




## REVIEW ARTICLE

# Skim sequencing: an advanced NGS technology for crop improvement

PARDEEP KUMAR<sup>1\*</sup>, MUKESH CHOUDHARY<sup>1</sup>, B. S. JAT<sup>1</sup>, BHUPENDER KUMAR<sup>1</sup>, VISHAL SINGH<sup>1</sup>, VIRENDER KUMAR<sup>2</sup>, DEEPAK SINGLA<sup>3</sup> and SUJAY RAKSHIT<sup>1</sup>

<sup>1</sup>ICAR-Indian Institute of Maize Research, PAU Campus, Ludhiana 141 004, India

<sup>2</sup>National Agri-food Biotechnology Institute (NABI), Mohali 140 308, India

<sup>3</sup>School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana 141 004, India

\*For correspondence. E-mail: pardeepkumar656@gmail.com.

Received 30 October 2020; revised 28 December 2020; accepted 18 January 2021

**Abstract.** High-throughput genotyping has become more convenient and cost-effective due to recent advancements in next-generation sequencing (NGS) techniques. Numerous approaches exploring sequencing advances for genotyping have been developed over the past decade, which includes different variants of genotyping-by-sequencing (GBS), and restriction-site associated DNA sequencing (RAD-seq). Most of these methods are based on the reduced representation of the genome, which ultimately reduces the cost of sequencing by many folds. However, continuously lowering the cost of sequencing makes it more convenient to use whole genome-based approaches. In this regard, skim sequencing, where low coverage whole-genome sequencing is used for the identification of large numbers of polymorphic markers cost-effectively. In the present review, we have discussed recent technological advancements, applicability, and challenges of skim sequencing-based genotypic approaches for crop improvement programmes. Skim sequencing is being extensively used for genotyping in diverse plant species and has a wide range of applications, particularly in quantitative trait loci (QTL) mapping, genomewide association studies (GWAS), fine genetic map construction, and identification of recombination and gene conversion events in various breeding programmes. The cost-effectiveness, simplicity, and genomewide coverage will increase the application of skims sequencing-based genotyping. The article summarizes the protocol, uses, bioinformatics tools, its application, and future prospects of skim sequencing in crop improvement.

**Keywords.** next-generation sequencing; genotyping-by-sequencing; skim sequencing; marker; genome.

## Introduction

The concept of forward and reverse genetics was augmented with the introduction of molecular markers. Molecular markers are used for the genetic characterization of individuals and to associate them with specific traits (Scheben *et al.* 2018). Genotyping data in association with phenotypic data have the potential to accelerate the pace of genetic research and breeding applications including genomic diversity, phylogenetic studies, DNA barcoding, marker-trait association, marker-assisted selection (MAS), and genomic selection (GS) (Scheben *et al.* 2017). Initially, genotyping was based on polymorphism in the restriction site of an enzyme such as restriction fragment length polymorphisms (RFLPs). With the discovery of polymerase chain reaction (PCR), PCR-based

markers such as random amplified polymorphic DNA (RAPD), simple sequence repeats (SSRs), inter-simple sequence repeats (ISSRs), and the hybrid of PCR and RFLP markers like amplified fragment length polymorphisms (AFLPs) has been designed. The sequenced characterized amplified region has been developed by sequencing specific polymorphic DNA bands from a gel. Among these markers, SSRs have been widely used for genetic mapping and diversity studies across a wide array of crops due to their relative abundance and codominant nature (Kumar *et al.* 2008; Rakshit *et al.* 2010; Kumar *et al.* 2014, 2015, 2019). However, most of these markers have low genome coverage; require large numbers of gel electrophoresis, time-consuming analysis, and high cost (Huang *et al.* 2009). In the present era, these problems to a large extent have been addressed by next-generation

sequencing (NGS) based single-nucleotide polymorphism (SNPs) markers. SNP markers are currently used for high-throughput genotyping because of their high abundance within genomes and their amenability to automation. Due to its numerous benefits, SNP became the marker of choice replacing PCR-based markers (Huang *et al.* 2009). Initially, SNP differences among individuals were assessed through microarray technology that is based on the hybridization of genomic DNA with oligonucleotides spotted on the gene chip. Later, the array-based genotyping method was further improved by adding more markers that enable the identification of thousands of markers in a single hybridization process (Winzeler *et al.* 1998). However, it has certain limitations as it involves more laborious and time, expensive to design the chip, and is suited for specific populations only.

Sequencing-based genotyping and genetic mapping to a larger extent has been used addressing the issues elaborated earlier. This is further strengthened through the genotyping of the large number of samples using a multiplex sequencing technique (Craig *et al.* 2008; Cronn *et al.* 2008). The multiplex sequencing technique allows using more number of libraries that are to be pooled and sequenced together at high density in a single instrument. The technique is more useful when researchers want to target a specific region in the genome or the tiny genome (Church and Kieffer-Higgins 1988). The availability of a reference genome for various plant species increases the applicability of high-throughput genotyping. Besides that, for crops without the reference genome, comparative genome alignment of the parental lines help to establish the relationship of a particular trait with genotypic variations. Presently, SNP-based genotyping offers several advantages in terms of cost-effectiveness, time-saving, genomewide coverage, high resolution, and establishment of syntenic relationships and has been successfully used for genotype screening, genetic mapping, purity testing, parent testing, haplotype map construction, association mapping, MAS, GS, etc. (Batley and Edwards 2007; Baird *et al.* 2008; Kirst *et al.* 2011; Metzker 2010). The availability of cheaper NGS based approaches like the restriction site associated genomic DNA (RAD), genotyping-by-sequencing (GBS), and reduced-representation sequencing (RRS) provide cost-effective genotyping, however, these techniques are still more expensive in terms of per marker cost compared to skim sequencing. The skim sequencing, which is based on whole-genome sequencing with low genome coverage provides large numbers of SNPs markers. Further refinement of the existing NGS technologies like Illumina Infinium assay, Ion torrent, Roche454, and the Affymetrix GeneChip has the potential to revolutionize the genome-based high-throughput genotyping and its utilization for association mapping studies in crop improvement programmes (Golicz *et al.* 2015). This review covers the basic principles, methodology, application for crop improvement and future perspectives of skim sequencing.

## Overview of GBS and skim sequencing

Genotyping by sequencing is an extensively used technology that relies on reduced genome via reduced-representation sequencing (RRS) or restriction-site associated DNA sequencing for detection of SNPs in population for diversity assessment and QTLs mapping (Baird *et al.* 2008; Elshire *et al.* 2011; Scheben *et al.* 2017). Previously, two different strategies of GBS have been developed (Poland *et al.* 2012). First is based on the restriction enzyme digestion with genome complexity reduction but the chance of missing important SNP is high, however, ideal for marker discovery for MAS. The reduced representation sequencing (RRS) approach of GBS includes cost-effectively genomewide coverage. The reduced-representation sequencing approach initially used restriction digestion, ligation of adaptors then pooling of DNA samples, and sequencing. The sequenced reads from a different individual can be isolated based on bioinformatic tools and based on an adaptor containing new barcode sequences (Davey *et al.* 2011). Similarly, restriction site associated genomic DNA (RAD) sequencing differs from RRS in the requirement of two adaptors and one extra random sharing of DNA. RAD sequencing-based GBS allows the reduction of genome complexity before sequencing and hence reduce the cost per sample and extensive efforts in data analysis. The second strategy, whole-genome resequencing (WGRS) allows low coverage of and involves multiplex enrichment PCR-based SNP identification in the genome, which is important for target traits-specific primer designing (Huang *et al.* 2009). Although, the popularity of these sequencing has increased in recent years, yet skim sequencing has emerged as a promising approach in different crops, namely Brassica and chickpea, rice, groundnut, beet and tobacco by Bayer *et al.* (2015), Kale *et al.* (2015), Kumar *et al.* (2019), Galewski and McGrath (2020), Tong *et al.* (2020), respectively. Skim sequencing addresses the limitations of the GBS approach related to missing regions of the genome due to its lower depth sequencing (Kale *et al.* 2015).

Skim sequencing is different from the RRS, RAD sequencing, and GBS in terms of avoidance of complexity reduction step and low depth sequencing (1-2x) with more coverage of the genome and cost-effective genotyping of large populations (Huang *et al.* 2009; Bayer *et al.* 2015). The most commonly used population in skim sequencing are recombinant inbred lines (RILs) and doubled haploid (DH) population. For simplifying the data analysis, heterozygous alleles need to be eliminated from data (Bayer *et al.* 2015). This will be an alternative to whole-genome sequencing in the future as most labs cannot afford throughout the world and would boost the modern breeding techniques like genomewide association study (GWAS) and genomic selection (GS).

**Table 1.** Comparisons between GBS and skim sequencing.

Description	GBS	Skim sequencing
Cost per sample	Low	High
Cost per marker data point	Moderate	Low
SNP discovery rate	Low to moderate	High
Restriction enzyme	Required	Not required, which reduced the cost of library preparation with a decreased number of steps
Analysis of complexity	Moderate	High
Prior genomic knowledge or reference genome	Not essential	Essential
Genome coverage	More coverage of genome with low resolution in reduced genome representative	Low genome coverage with high resolution on a whole-genome basis
Imputation	Required but not crucial	Very crucial step
Applications	De novo SNPs discovery and genetic mapping	SNPs discovery and high-resolution mapping and genetic mapping
Drawback	Labour intensive library preparation and high read depth variation	High cost and prior genomic information is required

### Comparisons between GBS and skim sequencing

The RAD sequencing-based GBS methods described by He *et al.* (2014) have been more popular since its discovery due to their high coverage of the large population. This approach helps to reduce genome complexity before sequencing through the use of a restriction enzyme. In contrast, WGRS based skim sequencing uses low-coverage whole-genome sequencing for the characterization of crossover and non-crossover distributions. Further, skim sequencing reduces the cost of library preparation by skipping the step of genome complexity reduction and potentially eliminates biases stemming from the use of restriction enzymes. The skim sequencing produces large numbers of SNPs markers as compared to RRS, RAD sequencing and GBS (a portion of markers for each individual are genotyped). The higher SNP discovery rate lowers the cost per SNP markers in skim sequencing as compared to GBS. The current era is shifting toward the WGRS approach due to the reduced cost of NGS data generation (Davey *et al.* 2011). However, it has a few limitations in terms of imputation accuracy in crops with nonavailability of the high-quality reference genome, high cost per sample (depending upon target coverage), and higher complexity in data analysis (Scheben *et al.* 2017). The genome content within the diverse accessions varies and a single reference genome is not enough for the imputation of polymorphic markers. The relative comparison between GBS and skim sequencing over various features have been provided in table 1. The skim sequencing provides opportunities in large polyploid genomes such as wheat as routine WGRS cost would be very high to sequence large populations for genomewide markers identification (Scheben *et al.* 2017).

The important point that needs to be considered during skim sequencing is that the SNP can be determined from

the existing SNPs list for the population or parents sample by mapping to the reference genome; however, high coverage is necessary for efficient detection of SNP from the parents (Bayer *et al.* 2015). In the case of nonavailability of a reference genome, it can be generated from the sequencing reads by *de novo* methods (Scheben *et al.* 2017).

### Concepts and methodology of skim sequencing

The development of the advanced NGS sequencing platforms has made it possible to generate the whole-genome data cost-effective and also resulted in the generation of reference maps in various crops. Skim sequencing is one of the prominent sequencing approaches that use low-coverage whole-genome sequencing for high resolution genotyping (Bayer *et al.* 2015). The sequence alignments of the reference genome along with population are prerequisite for SNP calling using hidden Markov models (Xie *et al.* 2010) and it is a good alternative to deep sequencing of parents which is quite costly. Rice was the first crop plant whose genome was sequenced (Goff *et al.* 2002; Yu *et al.* 2002; Matsumoto *et al.* 2005) after the model plant *Arabidopsis* (*Arabidopsis* genome 2000). The task of the generation of reference genome sequences in different crops accelerated with the adoption of advanced sequencing technologies like Roche 454 (Scheffler *et al.* 2009). Presently, Illumina-based (Illumina HiSeq) sequencing has been used for RRS, RAD sequencing and GBS sequencing. The availability of reference genomes in different crops opened new avenues for the development of sequence-based markers (SNPs) to accelerate crop improvement (Edwards and Batley 2010; Hayward *et al.* 2012; Edwards *et al.* 2013). The parental genome, if available, can also be used for further refinement of the results obtained in this approach, however, it is not

**Table 2.** Different programs/software required for skim sequencing.

Description/role	Program/software	URL	Reference
Sequence read alignment	BWA	<a href="http://bio-bwa.sourceforge.net/bwa.shtml">http://bio-bwa.sourceforge.net/bwa.shtml</a>	Li et al. (2009a, b)
	BWA-SW	<a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a>	Li and Durbin (2010)
SNP calling or variant calling	SOAP2	<a href="http://soap.genomics.org.cn/#">http://soap.genomics.org.cn/#</a>	Li et al. (2009a, b)
	Bowtie	<a href="http://bowtie.cbcb.umd.edu/">http://bowtie.cbcb.umd.edu/</a>	Langmead et al. (2009)
	bowtie2	<a href="http://bowtie-bio.sourceforge.net/bowtie2/index.shtml">http://bowtie-bio.sourceforge.net/bowtie2/index.shtml</a>	Langmead and Salzberg (2012)
	SAMTools	<a href="http://mamttools.sourceforge.net">http://mamttools.sourceforge.net</a> , <a href="http://htslib.org">http://htslib.org</a>	Li (2011)
	GATK	<a href="https://www.broadinstitute.org/gatk">https://www.broadinstitute.org/gatk</a>	De Pisto et al. (2011)
	Platypus	<a href="http://www.well.ox.ac.uk/platypus">http://www.well.ox.ac.uk/platypus</a>	Rimmer et al. (2014)
	Freebayes	<a href="http://github.com/ekg/freebayes">http://github.com/ekg/freebayes</a>	Garrison and Marth (2012)
	BreakDancer	<a href="http://breakdancer.sourceforge.net/">http://breakdancer.sourceforge.net/</a>	Fan et al. (2014)
	Dindel	<a href="https://sanger.ac.uk/resources/software/dindel">https://sanger.ac.uk/resources/software/dindel</a>	Albers et al. (2011)
	SGSautoSNP	On request basis	Lorenz et al. (2012)
Genotype visualize	Flapjack	<a href="https://fcs.hutton.ac.uk/flapjack/">https://fcs.hutton.ac.uk/flapjack/</a>	Milne et al. (2013)
	IGV	<a href="http://software.broadinstitute.org/software/igv/">http://software.broadinstitute.org/software/igv/</a>	James et al. (2011)
Cleaning or filtering of SNP	GATK	GATK offers a tool called VariantFiltration	De Pisto et al. (2011)
	IMPUTE2 v2.3.2	<a href="https://mathgen.stats.ox.ac.uk/impute/impute_v2.html">https://mathgen.stats.ox.ac.uk/impute/impute_v2.html</a>	Howie et al. (2009)
SNP imputation	MACH	<a href="http://csg.sph.umich.edu/abecasis/MACH/tour/imputation.html">http://csg.sph.umich.edu/abecasis/MACH/tour/imputation.html</a>	Li et al. (2006)
	BEAGLE	<a href="http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html">http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html</a>	Browning and Browning (2007, 2009)

mandatory. The skim sequencing can be effectively used for SNPs identification and utilization for QTL mapping, identification of recombination events, and genomewide association analysis as it has better genome coverage as compared to the GBS technique. A study of the recombination events can be employed if the skim sequencing is carried out at optimal coverage (Bayer et al. 2015). SNPs discovery for polyploid genomes is tricky due to the high repeat content and complexity of such genomes. The presence of duplicate regions/genes across polyploid genomes makes the process of development of an accurate genetic map tedious. Skim sequencing can be discussed in two parts for better understanding, i.e. pre-requisite and methodology.

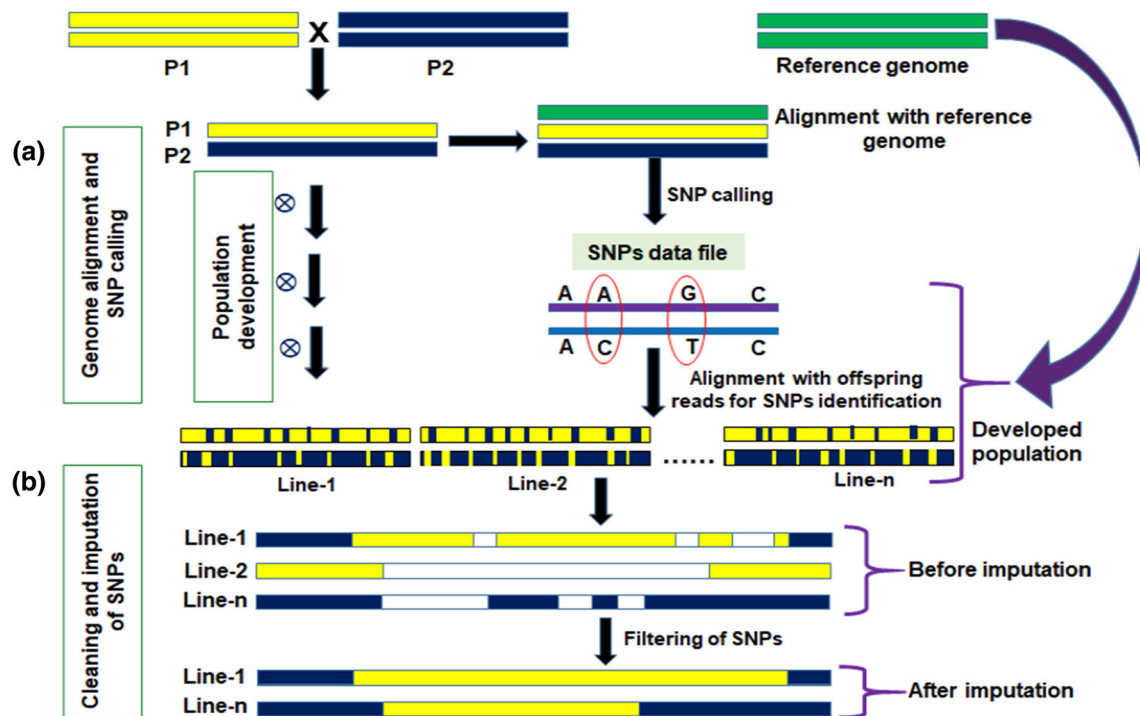
### Prerequisites for skim sequencing

The average genome size of model crops is around 600 Mbp highlighted that crops with similar genome size can be genotyped by SkimGBS (Gregory 2005). It is a two-step process that requires the reference genome and reads from parents as well as from the population. Further, these tasks could not be completed without the help of software, hence there were different software developed to complete the job (table 2). Thus, the requirement of a reference genome is a must for skim sequencing. With the decreasing cost of high-throughput data generation, the availability of a reference genome is continuously increasing. Chen et al. (2018) reviewed the genome sequencing data and observed that ~236 angiosperm and 181 species of horticultural crops have been sequenced of which most of them are of high economic importance. Due to large genome size, genome complexity such as polyploidy and heterozygosity sequencing and assembly of the genome at chromosome level is a daunting task. Thus many of the assemblies are in the draft form either at contig, scaffold, or pseudochromosome levels. Thus, creating the gap for in-depth analysis and highlighted the need for chromosomal assembly to truly utilize the power of skim sequencing for data analysis.

### Methodological steps of skim sequencing

The main steps of skim sequencing include (i) alignment of reads to the reference genome, (ii) variant calling, (iii) cleaning/filtering of SNPs, and (iv) SNP imputation. The above steps are described below and are shown in figure 1.

**Sequence read alignment:** The mapping of short sequence reads with a reference genome is referred to as alignment. The alignment is a difficult task that has been achieved using different softwares, e.g. BWA, BWA-SW, SOAP2, bowtie and bowtie2 based on the application for alignment to reference genome (Koren 2013; Pendleton 2015; Gordon 2016). The data generated through NGS techniques were



**Figure 1.** Fundamental steps for skim sequencing: (a) genome alignment and SNP calling, the genomes of parents or constructed population are aligned along with a reference genome for SNP identification (SNP calling). The yellow, blue, and green bars represent the genome (chromosome) of parents (P1, P2) or population and the reference genome, respectively. (b) Cleaning and imputation SNPs: cleaning/filtering of maximum SNP missing information and imputation leads to filling of such missing information in RILs through haplotype homology. White blocks represent missing information, where SNPs had to be imputed. Line 2 with maximum information missing is eliminated from imputation step.

filtered for adaptor sequenced, read length but low-quality reads have to be removed. A standard file format like sequence alignment map (SAM) and binary alignment map (BAM) were generated which is accepted by most of the softwares. The reads were aligned to the reference genome and used for further analysis. However, the multialigned reads (that align more than one genomic location) are often discarded to reduce the false SNP discovery (Lorenc *et al.* 2012). For NGS data analysis, various pipeline have been developed like Stacks (Catchen *et al.* 2011) or the Tassel GBS pipeline (Glaubitz *et al.* 2014), or improved GBS pipeline (Sonah *et al.* 2013) which is specifically used for the analysis of different types of reads produced through GBS technologies.

**SNPs/variant calling:** After the first step of alignment, the next step was to sequence reads comparison with a reference genome to identify the feasible variants. In low-depth sequencing like skim sequencing, only a few reads were aligned to the region, but was difficult to determine the true variant. Therefore, it is a complex phenomenon to differentiate between true variation and sequence errors (Baird *et al.* 2008) and these errors may generate false-positive variants. However, this will be addressed through an increase in the sensitivity of PCR, the updated version of

alignment software, and huge reference data (Kagale *et al.* 2016). Several software programs are available for SNPs calling, e.g. SAMTools, GATK, Platypus, Freebayes, BreakDancer, Dindel, SGSautoSNP for comparison of sequence reads with reference genome sequence to generate SNPs data for genotype/SNPs visualization.

**SNP genotyping:** After comparing the sequence read alignment with the reference genome, the variants having more than ~80–85% missing alleles have been discarded (Bayer *et al.* 2015). Based on the comparative read alignment of individuals of a population with reference genome, various variants are present at specific position which is called as SNP or allele.

**Visualizing genotypes:** The software tools are a very essential component for genomic data analysis. Further, graphical genotyping is required routinely to handle large datasets generated by current high-throughput sequencing platforms, GBS, and array-based profiling methods. Among the visualizing software, Flapjack is considered as one of the best softwares to provide real-time rendering with rapid navigation and comparisons between lines, markers and chromosomes, with visualization, sorting and querying based on associated data, such as phenotypes, quantitative-trait loci (Milne *et al.*

**Table 3.** Different applications of skim sequencing in crop plants.

Application	Crop	Cross (population)	Finding	Reference
Identification of crossovers or recombination events and gene conversion events	Canola	DH-(Westar × Zhonyou 821)	Total 1663 crossovers	Sun et al. (2007)
	Arabidopsis	F <sub>2</sub> population	Total 73 crossovers and 3000 gene conversion	Yang et al. (2012)
	Soybean	RILs-(Magellan × PI 438489B)	Total 3489 recombination events	Xu et al. (2013)
	Canola	DH-(Tapidor × Ningyou7)	An average of 151.18 and 115.53 crossovers, as well as 697.85 and 374.85 gene conversions per individual for A-genome and C-genome respectively were reported	Bayer et al. (2015)
Discovery of new SNPs	Chickpea	RILs-(PI489777 × ICC4958)	Reported 219 crossovers and 256 gene conversions	Bayer et al. (2015)
	Chickpea	RILs-(ICC 4958 × ICC 1882)	Total 1610 recombination	Kale et al. (2015)
	Rice	RILs ( <i>Oryza sativa</i> ssp. <i>indica</i> cv. 93-11 and ssp. <i>japonica</i> cv. Nipponbare )	Total of 123,302 SNPs with an average of 1 SNP every 3.15 kb	Huang et al. (2009)
	Soybean	RILs-(Magellan × PI 438489B)	109,273 SNPs	Xu et al. (2013)
	Chickpea	RILs-(PI489777 × ICC4958)	511,624 SNPs	Bayer et al. (2015)
	Canola	DH-(Tapidor × Ningyou7)	794,837 SNPs	Bayer et al. (2015)
	Chickpea	RILs-(ICC 4958 × ICC 1882)	53,169 SNPs	Kale et al. (2015)
	Pea	RILs-(Baccara × PI180693)	419,024 SNPs	Boutet et al. (2016)
	Cotton	CRISPR/cas9 edited plants	41,88,6404 SNPs	Li et al. (2019)
	Groundnut	RILs-(ICGV 00350 × ICGV 97045)	10,759 SNPs	Kumar et al. (2019)
QTL Mapping	Tobacco	RILs-(K 36 × Y 3)	100,57,282 SNPs	Tong et al. (2020)
	Rice	RILs ( <i>Oryza sativa</i> ssp. <i>indica</i> cv. 93-11 and ssp. <i>japonica</i> cv. Nipponbare )	Major QTL for plant height on chromosome 1 with phenotypic variation explained (PVE)= 31.3%	Huang et al. (2009)
	Rice	RILs-(Zhenshan 97 × Minghui 63)	Six QTLs for grain weight on chr. 1, 3, 5, 9 with PVE ranging from 7.5–21.8	Yu et al. (2011)
	Soybean	RILs-(Magellan × PI 438489B)	One major QTL for root-knot nematode resistance on chr. 10 with 23.6 PCV	Xu et al. (2013)
	Chickpea	RILs-(ICC 4958 × ICC 1882)	Total 71 major QTL were with two QTL hotspot regions ‘ <i>QTL-hotspot a</i> ’ (139.22 kb; 15 genes) and ‘ <i>QTL-hotspot b</i> ’ for drought tolerant	Kale et al. (2015)
	Rice	RILs-( <i>O. nivara</i> × <i>indica</i> )	Total 65 minor and major QTLs for drought tolerant	Ma et al. (2016)

2013). The other software which is considered as best is integrative genomic viewer (IGV), which handles a large heterogeneous datasets and focus on the integrative nature of genomic studies for both array-based and NGS data with clinical and phenotypic data to provide high-performance data and visualization (Thorvaldsdóttir et al. 2013).

**Cleaning of SNPs:** The cleaning/filtering of the SNP is an important step to remove the false positive SNPs, information generated by sequencing errors that do not fulfill the specific threshold for genotypic properties (Kagale et al. 2016) with minimum SNPs filtering quality of 30. There are multiple filtering/cleaning strategies used for SNPs, commonly based on reading depth, mapping quality, base quality, missing genotype calls and minor allele frequency (MAF) through currently available variant calling pipelines such as SAMtools and GATK (Kagale et al. 2016). The strategies parameters used of filtering/cleaning the SNPs are

minimum read depth of 5x for filtering the SNPs, maximum 50% missing data, MAF of 5%, and maximum heterozygosity of 40% (Malmberg et al. 2018). The missing data can be reduced through higher depth sequencing and lowering the level of the multiplex.

**SNPs imputation:** Imputation is used to increase the marker density of existing datasets toward the goal of integrating resources for downstream applications (Wang et al. 2018). Therefore, SNPs imputation is a crucial step of skim sequencing to fill the missing genotype information and can be done through open-access software like IMPUTE2 v2.3.2, MACH, and BEAGLE. Imputation depends on the haplotype segments present in the reference panel and that in turn represents the similarity with sequence reads (Wang et al. 2018). For example, it is based on the haplotype structure of the parents, which assumes that no recombination occurred (Golicz et al. 2015; Bayer et al. 2015). Imputation cannot

address too much missing information and under this situation, it is better to remove such genotypes before analysis.

### Application of skim sequencing for crop improvement

Skim sequencing has a large number of applications for crop improvement, namely molecular marker discovery, haplotype identification, recombination characterization, QTLs identification; GWAS, GS (Scheben *et al.* 2018), and organelle genome assembly studies (table 3). Reports on the application of skim sequencing in crop improvement are summarized in table 2. The use of skim sequencing has been successfully witnessed in wheat chromosomal lines for SNP discovery, diversity analysis, and marker-assisted selection (Huang *et al.* 2009). The skim sequencing will also play an important role in complementing the emerging breeding techniques such as speed breeding which significantly shortens the generations and helps to achieve six generations per year in wheat (Watson *et al.* 2018). Besides, it will also boost the QTL-seq approach that requires the WGRS data on partial mapping population (bulks) to facilitate the identification of the genomewide large number of SNPs and more specifically from the target candidate QTLs region controlling traits of interest (Pandey *et al.* 2016). It will also assist in refining the other WGS based candidate gene identification approaches such as MutMap, MutMap+, and BSR-seq. In nutshell, skim sequencing has the potential to serve as a revolutionary approach in crop improvement. Principal applications of skim sequencing are summarized below:

#### Identification of SNPs

The identification of SNP markers through skim sequencing is used for the construction of a genetic linkage map. Huang *et al.* (2009) detected a total of 123,302 SNPs with an average of one SNP for every 3.15 kb in rice crop. A part of RILs generated from *Indica* × *Japonica* crosses was randomly selected for sequencing, which resulted in 0.323x coverage of the genome. The SNP density of 1 per 50 kb in each RILs individual was reported to be sufficient for detecting recombination events in the population of 150 RILs. Similarly, in chickpea and canola, a total of 511,624 and 794,837 SNPs, respectively, were identified after filtering the overlapping segments (Bayer *et al.* 2015). In another study, a total of 109,273 SNPs were identified in 246 RILs of soybean through the WGRS approach (Xu *et al.* 2013). A total of 419,024 SNPs were discovered in four pea lines that remain 131,850 SNPs after filtering (Boutet *et al.* 2016). A total of 222 RILs chickpea population was generated from ICC 4958 × ICC 1882 and genotyped using a skim sequencing approach. The 53,169 SNPs were used for analysis after filtering from 84,963 SNPs (Kale *et al.* 2015). Li *et al.* (2019) identified a total of 41,88,6404 SNPs in a

CRISPR/cas9 edited plants of the cotton crop. A total of 100,57,282 SNPs were identified in a 271 RILs population developed from a cross between K 36 × Y 3 in *Nicotiana tabacum* (Tong *et al.* 2020).

#### Recombination and gene conversion events estimation

The recombination and gene conversion events are estimated based on the size of parental haplotype blocks. If the size of haplotype blocks ranges from 20 to 10,000 bp, then it can be defined as gene conversion and if it exceeds 10,000 bp then it is considered as recombination (Yang *et al.* 2012). In an F<sub>2</sub> population, generated from a cross of Columbia and Landsberg erecta of *A. thaliana*, 3000 gene conversion and 73 crossovers were estimated (Yang *et al.* 2012). Similarly, Bayer *et al.* (2015) used two reference sequences (A-genome from *B. rapa* and the C-genome from *B. oleracea*) of *B. napus* diploid progenitors for identification of crossovers and gene conversion (Bayer *et al.* 2015). The DH mapping population consisted of 92 individuals derived from Tapidor × Ningyou7 reported an average of 151.18 and 115.53 crossovers, as well as 697.85 and 374.85 gene conversions per individual for A-genome and C-genome, respectively. Further, the same research group has reported 219 crossovers and 256 gene conversion events in chickpea RILs (PI489777 × ICC4958) population (Bayer *et al.* 2015). Similarly, in chickpea, a total of 1610 recombination points have been identified through the skim approach in a RILs population of 232 lines for drought tolerance (Kale *et al.* 2015).

#### QTL mapping

The data generated through skim sequencing or identified SNPs are used for constructing the linkage map based on the recombination frequency between markers. The large numbers of SNPs have the potential of association mapping and gene-level resolution among the diverse set of the population as compared to the biparental mapping population (Sonah *et al.* 2015). In a recombinant inbred line population of rice (*Oryza sativa* ssp. *indica* cv. 93-11 and ssp. *japonica* cv. Nipponbare), SNPs identified a major QTL governing plant height with a phenotypic variance of 31.3%. Interestingly, this locus has been mapped in the same region that harbors the *sd1* gene (a semi-dwarfing gene in rice which is deployed in modern rice breeding) responsible for the ‘green revolution’ in rice (Huang *et al.* 2009). The WGRS using low coverage (0.06×) of the genome in combination with RFLP and SSR markers identified six QTLs for grain weight in 241 RILs population of rice (*indica* ssp.) (Yu *et al.* 2011). Similarly, three QTL for root-knot nematode resistance were identified in 246 RILs population generated through the crossing of Magellan (susceptible) and PI 438489B (resistant) variety of soybean at 0.2× depth

genome coverage to capture the recombination breakpoints (Xu *et al.* 2013). In another study, a set of 232 RILs population (ICC 4958 × ICC 1882) in chickpea was developed and genotyped through a skim sequencing approach. Subsequently, using 232 RILs sample paired read sequence data were generated with an average of 0.72x genome coverage to identify the SNPs (Kale *et al.* 2015). The skim sequencing would be helpful to boost various breeding programs by providing low coverage (Varshney *et al.* 2019) genomewide SNPs and help to understand the mechanism of various complex quantitative traits and breakage of linkage drag.

### Chloroplast and mitochondria assembly, and repeat quantification

There is continuous evolutionary development in NGS technologies and make a lead for the development of several assembly algorithms, however, only a few of them focus on the assembly of organelle genome. Further, this genome is used for phylogenetic studies and food identification and it is one of the most submitted eukaryotic genomes in GenBank. The WGS is the most authenticated and least laborious technology for producing organelle genome assembly because there is no specific tools designed for organelles genome assembly. Dierckxsens *et al.* (2017) developed a NOVOPlasty, an open-source software, which is specifically designed for organelle genomes assembling and capable of providing the whole-genome sequence. Further, the algorithm takes the benefit of more coverage of organelle genomes in available NGS data and a wide range of seed sequences without a reference genome. The above algorithm has been tested on one unknown mitochondrial genome (*Goniocotena intermedia*) and one unknown chloroplast genome (*Avicennia marina*) and two publically available data set (*Arabidopsis thaliana* and *Oryza sativa*) and found the best performance in terms of assembly accuracy and coverage. In another study, Kim *et al.* (2015) completed the sequences of chloroplast genome and 45S nuclear ribosomal DNA (45S nrDNA) of 11 *Panax ginseng* cultivars through WGS technology. They have completed the cp and 45S nrDNA sequences based on the representative barcoding target sequences for cytoplasm and nuclear genome, respectively, based on low coverage (0.1× to 0.3×) with NGS sequence of each variety. After the WGS sequencing, a total of 17 unique informative polymorphic sites were identified, further they have developed six reliable markers for analysis of ginseng diversity and cultivar purity. The first genomewide analysis using WGS low-coverage technology to explore the hidden genome components based on the characterization of major repeat families in the *B. rapa* (A genome) and *B. oleracea* (C genome) genomes was conducted by Perumal *et al.* (2017). In the present study, a total of 10 major repeats (MRs) including a new family, comprising about 18.8, 10.8, and 11.5% of the A, C and AC

genomes (*B. napus*), respectively. It may cause diversification between the A and C genomes.

### Perspectives and conclusion

The NGS techniques have proved a vital platform to overcome the challenges related to the short reads, particularly in complex plant genomes that are difficult to map. It is expected to have advanced genotyping methods capable of generating long-read sequences with higher accuracy in the future. The existing and emerging approaches of sequencing for the identification of SNP-trait associations have boosted genomics-assisted breeding (Boutet *et al.* 2016). Further, the dynamically emerging cheaper genotyping platforms will also help to transfer the complex traits into elite germplasm and hence strengthening food security in the era of climate change. Skim sequencing is a more flexible approach as it generates comparatively low volume sequence data for trait association. It also has the advantage that with the increase in sequence data volume, it facilitates the fine mapping of recombination events, gene conversion as well as structural variations (Bayer *et al.* 2015). Thus, with the dynamically evolving sequencing platforms, the sequencing costs are expected to be more cheaper in the future and hence boosting the application of skim sequencing for crop improvement.

### Acknowledgements

We thank Dr Prasad Bajaj, Dr Annapurna Chitikineni, and Dr Rajeev Varshney (International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad, India) for comments and valuable suggestions that greatly improved the manuscript. The first author is thankful to the Science and Engineering Research Board (SERB) Govt. of India for financial support through the EEQ (2018/001394) scheme.

### References

- Albers C. A., Lunter G., MacArthur D. G., McVean G., Ouwehand W. H. and Durbin R. 2011 Dindel: accurate indel calls from short-read data. *Gen. Res.* **21**, 961–973.
- Arabidopsis Genome Initiative 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796.
- Baird N. A., Etter P. D., Atwood T. S., Currey M. C., Shiver A. L., Lewis Z. A. *et al.* 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376.
- Batley J. and Edwards D. 2007 SNP applications in plants. In *Association mapping in plants* (ed. N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner and H. N. Silva) pp. 95–102. Springer, New York.
- Bayer P. E., Ruperao P., Mason A. S., Stiller J., Chan C. K. K., Hayashi S. *et al.* 2015 High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theor. Appl. Genet.* **128**, 1039–1047.



- Boutet G., Carvalho S. A., Falque M., Peterlongo P., Lhuillier E., Bouchez O. *et al.* 2016 SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RILs population. *BMC Genomics* **17**, 121.
- Browning S. R. and Browning B. L. 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097.
- Browning B. L. and Browning S. R. 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223.
- Catchen J. M., Amores A., Hohenlohe P., Cresko W. and Postlethwait J. H. 2011 Stacks: building and genotyping Loci de novo from short-read sequences. *G3* **1**, 171–182.
- Chen F., Dong W., Zhang J., Guo X., Chen J., Wang Z. *et al.* 2018 The sequenced angiosperm genomes and genome databases. *Front. Plant Sci.* **9**, 418.
- Church G. M. and Kieffer-Higgins S. 1988 Multiplex DNA sequencing. *Science* **240**, 185–188.
- Craig D. W., Pearson J. V., Szelinger S., Sekar A., Redman M., Comeveaux J. J. *et al.* 2008 Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887.
- Cronn R., Liston A., Parks M., Gernandt D. S., Shen R. and Mockler T. 2008 Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, 122.
- Davey J. W., Hohenlohe P. A., Etter P. D., Boone J. Q., Catchen J. M. and Blaxter M. L. 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510.
- De Pisto M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., Hartl C. *et al.* 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498.
- Dierckxsens N., Mardulyn P. and Smits G. 2017 NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18.
- Edwards D. and Batley J. 2010 Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* **8**, 2–9.
- Edwards D., Batley J. and Snowdon R. J. 2013 Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11.
- Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., Buckler E. S. and Mitchell S. E. 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, 19379.
- Fan X., Abbott T. E., Larson D. and Chen K. 2014 BreakDancer: Identification of genomic structural variation from paired-end read mapping. *Curr. Protoc. Bionform.* **45**, 15.6.1-151.6.11.
- Galewski P. and McGrath J. M. 2020 Genetic diversity among cultivated beets (*Beta vulgaris*) assessed via population-based whole genome sequences. *BMC Genomics* **21**, 1–14.
- Garrison E. and Marth G. 2012 Haplotype-based variant detection from short-read sequencing. arXiv preprint, [arXiv:1207.3907](https://arxiv.org/abs/1207.3907) [q-bio.GN].
- Glaubitz J. C., Casstevens T. M., Lu F., Harriman J., Elshire R. J., Sun Q. and Edward S. B. 2014 TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**, e90346.
- Goff S. A., Ricke D., Lan T. H., Presting G., Wang R., Dunn M. *et al.* 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* **296**, 92–100.
- Golicz A. A., Bayer P. E. and Edwards D. 2015 Skim-based genotyping by sequencing. In *Plant genotyping*, pp. 257–270. Humana Press, New York.
- Gordon D. 2016 Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344.
- Gregory T. R. 2005 The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* **95**, 133–146.
- Hayward A., Vighnesh G., Delay C., Samian M. R., Manoli S., Stiller J. *et al.* 2012 Second-generation sequencing for gene discovery in the Brassicaceae. *Plant Biotechnol. J.* **10**, 750–759.
- He J., Zhao X., Laroche A., Lu Z. X., Liu H. and Li Z. 2014 Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* **5**, 484.
- Howie B. N., Donnelly P. and Marchini J. 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529.
- Huang X., Feng Q., Qian Q., Zhao Q., Wang L., Wang A. *et al.* 2009 High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076.
- James T., Robinson H. T., Wendy W., Mitchell G., Eric S. L., Gad G. and Jill P. M. 2011 Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.
- Kagale S., Koh C., Clarke W. E., Bollina V., Parkin I. A. and Sharpe A. G. 2016 Analysis of genotyping-by-sequencing (GBS) data. *Methods Mol. Biol.* **1374**, 269–284.
- Kale S. M., Jaganathan D., Ruperao P., Chen C., Punna R., Kudapa H. *et al.* 2015 Prioritization of candidate genes in “QTL-hotspot” region for drought tolerance in chickpea (*Cicer arietinum* L.). *Sci. Rep.* **5**, 15296.
- Kim K., Lee S. C., Lee J., Lee H. O., Joh H. J., Kim N. H. *et al.* 2015 Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PLoS One* **10**, e0117159.
- Kirst M., Resende M., Munoz P. and Neves L. 2011 Capturing and genotyping the genome-wide genetic diversity of trees for association mapping and genomic selection. *BMC Proc.* **5**, 17.
- Koren S. 2013 Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101.
- Kumar B., Rakshit S., Singh R. D., Gadag R. N., Nath R. and Paul A. K. 2008 Genetic diversity of early maturing indian maize (*Zea mays* L.) inbred lines revealed by SSR markers. *J. Plant Biochem. Biotechnol.* **17**, 133–140.
- Kumar B., Talukdar A., Bala I., Verma K., Lal S. K., Sapra R. L. *et al.* 2014 Population structure and association mapping studies for important agronomic traits in soybean. *J. Genet.* **93**, 775–784.
- Kumar B., Talukdar A., Verma K., Bala I., Harish G. D., Gowda S. *et al.* 2015 Mapping of yellow mosaic virus (YMV) resistance in soybean (*Glycine max* L. Merr.) through association mapping approach. *Genetica* **143**, 1–10.
- Kumar P., Choudhary M., Hossain F., Singh N. K., Choudhary P., Gupta M. *et al.* 2019 Nutritional quality improvement in maize (*Zea mays*): progress and challenges. *Ind. J. Agric. Sci.* **89**, 895–911.
- Langmead B. and Salzberg S. L. 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359.
- Langmead B., Trapnell C., Pop M. and Salzberg S. L. 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.
- Li H. 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.
- Li H. and Durbin R. 2010 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

- Li Y., Cristen J., Willer J. D., Paul S. and Gonçalo R. A. 2006 Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet. S* **79**, 2290.
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N. et al. 2009a The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
- Li J., Manghwar H., Sun L., Wang P., Wang G., Sheng H. et al. 2019 Whole genome sequencing reveals rare off-target mutations and considerable inherent genetic or/and somaclonal variations in CRISPR/Cas9-edited cotton plants. *Plant Biotechnol. J.* **17**, 858–868.
- Li R., Yu C., Li Y., Lam T. W., Yiu S. M., Kristiansen K. and Wang J. 2009b SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967.
- Lorenc M. T., Hayashi S., Stiller J., Lee H., Manoli S., Ruperao P. et al. 2012 Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology* **1**, 370–382.
- Ma X., Fu Y., Zhao X., Jiang L., Zhu Z., Gu P. et al. 2016 Genomic structure analysis of a set of *Oryza nivara* introgression lines and identification of yield-associated QTLs using whole-genome resequencing. *Sci. Rep.* **6**, 27425.
- Malmberg M. M., Shi F., Spangenberg G. C., Daetwyler H. D. and Cogan N. O. 2018 Diversity and genome analysis of Australian and global oilseed *Brassica napus* L. germplasm using transcriptomics and whole genome re-sequencing. *Front. Plant Sci.* **9**, 508.
- Matsumoto T., Wu J. Z., Kanamori H., Katayose Y., Fujisawa M., Namiki N. et al. 2005 The map-based sequence of the rice genome. *Nature* **436**, 793–800.
- Metzker M. L. 2010 Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31.
- Milne I., Stephen G., Bayer M., Cock P. J. A., Pritchard L., Cardle L. et al. 2013 Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **14**, 193–202.
- Pandey M. K., Roorkiwal M., Singh V., Lingam A., Kudapa H., Thudi M., Chitkineni A. et al. 2016 Emerging genomic tools for legume breeding: current status and future perspectives. *Front. Plant Sci.* **7**, 455.
- Pendleton M. 2015 Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786.
- Perumal S., Waminal N. E., Lee J., Lee J., Choi B. S., Kim H. H. et al. 2017 Elucidating the major hidden genomic components of the A, C, and AC genomes and their influence on Brassica evolution. *Sci. Rep.* **7**, 1–12.
- Poland J., Endelman J., Dawson J., Rutkoski J., Wu S. Y., Manes Y., Dreisigacker S. J. et al. 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* **5**, 103–113.
- Rakshit A., Rakshit S., Singh J., Chopra S. K., Balyan H. S., Gupta P. K. and Bhat S. R. 2010 Association of AFLP and SSR markers with agronomic and fibre quality traits in *Gossypium hirsutum* L. *J. Genet.* **89**, 155–162.
- Rimmer A., Hang P., Iain M., Zamin I., Stephen R. F. T., Andrew O. M. W. et al. 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918.
- Scheben A., Batley J. and Edwards D. 2017 Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.* **15**, 149–161.
- Scheben A., Batley J. and Edwards D. 2018 Revolution in genotyping platforms for crop improvement. In *Plant genetics and molecular biology advances in biochemical engineering/biotechnology* (ed. R. Varshney, M. Pandey and A. Chitkineni), vol. 164. Springer, Cham.
- Scheffler B. E., Kuhn D. N., Motamayor J. C. and Schnell R. J. 2009 Efforts towards sequencing the Cacao genome (*Theobroma cacao*). Plant Anim. Genomes XVII conference 10–14 January, 2009, San Diego.
- Sonah H., Bastien M., Iquira E., Tardivel A., Legare G., Boyle B. et al. 2013 An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8**, e54603.
- Sonah H., O'Donoghue L., Cober E., Rajcan I. and Belzile F. 2015 Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* **13**, 211–221.
- Sun Z., Wang Z., Tu J., Zhang J., Yu F., McVetty P. B. and Li G. 2007 Anultradense genetic recombination map for *Brassica napus*, consisting of 13551 SRAP markers. *Theor. Appl. Genet.* **114**, 1305–1317.
- Tong Z., Zhou J., Xiu Z., Jiao F., Hu Y., Zheng F. et al. 2020 Construction of a high-density genetic map with whole genome sequencing in *Nicotiana tabacum* L. *Genomics* **112**, 2028–2033.
- Thorvaldsdóttir H., Robinson J. T. and Mesirov J. P. 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192.
- Varshney R. K., Pandey M. K., Bohra A., Singh V. K., Thudi M. and Saxena R. K. 2019 Toward the sequence-based breeding in legumes in the post-genome sequencing era. *Theor. Appl. Genet.* **132**, 797–816.
- Wang D. R., Agosto-Pérez F. J., Chebotarov D., Shi Y., Marchini J., Fitzgerald M. et al. 2018 An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9**, 3519.
- Watson A., Ghosh S., Williams M. J., Cuddy W. S., Simmonds J., Rey M. D. et al. 2018 Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plant* **4**, 23–29.
- Winzeler E. A., Richards D. R., Conway A. R., Goldstein A. L., Kalman S., McCullough M. J. et al. 1998 Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197.
- Xie W., Feng Q., Yu H., Huang X., Zhao Q., Xing Y. et al. 2010 Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 10578–10583.
- Xu X., Zeng L., Tao Y., Vuong T., Wan J., Boerma R. et al. 2013 Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. *Proc. Natl. Acad. Sci. USA* **110**, 13469–13474.
- Yang S., Yuan Y., Wang L., Li J., Wang W., Liu H. et al. 2012 Great majority of recombination events in Arabidopsis are gene conversion events. *Proc. Natl. Acad. Sci. USA* **109**, 20992–20997.
- Yu J., Hu S., Wang J., Wong G. K. S., Li S., Liu B. et al. 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296**, 79–92.
- Yu H., Yu W., Xie J., Wang Y., Xing C., Xu X. et al. 2011 Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One* **6**, e17595.