution (i.e., the number of SNPs that are assessed, which in turn can improve the estimations of genetic distance) and throughput (i.e., the number of samples that are processed simultaneously). Such improvements can be achieved by increasing the number of amplicons analysed, by varying amplicon sizes, by varying PCR conditions (e.g., multiplex PCR can reduce the processing time needed for each sample; Veeckman et al., 2019), or by adapting sequencing approaches (e.g., pool sequencing can be used to characterize larger plant populations).

The tools presented in this study provide the basis for cost-effective, multispecies genetic diversity assessments in grassland plants. In our view, these results represent a promising starting point for further improvements and adaptations. As awareness increases around the ecological significance of plant genetic diversity and as widespread monitoring of genetic diversity gains traction (Hoban et al., 2020; Laikre et al., 2020; Pärli et al., 2021), such multispecies approaches can be valuable additions to the toolset of genetic diversity monitoring efforts.

ACKNOWLEDGEMENTS

We thank Franco Widmer and Christoph Grieder from Agroscope for giving us access to laboratory space and greenhouses, respectively. Steven Yates (ETH Zurich) provided the list of SCOGs used for bait design. Thanks to Verena Knorst (ETH Zurich), Olivier Huguenin-Elie (Agroscope), and Roxane Muller (Agroscope) for giving us access to the additional species used to test the multispecies primers. Thanks to Damian Käch (ETH Zurich), who performed the SSR genotyping of this study as part of his Bachelor thesis. Library preparations and sequencing were performed at the Genetic Diversity Centre of ETH Zurich. We thank M. Hardegger and C. Kägi (Swiss Federal Office of Agriculture) for their valuable feedback regarding the design of the study. Open Access Funding provided by Eidgenossische Technische Hochschule Zurich.

CONFLICT OF INTEREST

The authors declare no conflict of interest. The funders had no role in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

AUTHOR CONTRIBUTIONS

Roland Kölliker and Bruno Studer conceived the study and provided insights on experimental design and data analysis. Miguel Loera-Sánchez performed the laboratory and bioinformatic analyses. All authors read and approved the final manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data, for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5061/dryad.gb5mkkwqw. The Python script "haplotyper.2.py" to generate raw SNP-based haplotypes from amplicon sequencing data can be found on GitHub (https://github.com/mloera/gendiv/commit/9f4d4a3606c8d192d8d7dd2680e5a4728994b7d6). A snapshot of it can also be found in the Dryad repository mentioned earlier.

DATA AVAILABILITY STATEMENT

Raw reads, bait sequences, loci assemblies, and the Python script "haplotyper.2.py" are available at: https://doi.org/10.5061/dryad.gb5mkkwqw

ORCID

Miguel Loera-Sánchez https://orcid.org/0000-0002-1395-9849
Bruno Studer https://orcid.org/0000-0001-8795-0719
Roland Kölliker https://orcid.org/0000-0003-1959-0402

REFERENCES

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., & Marshall, A. H. (2015). Achievements and challenges in improving temperate

perennial forage legumes. *Critical Reviews in Plant Sciences*, *34*, 327–380. https://doi.org/10.1080/07352689.2014.898462

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bengtsson, J., Bullock, J. M., Egoh, B., Everson, C., Everson, T., O'Connor, T., O'Farrell, P. J., Smith, H. G., & Lindborg, R. (2019). Grasslands-more important for ecosystem services than you might think. *Ecosphere*, *10*(2), e02582. https://doi.org/10.1002/ecs2.2582

Broad Institute (2019). *Picard Toolkit*, *Github Repository*. Retrieved from http://broadinstitute.github.io/picard/

Brown, C. T., & Irber, L. (2016). sourmash: A library for MinHash sketching of DNA. *The Journal of Open Source Software*, *1*(5), 27. https://doi.org/10.21105/joss.00027

Bushnell, B. (2014). *BBMap*. Retrieved from https://sourceforge.net/projects/bbmap/

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One*, *12*(10), 1–15. https://doi.org/10.1371/journal.pone.0185056

Caldu-Primo, J. L., Mastretta-Yanes, A., Wegier, A., & Piñero, D. (2017). Finding a needle in a haystack: Distinguishing Mexican maize landraces using a small number of SNPs. *Frontiers in Genetics*, *8*, 1–12. https://doi.org/10.3389/fgene.2017.00045

Carroll, E. L., Bruford, M. W., DeWoody, J. A., Leroy, G., Strand, A., Waits, L., & Wang, J. (2018). Genetic and genomic monitoring with minimally invasive sampling methods. *Evolutionary Applications*, *11*(7), 1094–1119. https://doi.org/10.1111/eva.12600

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cheng, Y., Zhou, K., Humphreys, M. W., Harper, J. A., Ma, X., Zhang, X., Yan, H., & Huang, L. (2016). Phylogenetic relationships in the Festuca-Lolium complex (Loliinae; Poaceae): New insights from chloroplast sequences. *Frontiers in Ecology and Evolution*, *4*, 1–12. https://doi.org/10.3389/fevo.2016.00089

Collins, R. P., Helgadóttir, Á., Frankow-Lindberg, B. E., Skøt, L., Jones, C., & Skøt, K. P. (2012). Temporal changes in population genetic diversity and structure in red and white clover grown in three contrasting environments in northern Europe. *Annals of Botany*, *110*(6), 1341–1350. https://doi.org/10.1093/aob/mcs058

Cropano, C., Place, I., Manzanares, C., Do Canto, J., Lübberstedt, T., Studer, B., & Thorogood, D. (2021). Characterization and practical use of self-compatibility in outcrossing grass species. *Annals of Botany*, *127*(7), 841–852. https://doi.org/10.1093/aob/mcab043

Cuyeu, R., Rosso, B., Pagano, E., Soto, G., Fox, R., & Ayub, N. D. (2013). Genetic diversity in a world germplasm collection of tall fescue. *Genetics and Molecular Biology*, *36*(2), 237–242. https://doi.org/10.1590/S1415-47572013005000021

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., Rognli, O. A., Jones, C., Swain, M., Geurts, R., Lang, C., Mayer, K. F. X., Rössner, S., Yates, S., Webb, K. J., Donnison, I. S., Oldroyd, G. E. D., Wing, R. A., Caccamo, M., … Skøt, L. (2015). Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Scientific Reports*, *5*(1), 17394. https://doi.org/10.1038/srep17394

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. https://doi.org/10.1093/bioinformatics/btv646

Fiil, A., Lenk, I., Petersen, K., Jensen, C. S., Nielsen, K. K., Schejbel, B., Andersen, J. R., & Lübberstedt, T. (2011). Nucleotide diversity and linkage disequilibrium of nine genes with putative effects on flowering time in perennial ryegrass (*Lolium perenne* L.). *Plant Science*, *180*(2), 228–237. https://doi.org/10.1016/j.plantsci.2010.08.015

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., Holderegger, R., & Widmer, A. (2017). Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, *18*(1), 1–15. https://doi.org/10.1186/s12864-016-3459-7

Hadincová, V., Skálová, H., Münzbergová, Z., & Fischer, M. (2020). Genotypic diversity and genotype identity of resident species drive community composition. *Journal of Plant Ecology*, *13*(2), 224–232. https://doi.org/10.1093/jpe/rtaa004

Hamrick, J. L., & Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1345), 1291–1298. https://doi.org/10.1098/rstb.1996.0112

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus Restriction Site Associated DNA sequencing for shallow systematics. *Systematic Biology*, *65*(5), 910–924. https://doi.org/10.1093/sysbio/syw036

Hoban, S., Bruford, M., D'Urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, *248*, 108654. https://doi.org/10.1016/j.biocon.2020.108654

Hodkinson, T. R., Perdereau, A., Klaas, M., Cormican, P., & Barth, S. (2019). Genotyping by sequencing and plastome analysis finds high genetic variability and geographical structure in *Dactylis glomerata* L. in Northwest Europe Despite Lack of Ploidy Variation. *Agronomy*, *9*(7), 342. https://doi.org/10.3390/agronomy9070342

Huber, R., & Finger, R. (2020). A meta-analysis of the willingness to pay for cultural services from grasslands in Europe. *Journal of Agricultural Economics*, *71*(2), 357–383. https://doi.org/10.1111/1477-9552.12361

Hysom, D. A., Naraghi-Arani, P., Elsheikh, M., Carrillo, A. C., Williams, P. L., & Gardner, S. N. (2012). Skip the alignment: Degenerate, multiplex primer and probe design using *k*-mer matching instead of alignments. *PLoS One*, *7*(4), e34560. https://doi.org/10.1371/journal.pone.0034560

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., & Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467. https://doi.org/10.1038/nature06148

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G.-S., Baker, W. J., & Wickett, N. J. (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. https://doi.org/10.1093/sysbio/syy086

Kalinowski, S. T. (2002). How many alleles per locus should be used to estimate genetic distances? *Heredity*, *88*(1), 62–65. https://doi.org/10.1038/sj.hdy.6800009

Kamvar, Z. N., Tabima, J. F., & Gr̈unwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2014*(1), 1–14. https://doi.org/10.7717/peerj.281

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt,

B., Sakai, H., Lee, S. S., Kim, J., Numa, H., … Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, *6*(1), 4. https://doi.org/10.1186/1939-8433-6-4

Knorst, V., Byrne, S., Yates, S., Asp, T., Widmer, F., Studer, B., & Kölliker, R. (2019). Pooled DNA sequencing to identify SNPs associated with a major QTL for bacterial wilt resistance in Italian ryegrass (*Lolium multiflorum* Lam.). *Theoretical and Applied Genetics*, *132*(4), 947–958. https://doi.org/10.1007/s00122-018-3250-z

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kölliker, R., Boller, B., Majidi, M., Peter-Schmid, M. K. I., Bassin, S., & Widmer, F. (2009). Characterization and utilization of genetic resources for improvement and management of grassland species. In G. Yamada, & T. Spangenberg (Ed.), *Molecular Breeding of Forage and Turf* (pp. 55–70). Springer New York. doi: https://doi.org/10.1007/978-0-387-79144-9_5

Laikre, L., Hoban, S., Bruford, M. W., Segelbacher, G., Allendorf, F. W., Gajardo, G., Rodríguez, A. G., Hedrick, P. W., Heuertz, M., Hohenlohe, P. A., Jaffé, R., Johannesson, K., Liggins, L., MacDonald, A. J., OrozcoterWengel, P., Reusch, T. B. H., Rodríguez-Correa, H., Russo, I.-R., Ryman, N., & Vernesi, C. (2020). Post-2020 goals overlook genetic diversity. *Science*, *367*(6482), 1083.2–1085. https://doi.org/10.1126/science.abb2748

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The *Arabidopsis* information resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*(D1), D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Last, L., Widmer, F., Fjellstad, W., Stoyanova, S., & Kölliker, R. (2013). Genetic diversity of natural orchardgrass (*Dactylis glomerata* L.) populations in three regions in Europe. *BMC Genetics*, *14*(1), 102. https://doi.org/10.1186/1471-2156-14-102

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J. Jr, & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Liu, S., Feuerstein, U., Luesink, W., Schulze, S., Asp, T., Studer, B., Becker, H. C., & Dehmer, K. J. (2018). DArT, SNP, and SSR analyses of genetic diversity in Lolium perenne L. using bulk sampling. *BMC Genetics*, *19*(1), 10. https://doi.org/10.1186/s12863-017-0589-0

Loera-Sánchez, M., Studer, B., & Kölliker, R. (2019). DNA-based assessment of genetic diversity in grassland plant species: Challenges, approaches, and applications. *Agronomy*, *9*(12), 881. https://doi.org/10.3390/agronomy9120881

Loera-Sánchez, M., Studer, B., & Kölliker, R. (2020). DNA barcode *trnH-psbA* is a promising candidate for efficient identification of forage legumes and grasses. *BMC Research Notes*, *13*(1), 1–6. https://doi.org/10.1186/s13104-020-4897-5

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L. E., Wang, Y. I., McGuire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van, S. L., Yuan, L., Wang, Z., Xia, Z., … Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*, *551*(7681), 498–502. https://doi.org/10.1038/nature24486

Malhotra, R., Prabhakara, S., Poss, M., & Acharya, R. (2013). Estimating viral haplotypes in a population using *k*-mer counting. In A. Ngom, E. Formenti, J.-K. Hao, X.-M. Zhao, & T. van Laarhoven (Eds.), *Pattern Recognition in Bioinformatics* (pp. 265–276). Springer Berlin Heidelberg.

Malyshev, A. V., Arfin Khan, M. A. S., Beierkuhnlein, C., Steinbauer, M. J., Henry, H. A. L., Jentsch, A., Dengler, J., Willner, E., & Kreyling, J. (2016). Plant responses to climatic extremes: Within-species variation equals among-species variation. *Global Change Biology*, *22*(1), 449–464. https://doi.org/10.1111/gcb.13114

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, *33*(7), 1875–1886. https://doi.org/10.1093/molbev/msw056

Meilhac, J., Durand, J.-L., Beguier, V., & Litrico, I. (2019). Increasing the benefits of species diversity in multispecies temporary grasslands by increasing within-species diversity. *Annals of Botany*, *123*(5), 891–900. https://doi.org/10.1093/aob/mcy227

Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Iii, D. L., Cornejo, O., Findley, S. D., Zheng, P., Utro, F., Royaert, S., Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B. E., Stack, J. C., Feltus, F. A., Mustiga, G. M., Amores, F., … Kuhn, D. N. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology*, *14*(6), r53. https://doi.org/10.1186/gb-2013-14-6-r53

Pärli, R., Lieberherr, E., Holderegger, R., Gugerli, F., Widmer, A., & Fischer, M. C. (2021). Developing a monitoring program of genetic diversity: What do stakeholders say? *Conservation Genetics*, *22*(5), 673–684. https://doi.org/10.1007/s10592-021-01379-6

Pérez-Cantalapiedra, C., Contreras-Moreira, B., Casas Cendoya, A. M., & Igartua Arregui, E. (2018). *Unmasking new intra-species diversity through k-mer count analysis. EUCARPIA Cereal Section/IWW2 Meetings (Polydome - Clermont-Ferrand, France)*. Retrieved from https://digital.csic.es/handle/10261/162254

Prieto, I., Violle, C., Barre, P., Durand, J.-L., Ghesquiere, M., & Litrico, I. (2015). Complementary effects of species and genetic diversity on productivity and stability of sown grasslands. *Nature Plants*, *1*(4), 15033. https://doi.org/10.1038/nplants.2015.33

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*(1), 239. https://doi.org/10.1186/1471-2105-13-239

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Reynolds, S. G. (2005). Chapter 1 Introduction. In J. M. Suttie, S. G. Reynolds, & C. Batello (Eds.), *Grasslands of the World*. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/y8344e00.htm

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. Boston, MA. Retrieved from http://www.rstudio.com/

Ruttink, T., Haegeman, A., van Parijs, F., Van Glabeke, S., Muylle, H., Byrne, S., Asp, T., & Roldán-Ruiz, I. (2015). Genetic diversity in candidate genes for developmental traits and cell wall characteristics in perennial ryegrass (*Lolium perenne*). In H. Budak, & G. Spangenberg (Eds.), *Molecular Breeding of Forage and Turf* (pp. 93–109). Springer International Publishing. https://doi.org/10.1007/978-3-319-08714-6_9

Sato, M., Hosoya, S., Yoshikawa, S., Ohki, S., Kobayashi, Y., Itou, T., & Kikuchi, K. (2019). A highly flexible and repeatable genotyping method for aquaculture studies based on target amplicon sequencing using next-generation sequencing technology. *Scientific Reports*, *9*(1), 6904. https://doi.org/10.1038/s41598-019-43336-x

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., Fujishiro, T., Katoh, M., Kohara, M., Kishida, Y., Minami, C., Nakayama, S., Nakazaki, N., Shimizu, Y., Shinpo, S., … Tabata, S. (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Research*, *15*(4), 227–239. https://doi.org/10.1093/dnares/dsn008

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–183. https://doi.org/10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*(7), 707–713. https://doi.org/10.1038/ng.3008

Seqtk-1.3 (2018). Retrieved from https://github.com/lh3/seqtk

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F. X., Schwartz, D. C., & Town, C. D. (2014). An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics*, *15*(1), 1–14. https://doi.org/10.1186/1471-2164-15-312

The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

Tkach, N., Schneider, J., Döring, E., Wölk, A., Hochbach, A., Nissen, J., Winterfeld, G., Meyer, S., Gabriel, J., Hoffmann, M. H., & Röser, M. (2020). Phylogenetic lineages and the role of hybridization as driving force of evolution in grass supertribe Poodae. *Taxon*, *69*(2), 234–277. https://doi.org/10.1002/tax.12204

Vatanparast, M., Powell, A., Doyle, J. J., & Egan, A. N. (2018). Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences*, *6*(3), e1036. https://doi.org/10.1002/aps3.1036

Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., Van Parijs, F. R. D., Byrne, S. L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., & Ruttink, T. (2019). Overcoming challenges in variant calling: Exploring sequence diversity in candidate genes for plant development in perennial ryegrass (*Lolium perenne*). *DNA Research*, *26*(1), 1–12. https://doi.org/10.1093/dnares/dsy033

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., & Baxter, I. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*(7282), 763–768. https://doi.org/10.1038/nature08747

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5), 534–540. https://doi.org/10.1038/s41588-020-0612-7

Xing, Y., Frei, U., Schejbel, B., Asp, T., & Lübberstedt, T. (2007). Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. *BMC Plant Biology*, *7*, 1–12. https://doi.org/10.1186/1471-2229-7-43

Zhao, Y., Liu, Z., & Wu, J. (2020). Grassland ecosystem services: A systematic review of research advances and future directions. *Landscape Ecology*, *35*(4), 793–814. https://doi.org/10.1007/s10980-020-00980-3

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, *18*(1), 186. https://doi.org/10.1186/s13059-017-1319-7

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Loera-Sánchez, M., Studer, B., & Kölliker, R. (2022). A multispecies amplicon sequencing approach for genetic diversity assessments in grassland plant species. *Molecular Ecology Resources*, 22, 1725–1745. https://doi.org/10.1111/1755-0998.13577