Received: 26 June 2018 | Revised: 28 August 2018 | Accepted: 30 August 2018

DOI: 10.1002/ece3.4533

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring

Petra Šarhanová[1*] | Simon Pfanzelt[2*] | Ronny Brandt[1] | Axel Himmelbach[1] | Frank R. Blattner[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]Institute of Biology and Environmental Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Correspondence
Petra Šarhanová, Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria.
Email: sarhanova@gmail.com

Present Address
Petra Šarhanová, Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

Simon Pfanzelt, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Ronny Brandt, Max Planck Genome Centre Cologne, Cologne, Germany

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: AL632/7-1 and BL462/11; Dr. Karl Wamsler Stiftung; Deutscher Akademischer Austauschdienst, Grant/Award Number: D/10/49464

Abstract

Microsatellites (or simple sequence repeats, SSR) are widely used markers in population genetics. Traditionally, genotyping was and still is carried out through recording fragment length. Now, next-generation sequencing (NGS) makes it easy to obtain also sequence information for the loci of interest. This avoids misinterpretations that otherwise could arise due to size homoplasy. Here, an NGS strategy is described that allows to genotype hundreds of individuals at many custom-designed SSR loci simultaneously, combining multiplex PCR, barcoding, and Illumina sequencing. We created three different datasets for which alleles were coded according to (a) length of the repetitive region, (b) total fragment length, and (c) sequence identity, in order to evaluate the eventual benefits from having sequence data at hand, not only fragment length data. For each dataset, genetic diversity statistics, as well as F_{ST} and R_{ST} values, were calculated. The number of alleles per locus, as well as observed and expected heterozygosity, was highest in the sequence identity dataset, because of single-nucleotide polymorphisms and insertions/deletions in the flanking regions of the SSR motif. Size homoplasy was found to be very common, amounting to 44.7%–63.5% (mean over all loci) in the three study species. Thus, the information obtained by next-generation sequencing offers a better resolution than the traditional way of SSR genotyping and allows for more accurate evolutionary interpretations.

KEYWORDS
genotyping, microsatellites, multiplex PCR, next-generation sequencing, single-nucleotide polymorphism, size homoplasy

1 | INTRODUCTION

Microsatellites (short tandem repeats, STR, or simple sequence repeats, SSR) are widely used markers in population genetics due to their ubiquitous occurrence in the nuclear and organellar genomes, high levels of polymorphism, and codominant character. Traditionally, allele information is extracted through recording fragment length, which serves as a proxy for the number of repetitive units and is used to calculate genetic and evolutionary distance between individuals. Nonetheless, single-nucleotide polymorphisms

[a]These authors contributed equally to this work.

(SNPs) or insertions/deletions (indel) polymorphisms in the nucleotide sequence of that fragment, either within the repetitive array or in the flanking regions (FR), remain undetected by length assessment alone. Moreover, indels in the flanking regions might be incorrectly confounded with size mutations of the SSR. Thus, using only length information, SSR alleles may appear identical in state (i.e., length/size), but actually they are not necessarily identical by descent in case of convergent mutation(s) to the same size ("size homoplasy", Estoup, Jarne, & Cornuet, 2002) or variability only in sequence but not in size. Estoup et al. (2002) used the term "molecularly accessible size homoplasy" to refer to the fraction of homoplasy that can be resolved by sequencing, which is only a subset of the size homoplasy that actually occurs at microsatellite loci. Still, sequencing cannot resolve homoplasy that arises from the convergence of two alleles to the same sequence and length.

As a consequence, the traditional assessment of fragment length may lead to underestimating genetic variability, inaccurate results, or even wrong evolutionary interpretations (Barthe et al., 2012; Blankenship, May, & Hedgecock, 2002; Peakall, Gilmore, Keys, Morgante, & Rafalski, 1998; summarized in Germain-Aubrey, Nelson, Soltis, Soltis, & Gitzendanner, 2016). To overcome such errors, information about the nucleotide sequence of each allele is needed. While using NGS data from different sequencing platforms for SSR marker development in non-model plant species is now a common practice (Weising, Wöhrmann, & Huettel, 2015), NGS is very rarely used for SSR scoring. In order to tackle the homoplasy problem and assess FR variation, some authors combined cloning or single-strand conformation polymorphism and sequencing (e.g., Germain-Aubrey et al., 2016; Lia, Bracco, Gottlieb, Poggio, & Confalonieri, 2007; Ortí, Pearse, & Avise, 1997; Šarhanová et al., 2017), but these methods are costly, time-consuming, and not easily applicable for polyploids. There are first forays among animals (Bradbury et al., 2018; De Barba et al., 2017; Vartia et al., 2016), but comparisons between traditionally scored fragment length data and the information obtained from sequencing are still missing.

Mutations in the SSR region (predominantly changes in the number of repeats) and FR (indels and SNPs) evolve at different rates: the fast-evolving repetitive region shows a mutation rate of about 10^{-6} to 10^{-2} per locus per generation (Schlötterer, 2000), whereas base substitutions occur at a much slower rate (depending on the genome size of the organism; Lynch, 2010), for example, in *Arabidopsis thaliana* at a rate of 7×10^{-9} mutations per nucleotide position per generation (Ossowski et al., 2010). The combined information from both regions can thus be used for the inference of evolutionary events at different timescales or at least indicate possible mutational saturation of the SSR region or its convergent evolution to the same size.

Here, an NGS strategy is described which allows to genotype hundreds of individuals at several, custom-designed SSR loci simultaneously, using multiplex PCR and barcoded primers to separate individual-specific Illumina sequence reads. Our objectives were (a) to generate nucleotide sequence data of several non-model plant species, for which prior genomic data did not exist, from both the SSR and the flanking regions, (b) to record the length of the repetitive region, as well as SNP and indel variation within the SSR and the FR, (3) to estimate the amount of molecularly accessible size homoplasy of each locus, and (4) to compare the degree of genetic variability between different datasets based on the number of repeat units, fragment length, and sequence identity.

2 | MATERIALS AND METHODS

2.1 | Study species

The method described here is based on three angiosperm species from southern South America: *Donatia fascicularis* (Stylidiaceae), *Mulguraea tridens* (Verbenaceae), and *Oreobolus obtusangulus* (Cyperaceae) (Table 1, Figure 1). In total, 859 individuals were genotyped at 58 nuclear SSR loci and statistically analyzed. For detailed information about two of the studied species (*D. fascicularis* and *O. obtusangulus*) and population sampling see results published elsewhere (Pfanzelt, Albach, & von Hagen, 2017; S. Pfanzelt, P. Šarhanová, D. C. Albach, & K. B. vonHagen, under review). The work is a part of a study that includes five further angiosperm species of a wide phylogenetic range, different ploidy levels, genome size, and reproductive system (*Astelia pumila*, Asteliaceae; *Berberis empetrifolia*, Berberidaceae; *Chuquiraga aurea*, Asteraceae; *Guaiacum sanctum*, Zygophyllaceae; *Rubus ulmifolius* agg., Rosaceae; in total, about 2,000 individuals were scored at 132 SSR and 3 chloroplast loci), although the data of these latter species are not included in the present study.

2.2 | SSR identification, primer design, and testing

Initial detection of SSR loci relied on assembled Illumina paired-end sequencing reads of cDNA transcripts, which in turn stemmed from RNA extracted from fresh plant or RNA-later (Qiagen) treated material using the RNeasy Mini Kit (Qiagen). RNA extraction followed a standard protocol and included subsequent DNA digestion and an RNA re-precipitation step. Libraries were prepared and sequenced on an Illumina HiSeq 2000 platform according to the manufacturer's instructions, using a TruSeq RNA Library Prep Kit v2 and 10% of the lane per library. De-novo assembly of RNA-Seq output data was done in Geneious 6.0.4 (Kearse et al., 2012) with medium sensitivity settings.

The resulting contigs from the de-novo assembly were screened for SSRs using Phobos 3.3.12 (as a plugin in Geneious; Mayer, 2010). The predefined repeat unit length was 3–6 bases (to avoid frequent PCR stuttering, SSRs with dinucleotide repeats were not considered; Miller & Yuan, 1997). The minimum length of the microsatellite region was at least 21 bp. One of the primers from a given primer pair was always selected to be close to the SSR to ensure that during SSR analysis, a single NGS read contains the entire repetitive region, thus assuring correct merging of paired reads. The target length of the amplicon was up to 450 bp.

Ninety-six primer pairs per species were designed employing Primer3 (as a plugin in Geneious; Rozen & Skaletsky, 2000) with the following settings: product size 250–400 bp; primer size = 18–22–30 bp

TABLE 1 List of study species

Species	Family, order	Ploidy level	Chromosome number	1C genome size (pg)	N SSR loci	N genotyped individuals	N analyzed individuals[a]
Donatia fascicularis J.R.Forst. & G.Forst.	Stylidiaceae, Asterales	4×, effectively 2×	2n = 48 (Moore 1983)	3.22	20	384	328
Mulguraea tridens (Lag.) N.O'Leary & P. Peralta	Verbenaceae, Lamiales	4×	Unknown	0.86	18	88	86
Oreobolus obtusangulus Gaudich.	Cyperaceae, Cyperales	4×, effectively 2×	2n = 48 (Moore 1967)	0.64	20	384	360

Note. Taxonomic and genome information on the three studied species with number (*N*) of individuals and loci used. Ploidy levels of *D. fascicularis* and *O. obtusangulus* were inferred from chromosome number (4×) and from number of alleles per locus (2×); ploidy level of *M. tridens* was estimated from number of alleles per locus and through flow cytometry, using genome size as a proxy.

[a]Individuals with missing data in more than 30% of the loci were excluded from statistical analyses.

FIGURE 1 Photographs of study species

(min–optimal–max); melting temperature (*Tm*) = 68–70–72°C; GC content = 40%–50%–60%; maximum T_m difference = 2°C; remaining settings as default. Of these originally 96 primer pairs, 68 successfully produced amplicons for *D. fascicularis*, 51 for *M. tridens*, and 61 for *O. obtusangulus*. Amplicons of four individuals per species (different species and all loci were pooled prior library preparation generating four pools) were sequenced on an Illumina MiSeq platform (2 × 250 bp paired-end, using MiSeq Reagent Kit v2 and 25% of a lane per library), following Meyer and Kircher (2010) and omitting fragmentation. Sequencing adapters were removed using CUTADAPT (Martin, 2011; minimum length 150, quality 15). Data were checked for read pairs in READTRIMMCHECKER (Beier, 2016). Assembly was done in CLC ASSEMBLY CELL 4.2.0 (using the clc_overlap_reads command, minimum overlap of 30), and contigs were imported into GENEIOUS. Based on intraspecific variability, up to 20 SSR loci (hereafter called target SSRs) for each species were selected.

2.3 | Barcoding of primers and multiplex PCR

DNA was extracted from silica-dried leaf material using the DNAeasy Plant Mini Kit (Qiagen) or the innuPREP Plant DNA Kit (Analytik Jena) following the respective manufacturer´s protocol. Individuals were assigned to sample sets with 96 individuals each (four sample sets in *D. fascicularis* and *O. obtusangulus*, one in *M. tridens*). To

allow for multiplexing during library construction, ten-nucleotide barcode sequences, specific for each SSR locus and each sample set of 96 individuals, were appended to the 5′-ends of both primers (forward and reverse) of the target SSRs (Supporting Information Appendix S1). In total, 2 (forward and reverse) × 20 (loci) × number of sets (1 or 4) primers per species have been ordered. This double-tagging allowed parallel sequencing of several conspecific samples through pooling after PCR (for a graphical description of the method see Figure 2). MULTIPLX 2.1 (Kaplinski, Andreson, Puurand, & Remm, 2005) was used to define primer groups within each of the sample sets in order to identify optimal primer compatibility and to avoid undesired primer pairing. The grouping was done for each species and barcoded primer set separately. The software was run with the default settings, and "Calculating scores" was set to "primer-primer any". Each multiplex group consisted of 2–5 loci (Supporting Information Appendix S2), as MULTIPLX 2.1 did not consider more loci to be appropriate for multiplexing due to the risk of primer dimerization.

Multiplex PCR reactions were tested on four individuals per species (amplicons of different species and all loci were pooled prior to library preparation, generating four pools) and sequenced on an Illumina MiSeq platform as described above, using 5% of a lane per library. Raw reads were processed as described in the previous step and separated by barcode (allowing single mismatches) in GENEIOUS. The numbers of reads per locus were recorded, and primer concentrations of multiplex PCRs were adjusted to achieve equal amounts (in terms of read output) of the products of each locus and multiplex reaction. The final PCR runs, with adjusted primer concentrations, were performed in 96-well microtiter plates for each of the multiplex groups separately, using Phusion Hot Start II High-Fidelity DNA polymerase (Thermo Fisher Scientific) or Multiplex PCR Plus Kit (Qiagen). PCR conditions are given in Supporting Information Appendix S2. The PCR products were then pooled, so that each of the 96 pools contained all target SSR loci of up to four individuals per species (individuals from different sample sets could be distinguished through the barcoded primers, Figure 2).

2.4 | Illumina paired-end sequencing of SSR amplicons

The 96 libraries (each including up to four individuals per species and all loci, see above) were prepared for paired-end sequencing (2 × 250 bp) on the Illumina MiSeq platform (using the MiSeq Reagent Kit v2 and the entire flow cell), following Meyer and Kircher (2010). Because of the length of SSR amplicons (<450 bp) targeted during primer development, there was no need to perform DNA fragmentation and size selection, which reduced costs and time for library preparation.

2.5 | Analysis pipeline

The data analysis pipeline included quality control, read merging, demultiplexing, de-novo assembly, and the construction of reference alignments. These steps are described in detail in the following.

TRIMGALORE 0.3.7 was used for adapter clipping and PEAR v0.9.5 (Zhang, Kobert, Flouri, & Stamatakis, 2014) for merging of paired-end reads (setting the p-value threshold for the statistical test to the strictest value, i.e., 0.001) and quality trimming (quality score threshold of 30). Demultiplexing was done using the PERL script fastx_barcode_splitter.pl from the FASTX Toolkit. The respective barcode file contained the specific 10 bp tag and the first 10 bp of the primer sequence, so a total length of 20 bp had to be matched. Two mismatches/partial matches were allowed. Forward and reverse merged reads from the split output carrying the same tags were subsequently concatenated. De-novo assembly was done using CAP3 (version date 12/21/07; Huang & Madan, 1999), with the overlap percent identity cutoff set to 99 and the maximum gap length in any overlap set to 2. The resulting contigs (specific for each individual and each locus, for all species) were imported into GENEIOUS 6.0.4, and a multiple alignment (consensus alignment, with the threshold set to 90%) was done together with a reference sequence (original sequence from cDNA transcripts) of the respective locus and sample set barcode in order to check for mis-tagging. Contigs were visually checked, and if variability was still present within a contig (indicating that CAP3 assembled two alleles into one), de-novo assembly was repeated in GENEIOUS (setting the maximum mismatches per read to 1%). Allele sequences (without tags and primer sequences) are deposited at NCBI GenBank (accession numbers MG322761–MG323307).

2.6 | Size homoplasy

The amount of size homoplasy was calculated as the ratio of the number of fragment length classes containing alleles with different sequences and the total number of fragment length classes. This was done for each species and locus separately.

2.7 | Individual error rate

Several individuals per species (seven for *D. fascicularis*, fourteen for *M. tridens*, and seven for *O. obtusangulus*) were sequenced two or more times at all loci, allowing for estimating the genotyping error rate. This was calculated based on whether identical genotypes (sequences) were observed when comparing the results of the different sequencing runs for each locus and individual.

2.8 | Ploidy level estimation

The individuals of *D. fasciularis* and *O. obtusangulus* had maximally two alleles per locus and were scored as effective diploids, although chromosome numbers point to tetraploidy. By contrast, *M. tridens* had up to four alleles per locus and individual, and the allele dosage for each individual could be determined based on read coverage ratios of the contigs (Figure 3). Therefore, each *M. tridens* individual was scored as tetraploid, with four alleles: (a) homozygous; (b) heterozygous with variable allele dosages (3:1, 2:1:1, 1:1:1:1, or 2:2). The ploidy level of *M. tridens* was also estimated through flow cytometry, using genome size as a proxy (data not provided).

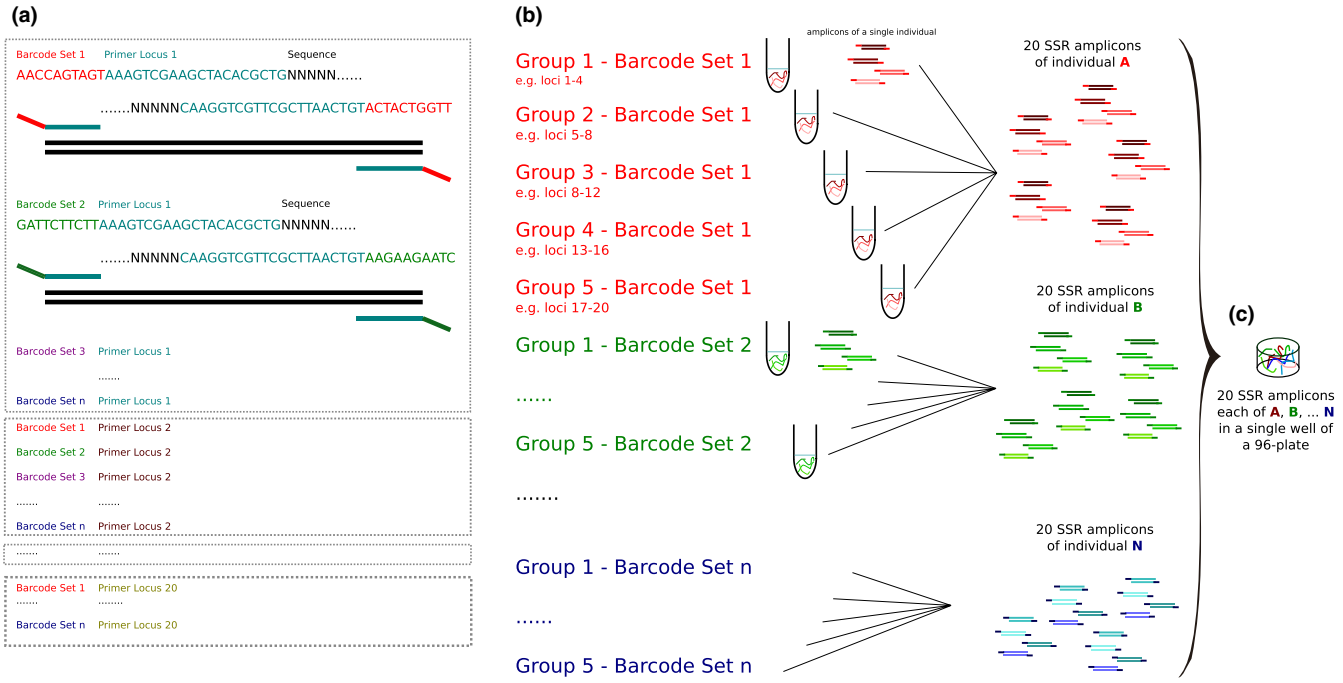

FIGURE 2 Flowchart of the laboratory procedure. (a) barcoding of primers: species A with $n \times 96$ individuals genotyped at 20 loci, n: number of sets of barcoded primers; (b) multiplex PCRs of several loci per multiplex group (depicted here are 4 loci per multiplex group; 2–5 loci in the actual study); (c) pooling of the SSR amplicons of up to n individuals per sequencing library

2.9 | Statistical analyses of population genetic structure and diversity

Data analyses were based on three different datasets, for which alleles were coded according to (a) length of the repetitive region (SSR-length dataset), (b) total fragment length (fragment-length dataset), and (c) sequence identity (sequence-identity dataset).

SPAGEDI 1.5a (Hardy & Vekemans, 2002) was used to calculate the total number of alleles N_A, gene diversity H_e (corrected for sample size, Nei, 1978), and observed heterozygosity H_o, as well as global *F*- and R-statistics for each of the three datasets separately (see above). To estimate the effects of the infinite allele (IAM; Kimura & Crow, 1964) versus stepwise mutation (SMM; Ohta & Kimura, 1973) models, we compared F_{ST} versus R_{ST} of the SSR-length and fragment-length datasets and tested whether observed R_{ST} was significantly higher than its value after permutation. P-values were obtained after 10,000 permutations of allele sizes among alleles within loci to test the null hypothesis that stepwise mutations do not contribute to genetic differentiation (Hardy, Charbonnel, Fréville, & Heuertz, 2003).

Additionally, we scored the number of variable sites, that is, SNPs and indels in the flanking and the repetitive regions plus the variation of the number of SSR units (if variable). Pearson coefficients were calculated to detect correlations between mean fragment length and the number of variable sites, number of SNPs, and degree

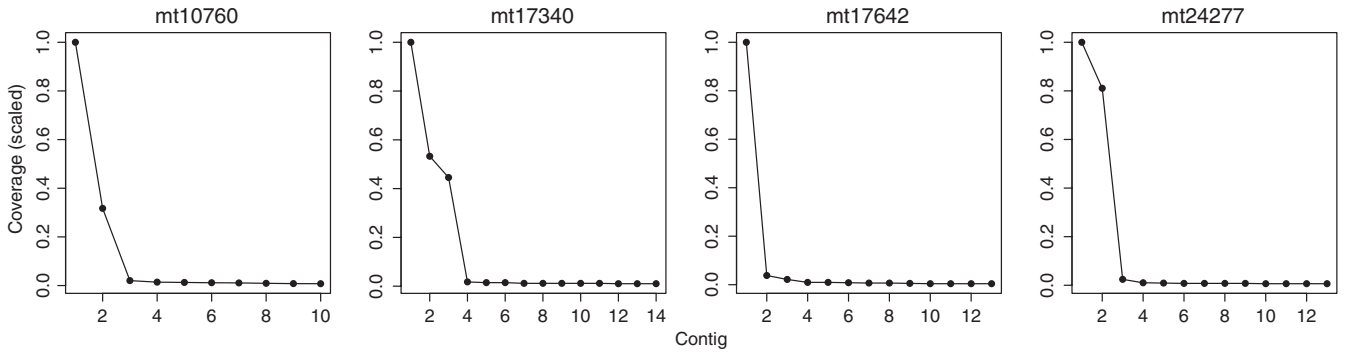

FIGURE 3 Selected coverage graphs of *Mulguraea tridens* individual Mt-033d. Exemplarily shown are the four loci mt10760, mt17340, mt17642, and mt24277. Read coverage of contigs is scaled to 1. Contigs are numbered according to read coverage, that is, the contig with the most reads is numbered contig 1. Heterozygosity and allele dosage become apparent when comparing relative read numbers: Individual Mt-033d is heterozygous at loci mt10760 (allele dosage 3:1), mt17340 (2:1:1), and mt24277 (2:2) and homozygous at locus mt17642

of size homoplasy per locus; and between degree of size homoplasy and number of variable sites and SNPs. Paired Student's *t* test was used to test whether H_e and H_o differed significantly between the fragment-length and the sequence-identity datasets.

3 | RESULTS

3.1 | Output statistics

Total number of raw reads from the Illumina MiSeq run was 41,990,310 (containing all eight species). Of these, 97.7% could be unambiguously assigned to the respective libraries based on sequencing adapters. Raw read numbers per library averaged 213,594 ± 57,455 (mean ± *SD*; forward and reverse libraries yet unmerged). PEAR successfully merged 99.6% of all read pairs passing quality control (mean over all 96 libraries), so that the total number of merged reads was 20,401,853 (i.e., 2.8% of the raw read output either did not pass quality control or lacked its respective mate).

With regard to the three species studied here, all loci could be recovered by demultiplexing, but the average number of reads per locus (within one species) differed among loci by up to three orders of magnitude (read coverage threshold ≥10). Five and seven loci of *D. fascicularis* and *O. obtusangulus* had low coverage (<10 reads per allele) in more than 10% of the individuals or contained putative null alleles and were excluded from all analyses. One locus of *D. fascicularis* and one of *M. tridens* contained highly divergent allele sequences, suggesting the existence of two or more paralogous copies. These loci were also excluded. Locus mt11151 of *M. tridens* contained two different repetitive regions and was separated into two loci in the SSR-length dataset. Individuals with missing data in more than 30% of the loci were excluded from the statistical analyses (Table 1).

All assemblies produced contigs with skewed read coverage distributions, that is, those contigs that represented the "true alleles" had much higher average coverage than the remaining contigs ("noise"; see Figure 3). Noise was caused by PCR recombinants (Meyerhans, Vartanian, & Wain-Hobson, 1990; recognizable as chimera of the most common alleles) and sequencing errors (SNPs with <1% occurrence among all reads of the specific allele and individual).

Several individuals per species were sequenced two or more times (during variation assessment, multiplex testing, and the final sequencing run) at all loci. This allowed for the estimation of the error rate. The same number of alleles was retrieved for each locus and individual, though sequence variation occasionally occurred. Overall error rate for *D. fascicularis* was 1.14%, 2.45% for *M. tridens*, and 1.71% for *O. obtusangulus*.

3.2 | Size homoplasy

The sequence data revealed that size homoplasy is very common (Table 2). It differed between species (mean over all loci 44.7%–63.5%) and—to a very high degree—between loci within one species, ranging from 20% to 100% with regard to the ratio of the number of fragment size classes with sequence variability/total number of size classes. Regarding the flanking regions, SNP variation was much more common than indel variation: SNP/indel ratios were 90/11, 97/7, and 71/7 for *D. fascicularis*, *M. tridens*, and *O. obtusangulus*, respectively. Many SSR loci also contained mutated SSR motifs (Table 3).

Size homoplasy was detected at all levels—among geographic populations, between different individuals of the same population, and even between alleles of the same individual (data not shown). The degree of size homoplasy was not correlated with mean fragment length, number of repetitive units, number of variable sites, or number of SNPs in the flanking regions (Pearson's correlation coefficient <0.05, data not shown). To check for erroneous SNP calls, the number of rare alleles (occurring just once in the respective dataset) was recorded for each species and locus (Table 4). Based on the sequence-identity dataset, the percentage of rare alleles in relation to the total number of alleles differed between species. The percentage of rare alleles was highest in tetraploid *M. tridens* (38.4%), followed by *D. fascicularis* (21%), and lowest in *O. obtusangulus* (3.5%).

3.3 | PCR recombination

Sequencing revealed the existence of PCR recombinants in all individual assemblies. Application of non-combinatorial barcoding enabled the detection of false alleles formed by recombination between individuals of different barcode sets. "Silent" recombinants (i.e., between individuals of the same barcode set) appeared as chimeric alleles composed of the most common "true" alleles and were detected because of lower read coverage in comparison to that of the alleles proper. The amount of recombinants within a library likely increases with increasing number of pooled individuals. Consequently, there is a trade-off between cost and time efficiency and the amount of noise in the data, produced by recombinant DNA sequences.

3.4 | Population genetic diversity and structure

The number of alleles N_A, H_e, and H_o differed between datasets (Tables 4 and 5). Whereas results were rather similar for the SSR-length and fragment-length datasets, genetic diversity estimates were generally higher for the sequence-identity dataset in all three species. Expected and observed heterozygosity differed significantly (paired Student's *t* test, $p < 0.05$) between the sequence-identity and fragment-length datasets of *D. fascicularis*, and also H_e differed significantly ($p < 0.01$) between these datasets in the case of *O. obtusangulus*. There were no significant differences between datasets in case of *M. tridens*. For most loci, F_{ST} values were similar between the datasets. However, for a few loci, F_{ST} differed markedly between the sequence-identity dataset as compared to the SSR-length dataset (df124143, oo20129; Table 5). Multilocus F_{ST} values of the sequence-identity dataset were lower than that of the other datasets in all species, but the difference was not significant ($p > 0.05$).

TABLE 2 Size homoplasy of the study species per locus

Species	Locus name	Total number of fragment size classes	Fragment size classes with homoplasy	% homoplasy
D. fascicularis	df14769	4	2	50
D. fascicularis	df123709	2	2	100
D. fascicularis	df124143	2	1	50
D. fascicularis	df126453	9	2	22.2
D. fascicularis	df137861	4	2	50
D. fascicularis	df138027	4	3	75
D. fascicularis	df142807	2	2	100
D. fascicularis	df22716	4	2	50
D. fascicularis	df51291	9	2	22.2
D. fascicularis	df61486	4	1	25
D. fascicularis	df79494	13	5	38.5
D. fascicularis	df80221	6	4	66.7
D. fascicularis	df80820	3	3	100
D. fascicularis	df91667	10	3	30
Total		76	34	44.7
M. tridens	mt10760	3	2	66.7
M. tridens	mt11151	13	4	30.8
M. tridens	mt14700	4	3	75
M. tridens	mt16240	6	3	50
M. tridens	mt16881	2	1	50
M. tridens	mt17340	4	1	25
M. tridens	mt17642	4	2	50
M. tridens	mt21753	5	1	20
M. tridens	mt23026	2	1	50
M. tridens	mt24277	4	1	25
M. tridens	mt25107	6	2	33.3
M. tridens	mt25266	7	5	71.4
M. tridens	mt27365	7	3	42.9
M. tridens	mt28267	1	1	100
M. tridens	mt30890	2	1	50
M. tridens	mt34724	1	1	100
M. tridens	mt57863	3	2	66.7
Total		74	34	45.9
O. obtusangulus	oo12746	5	2	40
O. obtusangulus	oo14265	3	3	100
O. obtusangulus	oo16914	3	2	66.7
O. obtusangulus	oo17752	3	2	66.7
O. obtusangulus	oo20129	4	2	50
O. obtusangulus	oo20553	3	2	66.7
O. obtusangulus	oo25879	4	2	50
O. obtusangulus	oo34170	4	3	75
O. obtusangulus	oo40886	8	4	50
O. obtusangulus	oo41307	8	7	87.5
O. obtusangulus	oo48962	10	5	50

(Continues)

TABLE 2 (Continued)

Species	Locus name	Total number of fragment size classes	Fragment size classes with homoplasy	% homoplasy
O. obtusangulus	oo56658	5	3	60
O. obtusangulus	oo59128	3	3	100
Total		63	40	63.5

Note. Homoplasy is defined as the number of fragment size classes containing hidden variation (i.e., alleles of the same length but differing in sequence) divided by the total number of fragment size classes. Per-locus percentages and the mean over all loci are given. Locus names carrying the prefix df correspond to *Donatia fascicularis*, whereas mt and oo stand for *Mulguraea tridens* and *Oreobolus obtusangulus*.

Mean permuted R_{ST} values were significantly lower than the observed values in some instances in *D. fascicularis* (df91667, SSR-length dataset) and *O. obtusangulus* (oo40886 and over all loci, both SSR- and fragment-length datasets). In *M. tridens*, there was no difference between observed and mean permuted R_{ST} values at any locus.

The higher resolution of the sequence-identity dataset became apparent in the higher number of alleles per locus for all loci and was itself highly variable between loci, that is, the number of alleles, as calculated from the sequence dataset, was 1.2–8.0, 1.3–13, and 1.6–4.0 times higher in *D. fascicularis*, *M. tridens*, and *O. obtusangulus*, respectively, than the number of alleles calculated from the SSR-length dataset (Table 4). In case of *M. tridens*, two loci (mt28267 and mt34724) had no variation in length of neither the SSR nor the whole fragment, but had five and thirteen alleles, respectively, based on sequence identity. There was also one locus in *D. fascicularis* (df142807) that was monomorphic in the SSR-length dataset, but in case of fragment-length and sequence-identity datasets, allele numbers were two and eight, respectively.

4 | DISCUSSION

Using Mark Twain's "the report of my death was an exaggeration" in their publication's title, Hodel et al. (2016) expressed the opinion that SSRs still represent a useful marker system because of its high mutation rates and cost-efficiency. They reviewed different NGS methods of SSR identification and primer development and discussed the pros and cons of using genotyping-by-sequencing (GBS) or restriction site-associated DNA sequencing (RAD-seq) in comparison to SSRs. That SSRs are not dead is reflected in the development of new analytical tools, for example, for the automatic inference of SSR genotypes (Zhan et al., 2017), and in a variety of publications that employed NGS for obtaining SSR sequence data directly from PCR amplicons (Bradbury et al., 2018; De Barba et al., 2017; Vartia et al., 2016). While these studies used animals as study systems, similar approaches have not yet been applied to plants. Our method based on barcoding of PCR primers combined with multiplexed PCR reactions and Illumina sequencing enabled us to obtain sequence data of many loci and individuals in parallel. We therefore could compare the sequencing output to what we would have obtained by recording fragment length.

4.1 | Output statistics and estimation of ploidy levels

Demultiplexing successfully recovered all loci of the three study species, although 12 out of 58 loci had low coverage (<10 reads per allele) and those were excluded from the analyses. The adjustment of relative primer concentrations within a given PCR multiplex group is rather approximate and cannot guarantee equal yield for each locus. The ploidy level of *M. tridens* has not been reported yet; however, the basic chromosome number of the genus *Mulguraea* is $x = 10$ and related species were reported to be di-, tetra-, and hexaploid (Botta & Brandham, 1993). Based on the number of retrieved alleles and graphs of the standardized number of reads per each allele of the respective locus and individual (Figure 3), we inferred tetraploidy for *M. tridens*. Thus, the method can be successfully applied also for tri- or tetraploid species. The applicability in higher polyploids has yet to be tested, but based on our data of two other species included in the original study (*Berberis microphylla*, *Chuquiraga aurea*; data not presented here) suggest that allele detection in octo- and higher ploids might be complicated due to the presence of PCR recombinants (Brassac & Blattner, 2015), PCR duplicates, sequencing errors, and the problem of missing single-copy alleles.

4.2 | Size homoplasy

All SSR loci used in our study contained SNP variation in the FR and sometimes also in the repetitive region (Table 3), that is, the true allele number per population was always higher than when only length information would have been recorded (Table 4). The mean amount of size homoplasy (see definition above) was similar between the three studied species (44.7%, 45.9%, and 63.5% for *D. fascicularis*, *M. tridens*, and *O. obtusangulus*, respectively; Table 2), which is surprisingly high at the species level considering that the degree of homoplasy increases with increasing time of divergence between populations and taxa (Estoup & Cornuet, 1999). On the other hand, high degrees of size homoplasy were observed in *Rubus* subgenus *Rubus* (based on cloning and sequencing of each SSR locus), detecting SNPs at all studied loci within and among the species (Šarhanová et al., 2017). Vartia et al. (2016) screened 16 individuals of Atlantic cod for homoplasy and detected that 71.7% of the analyzed loci carry sequence variation, which represented 32% of all genotypes.

TABLE 3 Information on the used SSR loci

Locus name	SSR motif	Mean SSR length (bp)	Mutated SSR motif	Fragment length (bp)	Mean fragment length (bp)	*N* variable sites	*N* SNPs	*N* indels	Indel length (bp)
df14769	$(CAC)_{5-7, 9, 10}$	22.2	Yes	367–382	374.5	6	5	0	0
df123709	$(AGG)_{6, 7}$	19.5	No	329–344	336.5	9	8	0	0
df124143	$(TGG)_{4, 7, 9}$	20	No	369–384	375.5	12	11	0	0
df126453	$(CTT)_{4, 7-10, 14}$	27.7	No	315–345	330	4	3	0	0
df137861	$(ACA)_{6-8}$	21	Yes	318–324	321	10	8	1	2
df138027	$(TTCTGA)_{1, 3-5}$	18	No	289–313	301	6	5	0	0
df142807	$(CTTTGC)_{3}$	18	No	288–301	294.5	9	8	1	13
df22716	$(CAAAC)_{2-5}$	14	No	329–344	336.5	6	5	0	0
df51291	$(CTT)_{4, 6, 7, 9-13}$	27.75	Yes	318–345	331.5	11	8	2	1, 8
df61486	$(CAAGG)_{4-6}$	25	No	289–299	294	6	4	1	1
df79494	$(CTCA)_{2, 4, 7, 9-13, 15}$	36.7	Yes	253–305	279	4	2	1	9
df80221	$(GGT)_{4-9}$	19.5	No	314–329	321.5	8	7	0	0
df80820	$(TTG)_{6, 8, 9}$	23	No	304–314	309	10	8	1	1
df91667	$(ATAG)_{3, 5-10}$	26.9	No	243–266	254.5	13	8	4	2, 2, 7, 15
					Total	114	90	11	
mt10760	$(AAG)_{6-9}$	22.5	No	337–346	341	2	1	0	0
mt11151[a]	$(TGA)_{3, 6-8, 10-13}/(TCC)_{4-6}$	26.25/15	Yes	330–363	350.5	19	18	1	3
mt14700	$(GAT)_{6, 7, 10}$	23	No	341–353	347	6	4	1	3
mt16240	$(GAA)_{6-11}$	25.5	Yes	301–316	308.5	7	6	0	0
mt16881	$(ATA)_{8, 11}$	28.5	No	379, 388	383.5	5	4	0	0
mt17340	$(TTGA)_{3-6}$	13.5	Yes	358–370	364	3	2	0	0
mt17642	$(GAG)_{6-9}$	22.5	Yes	310–319	314.5	6	4	1	3
mt21753	$(AGA)_{4, 6, 7}$	17	No	269–287	278.6	5	2	2	6, 15
mt23026	$(CAT)_{8, 10}$	27	Yes	369, 375	372	7	5	1	3
mt24277	$(ATC)_{10, 13, 14, 17}$	40.5	No	317–338	327.5	2	1	0	0
mt25107	$(GAA)_{5, 6, 8-10, 12}$	25	No	286–307	296	4	3	0	0
mt25266	$(ATG)_{5, 10-14}$	32.5	Yes	374–401	390.3	15	13	1	6
mt27365	$(ATC)_{5-11}$	24	No	275–293	284	5	4	0	0
mt28267	$(GAT)_{6}$	18	No	385	385	16	16	0	0
mt30890	$(CTT)_{6, 7}$	19.5	No	210, 213	211.5	6	5	0	0
mt34724	$(AAT)_{6}$	18	No	257	257	5	5	0	0
mt57863	$(GAA)_{7, 8, 10}$	25	Yes	331–340	335	5	4	0	0

(Continues)

TABLE 3 (Continued)

Locus name	SSR motif	Mean SSR length (bp)	Mutated SSR motif	Fragment length (bp)	Mean fragment length (bp)	*N* variable sites	*N* SNPs	*N* indels	Indel length (bp)
					Total	118	97	7	
oo12746	$(CAC)_{6-8,\ 10}$	23.25	No	315–327	321	7	6	0	0
oo14265	$(CGG)_{2,\ 6,\ 9}$	17	Yes	345–366	355.5	5	4	0	0
oo16914	$(TGG)_{7,\ 8,\ 12}$	27	Yes	412–418	415	9	8	0	0
oo17752	$(TTC)_{7-8}$	22.5	No	313–325	319	9	7	1	9
oo20129	$(CGCCTC)_{3-6}$	27	No	326–344	335	8	6	1	12
oo20553	$(GATTTG)_{4-6}$	30	Yes	356–368	362	4	3	0	0
oo25879	$(AACCAA)_{2-5}$	21	Yes	334–352	343	4	3	0	0
oo34170	$(GATT)_{3-6}$	18	No	307–319	313	7	5	1	4
oo40886	$(TTC)_{4-6,\ 9-12}$	24.3	Yes	337–364	350.5	5	3	1	3
oo41307	$(GAA)_{12-19}$	46.5	Yes	324–345	334.5	5	4	0	0
oo48962	$(GAT)_{7-16,\ 20}$	36.8	Yes	383–410	396.5	8	6	2	3, 12
oo56658	$(CAT)_{10-14}$	36	Yes	319–331	325	8	7	0	0
oo59128	$(TTCTT)_{4-6}$	20	No	301–313	307	11	9	1	11
					Total	90	71	7	

Note. SSR motif: sequence of the SSR motif, number indicates how many times the motif is present. Variable sites include sequence variation in the SSR region itself and in the flanking region plus the variation in number of repetitive motifs (if the variation was present). *N* indicates numbers of variable sites, SNPs and indels, respectively. Locus names carrying the prefix df correspond to *Donatia fascicularis*, whereas mt and oo stand for *Mulguraea tridens* and *Oreobolus obtusangulus*.

[a]Includes two different SSR regions.

TABLE 4 Total number of alleles and number of rare alleles (present only once in the respective dataset) for the three different datasets based on SSR-length, fragment-length (Fr length), and sequence-identity (Seq identity)

	N alleles			*N* rare alleles		
Locus	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity
df14769	4	4	13	1	1	3
df123709	2	2	11	–	–	3
df124143	2	2	6	–	–	–
df126453	9	9	11	–	–	1
df137861	3	4	12	1	1	5
df138027	4	4	8	–	–	–
df142807	1	2	8	–	–	–
df22716	4	4	8	1	1	3
df51291	8	9	13	2	3	4
df61486	3	4	6	–	–	–
df79494	12	13	20	1	2	3
df80221	6	6	17	–	–	5
df80820	3	3	9	–	–	1
df91667	7	10	15	–	1	5
Total	68	76	157	6	9	33
mt10760	4	4	5	2	2	3
mt11151a	8	13	20	1	5	10
mt11151b	3	–	–	1	–	–
mt14700	3	4	9	–	–	2
mt16240	6	6	10	1	1	5
mt16881	2	2	6	1	1	2
mt17340	4	4	6	2	2	2
mt17642	4	4	9	–	–	2
mt21753	3	5	7	–	–	1
mt23026	2	2	7	–	–	3
mt24277	4	4	5	–	–	–
mt25107	6	6	8	1	1	1
mt25266	6	7	15	2	2	9
mt27365	7	7	13	2	2	4
mt28267	1	1	13	–	–	7
mt30890	2	2	6	–	–	1
mt34724	1	1	5	–	–	1
mt57863	3	3	7	1	1	5
Total	69	75	151	14	17	58
oo12746	4	5	9	1	1	1
oo14265	3	3	6	–	–	–
oo16914	3	3	12	–	–	1
oo17752	2	3	8	–	–	–
oo20129	4	4	9	–	–	–
oo20553	3	3	10	–	–	1
oo25879	4	4	7	–	–	–
oo34170	4	4	8	–	–	–
oo40886	7	8	13	–	–	–

(Continues)

TABLE 4 (Continued)

	N alleles			*N* rare alleles		
Locus	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity
oo41307	8	8	23	–	–	1
oo48962	11	10	18	–	–	–
oo56658	5	5	12	–	1	1
oo59128	3	4	9	–	–	–
Total	61	64	144	1	2	5

Note. *N* indicates number of alleles. Locus names carrying the prefix df correspond to *Donatia fascicularis*, whereas mt and oo stand for *Mulguraea tridens* and *Oreobolus obtusangulus*.

Unfortunately, their way of calculating homoplasy is not fully clear, so that a direct comparison with our data is not possible. Nonetheless, we expect that the amount of homoplasy will increase with an increasing number of genotyped individuals, especially if these are geographically and evolutionarily more distant to each other.

In the FR, SNPs were around 10 times more frequent than indels (Table 3). This ratio is much higher than the one reported by Mogg et al. (2002) for *Zea mays*, in which the mean ratio over all loci between SNPs and indels was 2:1. High number of SNPs could be caused by PCR/sequencing errors, although it is not very likely for several reasons: (a) Such errors would appear as rare alleles, randomly occurring in the whole dataset; (b) rare alleles would not be present in a homozygous state; (c) if present as heterozygous, the coverage of the erroneous allele would tend to be lower than the coverage of the true allele(s) of the individual. These conditions are not met for the vast majority of rare alleles, which are species, population, and/or locus specific (Table 4). Nonetheless, negligible effects of sequencing errors cannot be ruled out.

In the case of *D. fascicularis*, the lengths of the indels of the FR were not congruent with the repeat motif length of the SSR, which was opposite to *M. tridens* and *O. obtusangulus*, where six and five out of six indels, respectively, could be confounded with tri-, tetra-, or hexanucleotide repeats if only fragment lengths were taken into account (Table 3).

The presented method allows using longer fragments (mean locus length was 329 bp over all loci and species with a maximum length of 418 bp), as compared to the traditional way of SSR scoring, increasing the likelihood that the FR contain genetic variation. This, in fact, does not prevent short fragments of already available primers and SSR loci to be successfully genotyped applying our method (Sochor, Šarhanová, Pfanzelt, & Trávníček, 2017). Nevertheless, the degree of size homoplasy did not correlate with mean fragment length, number of variable sites or number of SNPs or SSR units. Interestingly, the shortest locus of *D. fascicularis* (df91667) had the highest number of variable sites (Table 3). In the case of eventual correlations of size homoplasy and the number of variable sites, it should be considered that the way of calculating size homoplasy in the present study does not take into account how many alleles of the same fragment size class (differing in sequence, but not in length) are present. Therefore, the detection of a further SNP variant within a given fragment size class that already contains different sequences would not lead to a higher degree of size homoplasy. Although our SSR loci originated from RNA sequencing and we obtained BLAST hits for some of the loci, we could not test for a possible correlation between size homoplasy and the origin of sequences, that is, whether they stem from functional genes or the noncoding portion of the genome.

Although many workers have reported on size homoplasy and the problem of hidden variation earlier (for example in plants, see Adams, Brown, & Hamilton, 2004; Barkley, Krueger, Federici, & Roose, 2009; Curtu, Finkeldey, & Gailing, 2004; Kostia, Varvio, Vakkari, & Pulkkinen, 1995; Lia et al., 2007; Peakall et al., 1998), fragment length analysis was and still is carried out without considering FR polymorphism. Taking it into account does not eliminate all homoplasy in a dataset, because back mutation to the ancestral state still can occur. Nonetheless, besides revealing SNP variation, SSR sequencing avoids genotyping errors in case of indel polymorphism, like those that were detected in six of thirteen and seventeen SSR loci in *O. obtusangulus* and *M. tridens* and in seven of fourteen in *D. fascicularis*, respectively (Table 3).

4.3 | Estimation of genetic diversity

The statistical analyses confirmed that in all three study species, the sequence-identity dataset conveyed more information than the SSR-length and fragment-length datasets (Tables 4 and 5). The overall F_{ST} of *O. obtusangulus* reflects its relatively high degree of intraspecific genetic differentiation, whereas the low F_{ST} values of *D. fascicularis* and *M. tridens* suggest no population structure in any of the datasets. Only three loci (df124143, mt34724, and oo20129) showed markedly increased F_{ST} values in the sequence-identity dataset, but for most loci, F_{ST} values did not differ between the datasets. This implies that the additional information contained in the sequence-identity dataset does not necessarily influence overall population genetic and diversity statistics. The output can be affected by the nature of the study system, its genetic structure, selected loci, sampling, and other variables. *Donatia fascicularis*, for instance, has a low overall F_{ST} and shows almost no population genetic structure (S. Pfanzelt, P. Šarhanová, D. C. Albach, & K. B. von Hagen, under review). It has only one population (with ten individuals) that is genetically distinct, but the remaining individuals belong to a single, undifferentiated cluster. In such a case, where there is no structure

TABLE 5 Genetic diversity indices for the three different datasets based on SSR-length, fragment-length (Fr length), and sequence-identity (Seq identity)

	R_{st}			F_{st}			H_e			H_o		
	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity
all loci	0.205	0.164	–	0.177	0.178	0.159	0.368	0.388	0.564	0.275	0.287	0.438
df14769	0.136	0.136	–	0.142	0.142	0.157	0.401	0.401	0.545	0.318	0.318	0.465
df123709	0.155	0.155	–	0.155	0.155	0.162	0.395	0.395	0.763	0.301	0.301	0.610
df124143	0.007	0.007	–	0.007	0.007	0.154	0.012	0.012	0.459	0.006	0.006	0.318
df126453	0.147	0.147	–	0.200	0.200	0.202	0.639	0.639	0.642	0.508	0.508	0.511
df137861	0.137	0.128	–	0.142	0.146	0.146	0.499	0.509	0.636	0.381	0.381	0.488
df138027	0.106	0.106	–	0.094	0.094	0.093	0.514	0.514	0.739	0.424	0.424	0.622
df142807	–	0.200	–	–	0.200	0.137	–	0.258	0.638	–	0.163	0.482
df22716	0.095	0.095	–	0.259	0.259	0.186	0.248	0.248	0.382	0.175	0.175	0.317
df51291	0.193	0.175	–	0.156	0.155	0.129	0.346	0.348	0.418	0.314	0.314	0.378
df61486	0.188	0.187	–	0.217	0.212	0.237	0.418	0.422	0.438	0.323	0.326	0.326
df79494	0.171	0.173	–	0.106	0.106	0.111	0.856	0.857	0.873	0.675	0.675	0.706
df80221	0.219	0.219	–	0.346	0.346	0.119	0.183	0.183	0.654	0.076	0.076	0.462
df80820	0.360	0.367	–	0.403	0.403	0.368	0.204	0.204	0.249	0.085	0.085	0.110
df91667	0.566*	0.130	–	0.230	0.234	0.216	0.431	0.437	0.455	0.259	0.262	0.338
all loci	0.111	0.095	–	0.115	0.114	0.110	0.384	0.411	0.526	0.183	0.197	0.257
mt10760	0.087	0.087	–	0.065	0.065	0.046	0.354	0.354	0.359	0.227	0.227	0.234
mt11151a	0.056	0.039	–	0.060	0.068	0.084	0.777	0.802	0.826	0.457	0.465	0.481
mt11151b	0.093	–	–	−0.002	–	–	0.089	–	–	0.027	–	–
mt14700	0.098	0.088	–	0.097	0.067	0.080	0.357	0.377	0.511	0.276	0.283	0.356
mt16240	0.219	0.219	–	0.230	0.230	0.197	0.635	0.635	0.662	0.112	0.112	0.122
mt16881	0.101	0.098	–	0.101	0.098	0.107	0.023	0.023	0.181	0.000	0.000	0.102
mt17340	0.148	0.148	–	0.217	0.217	0.184	0.497	0.497	0.657	0.260	0.260	0.395
mt17642	0.171	0.171	–	0.127	0.114	0.121	0.261	0.271	0.458	0.185	0.193	0.24
mt21753	0.156	0.080	–	0.160	0.147	0.120	0.530	0.557	0.620	0.269	0.302	0.339
mt23026	0.008	0.008	–	0.008	0.008	0.007	0.435	0.435	0.637	0.261	0.261	0.362
mt24277	0.185	0.185	–	0.157	0.157	0.198	0.470	0.470	0.502	0.154	0.154	0.162
mt25107	0.041	0.041	–	0.087	0.087	0.127	0.508	0.508	0.562	0.220	0.220	0.22
mt25266	0.163	0.117	–	0.097	0.096	0.078	0.642	0.729	0.757	0.226	0.244	0.302

(Continues)

TABLE 5 (Continued)

	R_{st}			F_{st}			H_e			H_o		
	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity	SSR length	Fr length	Seq identity
mt27365	0.095	0.095	–	0.086	0.086	0.105	0.786	0.786	0.826	0.485	0.485	0.517
mt28267	–	–	–	–	–	0.035	0.000	0.000	0.489	0.000	0.000	0.24
mt30890	0.116	0.116	–	0.116	0.116	0.114	0.502	0.502	0.658	0.128	0.128	0.156
mt34724	–	–	–	–	–	0.266	0.000	0.000	0.170	0.000	0.000	0.091
mt57863	−0.052	−0.052	–	0.062	0.062	0.016	0.038	0.038	0.071	0.013	0.013	0.043
all loci	0.790***	0.785***	–	0.683	0.668	0.633	0.513	0.523	0.682	0.018	0.019	0.026
oo12746	0.803	0.734	–	0.887	0.784	0.740	0.401	0.574	0.603	0.003	0.006	0.006
oo14265	0.853	0.853	–	0.773	0.773	0.752	0.547	0.547	0.636	0.014	0.014	0.002
oo16914	0.632	0.593	–	0.663	0.599	0.642	0.512	0.396	0.689	0.003	0.006	0.008
oo17752	0.893	0.800	–	0.893	0.881	0.785	0.410	0.414	0.633	–	–	0.003
oo20129	0.191	0.513	–	0.121	0.466	0.447	0.081	0.152	0.603	–	–	0.042
oo20553	0.687	0.687	–	0.711	0.711	0.693	0.493	0.493	0.666	0.017	0.017	0.025
oo25879	0.834	0.834	–	0.814	0.814	0.760	0.585	0.585	0.629	0.017	0.017	0.019
oo34170	0.852	0.850	–	0.859	0.815	0.519	0.422	0.579	0.794	0.014	0.017	0.031
oo40886	0.851*	0.856*	–	0.741	0.709	0.692	0.637	0.642	0.649	0.014	0.014	0.014
oo41307	0.422	0.422	–	0.320	0.320	0.375	0.742	0.742	0.860	0.086	0.086	0.089
oo48962	0.725	0.489	–	0.597	0.501	0.600	0.761	0.600	0.775	0.019	0.019	0.019
oo56658	0.494	0.508	–	0.580	0.580	0.679	0.453	0.453	0.659	0.019	0.019	0.025
oo59128	0.794	0.829	–	0.707	0.703	0.662	0.621	0.623	0.674	0.028	0.028	0.031

Note. H_e: gene diversity corrected for sample size. H_o: observed heterozygosity. Locus names carrying the prefix df correspond to *Donatia fascicularis*, whereas mt and oo stand for *Mulguraea tridens* and *Oreobolus obtusangulus*.

*, **, and *** denote significance at α = 0.05, 0.01, and 0.001, respectively.

at all, the high resolution of the sequence-identity dataset is not informative either. Nonetheless, if additional information can be collected by sequencing of the studied loci, it is highly recommended to do so to ensure correct evolutionary interpretations. This mirrors Peakall et al.'s (1998: 1,283) earlier, but apparently often neglected, statement: "Consequently, DNA sequencing of SSR alleles will be essential to minimize the risk of misinterpretation and to maximize the genetic information that can be obtained." Now widely available NGS technologies make it possible to routinely score SSR alleles through sequencing.

On the other hand, genetic diversity parameters (H_e and H_o) were rather similar for the SSR-length and fragment-length datasets, but were generally higher for the sequence-identity dataset in all three species (Table 5). This was especially remarkable in one locus of *D. fascicularis* (df142807) and two loci of *M. tridens* (mt28267 and mt34724), which appear to be monomorphic if only fragment or SSR lengths are considered. However, these loci were highly variable based on sequence identity (8, 13, and 5 alleles, respectively; Table 4). In *D. fascicularis* and *O. obtusangulus*, there was a significant difference in H_e between the sequence-identity dataset compared to the traditional fragment-length dataset. This was not the case in *M. tridens*, which may simply reflect the heterozygous nature of this tetraploid species, which is visible already when analyzing the fragment size dataset.

4.4 | Microsatellite mutation models

Permuted R_{ST} values suggested for all but two studied loci (df91667, oo40886; Table 5) that stepwise mutations do not significantly contribute to genetic differentiation. Interestingly, the observed R_{ST} value of locus df91667 was higher than the permuted R_{ST}, indicating the fit to the SMM, but only for the SSR-length dataset. In fact, the four indels in the FR of that locus would mask this output in case of the traditional fragment length assessment. The second locus (oo40886) shows a bimodal distribution of the number of repeats (4–6 and 9–12) and thus fits rather to the two-phase model of microsatellite evolution (Di Rienzo et al., 1994), in which frequent single-step, but also rare large changes in repeat number occur. The IAM does not fit to the evolution of most of the studied loci, as it does not allow for the existence of homoplasy (Estoup et al., 2002). Other models like proportional slippage/point mutation (Kruglyak, Durrett, Schug, & Aquadro, 1998), the K-allele model (Crow & Kimura, 1970), or more complex stepwise models can better reflect the evolution of microsatellites and should be considered in future research.

5 | SUMMARY

Our multiplex SSR sequencing strategy produced useful information about the actual nucleotide sequences of SSR amplicons and allowed for the detection and quantification of hidden variation in a large dataset of non-model plant species. It was shown that size homoplasy is a very common phenomenon and that indel polymorphism in the FR can be erroneously confounded with length variation within the SSR region. The additional information allows for a better understanding of microsatellite mutation processes. Sequencing of SSR loci is a prospective method with the ability to detect variability on both the intra- and inter-species level and thus can be suitable for both wide- and fine-scale phylogeographic studies based on single marker types.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments to improve the quality of the manuscript. Funding from DFG (AL632/7-1, BL462/11), DAAD (D/10/49464) and Dr. Karl Wamsler Stiftung is acknowledged. CONAF (Chile), APN (Argentina), and the Falkland Islands Government issued collection permits and provided logistic help, as well as Centro EULA (Universidad de Concepción) and CEQUA (Universidad de Magallanes, Punta Arenas).

CONFLICT OF INTEREST

The authors have no conflict of interests to declare.

AUTHOR CONTRIBUTIONS

The first two authors contributed equally to this work. PS, SP, and FRB carried out the fieldwork. PS and SP performed laboratory work and data analyses. PS, AH, RB, and FRB developed the barcode approach and devised the sequencing strategy. PS, SP, and FRB wrote the manuscript. All authors contributed to and approved the final manuscript version.

DATA ACCESSIBILITY

Allele sequences (accession numbers MG322761-MG323307) and RNA-Seq raw read data (SRA accession numbers SRX4496448-SRX4496451) were submitted to NCBI GenBank.

ORCID

Petra Šarhanová http://orcid.org/0000-0002-1006-4003
Axel Himmelbach http://orcid.org/0000-0001-7338-0946
Frank R. Blattner http://orcid.org/0000-0001-7731-7741

REFERENCES

Adams, R. I., Brown, K. M., & Hamilton, M. B. (2004). The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (*Corythophora alta*) and an anadromous fish (*Morone saxatilis*). *Molecular Ecology*, *13*(9), 2579–2588. https://doi.org/10.1111/j.1365-294X.2004.02256.x

Barkley, N. A., Krueger, R. R., Federici, C. T., & Roose, M. L. (2009). What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles. *Plant Systematics and Evolution*, *282*(1–2), 71–86. https://doi.org/10.1007/s00606-009-0208-2

Barthe, S., Gugerli, F., Barkley, N. A., Maggia, L., Cardi, C., & Scotti, I. (2012). Always look on both sides: Phylogenetic information conveyed by simple sequence repeat allele sequences. *PLoS One*, *7*(7), e40699. https://doi.org/10.1371/journal.pone.0040699

Beier, S. (2016). ReadTrimmChecker. https://doi.org/10.5447/IPK/2016/14

Blankenship, S. M., May, B., & Hedgecock, D. (2002). Evolution of a perfect simple sequence repeat locus in the context of its flanking sequence. *Molecular Biology and Evolution*, *19*(11), 1943–1951. https://doi.org/10.1093/oxfordjournals.molbev.a004018

Botta, S., & Brandham, P. E. (1993). The taxonomic significance of chromosome number in *Junellia* (Verbenaceae). *Kew Bulletin*, *48*(1), 143–150. https://doi.org/10.2307/4115759

Bradbury, I. R., Wringe, B. F., Watson, B., Paterson, I., Horne, J., Beiko, R., … Bentzen, P. (2018). Genotyping-by-sequencing of genome-wide microsatellite loci reveals fine-scale harvest composition in a coastal Atlantic salmon fishery. *Evolutionary Applications*, *11*, 918–930. https://doi.org/10.1111/eva.12606

Brassac, J., & Blattner, F. R. (2015). Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and in silico cloning of multiple nuclear loci. *Systematic Biology*, *64*(5), 792–808. https://doi.org/10.1093/sysbio/syv035

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. New York, NY: Harper & Row.

Curtu, A. L., Finkeldey, R., & Gailing, O. (2004). Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). *Plant Molecular Biology Reporter*, *22*(4), 339–346. https://doi.org/10.1007/BF02772677

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2017). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low quantity and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. https://doi.org/10.1111/1755-0998.12594

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., & Freimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(8), 3166–3170. https://doi.org/10.1073/pnas.91.8.3166

Estoup, A., & Cornuet, J.-M. (1999). Microsatellite evolution: Inferences from population data. In D. B. Goldstein, & C. Schlötterer (Eds.), *Microsatellites: Evolution and applications* (pp. 49–65). Oxford, UK: Oxford University Press.

Estoup, A., Jarne, P., & Cornuet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, *11*, 1591–1604. https://doi.org/10.1046/j.1365-294X.2002.01576.x

Germain-Aubrey, C. C., Nelson, C., Soltis, D. E., Soltis, P. S., & Gitzendanner, M. A. (2016). Are microsatellite fragment lengths useful for population-level studies? The case of *Polygala lewtonii* (Polygalaceae). *Applications in Plant Sciences*, *4*(2), 1500115. https://doi.org/10.3732/apps.1500115

Hardy, O. J., Charbonnel, N., Fréville, H., & Heuertz, M. (2003). Microsatellite allele sizes: A simple test to assess their significance on genetic differentiation. *Genetics*, *163*(4), 1467–1482.

Hardy, O. J., & Vekemans, X. (2002). SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Resources*, *2*(4), 618–620. https://doi.org/10.1046/j.1471-8286.2002.00305.x

Hodel, R. G., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis, D. E. (2016). The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Applications in Plant Sciences*, *4*(6), 1600025. https://doi.org/10.3732/apps.1600025

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*, 868–877. https://doi.org/10.1101/gr.9.9.868

Kaplinski, L., Andreson, R., Puurand, T., & Remm, M. (2005). MultiPLX: Automatic grouping and evaluation of PCR primers. *Bioinformatics*, *21*(8), 1701–1702. https://doi.org/10.1093/bioinformatics/bti219

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Thierer, T. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kimura, M., & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, *49*(4), 725.

Kostia, S., Varvio, S. L., Vakkari, P., & Pulkkinen, P. (1995). Microsatellite sequences in a conifer, *Pinus sylvestris*. *Genome*, *38*(6), 1244–1248. https://doi.org/10.1139/g95-163

Kruglyak, S., Durrett, R. T., Schug, M. D., & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(18), 10774–10778. https://doi.org/10.1073/pnas.95.18.10774

Lia, V. V., Bracco, M., Gottlieb, A. M., Poggio, L., & Confalonieri, V. A. (2007). Complex mutational patterns and size homoplasy at maize microsatellite loci. *Theoretical and Applied Genetics*, *115*(7), 981–991. https://doi.org/10.1007/s00122-007-0625-y

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Martin, M. (2011). Cutadapt removes adapter sequences from high throughput sequencing reads. *Embnet.journal*, *17(1),pp*, 10. https://doi.org/10.14806/ej.17.1.200.

Mayer, C. (2010). *Phobos 3.3.12*. Retrieved from http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448. https://doi.org/10.1101/pdb.prot5448

Meyerhans, A., Vartanian, J. P., & Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Research*, *18*(7), 1687–1691. https://doi.org/10.1093/nar/18.7.1687

Miller, M. J., & Yuan, B. Z. (1997). Semiautomated resolution of overlapping stutter patterns in genomic microsatellite analysis. *Analytical Biochemistry*, *251*(1), 50–56. https://doi.org/10.1006/abio.1997.2234

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., & Edwards, K. (2002). Characterization of the flanking regions of *Zea mays* microsatellites reveals a large number of useful sequence polymorphisms. *Theoretical and Applied Genetics*, *105*(4), 532–543. https://doi.org/10.1007/s00122-002-0897-1

Moore, D. M. (1967). Chromosome numbers of Falkland Islands angiosperms. *British Antarctic Survey Bulletin*, *14*, 69–82.

Moore, D. M. (1983). The Flora of the Fuego Patagonian Cordilleras: its origins and affinities. *Revista Chilena de Historia Natural*, *56*(123–136), 0289.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*(3), 583–590.

Ohta, T., & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics Research*, *22*(2), 201–204. https://doi.org/10.1017/S0016672300012994

Ortí, G., Pearse, D. E., & Avise, J. C. (1997). Phylogenetic assessment of length variation at a microsatellite locus. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(20), 10745–10749. https://doi.org/10.1073/pnas.94.20.10745

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., … Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, *327*(5961), 92–94. https://doi.org/10.1126/science.1180677

Peakall, R., Gilmore, S., Keys, W., Morgante, M., & Rafalski, A. (1998). Cross-species amplification of soybean (*Glycine max*) simple

sequence repeats (SSRs) within the genus and other legume genera: Implications for the transferability of SSRs in plants. *Molecular Biology and Evolution*, *15*(10), 1275–1287. https://doi.org/10.1093/oxfordjournals.molbev.a025856

Pfanzelt, S., Albach, D. C., & von Hagen, K. B. (2017). Tabula rasa in the Patagonian Channels? The phylogeography of *Oreobolus obtusangulus* (Cyperaceae). *Molecular Ecology*, *26*(15), 4027–4044. https://doi.org/10.1111/mec.14156

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. In S. Misener, & S. A. Krawetz (Eds.), *Bioinformatics methods and protocols. Methods in molecular biology*, Vol. *132* (p. 365). Totowa, NJ: Humana Press. https://doi.org/10.1385/1-59259-192-2

Šarhanová, P., Sharbel, T. F., Sochor, M., Vašut, R. J., Dančák, M., & Trávníček, B. (2017). Hybridization drives evolution of apomicts in *Rubus* subgenus *Rubus*: Evidence from microsatellite markers. *Annals of Botany*, *120*(2), 317–328. https://doi.org/10.1093/aob/mcx033

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371. https://doi.org/10.1007/s004120000089

Sochor, M., Šarhanová, P., Pfanzelt, S., & Trávníček, B. (2017). Is evolution of apomicts driven by the phylogeography of the sexual ancestor? Insights from European and Caucasian brambles (*Rubus*, Rosaceae). *Journal of Biogeography*, *44*, 2717–2728. https://doi.org/10.1111/jbi.13084

Vartia, S., Villanueva-Cañas, J. L., Finarelli, J., Farrell, E. D., Collins, P. C., Hughes, G. M., … FitzGerald, R. D. (2016). A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science*, *3*(1), 150565. https://doi.org/10.1098/rsos.150565

Weising, K., Wöhrmann, T., & Huettel, B. (2015). The use of high-throughput DNA sequencing for microsatellite discovery in plants. In E. Hörandl & M. S. Appelhans (Eds.), *Next-generation sequencing in plant systematics* (pp. 205–270). Königstein, Germany: Koeltz Scientific Books.

Zhan, L., Paterson, I. G., Fraser, B. A., Watson, B., Bradbury, I. R., Nadukkalam Ravindran, P., … Bentzen, P. (2017). MEGASAT: Automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources*, *17*(2), 247–256. https://doi.org/10.1111/1755-0998.12561

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Šarhanová P, Pfanzelt S, Brandt R, Himmelbach A, Blattner FR. SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecol Evol*. 2018;8:10817–10833. https://doi.org/10.1002/ece3.4533