Chapter 2

Microsatellite Capture Sequencing

Keisuke Tanaka, Rumi Ohtake, Saki Yoshida and Takashi Shinohara

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72629

Abstract

Microsatellites (simple sequence repeats, SSRs) that consist of repetitive sequences of one to six bases are ubiquitous in most eukaryotic genomes. The use of molecular markers for this region is efficacious in molecular-assisted breeding, molecular phylogenetics, and population genetics. Recently, the detection of a number of SSRs using a high-throughput DNA sequencing assay has become possible. Particularly, microsatellite capture sequencing using our developed protocol can detect SSRs more effectively by enriching the DNA library using an SSR probe. Our protocol used in this study demonstrates the possibility of using low-input DNA (≥1 ng), and while the use of restriction enzymes was more suitable for identifying the heterozygous genotype than sonication was, sonication facilitated the detection of various SSR flanking regions with both species-specific and common characteristics more than restriction enzyme digestion did. Moreover, a simulation analysis using various scale reads estimated that a few thousand SSRs could be detected from 50 K reads per sample. Furthermore, we described an *in silico* polymorphic detection and phylogenetic analysis method based on microsatellite capture sequencing data.

Keywords: microsatellite, SSR, capture sequencing, molecular marker, non-model organisms, Myrtaceae

1. Introduction

Molecular markers for DNA were developed in the 1980s and have been used in a wide variety of research fields as a tool for detecting sequence polymorphism between individuals, cultivars, and lineages. In addition, various polymorphic detection methods using molecular markers have been devised based on the structural characteristics of DNA and molecular biological techniques. Among them, the molecular markers based on microsatellite (simple sequence repeat, SSR) regions enabled the development of robust assays with higher resolution and reliability than

those of conventional methods. Generally, SSR constructs consist of a repeat motif with one to six nucleotides, and SSR markers are useful for marker-assisted selection and construction of linkage maps as well as molecular phylogenetics and population genetics because they have various advantages such as high polymorphism, genomic specificity, abundance, and codominance [1, 2]. While SSR markers are very effective, the SSR detection required to construct the marker is often time-consuming. Each SSR detection technique uses colony hybridization, microsatellite enrichment, or both based on the biotin-streptavidin interaction [3–5]. Moreover, another technique was recently reported in which markers were developed in parallel with SSR detection using dual-suppression PCR [6]. However, these approaches have low throughput (a few samples or several tens of samples) since they depend on capillary sequencing.

A current high-throughput DNA sequencing technology, known as next-generation sequencing (NGS), allows the acquisition of huge amounts of data in a single assay. This technology facilitates exhaustive analyses such as whole-genome and RNA sequencing. Additionally, multiple samples can be analyzed at the same time since a specific sequence tag that identifies individuals is added to each library. Thus, the time-consuming assays required for traditional sequencing could be avoided by using such high-throughput DNA sequencing methods. Moreover, high-throughput DNA sequencing was recently used for SSR detection (**Table 1**) [7–42]. These previous studies report that high-throughput DNA sequencing can sufficiently analyze even non-model organisms. In agricultural research field, numerous global major crops such as rice, grapes, and poplar have provided abundant genomic information, whereas little has been reported on regional minor crops have including molecular

Target species	Library style	NGS platform	Number of reads	Sequences including SSR
Agkistrodon contortrix (Copperhead snake)	WG	GS-FLX	128,773	14,612
Anisogramma anomala	WG	GAIIX	26,036,313	44,247
Aristeus antennatus (Red shrimp)	WG	GS-FLX	165,507	247
Aristotelia chilensis (Maqui)	WG	GS-FLX	165,043	24,494
Artocarpus altilis	WG	MiSeq	2,341,465	47,607
Aspidistra saxicola	cDNA	HiSeq2000	13,133,336	4764
Brachiaria ruziziensis (Ruzigrass)	WG	GAII	186,764,108	139,098
Callosobruchus chinensis (Adzuki bean weevil)	WG	HiSeq2500	106,888,024	6593
Camelina sativa	cDNA	GAIIX	10,830,000	14,140
Camellia sinensis (Tea plant)	cDNA	HiSeq2000	26,874,116	5649
Carthamus tinctorius (Safflower)	WG	HiSeq2000	48,502,680	23,067
Catha edulis (Khat)	WG	GS-FLX	65,401	11,678
Catla catla (Catla)	WG	PGM	29,794	21,477
Chrysanthemum nankingense	cDNA	GAII	53,720,166	2813
Daphne kiusiana	WG	MiSeq	4,936,656	28,495
Handroanthus billbergii	WG	MiSeq	2,169,901	61,074
Hydropotes inermis (Water Deer)	WG	GS-FLX	260,467	20,101

Target species	Library style	NGS platform	Number of reads	Sequences including SSR
Hymenolaimus malacorhynchos (Blue duck)	WG	GS-FLX	17,215	231
Ipomoea batatas (Sweetpotato)	cDNA	GAII	59,233,468	4114
Lathyrus sativus (Grasspea)	MCS	GS-FLX	493,364	129,886
Mangifera indica (Mango)	WG	HiSeq2000	90,323,371	106,049
Miscanthus sinensis	cDNA	GS-FLX	241,051	381
Moa fossil	WG	GS-FLX	79,796	195
Panicum miliaceum (Broomcorn millet)	MCS	GS-FLX	1,087,428	223,894
Panicum virgatum (Switchgrass)	cDNA	GS-FLX	979,903	21,437
Pilosocereus genus	RAD	GS-FLX	2,282,266	54,420
Pisum sativum (Pea)	WG	HiSeq2500	173,245,234	8899
Prunus virginiana (Chokecherry)	WG	GS-FLX	145,094	405
Pseudosciaena crocea (Large yellow croaker)	WG	GS-FLX	207,246	2535
Python molurus bivittatus (Burmese python)	WG	GS-FLX	117,515	6616
Raja pulchra (Skate)	WG	GS-FLX	453,549	19,658
Raphanus sativus (Radish)	cDNA	HiSeq2000	71,950,000	11,928
Scabiosa columbaria	cDNA	GS-FLX, GAII	29,522,184	4320
Sesamum indicum (sesame)	cDNA	HiSeq2000	26,266,670	6276
Vicia faba (Faba bean)	MCS	GS-FLX	532,599	125,559
Viola mirabilis	WG	GS-FLX	443,935	36,670

WG: whole genome; MCS: microsatellite capture sequencing; RAD: restriction site associated DNA.

Table 1. Examples of simple sequence repeat (SSR) detection with high-throughput sequencing.

markers. Particularly, the development of genomic resources for tropical and subtropical fruits or underutilized fruit crops is limited [43]. We carried out microsatellite capture sequencing using a high-throughput DNA sequencing technology to obtain sequences with SSR regions of candidate SSR markers for five Myrtaceae plants (*Feijoa sellowiana*, *Myrciaria dubia*, *Psidium guajava*, *Psidium littorale*, and *Syzygium samarangense*) that are tropical and subtropical fruits [44].

In this chapter, we expound on the microsatellite capture sequencing method for detecting SSR regions using high-throughput DNA sequencing based on our developed protocol.

2. Overview of microsatellite capture sequencing method

The microsatellite capture sequencing method based on our protocol is explained in this section (**Figure 1**). Some procedures in this protocol are optional fragmentation, SSR enrichment, and data analysis using merge paired-end read with a short-read sequencer.

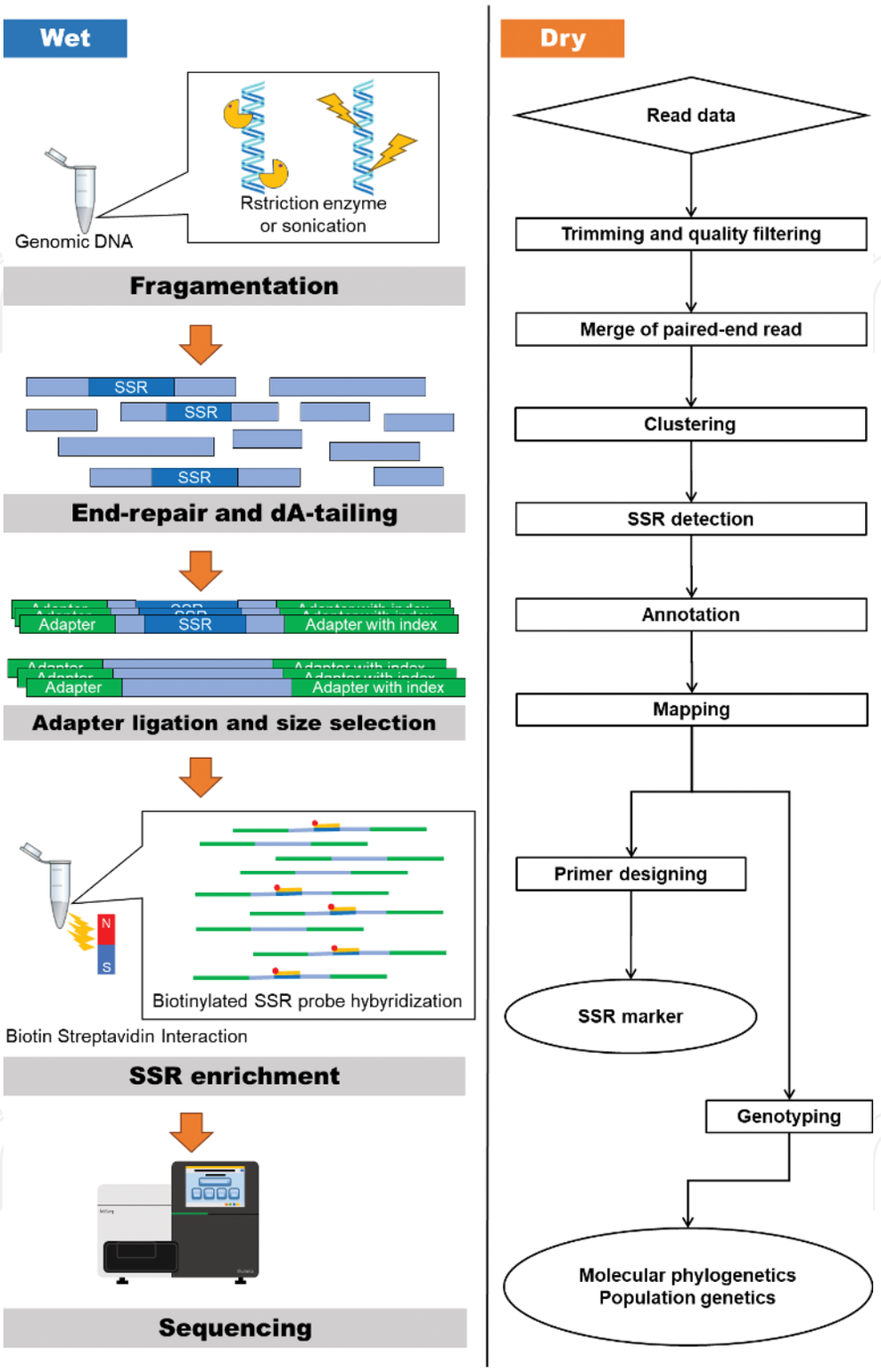

Figure 1. Workflow of our simple sequence repeat (SSR) detection method using microsatellite capture sequencing.

The purity and yield of extracted DNA are determined based on the absorbance ratios of 230/260 and 260/280 nm detected using a spectrometer and should be >1.5. In addition, a single band without smear should be obtained in 1% agarose gel electrophoresis. Subsequently, 1000 ng of the DNA is fragmented by digestion with the appropriate restriction enzyme or by

shearing to an average fragment size of 500 bp using an adaptive focused acoustics sonicator (Covaris, Woburn, MA, USA). The fragmented DNA is purified using a QIAquick PCR purification kit (Qiagen, Hilden, Germany), and then a standard Illumina NGS library is constructed using end-repair, dA-tailing, adaptor ligation, size selection, and PCR. We suggest using the NEBNext Ultra DNA Library Prep kit for Illumina (New England Biolabs, Ipswich, MA, USA) with DNA samples ≥10 ng, while the KAPA Hyper Prep kit for Illumina (Kapa Biosystems, Woburn, MA, USA) is recommended for DNA sample < 10 ng. Size selection is conducted using AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) with the approximate insert size set to 400–600 bp. The adaptor-ligated DNA is amplified through 15 high-fidelity PCR cycles. Subsequently, the PCR product is purified in a 20-μL volume via a cleanup stem using AMPure XP magnet beads (Beckman Coulter).

The purified product is mixed with 1 μL of a customized biotinylated SSR probe $(GA)_{10}$ from a 100 μM stock in TE buffer (one of the probes is typically used for SSR enrichment), incubated at 95°C for 10 min, and then placed on ice. The mixture is hybridized by incubating at 60°C for 60 min. After washing 20 μL of the Dynabeads MyOne streptavidin C1 beads (Life Technologies, Carlsbad, CA, USA), they are resuspended in 29 μL of 6× SSC buffer, added to each hybridized mixture, and incubated at 25°C for 30 min. The mixture is washed with 2× SSC buffer (once) and 1× SSC buffer (twice). Next, 23 μL of the SSR-enriched library is amplified using 15 high-fidelity PCR cycles with index primers. Subsequently, the amplified product is purified to a 20-μL volume via a cleanup step using AMPure XP magnetic beads (Beckman Coulter). The library quality and concentration are assessed using an Agilent Bioanalyzer 2100 (Agilent Technologies, Waldbronn, Germany) and Agilent DNA 1000 kit (Agilent Technologies). The specific concentration of each library is determined using quantitative real-time PCR using a KAPA library quantification kit (Kapa Biosystems). The library is first diluted to a concentration of 10 nM and then mixed in equal amounts. After denaturation with 0.2 N NaOH, the final concentration of the library mixture is diluted to 15 pM, including the 1% PhiX library (Illumina, CA, USA). The library mixture is sequenced using 2 × 300 bp paired-end sequencing using a MiSeq (Illumina). Reads in the FASTQ format were generated using a pipeline MiSeq reporter (version 2.5.1.3, Illumina).

Raw reads containing adaptors are removed using Trimmomatic version 0.32 [45]. Additionally, the FASTX-Toolkit version 0.0.13.2 [46] is used to clip uncertain bases called "N" and filter reads based on the quality score. The parameters of the quality filtering are as follows: (1) required minimum quality score is 20 and (2) minimum percentage of bases that must have [−q] quality is 80. The unpaired reads are then removed from the total remaining using a custom Perl script, and the preprocessed paired reads are integrated using the FLASh version 1.2.11 [47]. Furthermore, the integrated reads with similar sequences are clustered using the CD-HIT-EST version 4.6 [48], and the clustered reads including SSR regions are searched using the SSRIT for a stand-alone version (ftp://ftp.gramene.org/pub/gramene/archives/software/scripts/ssr.pl) [49]. The search parameters are (1) unit size, 2 and (2) minimum repeats, 10. Subsequently, sequences with a length of more than 20 bp flanking the SSR region are listed to enable the design of the primer set. The listed sequences are annotated with the top-hit description using the local BLAST program with the following settings: (1) execution program, BLASTn; (2) database, the NCBI nonredundant nucleotide database nt; and (3) e-value, 1e−4. Additionally, consensus sequences from these sequences are constructed using read mapping

in CLC Genomics Workbench 9.5 (CLC Bio-Qiagen, Aarhus, Denmark) with the following settings: (1) mismatch cost, 2; (2) insertion cost, 3; (3) deletion cost, 3; (4) length fraction, 0.5; and (5) similarity fraction, 0.8.

3. Available amount of DNA sample

Our protocol recommends using 1000 ng of the DNA sample. However, occasionally, only low amounts of DNA are obtained depending on the experiment design. Therefore, we adjusted the amounts to 1000, 100, 10, and 1 ng with rice (*Oryza sativa*) DNA as a test sample and investigated the feasibility of using these amount of DNA samples in our protocol (**Figure 2**). The above amounts (1000, 100, and 10 ng) of DNA samples were used with the NEBNext Ultra DNA Library Prep kit for Illumina (New England Biolabs), and 1 ng of the DNA sample was used with the KAPA Hyper Prep kit for Illumina (Kapa Biosystems). The processed DNA was quantified using an Agilent Bioanalyzer 2100 (Agilent Technologies) and Agilent DNA 1000 kit (Agilent Technologies) before and after SSR enrichment. As a result, the concentration of the processed DNA before SSR enrichment (after standard NGS library construction) was determined to be in the range of 226.9–14.7 nM in proportion to the input DNA, while that after SSR enrichment was gradually detected in the 23.5–3.4 nM range, although the concentration was reduced compared to the input DNA. Although the peaks after SSR enrichment with 10 and 1 ng input DNA could not be confirmed, they were checked using the Agilent high sensitivity DNA kit (Agilent Technologies). Therefore, our protocol can construct the SSR enriched library for high-throughput DNA sequencing if the prepared input DNA is ≥1 ng. If the constructed library does not meet the ≥1 nM required concentration for high-throughput DNA sequencing,

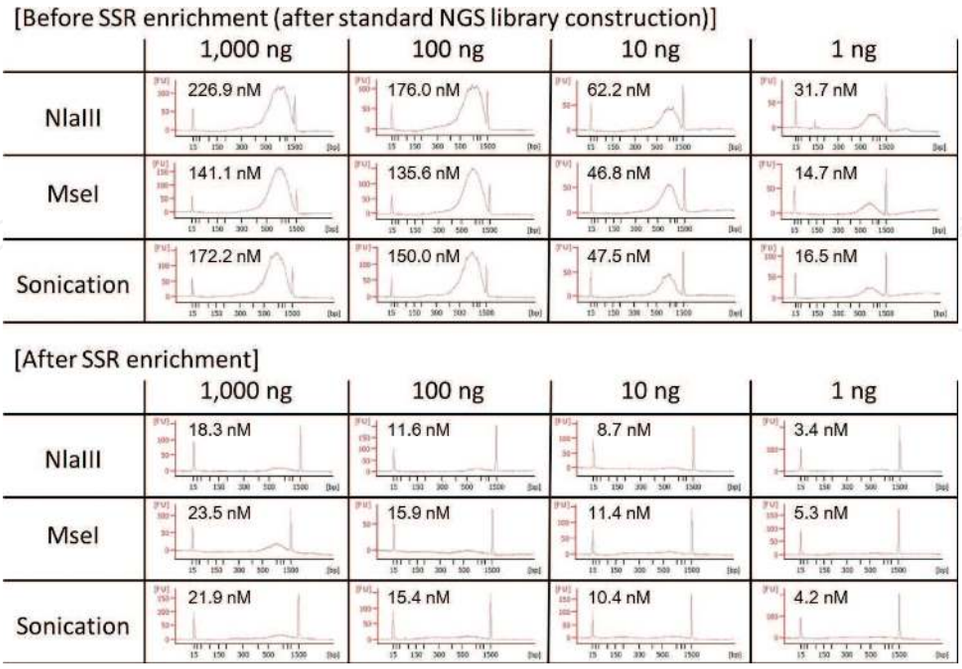

Figure 2. DNA profiling before and after simple sequence repeat (SSR) enrichment using Agilent Bioanalyzer 2100.

we suggest the following approaches: (1) elution with less buffer volume by re-performing the cleanup, (2) enrichment using an evaporator, and (3) several additional PCR cycles. The library constructed on a 1-ng scale may be useful for analyzing a few valuable samples such as cell masses with microsatellite instability and herbarium specimens.

4. Effect of fragmentation

Our protocol chooses between restriction enzyme digestion and sonication for the DNA fragmentation. In this experiment, the effect of data analysis on these different fragmentation methods was investigated using the sequence data (accession number DRA004725) for the Myrtaceae plants with the microsatellite capture sequencing.

During the integration process, the minimum overlapping length parameter was appropriated at 10-base intervals (min 10, 20, 30, 40, 50, and 60; **Figure 3**). The result showed that after integration and clustering, the integrated reads tended to be higher after fragmentation by sonication. Of the two restriction enzymes, MseI yielded more integrated reads than NlaIII did. The recognition site of MseI consists of only adenine and thymine (5′-T|TAA-3′), whereas that of NlaIII consists of all nucleotides (5′-CATG|-3′). Although the GC content in the whole genome is lower than 50% for many plants [50], Myrtaceae species also tend to exhibit low GC content [51, 52]. Thus, these results showed that the varying number of integrated reads obtained when different restriction enzymes were used was reasonable. Additionally, we compared variations in the integrated reads between the minimum overlapping parameters in the integration process. At min 10, the integrated reads constructed from the original paired-end reads were 25–43, 33–52, and 44–61% for NlaIII, MseI, and sonication, respectively while the corresponding values at min 60 were 19–32, 25–40, and 37–54%, respectively. Therefore, integrated reads ranging from several percentage points to approximately 10% at most could be varied by setting an arbitrary

		Feijoa sellowiana	*Myrciaria dubia*	*Psidium guajava*	*Psidium littorale*	*Syzygium samarangense*
NlaIII	min 10	364,353	429,099	440,662	597,423	452,422
	min 20	362,244	424,874	438,219	594,444	448,987
	min 30	360,190	420,444	435,790	591,219	445,546
	min 40	343,629	402,259	414,492	563,880	423,071
	min 50	309,712	366,615	372,452	506,293	378,749
	min 60	276,568	330,397	331,632	449,655	336,717
MseI	min 10	539,900	506,820	535,215	775,675	696,281
	min 20	537,164	504,452	532,637	773,343	693,006
	min 30	534,843	501,980	530,254	770,934	689,688
	min 40	513,778	487,128	513,130	748,055	665,606
	min 50	459,069	440,597	464,386	674,119	589,221
	min 60	406,394	390,403	413,059	594,978	516,492
Sonication	min 10	678,512	634,296	580,364	817,674	624,168
	min 20	677,383	632,166	578,307	815,020	622,443
	min 30	676,424	630,404	576,477	812,641	621,040
	min 40	660,327	619,743	568,532	799,495	611,014
	min 50	613,351	592,336	546,142	762,020	582,296
	min 60	559,465	559,277	517,957	715,415	547,505

Figure 3. Number of contigs after integration of paired-end reads and deletions of duplicated integrated-reads.

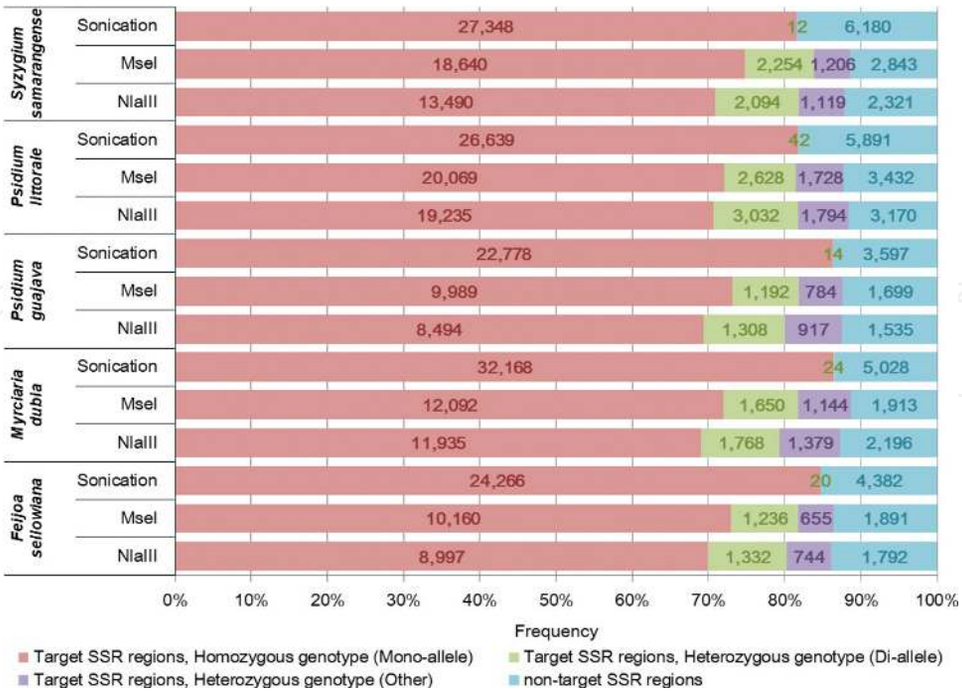

Figure 4. Simple sequence repeat (SSR) proportions of target and other regions.

parameter for the minimum overlapping length. We recommend using an overlapping length parameter of min 60 to select integrated reads with higher reliability.

After searching SSR regions of the clustered reads using the overlapping length parameter of min 60, most motifs (82–89%) were target SSR regions [$(CT)_n$, $(TC)_n$, $(GA)_n$, or $(AG)_n$; **Figure 4**]. This result could be attributed to the biotinylated probe used to enrich the SSR region. The various probe conditions required for capturing SSR regions have been reported previously [53–55]. Our protocol also showed that the target SSR region could be captured efficiently. Among the SSRs shown in **Figure 4**, probe-related SSR regions were characterized based on genotypic frequency. All species and the fragmentation conditions showed high and low rates of homozygous and heterozygous genotypes, respectively. The heterozygous genotype rate for all species was substantially higher after fragmentation using restriction enzymes than it was after fragmentation using sonication (18.75–20.86, 15.66–18.77, and 0.04–0.16% for NlaIII, MseI, and sonication, respectively). Additionally, fragmentation by NlaIII was more likely to detect the heterozygous genotype than that by MseI was and, thus, for our approach we recommend fragmentation using restriction enzyme digestion. Although the five Myrtaceae plants analyzed in this study are diploid, approximately one-third of heterozygous genotypes were more than tri-allelic in all species. This factor may be associated with the occurrence of multiple homologous copies or PCR error during library construction.

We confirmed the unique and common genes with SSR flanking regions in a family based on annotations (**Figure 5**). In *F. sellowiana*, 372 (NlaIII), 337 (MseI), and 887 (sonication) SSR regions were fragmentation-specific, whereas 440 (NlaIII), 474 (MseI), and 624 (sonication) SSR regions were common among other fragmentations. In *M. dubia*, 338 (NlaIII), 320 (MseI), and 1065 (sonication) SSR regions were fragmentation-specific, whereas 481 (NlaIII), 482 (MseI), and 724 (sonication) SSR regions were common among other fragmentations. In *P. guajava*, 227 (NlaIII),

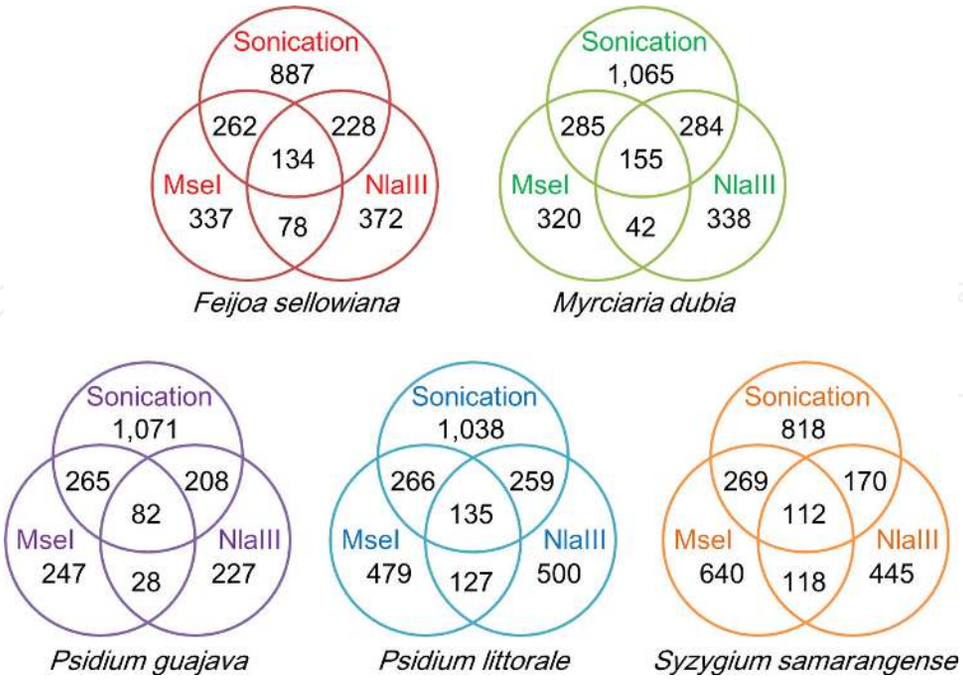

Figure 5. Venn diagram based on annotation within three fragmentations for each Myrtaceae plant.

247 (MseI), and 1071 (sonication) SSR regions were fragmentation-specific, whereas 318 (NlaIII), 375 (MseI), and 555 (sonication) SSR regions were common among other fragmentations. In *P. littorale*, 500 (NlaIII), 479 (MseI), and 1038 (sonication) SSR regions were fragmentation-specific, whereas 521 (NlaIII), 528 (MseI), and 660 (sonication) SSR regions were common among other fragmentations. In *S. samarangense*, 445 (NlaIII), 640 (MseI), and 818 (sonication) SSR regions were fragmentation-specific, whereas 400 (NlaIII), 499 (MseI), and 551 (sonication) SSR regions were common among other fragmentations. Therefore, the detected SSR flanking region had both fragmentation-specific and common characteristics. Notably, sonication yielded the most characteristics for all groups.

We constructed consensus sequences from the listed sequences including SSRs based on the read mapping. The percentage values of unsuited consensus sequences for molecular marker development determined by including the unknown nucleotide "N" were 4.4% (NlaIII), 5.3% (MseI), and 1.4% (sonication) in *F. sellowiana*; 5.7% (NlaIII), 5.4% (MseI), and 3.5% (sonication) in *M. dubia*; 9.7% (NlaIII), 10.7% (MseI), and 5.3% (sonication) in *P. guajava*; 6.5% (NlaIII), 5.7% (MseI), and 1.8% (sonication) in *P. littorale*; and 7.0% (NlaIII), 7.9% (MseI), and 1.3% (sonication) in *S. samarangense*. Conversely, approximately 90% of the consensus sequences could be candidate molecular markers for some gene and trait.

Fragmentation by restriction enzyme is limited in the restriction site flanking region, whereas fragmentation by sonication targets the whole genome. The comparison of different fragmentation methods for genomic DNA revealed that restriction enzymes were more suitable for identifying the heterozygous genotype than sonication was, whereas sonication facilitated the detection of various SSR flanking regions with both species-specific and common characteristics more than restriction enzyme digestion did. Therefore, the choice of the DNA fragmentation approach appears to depend on the ultimate research purpose. In particular, the effective

detection of a heterozygous genotype using DNA fragmentation by restriction enzymes is expected to contribute to the development of molecular markers for molecular-assisted breeding and population genetics that require the clear distinction of alleles.

5. Simulation analysis using various sequence data scales

The number of paired reads for the Myrtaceae plants ranged from 2 × 1,296,448 to 2 × 1,708,634. Here, various sequence data scales are simulated to demonstrate how SSRs can be detected (**Figure 6**). For the simulation, different scale data consisting of 1000, 500, 200, 100, and 50K reads were prepared by sampling without replacement of the original sequence data of the five Myrtaceae plants and three fragmentation approaches. Additionally, paired reads of each scale dataset were integrated using the minimum overlapping parameter of 10-base intervals (min 10, 20, 30, 40, 50, and 60), and sequences with SSR regions were detected from the integrated reads after clustering the same sequences. The simulation result showed that the 1000K reads scale detected tens of thousands of SSRs; 500 and 200K reads detected a few thousand to tens of thousands of SSRs; and 100 and 50K reads detected a few thousand SSRs. Thus, the 50K reads detected approximately 1/10 of the SSRs detected by the 1000K reads. Moreover, 384 samples could be used in a single assay when a sequence of 50K reads per sample is assumed, and the sequence cost is very reasonable.

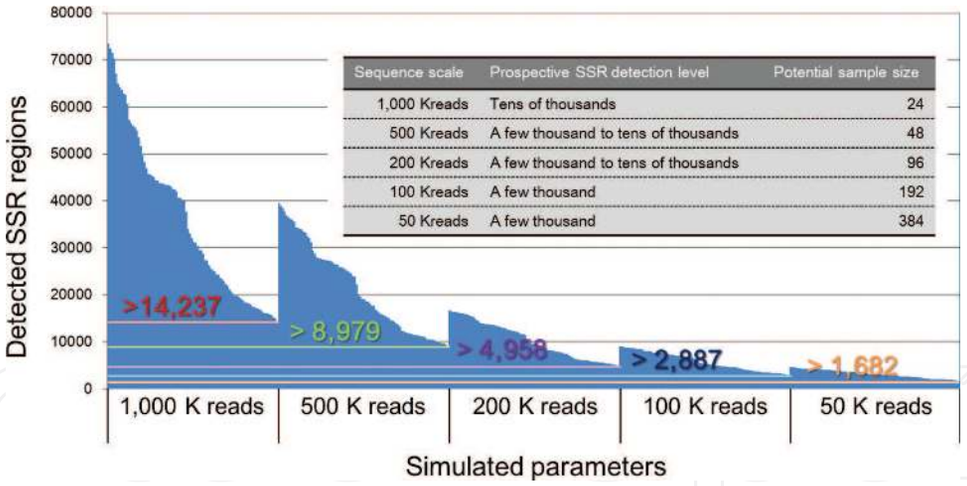

Figure 6. Simulation result based on assuming various read scales.

6. *In silico* polymorphic detection and phylogenetic analysis

SSRs detected using microsatellite capture sequencing are available as SSR markers by designing primer sets from the SSR flanking region. On the other hand, when common sequences with SSR regions among samples are prepared as reference sequences, the sequence data of each sample can be mapped to the reference, SSR polymorphisms can be detected among the

samples, and phylogenetic analysis is possible based on the polymorphic data. Here, we explain an *in silico* polymorphic detection and phylogenetic analysis method based on microsatellite capture sequencing data.

According to the protocol described above, the sequence set with the SSR region is prepared from analyzed data using paired-read integration, clustering of same sequences, and SSR detection. The sequence sets of each sample are merged into one file, and the merged sequence sets are re-clustered using CD-HIT-EST version 4.6 [48]. The sequence data of each sample are mapped to the clustered sequence as a reference by using the CLC Genomics Workbench 9.5 (CLC Bio-Qiagen, Aarhus, Denmark). The consensus sequences of each sample are created from the mapped data. The SSR repeat data of the consensus sequences are detected using SSRIT for a stand-alone version [49]. A polymorphic table merging the SSR detected data of each sample is constructed using the following script (merge_SSR.pl):

```
#!/usr/bin/perl.

use strict;

our @hashlist = ();

our @fnlist = ();

our %keyhash = ();

eval {.

        if($#ARGV <0) {

                print “usage: mergeSSR.pl [sample tsv files (ex. mergeSSR.pl *.txt)]\n”;

                exit −1;

         }

         ### header ########################################

         print “Locus”;

         for (my $i = 0; $i < = $#ARGV; $i++){

                my $filename = $ARGV[$i];

                print “\t”.$filename;

                push(@fnlist, $filename);

        }

        print “\n”;

        ### file to hash Locus ########################################

        for (my $i = 0; $i < = $#fnlist; $i++){

                my $filename = $fnlist[$i];
```

```
                my %hash = ();
                open (IN," < $filename") || die "cannot open $filename: $!";
                while (my $line = <IN>) {
                        chomp $line;
                        my @dt = split/\t/,$line;
                        my $key = $dt[0]."_".$dt[1];
                        my $val = $dt[4];
                        $hash{$key} = $val;
                        $keyhash{$key} = $key;
                }
                close (IN1);
                push(@hashlist, \%hash);
  }
  #### data ##############################################
  foreach my $key (keys %keyhash){
          print $key;
          foreach my $row (@hashlist) {
                  if(exists $row- > {$key}){
                          print "\t".$row- > {$key};
                  } else {.
                          print "\t";
                  }
          }
        print "\n";
  }
   exit 0;
};
```

The polymorphic table is edited to the input data of the Populations format. The genetic distance between samples is calculated using a distance matrix method using the Populations

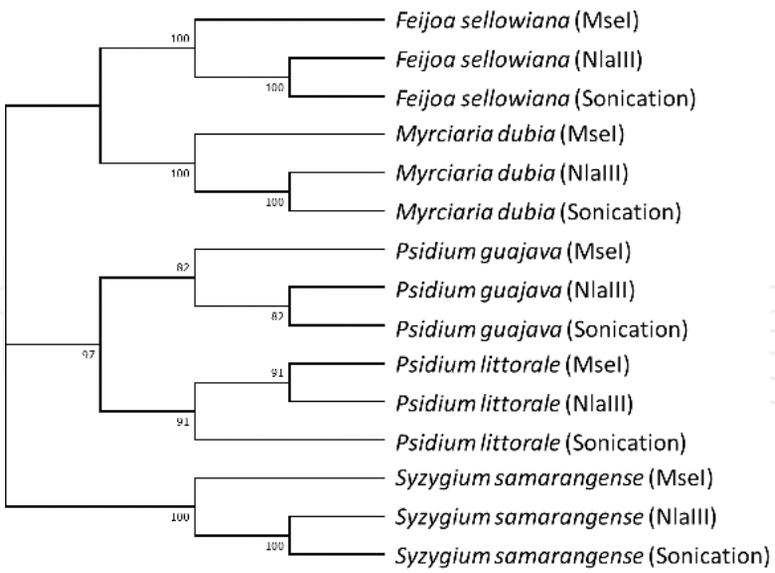

Figure 7. Neighbor-joining dendrogram (only topology) constructed from *in silico* polymorphic data of five Myrtaceae plants.

version 1.2.30 [56]. A dendrogram is drawn using the MEGA version 7 [57]. For example, we have shown the result of the phylogenetic analysis of the Myrtaceae plants (**Figure 7**). SSR polymorphisms in 38,636 loci were compared between samples, and the result showed that all organisms had a single clade even if the fragmentation differed.

7. Conclusion

Recently, the SSR detection approach using high-throughput DNA sequencing has been performed on various organisms (**Table 1**). Although SSRs are detected from the whole genome as a part of the data analyses in many cases, the microsatellite capture sequencing approach included in our protocol can detect numerous SSRs more effectively by enriching NGS library using an SSR probe than conventional approaches. Detected SSR data will considerably increase the spread of NGS in the future. Therefore, the construction of a database will be required to manage the massive amount of SSR data.

Acknowledgements

This work was supported by the Ministry of Education, Culture, Sports, Science, and Technology-Supported Program for the Strategic Research Foundation at Private Universities (S1311017). We thank Hironobu Uchiyama and Eri Kubota for their invaluable advice and Satoshi Sano, Hiroto Sekitoh, and Yu Hamaguchi for technical support.

Author details

Keisuke Tanaka[1], Rumi Ohtake[1], Saki Yoshida[2] and Takashi Shinohara[3]*

*Address all correspondence to: t3shinoh@nodai.ac.jp

1 NODAI Genome Research Center, Tokyo University of Agriculture, Tokyo, Japan

2 International Agricultural Development, Graduate School of Agriculture, Tokyo University of Agriculture, Tokyo, Japan

3 Junior College of Tokyo University of Agriculture, Tokyo University of Agriculture, Tokyo, Japan

References

[1] Ijaz S. Microsatellite markers: An important fingerprinting tool for characterization of crop plants. African Journal of Biotechnology. 2011;**10**:7723-7726

[2] Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: An overview of the recent progress in plants. Euphytica. 2011;**177**:309-334. DOI: 10.1007/s10681-010-0286-9

[3] Rassmann K, Schlötterer C, Tautz D. Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. Electrophoresis. 1991;**12**:113-118. DOI: 10.1002/elps.1150120205

[4] Karagyozov L, Kalcheva ID, Chapman VM. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research. 1993;**21**:3911-3912. DOI: 10.1093/nar/21.16.3911

[5] Connell JP, Pammi S, Iqbal MJ, Huizinga T, Reddy AS. A high through-put procedure for capturing microsatellites from complex plant genomes. Plant Molecular Biology Reporter. 1998;**16**:341-349. DOI: 10.1023/A:1007536421700

[6] Lian C, Hogetsu T. Development of microsatellite markers in black locust (*Robinia pseudoacacia*) using a dual-suppression-PCR technique. Molecular Ecology Resources. 2002;**2**:211-213. DOI: 10.1046/j.1471-8286.2002.00213.x-i2

[7] Castoe TA, Poole AW, Gu W, Jason de Koning AP, Daza JM, Smith EN, Pollock DD. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Molecular Ecology Resources. 2010;**10**:341-347. DOI: 10.1111/j.1755-0998.2009.02750.x

[8] Cai G, Leadbetter CW, Muehlbauer MF, Molnar TJ, Hillman BI. Genome-wide microsatellite identification in the fungus *Anisogramma anomala* using Illumina sequencing and genome assembly. PLoS One. 2013;**8**:e82408. DOI: 10.1371/journal.pone.0082408

[9] Heras S, Planella L, Caldarazzo I, Vera M, García-Marín JL, Roldán MI. Development and characterization of novel microsatellite markers by next generation sequencing for the blue and red shrimp *Aristeus antennatus*. PeerJ. 2016;**4**:e2200. DOI: 10.7717/peerj.2200

[10] Bastías A, Correa F, Rojas P, Almada R, Muñoz C, Sagredo B. Identification and characterization of microsatellite loci in maqui (*Aristotelia chilensis* [Molina] Stunz) using next-generation sequencing (NGS). PLoS One. 2016;**11**:e0159825. DOI: 10.1371/journal.pone.0159825

[11] De Bellis F, Malapa R, Kagy V, Lebegin S, Billot C, Labouisse JP. New development and validation of 50 SSR markers in breadfruit (*Artocarpus altilis*, Moraceae) by next-generation sequencing. Applications in Plant Sciences. 2016;**4**:e1600021. DOI: 10.3732/apps.1600021

[12] Huang D, Zhang Y, Jin M, Li H, Song Z, Wang Y, Chen J. Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *Aspidistra saxicola* (Asparagaceae). Molecular Ecology Resources. 2014;**14**:569-577. DOI: 10.1111/1755-0998.12197

[13] Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics. 2013;**14**:e17. DOI: 10.1186/1471-2164-14-17

[14] Duan CX, Li DD, Sun SL, Wang XM, Zhu ZD. Rapid development of microsatellite markers for *Callosobruchus chinensis* using illumina paired-end sequencing. PLoS One. 2014;**9**:e95458. DOI: 10.1371/journal.pone.0095458

[15] Mudalkar S, Golla R, Ghatty S, Reddy AR. *De novo* transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant Molecular Biology. 2014;**84**:159-171. DOI: 10.1007/s11103-013-0125-1

[16] Tan LQ, Wang LY, Wei K, Zhang CC, LY W, Qi GN, Cheng H, Zhang Q, Cui QM, Liang JB. Floral transcriptome sequencing for SSR marker development and linkage map construction in the tea plant (*Camellia sinensis*). PLoS One. 2013;**8**:e81611. DOI: 10.1371/journal.pone.0081611

[17] Ambreen H, Kumar S, Variath MT, Joshi G, Bali S, Agarwal M, Kumar A, Jagannath A, Goel S. Development of genomic microsatellite markers in *Carthamus tinctorius* L. (safflower) using next generation sequencing and assessment of their cross-species transferability and utility for diversity analysis. PLoS One. 2015;**10**:e0135443. DOI: 10.1371/journal.pone.0135443

[18] Curto MA, Tembrock LR, Puppo P, Nogueira M, Simmons MP, Meimberg H. Evaluation of microsatellites of *Catha edulis* (qat; Celastraceae) identified using pyrosequencing. Biochemical Systematics and Ecology. 2013;**49**:1-9. DOI: 10.1016/j.bse.2013.02.002

[19] Sahu BP, Sahoo L, Joshi CG, Mohanty P, Sundaray JK, Jayasankar P, Das P. Isolation and characterization of polymorphic microsatellite loci in Indian major carp, *Catla catla* using next-generation sequencing platform. Biochemical Systematics and Ecology. 2014;**57**:357-362. DOI: 10.1016/j.bse.2014.09.010

[20] Wang H, Jiang J, Chen S, Qi X, Peng H, Li P, Song A, Guan Z, Fang W, Liao Y, Chen F. Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale unigene assembly and SSR marker discovery. PLoS One. 2013;**8**:e62293. DOI: 10.1371/journal.pone.0062293

[21] Lee JH, Cho WB, Yang S, Han EK, Lyu ES, Kim WJ, Moon BC, Choi G. Development and characterization of 21 microsatellite markers in *Daphne kiusiana*, an evergreen broad-leaved

shrub endemic to Korea and Japan. Korean J Pl Taxon. 2017;**47**:6-10. DOI: 10.11110/kjpt.2017.47.1.6

[22] Morillo E, Buitron J, Limongi R, Vignes H, Argout X. Characterization of microsatellites identified by next-generation sequencing in the Neotropical tree *Handroanthus billbergii* (Bignoniaceae). Applications in Plant Sciences. 2016;**4**:e1500135. DOI: 10.3732/apps.1500135

[23] Yu JN, Won C, Jun J, Lim Y, Kwak M. Fast and cost-effective mining of microsatellite markers using NGS technology: An example of a Korean water deer *Hydropotes inermis argyropus*. PLoS One. 2011;**6**:e26933. DOI: 10.1371/journal.pone.0026933

[24] Abdelkrim J, Robertson B, Stanton JA, Gemmell N. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques. 2009;**46**: 185-192. DOI: 10.2144/000113084

[25] Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics. 2010;**11**:e726. DOI: 10.1186/1471-2164-11-726

[26] Yang T, Jiang J, Burlyaeva M, Hu J, Coyne CJ, Kumar S, Redden R, Sun X, Wang F, Chang J, Hao X, Guan J, Zong X. Large-scale microsatellite development in grasspea (*Lathyrus sativus* L.), an orphan legume of the arid areas. BMC Plant Biology. 2014;**14**:e65. DOI: 10.1186/1471-2229-14-65

[27] Ravishankar KV, Dinesh MR, Nischita P, Sandya BS. Development and characterization of microsatellite markers in mango (*Mangifera indica*) using next-generation sequencing technology and their transferability across species. Molecular Breeding. 2015;**35**:e93. DOI: 10.1007/s11032-015-0289-2

[28] Ho CW, TH W, Hsu TW, Huang JC, Huang CC, Chiang TY. Development of 12 genic microsatellite loci for a biofuel grass, *Miscanthus sinensis* (Poaceae). American Journal of Botany. 2011;**98**:e201-e203. DOI: 10.3732/ajb.1100071

[29] Allentoft M, Schuster SC, Holdaway R, Hale M, McLay E, Oskam C, Gilbert MT, Spencer P, Willerslev E, Bunce M. Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. BioTechniques. 2009;**46**:195-200. DOI: 10.2144/000113086

[30] Liu MX, Xu Y, Yang TY, Qiao ZJ, Wang RY, Wang YY, Lu P. Development of species-specific microsatellite markers for broomcorn millet (*Panicum miliaceum* L.) via high-throughput sequencing. Advances in Crop Science and Technology. 2017;**5**:e297. DOI: 10.4172/2329-8863.1000297

[31] Wang Y, Zeng X, Iyer NJ, Bryant DW, Mockler TC, Mahalingam R. Exploring the switchgrass transcriptome using second-generation sequencing technology. PLoS One. 2012;**7**: e34225. DOI: 10.1371/journal.pone.0034225

[32] Bonatelli IA, Carstens BC, Moraes EM. Using next generation RAD sequencing to isolate multispecies microsatellites for *Pilosocereus* (Cactaceae). PLoS One. 2015;**10**:e0142602. DOI: 10.1371/journal.pone.0142602

[33] Yang T, Fang L, Zhang X, Hu J, Bao S, Hao J, Li L, He Y, Jiang J, Wang F, Tian S, Zong X. High-throughput development of SSR markers from pea (*Pisum sativum* L.) based on next generation sequencing of a purified Chinese commercial variety. PLoS One. 2015;**10**: e0139775. DOI: 10.1371/journal.pone.0139775

[34] Wang H, Walla JA, Zhong S, Huang D, Dai W. Development and cross-species/genera transferability of microsatellite markers discovered using 454 genome sequencing in chokecherry (*Prunus virginiana* L.). Plant Cell Reports. 2012;**31**:2047-2055. DOI: 10.1007/s00299-012-1315-z

[35] Lü Z, Li H, Liu L, Cui W, Hu X, Wang C. Rapid development of microsatellite markers from the large yellow croaker (*Pseudosciaena crocea*) using next generation DNA sequencing technology. Biochemical Systematics and Ecology. 2013;**51**:314-319. DOI: 10.1016/j.bse.2013.09.019

[36] Hunter ME, Hart KM. Rapid microsatellite marker development using next generation pyrosequencing to inform invasive Burmese python—*Python molurus bivittatus*—Management. International Journal of Molecular Sciences. 2013;**14**:4793-4804. DOI: 10.3390/ijms14034793

[37] Kang JH, Park JY, Jo HS. Rapid development of microsatellite markers with 454 pyrosequencing in a vulnerable fish, the mottled skate, *Raja pulchra*. International Journal of Molecular Sciences. 2012;**13**:7199-7211. DOI: 10.3390/ijms13067199

[38] Zhai L, Xu L, Wang Y, Cheng H, Chen Y, Gong Y, Liu L. Novel and useful genic-SSR markers from *de novo* transcriptome sequencing of radish (*Raphanus sativus* L.). Molecular Breeding. 2014;**33**:611-624. DOI: 10.1007/s11032-013-9978-x

[39] Angeloni F, Wagemaker CA, Jetten MS, Op den Camp HJ, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Ouborg NJ. *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. Molecular Ecology Resources. 2011;**11**:662-674. DOI: 10.1111/j.1755-0998.2011.02990.x

[40] Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics. 2011;**12**:e451. DOI: 10.1186/1471-2164-12-451

[41] Yang T, Bao SY, Ford R, Jia TJ, Guan JP, He YH, Sun XL, Jiang JY, Hao JJ, Zhang XY, Zong XX. High-throughput novel microsatellite marker of faba bean via next generation sequencing. BMC Genomics. 2012;**13**:e602. DOI: 10.1186/1471-2164-13-602

[42] Kang JS, Lee B, Kwak M. Isolation and characterization of microsatellite markers for *Viola mirabilis* (Violaceae) using a next-generation sequencing platform. Plant Species Biology. 2017;**32**:448-454. DOI: 10.1111/1442-1984.12159

[43] Rai MK, Shekhawat NS. Genomic resources in fruit plants: An assessment of current status. Critical Reviews in Biotechnology. 2015;**35**:438-447. DOI: 10.3109/07388551.2014.898127

[44] Tanaka K, Ohtake R, Yoshida S, Shinohara T. Effective DNA fragmentation technique for simple sequence repeat detection with a microsatellite-enriched library and high-throughput sequencing. BioTechniques. 2017;**62**:180-182. DOI: 10.2144/000114536

[45] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**:2114-2120. DOI: 10.1093/bioinformatics/btu170

[46] Gordon A, Hannon GJ. FASTX-Toolkit. In: FASTQ/A Short-Reads Pre-Processing Tools. 2010. Available from: http://hannonlab.cshl.edu/fastx_toolkit/ Accessed: 2017–8-31

[47] Magoč T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;**27**:2957-2963. DOI: 10.1093/bioinformatics/btr507

[48] Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;**22**:1658-1659. DOI: 10.1093/bioinformatics/btl158

[49] Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Research. 2001;**11**:1441-1452. DOI: 10.1101/gr.184001

[50] Meister A, Barow M. DNA base composition of plant genomes. In: Doležel J, Greilhuber J, Suda J, editors. Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes. Weinheim: Wiley; 2007. pp. 177-215. DOI: 10.1002/9783527610921.ch8

[51] Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, Tibbits J. Chloroplast genome analysis of Australian eucalypts—Eucalyptus, corymbia, angophora, allosyncarpia, and stockwellia (Myrtaceae). Molecular Phylogenetics and Evolution. 2013;**69**:704-716. DOI: 10.1016/j.ympev.2013.07.006

[52] Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, Shimizu KK, Isagi Y. Genome sequencing of *Metrosideros polymorpha* (Myrtaceae), a dominant species in various habitats in the Hawaiian islands with remarkable phenotypic variations. Journal of Plant Research. 2016;**129**:727-736. DOI: 10.1007/s10265-016-0822-3

[53] Acquadro A, Portis E, Lanteri S. Isolation of microsatellite loci in artichoke (*Cynara cardunculus* L. Var. *scolymus*). Molecular Ecology Resources. 2003;***3***:37-39. DOI: 10.1046/j.1471-8286.2003.00343.x

[54] He G, Meng R, Newman M, Gao G, Pittman RN, Prakash CS. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biology. 2003;**3**:e3. DOI: 10.1186/1471-2229-3-3

[55] Stajner N, Jakse J, Kozjak P, Javornik B. The isolation and characterisation of microsatellites in hop (*Humulus lupulus* L.). Plant Science. 2005;**168**:213-221. DOI: 10.1016/j.plantsci.2004.07.031

[56] Langella O. Populations, 1.2.28. 1999. Available from: http://bioinformatics.org/populations/ [Accessed: 2017-8-31]

[57] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;**33**:1870-1874. DOI: 10.1093/molbev/msw054