

Analysis of SSR and SNP markers

Ankita Mishra¹, Pramod Kumar Singh², Abhishek Bhandawat¹, Vinay Sharma¹, Vikas Sharma³, Pradeep Singh⁴, Joy Roy¹ and Himanshu Sharma¹

¹Agri-Biotechnology Division, National Agri-Food Biotechnology Institute, Mohali, India, ²Department of Biosciences, Christian Eminent College, Indore, India, ³Department of Botany, Sant Baba Bhag Singh University, Khiala, India, ⁴Department of Biotechnology, Guru Nanak Dev University, Amritsar, India

8.1 Introduction

In the present scenario, genomic research has provided new tools and technologies that support the breeders for the improvement of important crops using molecular markers. DNA profiling has become a significant technique for genetic improvement in plant breeding and germplasm selection. Genetic diversity states the variability at the individual or multiple gene loci among individuals of a population or among populations (Datta et al., 2010). It is genuinely felt that there is a lack of information regarding the molecular markers' appropriateness, relevant marker combinations, and suitable marker utility parameters, which are the main constrain in breeding programs. Deeper knowledge about the genetic background of germplasms is essential for the selection of suitable material, conservation, and successful plant breeding. Higher diversity signifies the fitness of species and is more likely to adapt or survive the diverse environment (Hunter, 1996). There are different markers, that is, morphological, biochemical, and molecular markers used to evaluate genetic variation (Gwanme, Labuschangne, & Botha, 2000). Morphological markers are technically simple but they have several limitations and genome coverage is very small. DNA markers have been proved superior over other classes of markers for genotyping applications.

These markers provide a better estimate of diversity, genetic relationships, identification of clones, and establishment of core collection (Ghafoor & Ahmad, 2005). Molecular markers are well suited for plant breeding, genetic map construction, and conservation of endangered varieties (Renganayaki, Read, & Fritz, 2001). In recent years many polymerase chain reaction (PCR)-based markers are made available, which are used for crop improvements, such as random amplified polymorphic DNA (RAPD), simple sequence repeats (SSRs), amplified fragment length polymorphism (AFLP), inter-simple sequence repeats (ISSR), and intron length polymorphic primer (Sharma, Namdeo, & Mahadik, 2008). Each marker has one or more challenges. For example, RAPD produces less reproducible results. AFLP method is costly and difficult to score, while ISSRs are dominant in nature. SSRs also called microsatellites are locus-specific, codominant, and reproducible markers. SSRs are short tandem repeats of one to six bases classified as mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats. Due to the tandem repeat nature, often the SSR region accumulates variation due to replication slippage or mispairing, which is detected using PCR followed by gel separation. Primers of 20–25 nucleotide length are usually designed from repeat flanking region for PCR amplification.

Microsatellites may be identified using enrichment of specific repeat followed by sanger sequencing. Alternatively, SSRs may be identified at no cost by the mining of public genomic or expressed sequence data (Rahim, Sharma, Parveen, & Roy, 2018). In addition, degenerate SSR primers from closely related species may be designed due to the cross-transferable (conserved) nature of these locus-specific markers. Currently, SSR analyses have dominated plant genome analysis and breeding. The polymorphic markers used in plant breeding are generally fluorescence based and can be multiplexed where multiple primers producing the same allele size can be amplified and resolved from a single reaction. SSRs along with RFLP being reproducible are widely used in linkage studies and mapping important traits on the chromosome. The first linkage map of black gram was developed using morphological markers and three different molecular markers namely, SSR, RFLP, and AFLP (Han et al., 2005). SSRs have become the popular choice for genotyping, genetic mapping, QTL mapping, and marker-assisted selection in plants (Xu, Deb, & Mackill, 2004). SSRs were found to be

associated with a major QTL for submergence tolerance (Xu et al., 2004), sheath blight resistance (Che et al., 2003), bacterial blight resistance (Zhang et al., 2001), restorer gene for HL type cytoplasmic male sterile lines (Huang, Li, & Dai, 1999), and fragrance gene (Cordeiro, Christopher, Henry, & Reinke, 2002) in various plant species.

Besides other marker systems, the latest generation of markers is SNPs, SNP is the single base variation at the nucleotide level between two genotypes that might have arisen due to any transition or transversion mutation at the genomic level. They are the source of the simplest molecular markers available in plants and animals. Batley and Edwards (2007) have reported SNP marker frequency of 100–300 bp in plants and named them according to the position they occupied in the physical genome. At present, with the latest improvements in NGS technologies and available sequencing platforms, SNP-based genetic mapping has come up as a powerful tool for developing high-resolution genetic maps of plants even with the complex humongous genome (Elshire et al., 2011; Xu et al., 2012). SNP discovery with NGS is an easy, less time-consuming alternative along with amenability to the multiplexing approach. It provides satisfactory results at a lower cost with larger population size (Poland, Brown, Sorrells, & Jannink, 2012). So, based on the advantages of SSRs and SNPs have, this chapter highlights the SSRs and SNP analysis.

8.2 Analysis of SSR markers

8.2.1 Benefits and limitations of microsatellite markers

Microsatellites are tandem repeats of 1–6 bp, and which predominantly occur in both expressed and nonexpressed regions of the genome (Kalia et al., 2011). The mutation rates of SSR markers estimates to lie between 10^{-2} and 10^{-4} per generation. The primary benefit of SSR markers is that they are highly polymorphic and follow the Mendelian fashion of inheritance. Furthermore, they are multiallelic and distributed across the whole genome (Morgante, Hanafey, & Powell, 2002; Sahoo et al., 2019). SSRs have been helpful for the analysis of diversity, paternity, phylogenetic studies, genetic mapping, QTL mapping, marker-assisted breeding (Sharma et al., 2017; Verma et al., 2017; Verma et al., 2021), and establishing evolutionary relationships (Parida et al., 2009). Details of SSR analysis are given in Fig. 8.1. Earlier developmental costs and technical difficulties in the preparation of enriched libraries microsatellite-based markers were high. But later with the improvement of NGS, the SSR enrichment step can be omitted and the expense of

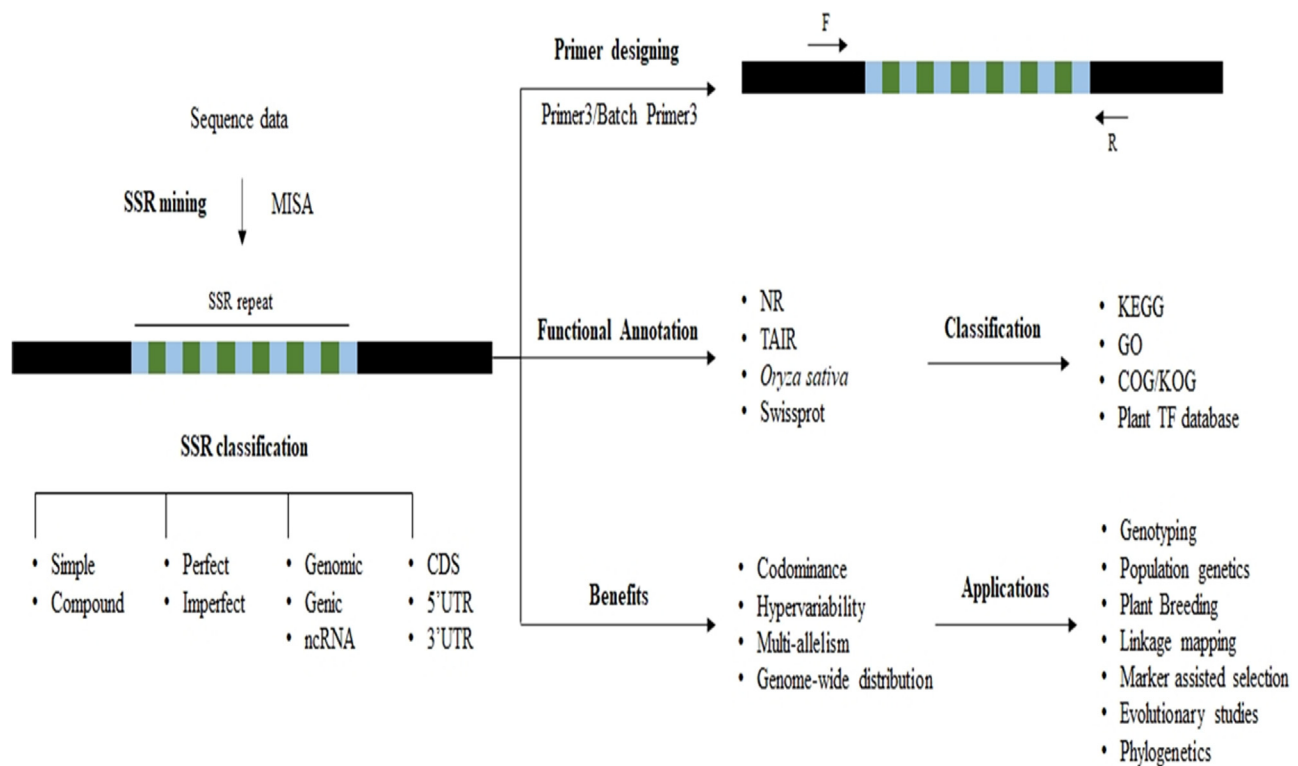


FIGURE 8.1 General workflow for simple sequence repeat mining, classification, primer development, function assignment, advantages, and application.

SSR identification is drastically reduced. Unlike random markers, such as RAPD and ISSRs, SSR markers are sequence specific and thus need to be designed for individual species.

8.2.2 SSR mining and primer designing

Sequences obtained after SSR enrichment or genomic or expressed data mining are utilized for investigating the occurrence of microsatellites by utilizing MISA script (<http://pgrc.ipk-gatersleben.de/misa/>; Thiel, Michalek, Varshney, & Graner, 2003) or SSRIT with repeat length 6 for di-repeats and 5 for remaining repeats. Monorepeats are usually not preferred for marker development due to higher chances of replication slippage during PCR generating stutter bands, which are hard to analyze. The SSR containing sequences are subsequently exploited for designing primers using Primer3 or Batchprimer3 software (<http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi>; Rozen & Skaletsky, 2000) with the following criteria: optimum primer length: 21 bp, amplicon size: 100–300 bp, optimum annealing temperature: 55°C, and GC content: 60%. SSR primers are designed from the sequences flanking SSRs. These sequences show substantial conservation across the species and sometimes across the genus. This makes SSR primers to amplify across genera.

8.2.3 Classification and genomic localization

SSRs may be classified as simple or compound based on the absence or presence of interruption by additional bases. Simple SSRs are uninterrupted while compound SSRs are interrupted by a nonrepetitive nucleotide sequence with a length of 100-bp sequence. SSRs are termed as perfect repeats if the SSR length is 20 bases or above. In the case of imperfect SSRs, the repeat length is 12–19 bases long. Based on their derivation from genomic regions, they are also classified as genomic or genic microsatellites. Furthermore, SSRs containing sequences are classified as CDS, 5'- or 3'-UTR dwelling based on their location with respect to ORF. OrfPredictor (Min, Butler, Storms, & Tsang, 2005) software is used to predict the ORF. Gene-based SSRs are sometimes assigned putative function hence also termed as functionally relevant SSRs (Bhardwaj, Sharma, Kumar, Sharma, & Ahuja, 2014; Bhandawat, Sharma, Nag, et al., 2015; Bhandawat, Sharma, Sharma, Sood, & Sharma, 2015; Sharma, Bhandawat, & Rawat, 2020; Sharma, Bhandawat, Kumar, et al., 2020). Recently a new class of SSRs encompassing noncoding RNAs and bZIP transcription factors have been reported (Bhandawat, Sharma, Pundir, Madhawan, & Roy, 2020; Sharma, Bhandawat, & Rawat, 2020; Sharma, Bhandawat, Kumar, et al., 2020).

8.2.4 Functional annotations of SSR containing sequences

Annotation of the SSRs containing sequences is done mainly based on the sequence homology with protein databases, such as TAIR, Rice, Swissprot, NR, and Kyoto Encyclopedia of Genes and Genomes (KEGG), by the use of blast search with a cut-off E -value of 10^{-5} . Gene ontology terms are allocated to SSR containing sequences and grouped in different categories, such as biological process, molecular function, and cellular component, by using WEGO software (Ye et al., 2006) to know the classification of the gene functions. Furthermore, the KEGG is used for the estimation of possible functional and molecular pathways (Kanehisa & Goto, 2000). Transcription factors are also predicted using a plant transcription factor database (PlantTFDB; Jin, Zhang, Kong, Gao, & Luo, 2014).

8.2.5 SSR amplification and evaluation of polymorphic potential

Evaluating genetic variation and diversity is an important aspect of germplasm characterization for crop improvement (Taheri et al., 2018). In recent years there has been great advancement in the area of genomics that allows measuring of genetic variability at the DNA level using various DNA molecular markers (Lightfoot & Iqbal, 2013). Microsatellite markers are the most commonly used markers because they are sequence specific and codominant, and the amount of variation per locus is large (Ghaffari & Hasnaoui, 2013). In SSR markers, allelic differences at a locus arise due to the variability in the length of the repeat motif. Allelic variation in the individual that occurs due to replication slippage and/or irregular crossing over in meiosis can be identified by the scoring of genotype profiles of various alleles through genotyping. Genotyping is mainly done by PCR amplification of the specific locus of an individual with a microsatellite marker. For the priming of microsatellite marker to the template, PCR annealing temperature has an important role. Previous studies reported that the amplification of microsatellite markers has been achieved with the temperature range of 45°C–60°C (Blair et al., 2011; Narina, d'Orgeix, & Sayre, 2011; Ravishankar, Mani, Anand, & Dinesh, 2011; Risterucci, Duval, Rohde, & Billotte, 2005; Singh, Sharma, Nag, Bhau, & Sharma, 2015). With the emergence of

technologies, primers with fluorophore labeling and automated electrophoresis system have significantly improved the detection efficiency and efficacy. For the amplification of microsatellite markers, three approaches are utilized.

8.2.5.1 Amplification by utilizing unlabeled microsatellite primers

The first step in the process is to prepare a master mix that contains all PCR components apart from the one variable component (either primer or template DNA). Then, the desired amount of the prepared master mix is added to the PCR plate. Then, the required quantity of variable constituent (primer or template DNA) is put into each PCR plate and spins it for mixing properly. The PCR plate is put in a thermocycler and run by using a standardized PCR program. Amplified PCR products have been detected on an agarose gel or polyacrylamide gel electrophoresis.

8.2.5.2 Amplification by utilizing labeled microsatellite primers

The protocol remains similar as stated above, except using labeled primer, either forward or reverse. In this approach, primers are initially modified at 5' end with fluorophores, TET, FAM, NED, etc., based on the electrophoresis platform used for the discernment of amplified PCR product. Automated capillary electrophoresis has been employed for the detection of amplified PCR products with the labeled primers for high-throughput genotyping.

8.2.5.3 Amplification by utilizing M13-tailed microsatellite primers and M13-labeled probes

For the cost savings in labeling primers with fluorophores, M13-tailed PCR has been introduced (Schuelke, 2000). This method is similar in steps as stated above in section (8.2.5.1), with the exception of using M13-labeled probe, i.e. M13 sequence labeled with different fluorophores FAM, NED, HEX, etc., at its 5' end in addition to unlabeled forward and reverse primers. PCR product and fluorescent dye labeling are achieved in a single reaction. Polymerase chain reaction is done utilizing three primers: a forward primer with M13 sequences, a reverse primer, and a primer with fluorescent-labeled M13 sequence. The ratio of primers to probes is an essential aspect in deciding the performance of the amplification and label incorporation (fluorophore).

Population genetics theories and tools are mainly used in answering various queries in the area of genetics, conservation biology, as well as to any background in which the function of dispersal and gene flow is significant (Kim & Sappington, 2013). Most population genetics deal with the study of genetic variation utilizing microsatellite markers (Kim & Sappington, 2006). Various software are presently used to analyze SSRs' genotypic data, but very few are widely employed for calculations of genetic variations, genetic structure, occurrences of spatiotemporal gene flow, and phylogenetic relationship between and within populations (Kim & Sappington, 2013). Microsatellite markers are a very informative marker for population genetic research due to their important feature presence, such as multiple alleles per locus and their ability to the distinction of heterozygous individuals. With the advancement in technologies the pace for advanced population genetic analysis also deployed: (1) bottleneck tests are used to determine the demography of a population (Cornuet & Luikart, 1996). Utilizing three distinct methods, bottlenecks of the population can be assessed. Wilcoxon tests are available in the software BOTTLENECK for assessing the degree of observed heterozygosity. Second, one might search for a change from the L-shaped scattering of alleles frequency predicted under mutation-drift equilibrium (Luikart, Allendorf, Cornuet, & Sherwin, 1998). Third, the M value of Garza and Williamson is determined using the software AGARST. This test can recognize bottlenecks that were encountered previously (Garza & Williamson, 2001).

The informativeness of a distinct SSR locus or a loci group mainly depends on the number of polymorphic alleles and their relative proportion in the genome (Botstein, White, Skolnick, & Davis, 1980). The informativeness of a marker is represented mainly in terms of Polymorphism Information Content (PIC) and heterozygosities. PIC is the expected ratio of informative descendants from the series of pedigree. It is always the case that the PIC statistic is always lower or equal to heterozygosity, with the two being closely correlated (Hildebrand, David, Torney, & Wagner, 1994).

8.2.6 Cross-transferability of SSR markers

The SSR markers developed are used to amplify DNA from different species, other than for which SSRs are designed. While calculating the cross-transferability in each case primer wise and species wise cross-species transferability can be measured based on positive amplifications at targeted markers loci. In the literature, cross-species transferability has been done in different plant species Ambrosia (Sharma, Hyvönen, & Poczai, 2020), Fern (Kumar, Bhandawat, Sharma, Nag, & Sharma, 2016), onion (Jayaswall et al., 2019; Jayaswall et al., 2021), bamboo (Bhandawat, Sharma, Nag, et al.,

2015; Bhandawat, Sharma, Sharma, et al., 2015), tea (Bhardwaj et al., 2013; Sharma et al., 2011), rose (Sharma et al., 2015), and rhododendron (Sharma, Bhandawat, & Rawat, 2020; Sharma, Bhandawat, Kumar, et al., 2020).

8.2.7 Future perspective

SSR markers are mainly designed and considered in varied crops and other economically important species. With more and more genomes sequenced and made freely available to the public, sequence-based markers, such as microsatellites, have become an efficient approach for DNA fingerprinting, linkage mapping, marker–trait association studies, diversity studies, population genetics, plant breeding, and marker-aided selection. The use of microsatellite markers in marker–trait association mapping offers significant potential for exploiting diverse germplasm, defining phenotypic variation, and characterizing marker–trait association in diverse germplasm for the mining of key alleles associated with the trait of interest. They are used so commonly due to their cost benefits, user-friendliness, and excellent performance. Microsatellite markers are mainly used majorly in genetics and genomics research, including analysis of genome and mapping of a gene. Still, many disadvantages prevent their use, such as stutter banding, null alleles, and heterologous amplicons. In summary, genomic technologies' pace of development in microsatellites will enable their usage more desirable in molecular breeding, genomics, and genetics, and ultimately, they will be of significant use in crop improvement.

8.3 Analysis of SNP markers

8.3.1 Approaches for sequence data generation

The beginning of high-throughput NGS technologies has facilitated a large amount of sequence data generation for genomic usage to study nonmodel organism and unravel the hidden variations in population genomics. For this purpose, the selection of the correct sequencing approach should be taken into account to address the biological question.

8.3.2 Sample preparation

Sample preparation for genomic sequence analysis has great importance and generally determines the quality and quantity of sequence data generated. The realization of the above-mentioned mainly is determined by the use of good quality high molecular weight nucleic acid sample isolation and its subsequent usage for the sequencing process. Ideally, when plant tissue samples are collected from the field, they should be immediately snap-frozen using liquid nitrogen or at -20°C using dry ice such that it freezes the cellular nucleic acid in frozen tissue and protects from degradation (Semagn, 2014). For obtaining a high molecular weight of good quality nucleic acid, the isolation protocol used is very crucial (Doyle & Doyle, 1990). Prior ethical permission and documentation regarding the collection of tissue samples should be taken from legal sources.

8.3.3 Sequencing of complex genomes

Based on research requirements, one can choose the sequencing strategies to generate a large amount of sequence data. In recent years more frequently genome intricacy reduction approach has been used to sequence complex genomes of large sizes. The above approach works in the following two ways; (1) restriction enzyme-based genomic DNA digestion method and (2) capturing the genomic DNA fragment of interest with synthetic baits.

8.3.3.1 Restriction enzyme-based DNA sequencing

The creative method for restriction-site-associated DNA sequencing (RAD-Seq) was prescribed by Baird et al. (2008). This involves the digestion of genomic DNA using a single restriction enzyme. The sequencing method offers flexibility in the case of a restriction enzyme to be selected to achieve the preferred reduction factor. Therefore sequencing of either a few or many loci at high or low coverage can be achieved using RAD-Seq. The RAD-Seq principle has been commonly used in genotyping by sequencing (Elshire et al., 2011) and double-digest RAD-Seq (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Both the above-mentioned technologies have been frequently used for population genomics in a wide variety of organisms, mainly due to the flexible nature and cost-effectiveness (Rheindt, Fujita, Wilton, & Edwards, 2014).

8.3.3.2 Sequence capture

Another alternative method to reduce genome complexity is sequence capture. The process involves identifying the region of interest and synthesizing oligonucleotide baits complementary to the region of interest. By using these synthetic oligonucleotide baits, one can recover the genomic regions of interest by the use of hybridization. The nature and number of baits to be used in the process of sequence capture could be measured by the research needs and may include any number of genes to the genomic intervals of interest (Evans et al., 2014).

8.3.3.3 Whole genome sequencing

Whole genome sequencing is used frequently in population genomics to identify the possible different sites in an individual's genome, which are organizationally embodied in the available reference genome. However, a sequencing depth preferably greater than $50\times$ per individual is recommended (Schreiber, Stein, & Mascher, 2018). The key disadvantage of this approach is the cost of sequencing per individual, which is significantly higher than other approaches, with large and complex genome sizes. Therefore this leads the researchers to compromise between the required sequencing depth and the number of individuals to be sequenced. A study suggested that more reliable results could be obtained by sequencing a larger number of individuals at a lower sequencing depth (Fumagalli, 2013).

8.3.3.4 Transcriptome sequencing

Yet another approach to obtain data of transcriptionally active genes within the genome is through transcriptome sequencing. The processed sequencing data can be mapped alongside the reference genome to detect the genic variants. Apart from reference-based mapping, de novo assembly and functional annotation of the transcriptome data are comparatively easy (Haas et al., 2013). However, the relative abundance of the transcripts and genes is highly dependent on the tissue or organs sequenced for the organism. However, still, it is a good option to sequence organisms with large and complex genomes (Schreiber et al., 2018).

8.3.3.5 Sequencing platforms

Based on the research requirements, one can opt for the sequencing platform better suitable match for the project. Organisms with already available reference genomes should preferably use for resequencing using short-read technology (Chia et al., 2012). On the other hand, short reads could be combined with longer reads technology to identify genetic variants in an organism whose reference genome is not available (Berlin et al., 2015; Denoeud et al., 2014; Ummat and Bashir, 2014). At present Illumina offers a wide range of sequencing platforms for high-throughput sequence data generation. The currently available suit of sequences, that is, HiSeq, MiSeq, MiniSeq, NextSeq, and NovaSeq, can yield up to 6 Gb of sequencing data in a few hours to a minimum of 4 days. Here, we summarize the characteristics of different sequencing technologies (Table 8.1).

8.3.4 Assessment of sequence quality

The quality of the raw reads is very important for downstream processing of the sequencing data in the next-generation high-throughput sequencing analysis. Numerous bioinformatics tools are existing to measure the quality of the raw sequencing reads and are adopted as the NGS QC toolkit (Patel & Jain, 2012) and FastQC toolkit (Andrews, 2010). These tools provide the initial analysis report on the basic statistics of the raw reads (per base quality, sequence quality score, GC content, per base N content, sequence length distribution level, etc.). Trimming of the poor sequence reads at the 5' and 3' ends is performed for the removal of adaptor sequence contamination, several tools are available, such as AdaptorRemoval (Lindgreen, 2012), Trimmomatic (Bolger, Lohse, & Usadel, 2014), Trim Galore (Krueger, 2015), and Cutadapt (Martin, 2011) for trimming. The usage of the tools is not specific to the sequencing platform rather it extremely depends on the dataset and the parameters used for downstream analysis.

8.3.5 Reads assembly and mapping to reference genome

Bioinformatics analysis proceeds toward the second step, which aligned or maps the assembled reads to the model genome. In the process, a mapping algorithm will try to map the assembled reads to the reference genome at the exact matching position using the specific conditions that allow mismatches to a certain number for subsequent sequence variation. Around hundreds of tools are available to map the assembled NGS data to the reference genome. The mainly and frequently used is the Bowtie and Burrows–Wheeler Aligner (BWA). Bowtie is mainly used for mapping reads

TABLE 8.1 List of various sequencing platforms for next-generation sequencing and their application.

Sequencing platform	Library type	Maximum read length (bp)	Maximum reads per run	Error rate (%)	Advantages	Disadvantages	References
454/GS FLX titanium XL	Single end/ paired end	1000	1,000,000	0.2	Medium size reads. Errors are well characterized	Inaccurate homopolymers' detection	Shao et al. (2013)
Illumina-Solexa/HiSeq 4000	Single end/ paired end	2 × 150	5,000,000,000	0.2–0.8	Widely used. Flexible library preparation methods	Not optimal for de novo assembly	Quail et al. (2012)
Illumina MiSeq	Single end/ paired end	2 × 300	25,000,000	–	High throughput is well suited for resequencing projects. Good characterization of biases	–	Schirmer et al. (2015)
Life Technologies/SOLiD	Single end/ paired end	1 × 75/2 × 50	1,400,000,000	0.01	Second most used. Second highest throughput. Each base is read twice, thus decreasing the error rate	Color space is not supported by many mappers. Short reads	Van Dijk, Auger, Jaszczyszyn, and Thernes (2014)
PacBio RS II/SMRT	Single end	20,000	55,000	13	Longest reads. Good for improving de novo assemblies. Single molecule sequencing	Low throughput. High cost-throughput ratio. High error rate	Quail et al. (2012)
Life Technologies/Ion Torrent	Single end/ paired end	400	5,500,000	1.8	Short running times	Bias against AT-rich regions. Inaccurate homopolymers detection	Quail et al. (2012)
Oxford Nanopore/MinION	Single end	2000	60,000	30	Long reads. No GC bias. Portable, scalable, real-time data. It is possible to read both strands of the DNA sequence	High error rates. Quality scores are only defined by the quality of the alignment to a reference sequence	Cherf et al. (2012)

that have been generated in RNA-Seq or transcriptome sequencing, whereas BWA can also be used efficiently to map both short reads and long reads sequencing data generated by various NGS platforms (Langmead, Trapnell, Pop, & Salzberg, 2009; Li & Durbin, 2010; Miller, Koren, & Sutton, 2010; Ruffalo, LaFramboise, & Koyutürk, 2011).

8.3.6 Postprocessing of mapped reads

Postprocessing of mapped reads is essential for the escalation of the accuracy and quality of mapped data before performing variant calling. In this step, we perform filtering of duplicate reads and improve base quality score (Robinson, Piro, & Jager, 2017; Tian, Yan, Kalmbach, & Slager, 2016). To do so, one can use the Picard tools (Zhao, 2018) and SamTools (Li et al., 2009). The postprocessing of mapped reads is required because the variant calling algorithm considers that the mapped reads are free of fragmentation-based libraries and PCR duplicates, therefore assuming that all sites mapped are unique (Robinson et al., 2017).

8.3.7 Variant calling and filtration

Identifying variants in the postprocessed mapped file is the main objective of the NGS bioinformatics pipeline. Several tools are available that can recognize variations based on high confidence base calls, likelihood, Bayesian methods using mapping quality scores to identify variants. Mainly these tools use SAM/BAM file format as input and produce a variant call format file as output (Danecek et al., 2011). The initial process of variant calling identifies hundreds–thousands of variants that require further filtering to lessen untrue positive variant calls. This requires the user to define the threshold line on the observed data and remove variants that might have occurred due to sequencing bias or low coverage. Therefore the user can apply stringent filtering parameters to identify high-quality variants. The most commonly used filter is minor allele frequency (Fig. 8.2); this can sort the variants based on rare allele into three classes: (1) rare variant (<0.5%), (2) low frequency (0.5%–5%), and (3) common variants (> 5%) (Li et al., 2009). The list of bioinformatics tools and software used in the NGS downstream data analysis is described in Table 8.2.

8.3.8 Functional annotation of variant

Defining biological context to the identified variants is of utmost importance and the key step in NGS analysis. The VCF file obtained after variant analysis contains information about the biological consequence of the identified variant. To predict the effect (synonymous, nonsynonymous, stop-gain, start-lost, and frame-shift) of identified variant, several tools are available, such as ANNOVAR (McLaren et al., 2010), variant effect predictor (Ng et al., 2009), and SnpEff (Cingolani et al., 2012). The ANNOVAR and SNPEff are command line-based tools that can annotate variants as SNPs

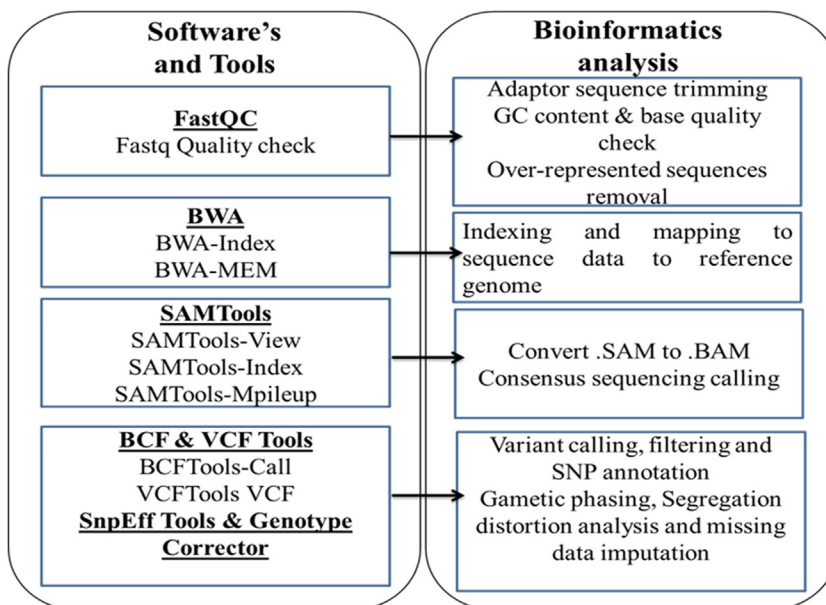


FIGURE 8.2 The flow diagram showing the bioinformatics workflow for sequence data processing and variant calling.

TABLE 8.2 List of various bioinformatics tools used in next-generation sequence data processing.

Preprocessing of raw reads					
Tools	Library type	NGS platform	Input format	Output format	References
NGS QC toolkit	Single, pair end	454, Illumina1	FastQ	FastA, FastQ	Patel and Jain (2012)
CutAdapt	Single, pair end	454, Illumina, SOLID	FastQ	FastQ, FastA	Martin (2011)
SolexaQA	Single, pair end	Illumina	FastQ	FastQ	Cox, Peterson, and Biggs (2010)
Sickle	Single, pair end	Illumina	FastQ	FastQ	
Trimmomatic	Single, pair end	Illumina	FastQ	FastQ	Bolger et al. (2014)
FastQC	Single, pair end	Illumina	FastQ	FastQ	Andrews (2010)
Adaptor removal and trimming					
Tools	Library type	NGS platform	Input format	Output format	References
Cutadapt	Single, pair end	454, Illumina, SOLID	FastQ	FastQ	Martin (2011)
Trimmomatic	Single, pair end	Illumina	FastQ	FastQ	Bolger et al. (2014)
FastX toolkit	Single, pair end	454, Illumina, SOLID	FastQ, FastA	FastQ, FastA	Gordon and Hannon (2010)
Aligning reads to a reference					
Tools	Library type	Algorithm	Input format	Output format	References
Bowtie2	Paired	Burrows–Wheeler transform	Fasta	Bam	Langmead and Salzberg (2012)
Burrows–Wheeler Aligner	Paired	Burrows–Wheeler transform	Fastq	Bam	Li et al. (2009)
Stampy	Paired	Burrows–Wheeler transform	Fastq	Sam	Lunter and Goodson (2011)
Variant annotation					
Tools	Description	Variant types	Input format	Output format	References
ANNOVAR	Command line tool	SNPs, INDELs, CNVs, and block substitutions	VCF	VCF	Wang, Li, and Hakonarson (2010)
PolyPhen-2	Web application	Predicts the impact of amino acid substitution on protein	FASTA	–	Adzhubei, Jordan, and Sunyaev (2013)
SnPEff	Command line tool	SNPs, INDELs, and multiple-nucleotide polymorphisms	VCF, BED	VCF	Cingolani et al. (2012)

(Continued)

TABLE 8.2 (Continued)

Preprocessing of raw reads					
Tools	Library type	NGS platform	Input format	Output format	References
SnpsSIFT	Command line tool	Filter and manipulate variant	VCF	VCF	Ruden et al. (2012)
VEP	Web application	Determines the effect of variants on genes	Coordinates of variants and nucleotide changes; whitespace-separated format, VCF, pileup, HGVS	VCF, JSON, Statistics	McLaren et al. (2016)

NGS, Next-generation sequencing; *SNPs*, single-nucleotide polymorphisms; *VEP*, variant effect predictor.

and INDELs, major difference between ANNOVAR and SNPEff is that ANNOVAR can also annotate CNVs but same can not be done by SnpEff. The variant effect predictor is the core database from Ensemble that determines the variant based on genomic structure and predicts the consequence. Apart from predicting the consequences of the identified variant, these tools also determine the genomic location (intergenic, intron, exon, 5'- and 3'-UTR variant, and upstream and downstream of gene) for the variant and determine the consequences on the encoded protein.

8.4 Conclusion

With the advancement in sequencing technologies, reduction in per base sequencing cost and generation of humongous sequencing data for population genomics studies have become a trivial question nowadays. The availability of portable sequencing platforms has answered the question that can even sequence a large number of genomes to a certain depth ($10\times$ or $20\times$) to identify structural variants. Thus NGS is growing speedily not to deal with the conventional genomic approach but getting more applicability. The target of characterizing genome diversity particularly for the species with complex repetitive genome remains elusive. The above problem could be addressed using long read sequencing technologies, such as Nanopore and PacBio, to address the problem of gap filling. This will help scientists and researchers to unravel the multifarious biological challenges, thus improving the genetic diversity among species to improve a given trait.

Acknowledgments

National Agri-Food Biotechnology Institute (NABI), Mohali Punjab, Department of Biotechnology, Govt. of India, is acknowledged for support. DeLCON (DBT-Electronic Library Consortium), Gurugram, India, is acknowledged for access to the e-resources.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1), 7–20.
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, 3(10), e3376.
- Batley, J., & Edwards, D. (2007). SNP applications in plants. In N. Oraguzie, E. Rikkerink, S. Gardiner, & H. Nihal De Silva (Eds.), *Association mapping in plants* (pp. 95–102). New York: Springer.
- Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6), 623–630.

- Bhandawat, A., Sharma, H., Nag, A., Singh, S., Ahuja, P. S., & Sharma, R. K. (2015). Functionally relevant novel microsatellite markers for efficient genotyping in *Stevia rebaudiana* Bertoni. *Journal of Genetics*, *94*(1), 75–81.
- Bhandawat, A., Sharma, H., Pundir, N., Madhawan, A., & Roy, J. (2020). Genome-wide identification and characterization of novel non-coding RNA-derived SSRs in wheat. *Molecular Biology Reports*, *47*(8), 6111–6125.
- Bhandawat, A., Sharma, V., Sharma, H., Sood, A., & Sharma, R. K. (2015). Development and cross-transferability of functionally relevant microsatellite markers in *Dendrocalamus latiflorus* and related bamboo species. *Journal of Genetics*, *94*(1), 48–55.
- Bhardwaj, P., Kumar, R., Sharma, H., Tewari, R., Ahuja, P. S., & Sharma, R. K. (2013). Development and utilization of genomic and genic microsatellite markers in Assam tea (*Camellia assamica* ssp. *assamica*) and related *Camellia* species. *Plant Breeding*, *132*(6), 748–763.
- Bhardwaj, P., Sharma, R. K., Kumar, R., Sharma, H., & Ahuja, P. S. (2014). SSR marker based DNA fingerprinting and diversity assessment in superior tea germplasm cultivated in Western Himalaya. *Proceedings of the Indian National Science Academy*, *80*(1), 157–162.
- Blair, M. W., Hurtado, N., Chavarro, C. M., Muñoz-Torres, M. C., Giraldo, M. C., Pedraza, F., . . . Wing, R. (2011). Gene-based SSR markers for common bean (*Phaseolus vulgaris* L.) derived from root and leaf tissue ESTs: An integration of the BMC series. *BMC Plant Biology*, *11*(1), 50.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314.
- Che, K., Zhan, Q., Xing, Q., Wang, Z., Jin, D., He, D., & Wang, B. (2003). Tagging and mapping of rice sheath blight resistant gene. *Theoretical and Applied Genetics*, *106*(2), 293–297.
- Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, *30*(4), 344–348.
- Chia, J. M., Song, C., Bradbury, P. J., Costich, D., De Leon, N., Doebley, J., . . . Gore, M. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, *44*(7), 803–807.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.
- Cordeiro, G. M., Christopher, M. J., Henry, R. J., & Reinke, R. F. (2002). Identification of microsatellite markers for fragrance in rice by analysis of the rice genome sequence. *Molecular Breeding*, *9*(4), 245–250.
- Cornuet, J. M., & Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, *144*(4), 2001–2014.
- Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, *11*(1), 1–6.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158.
- Datta, S., Mahfooz, S., Singh, P., Choudhary, A. K., Singh, F., & Kumar, S. (2010). Cross-genera amplification of informative microsatellite markers from common bean and lentil for the assessment of genetic diversity in pigeonpea. *Physiology and Molecular Biology of Plants*, *16*(2), 123–134.
- Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., . . . Aury, J. M. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science (New York, N.Y.)*, *345*(6201), 1181–1184.
- Doyle, J. J., & Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus (San Francisco, Calif.)*, *12*(13), 39–40.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*(5), e19379.
- Evans, J., Kim, J., Childs, K. L., Vaillancourt, B., Crisovan, E., Nandety, A., . . . Buell, C. R. (2014). Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*. *The Plant Journal*, *79*, 993–1008.
- Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, *8*(11), e79667.
- Garza, J. C., & Williamson, E. G. (2001). Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, *10*(2), 305–318.
- Ghaffari, S., & Hasnaoui, N. (2013). *Microsatellite amplification in plants: Optimization procedure of major PCR components*. *Microsatellites* (pp. 139–146). Totowa, NJ: Humana Press.
- Ghafoor, A., & Ahmad, Z. (2005). Diversity of agronomic traits and total seed protein in black gram *Vigna mungo* (L.) Hepper. *Acta Biologica Cracoviensia Series Botanica*, *47*(2), 69–75.
- Gordon, A., & Hannon, G. J. (2010). *Fastx-toolkit. FASTQ/A short-reads preprocessing tools* (unpublished) (p. 5). http://hannonlab.cshl.edu/fastx_toolkit.
- Gwanne, C., Labuschangne, M. J., & Botha, A. M. (2000). Analysis of genetic variation in *Cucurbita moschata* by random amplified polymorphic DNA (RAPD) markers. *Euphytica*, *113*, 19–24.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., . . . MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512.
- Han, O. K., Kaga, A., Isemura, T., Wang, X. W., Tomooka, N., & Vaughan, D. A. (2005). A genetic linkage map for azuki bean [*Vigna angularis* (Willd.) Ohwi & Ohashi]. *Theoretical and Applied Genetics*, *111*(7), 1278–1287.
- Hildebrand, C. E., David, C., Torney, C., & Wagner, P. (1994). Informativeness of polymorphic DNA markers. *Los Alamos Science/Los Alamos Scientific Laboratory*, *20*, 100–102.

- Huang, Z. P., Li, J. K., & Dai, O. H. (1999). Preliminary study of male sterile mutant *Wh921* and its heterosis in soybean. *Chinese Journal of Oil Crop Sciences*, 21, 20–23.
- Hunter, M. L. (1996). *Fundamentals of conservation biology*. London: Blackwell Sciences.
- Jayaswall, K., Sharma, H., Bhandawat, A., Kumar Yadav, V., Mahajan, V., & Singh, M. (2019). Characterization of *Allium* germplasm for conservation and sustainable management using SSR markers. *Indian Journal of Traditional Knowledge*, 18(1), 193–199.
- Jayaswall, K., Sharma, H., Bhandawat, A., Sagar, R., Jayaswal, D., Kumar, A., & Singh, M. (2021). Chloroplast derived SSRs reveals genetic relationships in domesticated alliums and wild relatives. *Genetic Resources and Crop Evolution*. Available from <https://doi.org/10.1007/s10722-021-01235-z>.
- Jin, J., Zhang, H., Kong, L., Gao, G., & Luo, J. (2014). PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*, 42(D1), D1182–D1187.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kim, K. S., & Sappington, T. W. (2006). Molecular genetic variation of boll weevil populations in North America estimated with microsatellites: Implications for patterns of dispersal. *Genetica*, 127(1–3), 143.
- Kim, K. S., & Sappington, T. W. (2013). *Microsatellite data analysis for population genetics. Microsatellites* (pp. 271–295). Totowa, NJ: Humana Press.
- Krueger, F. (2015). *Trim galore*. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files (Vol. 516, p. 517).
- Kumar, V., Bhandawat, A., Sharma, H., Nag, A., & Sharma, R. K. (2016). Novel microsatellite markers identification and diversity characterization in *Pteris cretica* L. *Journal of plant biochemistry and biotechnology*, 25(1), 104–110.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), 589–595.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079.
- Lightfoot, D. A., & Iqbal, M. J. (2013). *Molecular mapping and breeding with microsatellite markers. Microsatellites* (pp. 297–317). Totowa, NJ: Humana Press.
- Lindgreen, S. (2012). AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Research Notes*, 5(1), 337.
- Luikart, G., Allendorf, F. W., Cornuet, J. M., & Sherwin, W. B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity*, 89(3), 238–247.
- Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6), 936–939.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17(1), 10–12.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, 17(1), 122.
- McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics (Oxford, England)*, 26(16), 2069–2070.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315–327.
- Min, X. J., Butler, G., Storms, R., & Tsang, A. (2005). OrfPredictor: Predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Research*, 33(Suppl. 2), W677–W680.
- Morgante, M., Hanafey, M., & Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics*, 30(2), 194–200.
- Narina, S. S., d’Orgeix, C. A., & Sayre, B. L. (2011). Optimization of PCR conditions to amplify microsatellite loci in the bunchgrass lizard (*Sceloporus slevini*) genomic DNA. *BMC Research Notes*, 4(1), 26.
- Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., ... Bamshad, M. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272–276.
- Parida, S. K., Kalia, S. K., Kaul, S., Dalal, V., Hemaprabha, G., Selvi, A., ... Srivastava, P. S. (2009). Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. *Theoretical and Applied Genetics*, 118(2), 327–338.
- Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*, 7(2), e30619.
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, 7(5), e37135.
- Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, 7(2), e32253.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1), 1–13.
- Rahim, M. S., Sharma, H., Parveen, A., & Roy, J. K. (2018). Trait Mapping Approaches Through Association Analysis in Plants. In R. Varshney, M. Pandey, & A. Chitkineni (Eds.), *Plant Genetics and Molecular Biology. Advances in Biochemical Engineering/Biotechnology*. In: (164). Springer, Cham. Available from https://doi.org/10.1007/10_2017_50.

- Ravishankar, K. V., Mani, B. H. R., Anand, L., & Dinesh, M. R. (2011). Development of new microsatellite markers from Mango (*Mangifera indica*) and cross-species amplification. *American Journal of Botany*, 98(4), e96–e99.
- Renganayaki, K., Read, J. C., & Fritz, A. K. (2001). Genetic diversity among Texas bluegrass genotypes (*Poa arachnifera* Torr.) revealed by AFLP and RAPD markers. *Theoretical and Applied Genetics*, 102(6–7), 1037–1045.
- Rheindt, F. E., Fujita, M. K., Wilton, P. R., & Edwards, S. V. (2014). Introgression and phenotypic assimilation in *Zimmerius* flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide SNPs. *Systematic Biology*, 63(2), 134–152.
- Risterucci, A. M., Duval, M. F., Rohde, W., & Billotte, N. (2005). Isolation and characterization of microsatellite loci from *Psidium guajava* L. *Molecular Ecology Notes*, 5(4), 745–748.
- Robinson, P. N., Piro, R. M., & Jager, M. (2017). *Computational exome and genome analysis*. CRC Press.
- Rozen, S., & Skaletsky, H. (2000). *Primer3 on the WWW for general users and for biologist programmers*. *Bioinformatics methods and protocols* (pp. 365–386). Totowa, NJ: Humana Press.
- Ruden, D. M., Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., & Lu, X. (2012). Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, 3, 35.
- Ruffalo, M., LaFramboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics (Oxford, England)*, 27(20), 2790–2796.
- Sahoo, J. P., Sharma, V., Verma, R. K., Chetia, S. K., Baruah, A. R., Modi, M. K., & Yadav, V. K. (2019). Linkage analysis for drought tolerance in kharif rice of Assam using microsatellite markers. *Indian Journal of Traditional Knowledge*, 18(2), 371–375.
- Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6), e37.
- Schreiber, M., Stein, N., & Mascher, M. (2018). Genomic approaches for studying crop evolution. *Genome Biology*, 19(1), 140.
- Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology*, 18(2), 233–234.
- Semagn, K. (2014). *Leaf tissue sampling and DNA extraction protocols*. *Molecular plant taxonomy* (pp. 53–67). Totowa, NJ: Humana Press.
- Shao, W., Boltz, V. F., Spindler, J. E., Kearney, M. F., Maldarelli, F., Mellors, J. W., ... Coffin, J. M. (2013). Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology*, 10(1), 18.
- Sharma, A., Namdeo, A., & Mahadik, K. (2008). Molecular markers: New prospects in plant genome analysis. *Pharmacognosy Reviews*, 2(3), 23.
- Sharma, H., Bhandawat, A., Kumar, P., Rahim, M. S., Parveen, A., Kumar, P., ... Roy, J. (2020). Development and characterization of bZIP transcription factor based SSRs in wheat. *Gene*, 756, 144912.
- Sharma, H., Bhandawat, A., & Rawat, S. (2020). Cross-transferability of SSR markers developed in *Rhododendron* species of Himalaya. *Molecular Biology Reports*, 47(8), 6399–6406.
- Sharma, H., Hyvönen, J., & Pocza, P. (2020). Development of chloroplast microsatellite markers for giant ragweed (*Ambrosia trifida*). *Applications in Plant Sciences*, 8(1), e11313.
- Sharma, H., Kumar, R., Sharma, V., Kumar, V., Bhardwaj, P., Ahuja, P. S., & Sharma, R. K. (2011). Identification and cross-species transferability of 112 novel unigene-derived microsatellite markers in tea (*Camellia sinensis*). *American Journal of Botany*, 98(6), e133–e138.
- Sharma, R. K., Chaudhary, A., Sharma, H., Bhardwaj, P., Sharma, V., Kumar, R., & Ahuja, P. S. (2015). Identification and cross-species amplification of microsatellite markers derived from expressed sequence data of rose species. *Journal of Plant Biochemistry and Biotechnology*, 24(3), 359–364.
- Sharma, V., Verma, R. K., Dey, P. C., Chetia, S. K., Baruah, A. R., & Modi, M. K. (2017). QTLs associated with yield attributing traits under drought stress in upland rice cultivar of Assam. *ORYZA-An International Journal on Rice*, 54(3), 253–257.
- Singh, P., Sharma, H., Nag, A., Bhau, B. S., & Sharma, R. K. (2015). Development and characterization of polymorphic microsatellites markers in endangered *Aquilaria malaccensis*. *Conservation Genetics Resources*, 7(1), 61–63.
- Taheri, S., Lee Abdullah, T., Yusop, M. R., Hanafi, M. M., Sahebi, M., Azizi, P., & Shamshiri, R. R. (2018). Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. *Molecules (Basel, Switzerland)*, 23(2), 399.
- Thiel, T., Michalek, W., Varshney, R., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, 106(3), 411–422.
- Tian, S., Yan, H., Kalmbach, M., & Slager, S. L. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics*, 17(1), 403.
- Ummat, A., & Bashir, A. (2014). Resolving complex tandem repeats with long reads. *Bioinformatics (Oxford, England)*, 30(24), 3491–3498.
- Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9), 418–426.
- Verma, R. K., Chetia, S. K., Dey, P. C., Rahman, A., Saikia, S., Sharma, V., & Modi, M. K. (2021). Genome-wide association studies for agronomical traits in winter rice accessions of Assam. *Genomics*, 113(3), 1037–1047.
- Verma, R. K., Chetia, S. K., Dey, P. C., Sharma, V., Baruah, A. R., & Modi, M. K. (2017). Development of advanced breeding lines for high grain yield under drought stress in elite rice genetic background. *Research on Crops*, 18(4), 705–710.
- Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), e164.
- Xu, K., Deb, R., & Mackill, D. J. (2004). A microsatellite marker and a codominant PCR-based marker for marker-assisted selection of submergence tolerance in rice. *Crop Science*, 44(1), 248–253.
- Xu, P., Wu, X., Wang, B., Luo, J., Liu, Y., Ehlers, J. D., ... Li, G. (2012). Genome wide linkage disequilibrium in Chinese asparagus bean (*Vigna unguiculata* ssp. *sesquipedalis*) germplasm: Implications for domestication history and genome wide association studies. *Heredity*, 109(1), 34–40.

- Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., . . . Wang, J. (2006). WEGO: A web tool for plotting GO annotations. *Nucleic Acids Research*, *34*(Suppl. 2), W293–W297.
- Zhang, Q., Wang, C. L., Zhao, K. J., Zhao, Y. L., Caslana, V. C., Zhu, X. D., & Jiang, Q. X. (2001). The effectiveness of advanced rice lines with new resistance gene Xa23 to rice bacterial blight.
- Zhao, Q. (2018), A study on optimizing markduplicate in genome sequencing pipeline. In *Proceedings of the 2018 5th international conference on bioinformatics research and applications* (pp. 8–15).

Further reading

- Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, *177*(3), 309–334.