We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

4,200
Open access books available

116,000
International authors and editors

125M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
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

<table>
<thead>
<tr>
<th>Target species</th>
<th>Library style</th>
<th>NGS platform</th>
<th>Number of reads</th>
<th>Sequences including SSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agkistrodon contortrix (Copperhead snake)</td>
<td>WG</td>
<td>GS-FLX</td>
<td>128,773</td>
<td>14,612</td>
</tr>
<tr>
<td>Anisogramma anomala</td>
<td>WG</td>
<td>GAIIX</td>
<td>26,036,313</td>
<td>44,247</td>
</tr>
<tr>
<td>Aristaeus antennatus (Red shrimp)</td>
<td>WG</td>
<td>GS-FLX</td>
<td>165,507</td>
<td>247</td>
</tr>
<tr>
<td>Aristotelia chilensis (Maqui)</td>
<td>WG</td>
<td>GS-FLX</td>
<td>165,043</td>
<td>24,494</td>
</tr>
<tr>
<td>Artocarpus altis</td>
<td>WG</td>
<td>MiSeq</td>
<td>2,341,465</td>
<td>47,607</td>
</tr>
<tr>
<td>Aspidistra saxicola</td>
<td>cDNA</td>
<td>HiSeq2000</td>
<td>13,133,336</td>
<td>4764</td>
</tr>
<tr>
<td>Brachiaria ruzicizensis (Ruzigrass)</td>
<td>WG</td>
<td>GAII</td>
<td>186,764,108</td>
<td>139,098</td>
</tr>
<tr>
<td>Callosobruchus chinensis (Adzuki bean weevil)</td>
<td>WG</td>
<td>HiSeq2500</td>
<td>106,888,024</td>
<td>6593</td>
</tr>
<tr>
<td>Camelina sativa</td>
<td>cDNA</td>
<td>GAII</td>
<td>10,830,000</td>
<td>14,140</td>
</tr>
<tr>
<td>Camellia sinensis (Tea plant)</td>
<td>cDNA</td>
<td>HiSeq2000</td>
<td>26,874,116</td>
<td>5649</td>
</tr>
<tr>
<td>Carthamus tinctorius (Safflower)</td>
<td>WG</td>
<td>HiSeq2000</td>
<td>48,502,680</td>
<td>23,067</td>
</tr>
<tr>
<td>Catha edulis (Khat)</td>
<td>WG</td>
<td>GS-FLX</td>
<td>65,401</td>
<td>11,678</td>
</tr>
<tr>
<td>Catla catla (Catla)</td>
<td>WG</td>
<td>PCM</td>
<td>29,794</td>
<td>21,477</td>
</tr>
<tr>
<td>Chrysanthemum hankingense</td>
<td>cDNA</td>
<td>GAII</td>
<td>53,720,166</td>
<td>2813</td>
</tr>
<tr>
<td>Daphne kiusiana</td>
<td>WG</td>
<td>MiSeq</td>
<td>4,936,656</td>
<td>28,495</td>
</tr>
<tr>
<td>Handroanthus bilbergii</td>
<td>WG</td>
<td>MiSeq</td>
<td>2,169,901</td>
<td>61,074</td>
</tr>
<tr>
<td>Hydropotes inermis (Water Deer)</td>
<td>WG</td>
<td>GS-FLX</td>
<td>260,467</td>
<td>20,101</td>
</tr>
</tbody>
</table>
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 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.
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 step using AMPure XP magnetic 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 0.13.2 [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 preprocessing 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,
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

![Figure 3](http://dx.doi.org/10.5772/intechopen.72629)
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 4. Simple sequence repeat (SSR) proportions of target and other regions.
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 $2 \times 1,296,448$ to $2 \times 1,708,634$. 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.

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):

```perl
#!/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];
    print "\nfilename = $filename;"
    push(@fnlist, $filename);
  }

  print "\n";
```

Microsatellite Capture Sequencing
http://dx.doi.org/10.5772/intechopen.72629
my %hash = ();

open (IN, "<$filename") || die "cannot open $filename: $!";
while (my $line = <IN>) {
    chomp $line;
    my @dt = split/	/,$line;
    my $key = $dt[0] . "_." . $dt[1];
    my $val = $dt[4];
    $hash{$key} = $val;
    $keyhash{$key} = $key;
}
close (IN);
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
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\textsuperscript{1}, Rumi Ohtake\textsuperscript{1}, Saki Yoshida\textsuperscript{2} and Takashi Shinohara\textsuperscript{3}\textsuperscript{*}

\*Address all correspondence to: t3shinoh@noda.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


