REVIEW

MOLECULAR MARKERS FOR ANALYSIS OF PLANT GENETIC DIVERSITY

Vu Tuan Nam¹, Pham Le Bich Hang¹, Nguyen Nhat Linh¹, Luu Han Ly¹, Huynh Thi Thu Hue¹,², Nguyen Hai Ha¹,², Ha Hong Hanh¹, Le Thi Thu Hien¹,²

¹Institute of Genome Research, Vietnam Academy of Science and Technology
²Graduate University of Science and Technology, Vietnam Academy of Science and Technology

To whom correspondence should be addressed. E-mail: hienlethu@igr.ac.vn

Received: 27.4.2020
Accepted: 22.6.2020

SUMMARY

Genetic diversity plays an important role in diversity conservation at multiple levels and supports to monitor and assess genetic variation. In plants, genetic diversity provides the ability to adapt and respond to environmental conditions that helps plants to survive through changing environments. Genetic diversity analyses based on molecular genetic markers are effective tools for conservation and reintroduction of rare and endangered species. In recent years, the development of various chemical and molecular techniques for studying genetic diversity has received great attention. While biochemical markers are primarily used in the diagnosis of pathogens, DNA markers have been developed and widely applied for identification of species and population based on the genotype of an organism that is more stable and not easily affected by the environmental factors. PCR-based molecular marker tools, such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), simple sequence repeats (SSRs) are used for analysing the difference in the targeted DNA sequences. With the rapid and robust development of genomic sequencing technology it is now possible to obtain and analyse DNA sequences of the whole genome of studied organisms. However, each type of DNA markers has different principles, as well as the pros and cons of specificity. In this article, we review methods and point out DNA markers, which are considered as reliable and widely used tools for the detection of genetic variation. In addition, we present the application of DNA marker in analysing genetic diversity of wild, domestic and medicinal plants, as well as some perspectives on the future of DNA marker’s application in the analysis of genetic diversity.

Keywords: DNA sequencing analysis, genetic diversity, molecular markers, SNPs, SSRs.

INTRODUCTION

Genetic diversity is a part of biological diversity, which is referred to as any variation of genetic materials (nucleotides, genes, chromosomes, or genomes) in all organisms in the ecosystems that they occur for a given time (Ramanatha, Hodgkin, 2002; Fu, 2015). Genetic diversity is a crucial characteristic of any population or species, which included the total number of genetic characteristics in the genetic makeup of a species, as genes coding for biological traits adapting to environmental conditions (Nevo, 2001). Genetic diversity plays an important role in ecology because of several reasons. Firstly, genetic diversity within a single species may correlate with species diversity both within and between trophic levels (Wimp et al., 2004; Vellend, Geber, 2005; Crutsinger et al., 2006; Johnson et al., 2006; Vellend, 2006).
Secondly, in communities overwhelmed by a single species, hereditary variety inside that species may be as (or more) imperative as variety between species (Hughes, Stachowicz, 2004). One more reason, diversity may serve as a proxy for a complex multivariate phenotypic space that is often difficult to measure (Hughes et al., 2008). Indeed, for a run of plant species, test controls of the number of genotypes have been appeared to influence population, community, and biological system properties (Bailey et al., 2009).

Genetic diversity provides plants with the ability to adapt and response to environmental conditions, especially adverse factors as well as develop new and improved traits relating to yield and production, pest and disease resistance, and so on (Evenson, Gollin, 2003). This can play a very important role in providing adaptive genes that ultimately leads to long-term increase in food productivity related to environmental harm and also overcomes other types of phenomena such as genetic erosion (loss of genetic diversity), extinction of primitive, and adaptive genes (soil loss) caused by long-term exploitation of changes (Holderegger et al., 2006). New strains can be found to be more stable in adaption over time. Each individual is genetically unique, unlike any individual in the same community or population. Genetic diversity also contributes in diversity conservation at multiple levels and supports to monitor and assess genetic variation.

Serious agents relating to environment, such as climate change, environmental pollution, as well as population growth, along with human activities are the dominant pressures for genetic change. Climate change is one of the most significant factors that immediately affects to biodiversity at all levels including genetic diversity and particularly in agriculture (Schmidhuber, Tubiello, 2007). Food production systems rely on highly selective varieties in a better environment but it may be increasingly susceptible to the effects of climate change, such as pests and disease spread (Gornall et al., 2010). The current genetic composition of a crop affects how well its members will adapt to the physical and biological environments of the future. Rapid world population growth, along with migration in the world, are also some of the pressures to reduce the genetic diversity of living things, especially plants. Due to the need for living and farming space, humans have made the genome range narrower and more homogeneous leading to the life conditions become less favorable to exist and reproduce. Such populations are more likely to become locally extinct, and in extreme cases, all plant species may be at risk of extinction. Reducing genetic variation is a major step in the extinction process. Migration as well as population growth and unsustainable food production systems have been shown to cause a decrease in the variation of genetic loci (gene alleles) that controls physical and phenotypic responses to climate change (Jump, Peñuelas, 2005). Therefore, hereditary variations are considered as the factors that help plants to survive through changing environments. This is also the fundamental for genetic breeding (Xia et al., 2019).

Genetic diversity analysis is necessary for conservation and reintroduction of rare and endangered species (Zhuravlev et al., 2010). A species with a small hereditary changeability may endure from decreased wellness in its current environment. It may not have the developmental potential essential for a changing environment. Hereditary inconstancy is basic for a species to adjust to natural changes and survive in the long term (Reed, Frankham, 2003). Knowledge of genetic diversity within and among populations is imperative for preservation administration, particularly for distinguishing hereditarily interesting basic units inside a species (Reunova et al., 2014). In recent years, the development of various chemical and molecular techniques for studying genetic diversity has received great attention. In this article, we review some popular methods for the detection of genetic variation and point out DNA marker systems, which are considered as reliable and widely available tools used to distinguish variations.

**DNA markers**

DNA markers are the foremost broadly
utilized markers basically due to their plenitude (Yang et al., 2015). They emerge from distinctive types of DNA mutations, such as point mutations, rearrangements (insertion or deletion) or repeated DNA replication mistakes (Mammadov et al., 2012). These markers are specifically unbiased since they are found in non-coding regions of the DNA within the chromosomes. Unlike others, DNA markers are boundless in number and unaffected by environmental factors as well as plant growth stage (Govindaraj et al., 2015). They can be classified into three categories based on their detection strategy: hybridization, polymerase chain reaction (PCR), and DNA sequencing (Nadeem et al., 2018).

Hybrid-based DNA marker

Restriction fragment length polymorphisms, or RFLPs, was the first DNA-based marker and also the typical hybrid-based marker widely used in biology including evaluating genetic diversity (Petersen et al., 1994). This is a collection of DNA fragments with different lengths cut by restriction enzymes, also known as polymorphisms, which can be visualized separately on agarose gels based on its size, or transferred to a membrane via Southern blot. It was an important marker for genome mapping, localization of genes for genetic disorders, determination of risk for disease, and paternity testing (Ben-Ari, Lavi, 2012). The advantages of RFLPs include unlimited locus detection and are primarily robust, reliable and results can be transferred across populations. However, it is very expensive, time-consuming, labor-intensive, and it requires a larger amount of DNA, and limited polymorphism, when comparing to PCR-based markers developed later (Tabit, 2016).

PCR-based markers

The discovery of PCR technique in 1984 by Kary Mullis opened a second-generation of PCR-based DNA markers (Singh, Singh, 2015). Comparing to the first-generation hybridization-based markers, PCR-based DNA markers were considered as simpler, less expensive, and more convenient marker systems. They requires much smaller quantity of DNA of relatively lower quality and are much more user friendly and amenable to automation (Srivastava et al., 2012). Simple sequence repeats (SSRs), amplified fragment length polymorphisms (AFLPs), and randomly amplified polymorphic DNA (RAPD) markers are some of the widely used PCR-based markers. Even RFLPs, one of former DNA markers has been renewed and improved its efficiency when combined with PCR to form PCR-RFLPs (Singh, Singh, 2015).

Random amplified polymorphic DNA (RAPD)

RAPD is a PCR-based marker generated by PCR machines that use genomic DNA and random primers to make multiple copies of DNA chains (Simner et al., 2015). Less stringent conditions allow primers binding to both sample DNA strand of template where it is matched or partially matched, resulting in strain-specific heterogeneous DNA products (Sorof, Cheng, 2015). The advantages of RAPD include being fast, simple, inexpensive, and multiple loci can be amplified from a single pair of primers with a small amount of DNA (Sharaf-Eldin et al., 2018). However, the results from RAPD may not be reproduced in different laboratories and can only detect key features of interest (Ramos et al., 2008).

Amplified fragment length polymorphisms (AFLPs)

AFLPs are generated by digesting PCR amplified fragments using specific restriction enzymes to cut DNA at or near a specific nucleotide-specific site. AFLPs are highly reproducible and this enables the creation of fast and high-frequency identifiable AFLPs, making it an attractive technique for polymorphic identification and interconnection identification. AFLPs system detects polymorphism due to the sequence variation in and around the recognition sites of restriction endonucleases and uses PCR for marker assay (Paun, Schönswetter, 2012). AFLPs possessed high degrees of polymorphism and good reproducibility, have been widely used in population genetics studies (Mba, Tohme, 2005).
AFLP markers are dominant, so it cannot distinguish heterozygotes from homozygotes, but they can enhance the resolution achieved in population assignment, especially among weakly differentiated populations (Campbell et al., 2003; Mba, Tohme, 2005). Primers for AFLP markers are universal among species, which can be easy to design for any experiments. In order to detect genetic diversity, AFLPs were obviously good in terms of the success rates they achieved and the polymorphism detection (Song et al., 2015). Thus, AFLP markers are still practical markers to use when evaluating the genetic diversity and population structure of plants (Huang et al., 2019).

**Microsatellites**

Microsatellites are short (1-8 bp) repeat motifs usually associated with a high level of frequency of length polymorphism, which are considered among the best markers for genetic research (Hosseinzadeh-Colagar et al., 2016; Vieira et al., 2016). Microsatellites are also understood as simple sequence repeats (SSRs), short tandem repeats (STRs) or simple sequence length polymorphisms (SSLPs) (Vieira et al., 2016). SSRs is very diverse and uniformly distributed in the genome and common in eukaryotes. The number of repetition units varies greatly among plant species. The repetition sequence is usually simple, consisting of two, three or four nucleotides, also known as di-, tri- and tetranucleotide, respectively, in which, the most common for microsatellites is repeat dinucleotide (Fan, Chu, 2007). The inter-space areas tend to be conserved within species, although sometimes they can also be preserved at a higher taxonomic level.

SSRs are usually isolated on polyacrylamide gel in combination with AgNO₃ staining, autoradiography, or fluorescence detection systems (Hosseinzadeh-Colagar et al., 2016). Agarose gels with ethidium bromide may also be used when the difference in allele size between samples is greater than 10 bp (Li et al., 2017). Electrophoresis bands can be scored in a denominator either as it is present or absent. SSRs is easy to automate, highly polymorphic, along with good resolution and reproducibility. This is most widely used for individual genotypes, germ cell evaluation, genetic diversity research, genome mapping, phylogenetic and evolutionary research (Vieira et al., 2016). However, the development of microsatellites required extensive comprehension of the DNA sequence. Thus, they are developed primarily for agricultural species, instead of the wild (Mondini et al., 2009).

**Inter-simple sequence repeats (ISSRs)**

Inter-simple sequence repeats (ISSRs) are used to assess genetic diversity and identify closely related cultivars in many plant species (Verma et al., 2017). The ISSR marker system detects polymorphisms in inter-microsatellite DNA regions without any prior sequence knowledge (Zietkiewicz et al., 1994). Primers are based on a repeat sequence, often with a degenerate 3’anchor, and amplify plenty number of amplicons, providing high reproducibility at a low cost. ISSRs have been used for a wide range of organisms in DNA fingerprinting, diversity analysis, and genome mapping (Rahimi et al., 2019). ISSR technique has demonstrated as a powerful, rapid, simple, reproducible, and inexpensive way, especially ISSR markers are more reproducible than RAPD for the same purpose (Verma et al., 2017). However, a major disadvantage of ISSR markers is that they are not highly reproducible, and some primers generate poorly reproducible band patterns (Singh, Singh, 2015).

**Single nucleotide polymorphisms (SNPs)**

SNPs are single base substitutions inside DNA sequence of the individuals in the population that occur when a single nucleotide in the genome sequence is changed. SNPs occur more often than any other type of marker and are very close or even in the gene of interest. SNPs, the most abundant in the genome, are widely distributed across genomes with different distributions among species of plants and other organisms (Govindaraj et al., 2015). They provide valuable markers for fast identification
of plant varieties, the construction of ultra-high-density genetic maps, and developing adaptive traits in plant species (Fischer et al., 2017; De Lorenzis et al., 2019; Xia et al., 2019). SNP markers represented to third-generation molecular markers which possessed many significant benefit as high genetic stability and diversity, but their use requires advanced technology and has high costs (Xu et al., 2015).

Cleaved amplified polymorphic sequence (CAPS)

CAPS is a combination of the RFLPs and PCR and it was originally named PCR-RFLPs, which using amplified DNA fragments digested with a restriction endonuclease to display restriction site polymorphisms (Singh, Hazarika, 2020). The technique involves amplification of a target DNA through PCR, followed by digesting with restriction enzymes (Elangbam, Misra, 2016). Therefore, CAPS markers rely on differences in restriction enzyme digestion patterns of PCR fragments caused by nucleotide polymorphism between samples. Critical steps in the CAPS marker approach included DNA extraction, and the number or distribution of polymorphic sites, and PCR conditions. The restriction fragment length polymorphisms are based on PCR amplification, it is much easier and less time-consuming than analysing alternative types of markers that require southern hybridizations. Next, CAPS primers are more useful as genetic markers for comparative mapping study than those markers derived from non-functional sequences such as genomic microsatellite markers (Matsumoto, Tsumura, 2004). In genetic diversity analysis, their genotypes which are easily scored and interpreted, and only a small quantity of DNA is needed for one assay. Also, the cleaved and un-cleaved amplification products can be adjusted arbitrarily by the appropriate placement of the PCR primers. The procedure is technically simple with robust results because the amplification product is always obtained (Idrees, Irshad, 2014).

Expressed sequence tags (ESTs)

DNA markers could be developed from whole nuclear genome or expressed sequence tags (ESTs) (Singh et al., 2016). ESTs are short cDNA sequences reversely transcribed from mRNA. The identification of ESTs has proceeded rapidly, with over 6 million ESTs now available in computerized databases. In general, ESTs were instrumental in gene discovery, for obtaining data on gene expression and regulation, sequence determination, and for developing highly valuable molecular markers, such as EST-based RFLPs, SSRs, SNPs, and CAPS (Idrees, Irshad, 2014). Using EST-derived primer pairs to amplify nuclear genome, the amplicons may consist of intron sequences that displayed higher variation to develop informative markers for variety identification (Shu et al., 2010).

Cytoplasmic genome-derived-DNA markers

DNA markers could be also derived from the cytoplasmic genome, such as the mitochondrial genome (mtDNA) and chloroplast genome (cpDNA). The cytoplasmic CAPS markers are not only maternal inherited from haploid genome but also have a slower nucleotide substitution rate than the nuclear DNA (Kaundun, Matsumoto, 2011). Because of conservative evolution, they have been widely used in detecting geographical origins of plant species (Kaundun, Matsumoto, 2002; Katoh et al., 2003) and population differentiation (Schaal, Olsen, 2000).

The studies often use genetic information contained in cpDNA and mtDNA, as they contain two necessary sets of genes in plants (Skuza et al., 2019). They encode many key proteins for basic cell bioenergy processes. Additionally, they encode many components necessary for the proper expression of their own genes. Considering the essential importance of these two sets of organelle genes, one can expect them to change very slowly during evolution (Bendich, 2010). In addition, organelle DNA is a useful tool in the search for species-specific molecular markers.

Interspecies varieties can be effectively and quickly recognized utilizing chloroplast genome than by utilizing existing widespread DNA standardized tag, and subsequently, the
chloroplast genome has been utilized as the premise for creating markers for species distinguishing proof and classification (Chen et al., 2010). Chloroplast-based SNP markers have as of now been created for the major restorative crops, counting ginseng, and they have been effectively utilized for species recognizable proof, phylogenetic investigation, and hereditary differing qualities evaluation (Jo et al., 2016). Additionally, chloroplast-derived CAPS markers have moreover been detailed to identify the intra- and interspecies variety of diverse species (Kim et al., 2018).

The basic characteristics of some common DNA markers are shown in Table 1 and their pros and cons are showed in Table 2.

Table 1. Important feature of different types of DNA markers (Miah et al., 2013).

<table>
<thead>
<tr>
<th>Feature</th>
<th>RFLPs</th>
<th>RAPDs</th>
<th>AFLPs</th>
<th>SSRs</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Require (µg)</td>
<td>10</td>
<td>0.02</td>
<td>0.5–1.0</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>PCR based</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DNA quality</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>No. of polymorph loci analyzed</td>
<td>1–3</td>
<td>1.5–50</td>
<td>20–100</td>
<td>1–3</td>
<td>1</td>
</tr>
<tr>
<td>Type of polymorphism</td>
<td>Single base change, insertion, deletion</td>
<td>Single base change, insertion, deletion</td>
<td>Single base change, insertion, deletion</td>
<td>Change in repeat length</td>
<td>Single nucleotide change, insertion, deletion</td>
</tr>
<tr>
<td>Dominance</td>
<td>Co-dominant</td>
<td>Dominant</td>
<td>Dominant/Co-dominant</td>
<td>Co-dominant</td>
<td>Co-dominant</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>High</td>
<td>Unreliable</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Ease of use and development</td>
<td>Not easy</td>
<td>Easy</td>
<td>Easy</td>
<td>Easy</td>
<td>Easy</td>
</tr>
<tr>
<td>Automation</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Cost per analysis</td>
<td>High</td>
<td>Low</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Developmental cost</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Need for sequence data</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Very high</td>
<td>Very low</td>
<td>Medium</td>
<td>High</td>
<td>Very high</td>
</tr>
<tr>
<td>Radioactive detection</td>
<td>Usually yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Genomic abundance</td>
<td>High</td>
<td>Very high</td>
<td>Very high</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Part of genome surveyed</td>
<td>Low copy coding regions</td>
<td>Whole genome</td>
<td>Whole genome</td>
<td>Whole genome</td>
<td></td>
</tr>
<tr>
<td>Level of polymorphism</td>
<td>Low</td>
<td>Low to moderate</td>
<td>Low to moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Effective multiplex ratio</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Marker index</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Inheritance</td>
<td>Codominant</td>
<td>Dominant</td>
<td>Dominant</td>
<td>Codominant</td>
<td>Codominant</td>
</tr>
<tr>
<td>Detection of alleles</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Utility for genetic mapping</td>
<td>Species specific</td>
<td>Cross specific</td>
<td>Cross specific</td>
<td>Species specific</td>
<td>Species specific</td>
</tr>
<tr>
<td>Utility in marker assisted selection</td>
<td>Moderate</td>
<td>Low to moderate</td>
<td>Low to moderate</td>
<td>High</td>
<td>Low to moderate</td>
</tr>
<tr>
<td>Feature</td>
<td>RFLPs</td>
<td>RAPDs</td>
<td>AFLPs</td>
<td>SSRs</td>
<td>SNPs</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Cost and labor involved in generation</td>
<td>High</td>
<td>Low to moderate</td>
<td>Low to moderate</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

Table 2. Some potential advantage and weakness of the most commonly used DNA markers (Miah et al., 2013)

<table>
<thead>
<tr>
<th>Marker type</th>
<th>Benefits</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFLPs</td>
<td>Co-dominant</td>
<td>Need high-quality DNA</td>
</tr>
<tr>
<td></td>
<td>Genomic abundance high</td>
<td>Laborious (compared to RAPD)</td>
</tr>
<tr>
<td></td>
<td>Highly reproducible</td>
<td>Complex to automate</td>
</tr>
<tr>
<td></td>
<td>Better genome exposure</td>
<td>Radioactive labeling essential</td>
</tr>
<tr>
<td></td>
<td>Applicable across the species</td>
<td>Characterization of probe is essential</td>
</tr>
<tr>
<td></td>
<td>No need for sequence information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliably used in plants</td>
<td></td>
</tr>
<tr>
<td>RAPDs</td>
<td>Genomic abundance high</td>
<td>No need of probe information</td>
</tr>
<tr>
<td></td>
<td>Better genome coverage</td>
<td>Dominant markers</td>
</tr>
<tr>
<td></td>
<td>Sequence information unneeded</td>
<td>Not reproducible</td>
</tr>
<tr>
<td></td>
<td>Perfect for automation</td>
<td>Not suitable for across species</td>
</tr>
<tr>
<td></td>
<td>Requires less DNA</td>
<td>Not well tested</td>
</tr>
<tr>
<td></td>
<td>No radioactive labeling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More rapid</td>
<td></td>
</tr>
<tr>
<td>SSRs</td>
<td>Easy to automate</td>
<td>Not well-examined</td>
</tr>
<tr>
<td></td>
<td>Genomic abundance high</td>
<td>Not suitable for across species</td>
</tr>
<tr>
<td></td>
<td>Highly reproducible</td>
<td>Sequence information needed</td>
</tr>
<tr>
<td></td>
<td>High polymorphism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple alleles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderately genome coverage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No radioactive labeling</td>
<td></td>
</tr>
<tr>
<td>AFLPs</td>
<td>High polymorphism</td>
<td>Very tricky due to changes in materials use</td>
</tr>
<tr>
<td></td>
<td>Genomic abundance high</td>
<td>Not reproducible</td>
</tr>
<tr>
<td></td>
<td>Can be used across species</td>
<td>Very good primers needed</td>
</tr>
<tr>
<td></td>
<td>No need for sequence information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Useful in preparing counting maps</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Works with smaller RFLPs fragments</td>
<td></td>
</tr>
<tr>
<td>Sequence-tagged site (STS)</td>
<td>Helpful in preparing counting maps</td>
<td>Need sequence information</td>
</tr>
<tr>
<td></td>
<td>Highly reproducible</td>
<td>Out of the target sites, mutation detection not possible</td>
</tr>
<tr>
<td></td>
<td>No radioactive labeling</td>
<td>Laborious</td>
</tr>
<tr>
<td></td>
<td>Can use filters many times</td>
<td>Cloning and probe characterization required</td>
</tr>
<tr>
<td></td>
<td>Moderate genome coverage</td>
<td></td>
</tr>
</tbody>
</table>

DNA markers based on the sequencing

DNA sequencing was first introduced by Sanger and then developed as method allowing rapid identification of DNA sequences based the DNA polymerase activity during DNA synthesis (Sanger, Coulson, 1975). The DNA sequencing technique using dideoxy nucleotide to stop the random DNA synthesis; or by chemical using marked DNA later have widely applied around the world (Metzker, 2005). Success in sequencing the human genome in 2003 has opened a new era of development in life science research with many new, complex and modern techniques, developed and applied (Hood, Rowen, 2013). New generation DNA sequencing did not rely on Sanger chemistry, enables efficient and rapid decoding of the entire genome sequencing -WGS). The first of this kind of 2nd generation of sequencing technique appeared in 2005 with the landmark publication of the sequencing-by-synthesis technology developed by 454 Life Sciences based on pyrosequencing (Margulies et al., 2005; Rothberg, Rothberg, 2015). Commercial 2nd generation sequencing methods included four main platforms such as Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System and Polonator.
G.007 (Lerner, Fleischer, 2010). They were all amplification-based and could be distinguished by the role of PCR in library preparation.

The single-molecule sequencing method, also known as 3rd generation or next-next generation, was recently developed by Helicos Genetic Analysis System using the technology developed by Braslavsky et al. (Braslavsky et al., 2003; Rothberg, Rothberg, 2015). There was high-throughput and low-cost sequencing method which provides the first universal genetic analysis platform that requires no amplification. It pursues a molecular sequencing strategy that simplifies the process of DNA sample preparation, avoids errors and deviations caused by PCR, simplifies data analysis, and tolerates degraded samples (Ozsolak et al., 2009). With the advantage of fast, high-throughput and accuracy, the new gene sequencing systems are widely used in most modern biology research (Gasperskaja, Kučinskas, 2017). However, the traditional sequencing method (Sanger method) with long readings and high accuracy continues to be used in many cases, especially in small-scale projects with limited samples (Shendure, Ji, 2008). The features of each sequencing technology platforms are summarized in Table 3.

| Table 3. Summary of the output per sequencing technology platforms (Buermans, den Dunnen, 2014). |
|---|---|---|---|---|---|---|---|
| **Sequence by** | **Detection** | **Run types** | **Run time** | **Read length (bp)** | **# reads per run** | **Output per run** | **Remarks** |
| Roche | GS FLX Titanium XL + | Pyrophosphate detection | Single end | 23 h | 700 | 1 million | 700 Mb |
| | GS Junior System | Pyrophosphate detection | Single end | 10 h | 400 | 0.1 million | 40 Mb |
| LifeTechnologies | Ion torrent | Proton release | Single end | 4 h | 200–400 | 4 million | 1.5–2 Gb ion318 Chip |
| | Proton | Proton release | Single end | 4 h | 125 | 60–80 million | 8–10 Gb IonP1 chip |
| Ab/solid | Ligation | Fluorescence detection of di-base probes | Single and paired-end | 10 days | 75 + 35 | 2.7 billion | 300 Gb |
| Illumina/solexa | HiSeq2000/2500 | Fluorescence; reversible terminators | Single and paired-end | 12 days | 2 × 100 | 3 billion | 600 Gb High output mode |
| | MiSeq | Fluorescence; reversible terminators | Single and paired-end | 65 h | 2 × 300 | 25 million | 15 Gb |
| Pacific biosciences | RSII | Single molecule synthesis | Fluorescence; terminally phosphorylated | Single end | 2 days | 50% of reads > 10 kb | 0.8 million | 5 Gb 16 SMRT cells |
| Helicos | Heliscope | Single molecule synthesis | Fluorescence; virtual terminator | Single end | 10 days | ~ 30 | 500 million | 15 Gb Two flow cells in parallel |

**Target-enrichment strategies for next generation sequencing**

In plant investigation, targeted re-sequencing of enriched genomic DNA regions has ended up a versatile and cost-effective strategy for the disclosure of genome-wide sequence varieties to be misused to address diverse organic questions (Kaur, Gaikwad, 2017). Plants have expansive or polyploid genomes, therefore entire genomes may not be promptly gathered and the examination still exceptionally costly, that
requires an elective procedure to WGS and to produce a diminished representation of the genome (Fuller et al., 2009). Genetic variations can be obtained by utilizing target enrichment methodologies, which comprises within the separation of particular genomic loci (e.g., qualities, atomic markers, bigger genomic locales, and organelle genomes) coupled with NGS (Gnirke et al., 2009). Compared to WGS, the decrease in sequencing space involves three primary focal points, test multiplexing that embroils a generally lessening of the sequencing fetched per sample; critical decrease within the complexity of the analysis (Mertes et al., 2011).

Re-sequencing the genomic locales that are held is essentially more time- and cost-effective, and the coming about information are significantly less awkward to examine. Few approaches to target enrichment have been created and there are some parameters by which the execution of each can be measured, which shift from one approach to another such as affectability, specificity, consistency, reproducibility, fetched; ease of utilizing; and sum of DNA required. Currently, transcriptome-based, confinement enzyme-based, PCR-based, and hybridization-based strategies, all consistent with the foremost well-known NGS stages, have been created to improve particular targets (Cronn et al., 2012).

Transcriptome-based enrichment is one of the foremost broadly utilized methodologies to diminish genome complexity, since it focuses only on the transcribed portion of the genome. The key point of transcriptome sequencing, moreover known as RNA-seq, is to determine quality expression profiles of each transcript amid advancement and beneath diverse conditions (Martin et al., 2013). SNP discovery and molecular marker development via RNA-seq are often performed, especially in organisms with large genomes (Egan et al., 2012). Essentially, because RNA-seq is autonomous from any a priori comprehension on the genome sequence of the species beneath analysis, it permits the investigation of ineffectively characterized species.

Restriction enzyme-based enrichment makes use of the discriminatory power of the restriction endonucleases to produce restriction fragments among individuals in a population. Three main techniques have been developed so far: RAD-seq (restriction-site associated DNA sequencing), GR-RSC (genomic reduction based on restriction site conservation), and GBS (genotyping-by-sequencing). All these methods are flexible and quite inexpensive and have been used to identify and score, in a group of individuals, thousands of genetic markers randomly distributed along the genome enabling SNP discovery, genotyping as well as quantitative genetic and phylogeographic studies (Baird et al., 2008; Maughan et al., 2009; Rowe et al., 2011; Cronn et al., 2012; He et al., 2014).

PCR-based target enrichment includes the direct sequencing of small and long PCR products. NGS of PCR fragments has been preferentially applied to chloroplast genomes in systematic studies and in some cases, to nuclear genomic regions despite their complexity. The main disadvantages associated with this method are the high level of failed target amplifications and/or non-specific amplifications as well as the difficulty in obtaining an accurate pooling of samples for NGS multiplexing. Anyway, PCR-based enrichment remains feasible for targeting small to medium-sized regions of the genome, but for high-throughput sequencing of tens of thousands of PCR amplicons, its efficiency falls off, given the initial cost per sample and challenges in sample multiplexing. Microfluidic-based multiplexing PCR can reduce costs but continues to be more expensive than other enrichment methods (Durstewitz et al., 2010; Mamanova et al., 2010; Cronn et al., 2012; Uribe-Convers et al., 2016).

Hybridization-based enrichment or sequence capture methods exploit the high specificity of DNA or RNA probes (also called baits), which are designed to be complementary to target genomic regions. RNA baits have significant advantages over DNA probes because RNA-DNA hybrids have a higher affinity and melting temperature than DNA-DNA hybrids. Two main
technologies have been developed for hybrid-capture applications: on-array- or solid-based hybridization, which implies sample hybridization on a solid support and in-solution- or liquid-based hybridization, where pooled baits are used in reaction tubes. Due to their moderate costs and high specificity, low amounts of required DNA per sample and power to simultaneously target large numbers of markers, several protocols and commercial kits have been developed. The most widespread ones and reliable in studies on plant species were provided by Agilent Technologies (SureSelect), Roche NimbleGen (SeqCap EZ), MYcroarray (MYbaits), and Ion Torrent (TargetSeq) (Okou et al., 2007; Gnirke et al., 2009).

GENETIC DIVERSITY ANALYSIS IN WILD, CROP, AND MEDICINAL PLANTS

Genetic diversity analysis in wild plants

DNA markers were widely applied for genetic diversity analysis of wild plant. For instance, a series of DNA markers as AFLPs, ISSRs, SSRs were used to analyse the diversity of 389 accessions of 18 wild almond species, which were considered as crucial importance in breeding. The result showed highest polymorphic information for SSR markers (Sorkheh et al., 2017). SSRs were also applied to analyse the genetic diversity of 86 almond accessions of diverse geographic origin, ranging from Central Asia to the USA. The results showed that there were slight losses of genetic diversity by geographic isolation, human selection but no indication of a major decrease in genetic variability in almond germplasm from Asia to Europe. The results also suggested the need to avoid the decline in genetic diversity during the almond domesticating even though its risk is present in many subpopulations (Halász et al., 2019). Genetic diversity in Chinese wild apple species along with cultivars was evaluated using SSR markers (Zhang et al., 2012), zombi pea (Vigna vexillata) (Dachapak et al., 2017), wild soybean (Glycine soja) (Nawaz et al., 2017), wild date palm (Phoenix sylvestris) (Huda et al., 2018), wild kiwifruit (Actinidia eriantha) (Huda et al., 2018; Liao et al., 2019).

RAPD is also a traditional marker commonly used in investigation the genetic diversity of wild plants until now. In the past, RAPD markers were utilized in the analysis genetic variation of Saxifraga cernua (Bauert et al., 1998), Calamagrostis porteri (Poaceae) (Esselman et al., 1999) or to reveal genetic diversity within and between populations of cashew (Anacardium occidentale L.) (Mneney et al., 2001). Genetic variability and species identification within 22 species of the genus Encephalartos, the second largest genus with 65 species and 2 sub-species were analysed using RAPD (Prakash, Van-Staden, 2008). The result showed that most species of Encephalartos are morphologically very similar and therefore could benefit from additional tools for their correct identification. RAPD markers were also applied to assess genetic diversity and structure of natural Calophyllum brasiliense (Clusiaceae) populations in Riparian forests (Mendonça et al., 2014). The genetic differentiations among Jacaraitia mexicana populations, a native tree in Mexico, also be discovered at even small geographic scale by RAPD method that considered in conservation actions for this genetic resource (Arias et al., 2012). These studies above suggested the potential of the RAPD for correct genetic identification of individual species (Prakash, Van-Staden, 2008).

Some other wild plants were also estimated the level of genetic diversity, such as Phragmites (Poaceae) by AFLP fingerprinting, Roseroot (Rhodiola rosea L.) by SSRs and ISSRs, Saruma henryi, an endangered plant endemic to China by SSRs, Galium cracoviense, G. oelandicum, G. sudeticum (Rubiaeaceae) and Caragana species by AFLPs or genetic diversity of Calamagrostis porteri (Poaceae), Amentotaxus argotaenia (Taxaceae) were determined by ISSRs (Esselman et al., 1999; Ge et al., 2005; Cieslak, Szelag, 2010; György et al., 2012; Zhou et al., 2012). Other markers such as AFLPs, ISSRs have been commonly used in the past but are not
currently used frequently for analysis of genetic diversity.

**Genetic diversity analysis in crop plants**

Plants are very important to the survival and development of humans. Plant domestication has been done for a long time based on desirable traits. The interest in genetic diversity has been paid attention and especially with the helpful support of DNA markers. DNA markers were used in plant breeding for support the evaluation of propagation materials and cross-breeding (Govindaraj et al., 2015). Some supported effects of DNA markers can be listed here for evaluation of genetic diversity, parental selection, identification and purity evaluation of varieties, research on heterozygosity, and identifying the currently selected genomic regions (Nadeem et al., 2018).

Traditional markers as RFLPs have been widely used to compare genomes in major cereal families such as rye, wheat, maize, sorghum, barley, and rice or other plant as pea (Dijkhuizen et al., 1996; Lu et al., 1996). Another one, ISSR marker was used for cultivar identification, genetic diversity analysis and genotypes validation in many species including banana, sorghum, Arabidopsis, sunflower, tomato (Godwin et al., 1997; Bornet, Branchard, 2001; He et al., 2003), and mango varieties in different mango growing regions including Australia (González et al., 2002), China (He et al., 2005), and India (Damodaran et al., 2012).

SSRs seemed to be a DNA marker which is most widely used in analysis of genetic biodiversity. Twenty-four SSR markers were utilized to evaluate the population structure and genetic diversity of thirty-six guava varieties (*Psidium guajava* L.) including wild species to develop varieties with better fruit yield and nutritional quality which assisted assessment for morpho-physiological traits in guava (Kherwar et al., 2018). *Chrysophyllum albidum* is endangered tree species among the forest tree of which fruit is widely consumed and play a significant role in food security in many parts of tropical Africa. A very low genetic variation in *C. albidum* population was discovered by SSR markers suggested the need for preserving the remnant inherent diversity towards capturing the existing local adaptation (Boboye et al., 2018).

To assess the genetic diversity of 109 newly introduced accessions of narrow-leafed lupin (*Lupinus angustifolius* L.), a great potential to be a new crop in China (Ji et al., 2020), 76 genomic SSR markers were utilized. SSR markers were also applied for evaluating the genetic structure and gene flow pattern of 285 domestication cotton to provide useful information for understanding the genetic base of upland cotton superior varieties and will also promote future high yield and excellent fiber quality breeding (Zhang et al., 2020).

Rice (*Oryza sativa* L.), is an important food crop for human life demand. Genetic diversity analysis of rice germplasm was investigated using PCR-based SNP markers to assess genetic diversity among indigenous rice varieties to identify the different blast resistance genes in a collection of 74 rice germplasms from Tripura (India) (Anupam et al., 2017). Another study using the 32 SSR markers to investigate the genetic diversity of 50 aromatic rice accessions from three regions (Peninsular Malaysia, Sabah, and Sarawak) to determine the high yielding aromatic rice genotypes. This was needed to identify the potential diverse genotypes for use as a parent in future rice breeding program (Jasim Aljumaili et al., 2018).

Rye (*Secale cereale* L.) is also a cereal grass that is an important food crop in Central and Eastern Europe but the genomic resources in rye are underdeveloped because few population genetic studies using genome-wide markers have been published to date. A collection of 603 individuals from 101 genebank accessions of domesticated rye and its wild progenitor *S. cereale* subsp. *vavilovii* and related species in the genus *Secale* was done. The analysis of SNPs data of rye and its wild relatives obtained through sequencing detected 55,744 SNPs relating to cultivated rye yield with present genotype calls in 90% of samples (Schreiber et al., 2019).
Genetic diversity analysis in medicinal plants

Medicinal plants are defined as plants that have medicinal properties due to complex chemical substances of different composition such as alkaloid, glycoside, corticosteroid, and essential oils, which are basic ingredients for traditional and modern medicines and therapies for health care and also for a number of other purposes (pharmaceuticals, herbs, herbal foods) (Paramanik, Chikkaswamy, 2014). The World Health Organization (WHO) considers wild plants and plants used for medicinal purposes to be medicinal plants and they are highly significant for their role in the traditional therapies used by 70-80% of citizens of developing countries (Aziz et al., 2018). The number of medicinal plants being used globally varies from 70,000 to 125,000 (Schippmann et al., 2006). In India, from 7,000 to 8,000 plants are considered as medicinal plants (Rajeswara et al., 2012). In Vietnam, the number of medicinal plants used is 4,700 while the resources of medicinal plants can be much larger, estimated to reach 6,000 species (Chi, 2012). However, the breakthrough of modern humans’ activities in forests for exploitation, farms, timber, transportation, industrial housing etc., leads to loss/fragmentation/degradation of habitats of plants species including to medical plant, leading to loss of their genetic diversity that cannot be recovered (Haddad et al., 2015).

The efforts to assess genetic diversity and erosion through molecular marker techniques results in positive genetic diversity that exists in existing genetic resources. A typical example, RAPD markers were used to investigate the genetic diversity relationship of 18 medicinal plant species (Paramanik, Chikkaswamy, 2014). In another study, 479 samples of A. austrokoreense from 7 populations collected in South Korea were analysed the genetic diversity and structure by microsatellite amplification with the size of the amplified products ranged from 300-5000 bp depended on the sequence of random primers and medicinal plants (Lee et al., 2018). ISSR, PCR-RFLP and SSR markers were applied for analysing the genetic diversity of Peganum harmala L. (Peganaceae), a perennial herbaceous plant for medicinal purposes to treat hypertension, cardiac disease, some nervous system disorders such as Parkinson's disease, Lumbago asthna, colic, jaundice and as a stimulant emmenagogue (El-Bakatoushi, Ahmed, 2018). The genetic diversity of the endemic and medicinally important plant Rheum officinale also revealed by ISSR markers (Wang et al., 2012).

Genetic diversity analysis of Panax

Panax (Panax spp.) is one of the most widely used medicinal herbs in the world that contains about 11 species of very slow-growing perennial plants with tubers belonging to the Araliaceae (Shahrajabian et al., 2019). They distributed mostly in Korea, China, Eastern Siberia, and North American distribution and 35 other countries around the world, where characterized by cold countries (Li, Mazza, 1999). The two most important species grown commercially are Panax quinquefolius L. (North American ginseng), which is native to eastern North America, and Panax ginseng C. A. Meyer (Korean ginseng), which is native to northeastern China and the Korean peninsula (Proctor, Bailey, 2011). In North America, the harvesting of P. quinquefolius began over 100 years ago from wild populations in Canada and the new cultivating primarily in British Columbia and Ontario in recent 30 years (Bai et al., 1997; Li, Mazza, 1999). Recently, a new species of Panax was named Panax vietnamensis Ha et Grushv., discovered in Vietnam, is a true ginseng found in the south pole and to be considered as a newest subject to be studied. P. vietnamensis is the most recently identified Panax species. The plants are rarely identified at 1700-2000 meters high-elevation forests with cool weather all year round in the tropical Central Vietnam (Dung, Grushvisky, 1985; Court, 2006). The plants were known for their medicinal value in Vietnam like other Panax species. The diversity of secondary metabolites and pharmacological effects of P. vietnamensis were revealed in recent years (Court, 2006; Le et al., 2015) beside the studies of phylogenetic analysis and transcriptomics in P. vietnamensis and related natural variant were
performed (Komatsu et al., 2001; Zhang et al., 2015; Manzanilla et al., 2018). However, the comprehension of current genomic resource of *P. vietnamensis* is very limited.

Ginseng has a relatively high level of genetic diversity, similar to cultivated and wild groups. Ninety-two polymorphic microsatellite markers were developed in 147 individual plants, including cultivars, breeding lines, and wild populations in Korea and neighbouring countries to elucidate genetic diversity. There was no statistically significant difference in genetic diversity between cultivated and wild ginseng groups, but high level of genetic diversity in current breeding populations for breeding of elite ginseng cultivars (Jang et al., 2019). It is limited research effort to analyse the genetic diversity and population structure of ginseng because of its growth habits. Even though, recent studies tried to analyse diversity level by regional distribution. Using SSR markers approach, 94.0%, 5.5% and 0.5% of 1109 accessions of Ginseng (*Panax ginseng* C.A. Meyer) in South Korea were landraces, breeding and cultivars lines, respectively, that were revealed from 56 different alleles from 12 clusters and indicated that average gene diversity was 0.49 (Lee et al., 2020). RAPD-PCR was utilized for investigating the genetic diversity of North American ginseng (*P. quinquefolius* L.) population in Ontario (Canada). The similarity coefficients among the DNA of ginseng plants analysed were low, indicating that a high degree of genetic diversity existed in the ginseng population. However, lower levels of genetic diversity were detected among 3-year-old ginseng plants selected on the basis of greater plant height than among the plants randomly selected from the same subpopulation or over the whole population. This research result suggested that genetic factors at least partly contribute to morphological variation within the ginseng population and that visual selection can be effective in identifying the genetic differences. The significance of a high degree of genetic variation in the ginseng population on its potential for improvement by breeding (Bai et al., 1997). Genetic diversity of *P. stipuleanatus* Tsai in North Vietnam detected by ISSR markers showed the higher population genetic diversity in the Lao Cai than in Lai Chau region and a highly gene flow within *P. stipuleanatus* populations in Vietnam (Trieu et al., 2016).

**CONCLUSION**

A variety of plant genetic sources are essential possessions for human. Natural genetic variant in vegetation is extensive, and stays to be completely described and harnessed in plant species. This may additionally be accomplished through the phenotypic and molecular markers of plant genetic assets. In the field of molecular genetics studies, molecular markers have been developed and utilized extensively for the exploitation and identification of plant genetic diversity. Various techniques are used to estimate genetic diversity, in which, the most popular DNA markers for genetic diversity in plants include RAPDs, ISSRs and AFLPs, or RFLPs, SSRs, CAPS, ESTs, SNPs. Today, new techniques are being developed frequently. No such techniques are ideal yet, each technique has its own advantages and limitations. However, low cost, inexpensive equipment, throughput, alleviation, and ease of check advancement and computerization are indispensable components when deciding on innovation. For example, AFLPs and RFLPs were common techniques for analyzing genetic diversity, however, the large requirement of time and amount of DNA required for analysis makes them no longer effective. Instead, newer markers such as SNPs or CAPS with more advantages become more popular. Some markers that still show efficacy in such analyzes, such as RAPDs or SSRs, continue to be widely used in genetic diversity studies. New methods are frequently being developed to improve their effectiveness and relevance to use although no method is ideal for all applications so scientist should look at both pros and cons of methods when starting a new project. As science advances, it will be feasible to gather near-complete descriptions of genetic variant in natural populations.
Acknowledgements: This work was supported by the Ministry of Science and Technology (Project No. 16/2017-HD-NVQG on “Transcriptome sequencing and analysis of Panax vietnamensis Ha et Grushv.”).

REFERENCES


Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100: 3960-3964.


Vu Tuan Nam et al.


Maughan PJ, Yourstone SM, Jellen EN, Udall JA (2009) SNP discovery via genomic reduction,
barcoding, and 454-Pyrosequencing in Amaranth. Plant Genome 2: 260-270.


Trieu LN, Mien NT, Van Tien T, Van Ket N, Van Duy N (2016) Genetic diversity of Panax stipuleanatus


