frontiers in Plant Science

REVIEW
published: 06 February 2018
doi: 10.3389/fpls.2018.00104

Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array

Qian You[1,2], *Xiping Yang*[2], *Ze Peng*[2], *Liping Xu*[1]* and *Jianping Wang*[2,3,4]*

[1] Key Laboratory of Sugarcane Biology and Genetic Breeding Ministry of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, China, [2] Agronomy Department, University of Florida, Gainesville, FL, United States, [3] Plant Molecular and Cellular Biology Program, Genetics Institute, University of Florida, Gainesville, FL, United States, [4] Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, China

OPEN ACCESS

Edited by:
Jacqueline Batley, University of Western Australia, Australia

Reviewed by:
Awais Rasheed, International Maize and Wheat Improvement Center (Mexico), Mexico
Samantha Baldwin, The New Zealand Institute for Plant & Food Research Ltd., New Zealand
Joerg Guenter Plieske, TraitGenetics GmbH, Germany

***Correspondence:**
Liping Xu
xlpmail@126.com
Jianping Wang
wangjp@ufl.edu

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *25 July 2017*
Accepted: *19 January 2018*
Published: *06 February 2018*

Citation:
You Q, Yang X, Peng Z, Xu L and Wang J (2018) Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. Front. Plant Sci. 9:104. doi: 10.3389/fpls.2018.00104

Polyploid species play significant roles in agriculture and food production. Many crop species are polyploid, such as potato, wheat, strawberry, and sugarcane. Genotyping has been a daunting task for genetic studies of polyploid crops, which lags far behind the diploid crop species. Single nucleotide polymorphism (SNP) array is considered to be one of, high-throughput, relatively cost-efficient and automated genotyping approaches. However, there are significant challenges for SNP identification in complex, polyploid genomes, which has seriously slowed SNP discovery and array development in polyploid species. Ploidy is a significant factor impacting SNP qualities and validation rates of SNP markers in SNP arrays, which has been proven to be a very important tool for genetic studies and molecular breeding. In this review, we (1) discussed the pros and cons of SNP array in general for high throughput genotyping, (2) presented the challenges of and solutions to SNP calling in polyploid species, (3) summarized the SNP selection criteria and considerations of SNP array design for polyploid species, (4) illustrated SNP array applications in several different polyploid crop species, then (5) discussed challenges, available software, and their accuracy comparisons for genotype calling based on SNP array data in polyploids, and finally (6) provided a series of SNP array design and genotype calling recommendations. This review presents a complete overview of SNP array development and applications in polypoid crops, which will benefit the research in molecular breeding and genetics of crops with complex genomes.

Keywords: polyploidy, SNP, SNP array, genotype calling, high throughput genotyping

INTRODUCTION

Polyploid species contain more than two sets of chromosomes (Comai, 2005). It is estimated that polyploid species account for up to 80% of living plants (Otto, 2007; Rieseberg and Willis, 2007). The occurrence of polyploidy can be explained by two major mechanisms: (1) the failure of meiosis or mitosis, and the fusion of unreduced gametes (Comai, 2005), which results in genome doubling in a single cell and produces autopolyploid species, such as cultivated potato (Watanabe, 2015); and (2) the combination of two or more genomes from different but related species through

hybridization and subsequent chromosome doubling, which produces allopolyploid species, such as bread wheat (Chen and Ni, 2006; Marcussen et al., 2014; Borrill et al., 2015). Polyploidization has become a common approach to overcome the sterility of interspecific hybrids during plant breeding. For example, triticale is a hybrid derived from the cross between wheat (*Triticum turgidum*) and rye (*Secale cereal*) (Gupta and Priyadarshan, 1982). After polyploidization, triticale is able to be propagated from seed, and is now widely grown as a forage in 37 countries across the world (http://www.fao.org/faostat/en/#data/QC). Based on time of origin, polyploids can be classified as either neopolyploids or paleopolyploids. Neopolyploids are any newly formed polyploid without diploidized genomes, such as potato, cotton, canola, wheat, and sugarcane. On the other hand, paleopolyploids are formed at least several million years ago with ancient and diploidized genomes, such as maize and soybean (Hilu, 1993; Tayalé and Parisod, 2013). The chromosome numbers of gametes of neopolyploids are usually multiples of the basic chromosome number in the genera (Hilu, 1993), while paleopolyploids usually have large basic chromosome numbers (mostly larger than $n = 13$) (Guerra, 2008).

Polyploidy is an important force during plant evolution. The duplication or addition of the genomes involves molecular and physiological adjustments in plants, such as changing gene expressions, and generating novel phenotypes (Adams and Wendel, 2005). Moreover, new species can be generated due to polyploidization (Soltis et al., 2015). It was estimated that 15 and 31% of speciation events in angiosperms and ferns, respectively, are involved in ploidy changes (Wood et al., 2009). Different phenotypic consequences were observed for polyploids. For example, in allopolyploid wheat, there are various phenotypic consequences of polyploidy depending on the relationship of homoeologs, such as accelerated senescence (dosage effects), changing growth habit (homoeolog dominance), and changing specific domestication traits (interaction of homoeologs) (Borrill et al., 2015). The combination of two or more sets of chromosomes from different species in allopolyploids through hybridization can lead to heterosis (Chen, 2013). This may be explained by the allelic interactions or epigenetic modifications of some key genes. By contrast, autopolyploids, in comparison with their respective diploids, experience neither a significant genome restructuring nor a strong alteration of gene expressions (Parisod et al., 2010). However, A few genes whose expressions and final phenotypes, such as cell size and organ thickness, are dramatically changed and linearly correlated with the ploidy (Stupar et al., 2007). The role of polyploidy in plant evolution has been reviewed elsewhere (Adams and Wendel, 2005; Madlung, 2013; Soltis et al., 2015), and thus not discussed in details in this review.

Despite of the popularity and importance of polyploid species in agriculture, in the genomic era, improvement and molecular breeding of polyploid crops has lagged behind many diploid crop species largely due to the complexity of the genome composition. As a result, polyploid genetic studies have also fallen behind. Over the past decade, crop genetic studies and molecular breeding particularly for diploid crops, have achieved remarkable success owing to the application of molecular marker technologies (Würschum et al., 2013; García-Pereira et al., 2014; Ergül et al., 2015). With the advantages of abundance, cost-efficiency, and high-throughput assays, single nucleotide polymorphism (SNP) has become increasingly important in crop genetic studies (Seeb et al., 2011), such as association mapping and genomic selection (Cavanagh et al., 2013; Huang and Han, 2014; Allwright and Taylor, 2016).

As large amounts of SNPs have been discovered, the demand of high-throughput SNP genotyping has increased. SNP genotyping technologies include the low-throughput gel-based approach, the cleaved amplified polymorphic sequence (CAPS) marker approach (Thiel et al., 2004), PCR-based fluorescently-labeled high-throughput methods, high-resolution melting (HRM) curve analysis, TaqMan® and KASP™ assay (Martino et al., 2010), fixed array systems such as Illumina Infinuium (Mason et al., 2017), Affymetrix Axiom (Allen et al., 2017), and next generation sequencing (NGS) enabled approaches such as restriction-enzyme-based genotyping by sequencing (GBS) (Thomson, 2014). The time, flexibility, and cost-effectiveness of the above SNP genotyping technologies has been well-discussed by Thomson (2014).

Currently, the most popular high throughput genotyping platforms are the hybridization based SNP array and various NGS enabled genotyping such as GBS, which refers to any platform that uses sequencing to obtain genotypes and a total of 13 different GBS methods have been summarized (Scheben et al., 2017). GBS has been successfully utilized in crop species mainly due to the rapid developments of sequencing technologies, increasing read length, and more available reference genomes. The utilization of GBS in plants as well as its advantages and disadvantages have been well-summarized and reviewed (Deschamps et al., 2012; He et al., 2014; Rasheed et al., 2017). SNP array is a type of DNA microarray containing designed probes harboring the SNP positions, which is hybridized with fragmented DNA to determine the specific alleles of all SNPs on the array for the hybridized DNA sample (LaFramboise, 2009). Many SNP arrays have been successfully applied in diploid species genotyping, such as the Apple 480K SNP array (Bianco et al., 2016), the Maize 600K SNP array (Unterseer et al., 2014), and the Rice 700K SNP array (McCouch et al., 2016). However, SNP identification and utilization for genotyping in polyploid species has progressed slowly due to the complex nature of genomes and inheritance of polyploids. Although SNP identification in highly polyploid species has more obstacles than in diploids as reviewed by Clevenger et al. (2015), the development of SNP arrays in polyploids has achieved noticeable progress (Bassil et al., 2015; Aitken et al., 2016; Clevenger et al., 2017).

While SNP array and GBS have their own pros and cons, they could complement each other. To date, SNP identification in polyploids (Kaur et al., 2012; Clevenger et al., 2015) and SNP array development mainly in diploid species (Rasheed et al., 2017) have been extensively reviewed. However, the summarization and discussion of SNP array in polyploid crop species is lacking. The review from Kaur et al. (2012) mainly focused on SNP identification and validation in allopolyploids. Little attention was paid to autopolyploids or genotyping using

SNP array technology. As the development of SNP array for polyploid crops is more complicated than diploids, it would be helpful to review the current progress of SNP array development and applications in polyploid crops. The current review focuses on following aspects in both autopolyploid and allopolyploid crops: (1) comparing SNP array technology with other methods, especially NGS-based technologies for genotyping polyploids; (2) explaining SNP selection criteria for array design for polyploids; (3) summarizing currently available arrays and their applications in polyploid crops; (4) discussing available SNP array genotype calling software; and (5) providing suggestions and future perspectives of SNP array application in polyploid crops.

SNP IDENTIFICATION IN POLYPLOIDS

In this section we briefly describe how the SNPs are identified in polyploid species. The first step for SNP array development is SNP identification from the DNA or dDNA sequences. Over the past decade, the cost and running time for NGS technologies have dramatically reduced, thus they have been extensively used for genome and transcriptome sequencing for a wide range of species including polyploids. The substantial sequences generated from the NGS served as a valuable reservoir for SNP identification. Several NGS enabled approaches were also developed specifically for high throughput genotyping such as GBS, exome sequencing (Exom-seq), restriction site associated DNA sequencing (RAD-seq), and the double-digest RAD-seq (ddRAD-seq) (Peterson et al., 2012; Borrill et al., 2015). Therefore, these NGS enabled genotyping methods have identified a substantial number of SNPs in many crop species including polyploid crops. For example, as the most widely used NGS-enabled genotyping method, GBS (Elshire et al., 2011) was reported in identifying and assaying SNPs in several polyploids: 20K SNPs from 164 wheat DH lines (Poland et al., 2012), 76K SNPs from 14 sugarcane accessions (Yang et al., 2017), and 84K SNPs from 151 sugarcane clones (Balsalobre et al., 2017).

SNP identification pipelines based on NGS data generally include NGS reads mapping or alignment, SNP calling, filtering, and validation, which have been reviewed for polyploids (Clevenger et al., 2015). One notable challenge of using these NGS enabled approaches for genotyping polyploid species is that the required read depth is much higher than that of diploid species. For example, diploids require 4∼6x read depth for sequencing large number samples (around 400) (Le and Durbin, 2011) or 7.7x depth for sequencing 99% of the alleles (Clevenger et al., 2015); while the suggested depth for polyploids was up to 48.7x in octoploid strawberry (Bassil et al., 2015), 60x in tetraploid potato (Hamilton et al., 2011), and 100x in sugarcane (Song et al., 2016), which would require a high depth of sequencing (Hamilton et al., 2011; Wang et al., 2014; Bassil et al., 2015; Aitken et al., 2016; Clevenger et al., 2017). The depth can be even higher if different SNP dosages need to be called. For allopolyploids, distinguishing homologous SNPs between genotypes from homoeologous SNPs between subgenomes or paralogous SNPs between duplicated gene copies is a daunting task due to the high sequence similarity between the subgenomes (Kaur et al., 2012; Dufresne et al., 2014; Clevenger et al., 2015). Take peanut (AABB genome) as an example, the two sub-genomes of peanut are highly similar (96% median identity; Bertioli et al., 2016), thus the homoeologous SNPs account for a large proportion of identified SNPs (Clevenger et al., 2015; Peng et al., 2016). While homoeologous variations may hinder genomic analyses, the understanding of homoeologs can be helpful for adjusting the response of quantitative traits in polyploid crops. Since the functional redundancy of homoeologs can "lock up" some phenotypic variation, as was reported in wheat, the manipulation of homoeologs will likely unlock the full polyploidy potential of the crop (Borrill et al., 2015). The factors influencing the discrimination of these different types of SNP classes and the status of SNP discovery in allopolyploid crops have been reviewed elsewhere (Kaur et al., 2012; Clevenger et al., 2015).

Beside the NGS enabled genotyping approaches, datasets of published DNA sequences or expressed sequence tags (ESTs) also are valuable sources for SNP discovery in polyploids. For example, 8,327 SNP were identified from the Sanger EST datasets of three potato cultivars and finally filtered to 2,358 high quality SNPs by Hamilton et al. (2011). Similarly, Tinker et al. discovered 7,680 SNPs from four genomic DNA sequence sources in NCBI, which were derived from more than 20 diverse oat cultivars (Tinker et al., 2014). In practice, selection of SNP discovery technology may largely rely on the availability of SNP resources, and the specific biological question of interest.

FEATURES OF SNP ARRAY TECHNOLOGY

SNP array is a high-throughput, relatively cost-efficient, and automatically genotyping assay. It has been widely used in genetic studies of crops, including genome-wide association studies (GWAS) (Chen et al., 2014; McCouch et al., 2016), linkage map construction (Ganal et al., 2011; Felcher et al., 2012), genomic selection (Yu et al., 2014; Clarke et al., 2016), population structure analysis (Unterseer et al., 2014; Hulse-Kemp et al., 2015; Wang et al., 2016), and gene mapping (Dalton-Morgan et al., 2014). The capacity of currently available SNP arrays is up to 700K in diploids (rice) (McCouch et al., 2016), 487K in triploids (apple) (Bianco et al., 2016), 58K in tetraploids (peanut) (Clevenger et al., 2017; Pandey et al., 2017), 820K in hexaploids (wheat) (Winfield et al., 2016), 90K in octoploids (strawberry) (Bassil et al., 2015), and 345K in dodecaploids (sugarcane) (Aitken et al., 2016).

SNP array technology, similar to many other biotechnologies, has its pros and cons. For high-throughput genotyping, SNP array has several advantages over other genotyping approaches. First, SNP array data is relatively easy to analyze compared to data generated using NGS-based methods. Particularly when considering labor-intensive NGS library preparation (Garvin et al., 2010) and downstream bioinformatics data analysis investment for accurate SNP calling (Liu et al., 2012; Torkamaneh et al., 2017). To be more specific, the genotypes of SNP markers can be called and provided by Affymetrix or Illumina, or researchers also can call genotypes following the Affymetrix or Illumina genotype calling pipeline according to the user

guide. However, it is more difficult and time-consuming to call genotypes using NGS-based data in polyploid species. As mentioned above, SNP genotype calling includes reads trimming, reads alignment, SNP calling, SNP filtering, etc. (Clevenger et al., 2015), which requires the background of bioinformatics. Second, SNPs from genomic regions of interest can be specifically included on the array. In addition, the number of interested SNPs to be placed on the array is flexible for Illumina and Affymetrix platforms. Third, SNP array is considered as low to moderate costs of per sample, despite the significant decrease in costs associated with NGS. It costs \$28∼\$90 (USD) per sample for Affymetrix Axiom SNP array (personal communication), while NGS approaches have the lowest price for GBS at \$35 per sample, followed by RNA-seq at \$260 per sample, and whole genome sequencing (WGS) at >\$500 per sample (Peng et al., 2017). As polyploid crops usually have a large genome size, the NGS-based methods would require a huge amount of sequencing data for adequate coverage.

SNP array and NGS can complement each other. Though NGS has absolute advantages in generating sequence and identification of variants, its genotyping in polyploids is still hindered. SNP array is still a good option as a genotyping platform based on the increasing variants discovered by NGS methods. However, SNP array has its own shortcomings, such as required prior genomic information, only genotyping known SNP locations, and manual dosage scoring (in some case) (Wang et al., 2014; Vos et al., 2015). In addition, its design and further optimization can be time consuming. Ascertainment bias is a common issue for genotyping arrays, which is due to non-random sampling of polymorphisms in the population of interest (Heslot et al., 2013) or due to a small number of samples used as SNP discovery panels (Albrechtsen et al., 2010). For example, a small sample size may mainly capture common alleles, excluding rare alleles (Gravel et al., 2011). This would likely distort subsequent genetic inferences. Efforts have been taken to reduce ascertainment bias such as adopting whole genome sequencing with high coverage, updating the markers on the SNP array, and combining markers from multiple arrays.

The comparison of SNP validation rate between SNP array and NGS-based methods would be difficult, since factors like population type (bi-parental or natural population) and population size used for validation will have a great impact on the validation rate. For example, when genotyping a total of 108 hexaploid wheat varieties, a 900K SNP array in wheat showed 99.8K polymorphic SNP probes. However, when genotyping a total of 475 accessions including the relatives of those varieties and progenitors, the number of polymorphic probes increased to 453.1K (Winfield et al., 2016). Thus, only the validation rate of SNP array has been summarized and discussed as follows. Currently, SNP array has been applied in many polyploid crops and most of the arrays achieved decent polymorphic rates, as summarized in **Table 1**. Among the 16 SNP arrays in nine polyploid crop species, eight (50%) SNP arrays had a polymorphic rate ≥80%, and four (25%) had a polymorphic rate between 60 and 70%. Only two SNP arrays had a polymorphic rate of less than 60%, to be specific, 42.7% in oat (Oliver et al., 2013), and 56.8% in sugarcane (Aitken et al., 2016).

SNP SELECTION FOR SNP ARRAY DESIGN AND GENOTYPE CALLING IN POLYPLOID SPECIES

Technically, two platforms have been used for SNP array in polyploid species, Illumina (Dalton-Morgan et al., 2014; Vos et al., 2015; Clarke et al., 2016) and Affymetrix (Bassil et al., 2015; Clevenger et al., 2017; Pandey et al., 2017). Different chemistries and computational algorithms are used for assay and genotype calling between Affymetrix and Illumina SNP arrays, the hybridization principle (complementary base pairing) and captured signal intensity principle (calculation of amount of target DNA and the affinity between target DNA and probes) are alike (LaFramboise, 2009).

For SNP marker selection in development of the array, most of the criteria are the same between Illumina and Affymetrix, except for two parameters: (1) sequence length on either side of the SNP (50 bp for Illumina vs. 20 bp for Affymetrix), and (2) evaluation parameters (Illumina Assay Design Tool (ADT) value vs. Affymetrix P-convert value). The general considerations for array SNP selection include SNP depth, SNP types, SNP frequency, additional variations within probe sequence of target SNPs, and probe sequence parameters. Specifically, (1) The average SNP read depth, or single genotype SNP depth is considered as it is related to the accuracy of SNPs called. If the depth is too low, the SNPs could be called due to sequence errors. If the depth is too high, the SNPs may be called from repetitive sequences. Thus, a range of depths for different crop species is recommended (Deulvot et al., 2010; Chagné et al., 2012; Bianco et al., 2014). (2) There are two types of SNPs: transition SNPs such as A/G, T/C, and transversion SNPs such as A/T, C/G, A/C, and T/G. For SNP array development, the transition SNP type is preferred and transversion SNPs, INDELs, or multiple allelic SNPs are typically excluded (Bianco et al., 2016; Clarke et al., 2016). Particularly, A/T or C/G SNPs are eliminated, as these types require two probes, while other SNP types require just one probe for genotyping (Dalton-Morgan et al., 2014; Clevenger et al., 2017). (3) SNPs with very rare alleles (present in less than two genotypes), which most likely could be due to sequence or alignment errors, are eliminated from the array (Clarke et al., 2016). (4) Additional variants within the probe sequence is a definite consideration for exclusion. The Affymetrix Axiom technology is more tolerant to any additional variants present in the probe sequence than Illumina Infinium, since Affymetrix removes SNPs (labeled as not recommended) with additional variants within 20 bp up or downstream of target SNPs (Bassil et al., 2015), while it's 50 bp for Illumina (Dalton-Morgan et al., 2014; Clarke et al., 2016). (5) Before probe design for the SNP array, the flanking sequences of the target SNPs are thoroughly evaluated by Affymetrix P-convert value or Illumina ADT value. The P-convert value predicts the probability of SNP converting on the array, which is used to assign forward or reverse probes for each SNP by considering probe sequence, binding energies, impacts from adjacent SNPs, etc. (Liu et al., 2014; Qi et al., 2017). Similarly, the ADT value predicts a likelihood of success for requested SNP loci (https://www.illumina.com/literature.html). SNPs are kept if the

TABLE 1 | Summary of SNP arrays developed in polyploid crops.

Species	SNP source	Array size	Genotyping sample size	SNP efficiency	Application	Reference
Peanut (allo-tetraploid, $2n = 4x = 40$)	163K SNPs from DNA re-sequencing of 38 accessions and RNA-sequencing of 3 accessions	58K (58,233) Axiom (Affymetrix)	297 accessions from 48 countries, including 36 wide species	44,424 (73.3%) polymorphism	Genetic diversity and genetic architecture	Pandey et al., 2017
Potato tetraploid (auto-tetraploid, $2n = 4x = 48$)	Two million SNPs from RNA-seq of 3 commercial cultivars and 8K SNPs from Sanger EST of 3 additional cultivars	96 Illumina BeadXpress (Illumina)	248 lines	82 (85.4%) reliably scored	Genetic diversity analysis (population structure)	Hamilton et al., 2011
Potato tetraploid (auto-tetraploid, $2n = 4x = 48$)	69K high confidence SNPs from previous identification (Hamilton et al., 2011)	8K (8,303) Infinium (Illumina)	184 progeny, 92 from population D84 and 92 from population DRH	Over 4,400 (53.0%) markers were mapped	Development of linkage maps	Felcher et al., 2012
Potato tetraploid (auto-tetraploid, $2n = 4x = 48$)	20K SNPs from previous identification (Hamilton et al., 2011; Uitdewilligen et al., 2013)	18K (17,987) Infinium (Illumina)	569 accessions, including 537 tetraploids and 32 diploids	14,530 (80.8%) successfully scored with fitTetra	Reconstruction of the breeding history, shaping the genetic composition	Vos et al., 2015
Cotton (allo-tetraploid, $2n = 4x = 52$)	50K SNPs from 9 intra-specific data sets, and 20K SNPs from 4 inter-specific data sets (11 previous studies and 2 unpublished studies)	63K (63,058) Infinium (Illumina)	1,156 individual samples	38,822 (61.6%) polymorphic markers	Development of high density genetic map	Hulse-Kemp et al., 2015
Alfalfa (auto-tetraploid, $2n = 4x = 32$)	900K SNPs from RNA-sequence of 27 alfalfa genotypes (including 23 tetraploid and 4 diploid) by previous reported (Li et al., 2012)	9K (9,277) Infinium (Illumina)	280 diverse genotypes including related species	7,476 (81%) polymorphic markers	Evaluation population structure and linkage disequilibrium	Li et al., 2014b
Brassica (allo-tetraploid, $2n = 4x = 38$)	54K SNPs identified and evaluated by previous reports (Bus et al., 2012 and Dalton-Morgan et al., 2014) and unpublished data; 24M SNPs discovered from published sequence data sets (Harper et al., 2012 and Clarke et al., 2013)	52K (52,157) Infinium (Illumina)	437 diverse genotypes (432 diverse genotypes were generated independently in two laboratories)	About 60% genome-specific markers	Genetic map generation	Clarke et al., 2016
Wheat (allo-hexaploid, $2n = 6x = 42$)	25K SNPs from RNA-sequence of 26 hexaploid accessions	9K (9,000) Infinium (Illumina)	2,994 hexaploid accessions, including landraces and modern cultivars	7,733 (85.9%) successfully genotyped	Genetic diversity and population structure, selection scans	Cavanagh et al., 2013
Wheat (allo-hexaploid, $2n = 6x = 42$)	128K SNPs from RNA-sequencing of 19 hexaploid and 18 tetraploid accessions	90K (91,829) Infinium (Illumina)	646 accesions, including 55 tetraploid cultivars, 447 hexaploid cultivars, and 144 hexaploid landraces	81,587 (89%) produced functional assays	Characterization of genomic diversity	Wang et al., 2014
Wheat (allo-hexaploid, $2n = 6x = 42$)	921K SNPs from Exom sequencing of 14 diploid, 5 tetraploid, 23 hexaploid, and 1 decaploid accessions	820K (819,571) Axiom (Affymetrix)	475 accessions, including diploid, tetraploid, and hexaploid wheat accessions and wheat relatives	546,299 (66.7%) polymorphic SNPs	Physical and genetic mapping, genetic characterization	Winfield et al., 2016
Wheat (allo-hexaploid, $2n = 6x = 42$)	35K SNPs from previous study (Winfield et al., 2016)	35K (35,143) Axiom (Affymetrix)	1,843 DNA samples, including 1,779 unique hexaploid wheat accessions and 64 replicates	33,326 (94.8%) polymorphic SNPs	Genetic mapping, and characterization of genetic diversity	Allen et al., 2017

(Continued)

TABLE 1 | Continued

Species	SNP source	Array size	Genotyping sample size	SNP efficiency	Application	Reference
Oat (allo-hexaploid, $2n = 6x = 42$)	11K high-confidence SNPs from RNA-sequencing of 20 genotypes (Oliver et al., 2011)	3K (3,072) GoldenGate (Illumina)	390 recombinant inbred lines	1,311 (42.7%, success rate) robust markers	Development of physically anchored consensus map	Oliver et al., 2013
Oat (allo-hexaploid, $2n = 6x = 42$)	8K SNPs from 4 DNA sequence data sets. (Tinker et al., 2009; Poland et al., 2012; Oliver et al., 2013)	6K (5,743) Infinium (Illumina)	1,110 hexaploid samples, including 109 diverse cultivars, 390 progeny, and 595 breeding lines	4,975 (86.6%) SNPs produced successfully assays	SNP discovery and annotation, population genetic characteristics	Tinker et al., 2014
Strawberry (allo-octploid, $2n = 8x = 56$)	160K di-allelic SNPs from DNA-sequencing of 19 octoploid and 6 diploid strawberry accessions	90K (95,062) Axiom (Affymetrix)	384 samples, including 357 octoploid accessions and cultivars, 4 diploid accessions, and 23 progeny.	60,473 (64%) polymorphic SNPs, including 23,355 (24.6%, success rate) markers in *PHR*	High density linkage maps, QTL identification	Bassil et al., 2015
Sugarcane (auto-dodecaploid, $2n = 12x = 100\sim120$)	2.6M SNPs from target gene-rich regions sequencing of 16 lines	345K (345,704) Axiom (Affymetrix)	367 clones, including parental clones, cultivars, and unselected families	48,802 (14.1%) validated polymorphic markers and 11,443 (3.3%, success rate) markers in *PHR*	Association analysis of cane yield and sugar content, genetic mapping	Aitken et al., 2016

PHR, Poly High Resolution, which were polymorphic and passed all quality control (Bassil et al., 2015).

Affymetrix P-convert value or Illumina ADT value is higher than 0.6 for both forward and reverse probes. Sometimes, SNPs distribution, such as spreading over all chromosomes evenly or in exonic region (Deulvot et al., 2010; Chagné et al., 2012) are taken into consideration. For allopolyploids, the chosen SNPs should be evenly distributed across the sub-genomes. In addition, genome-specific SNPs should be included, since their segregation patterns follow that of a diploid, as in peanut (Clevenger et al., 2017) and wheat (Cavanagh et al., 2013). Moreover, inter-specific SNPs can be included to reduce ascertainment bias for array design, which has been done in cotton (Hulse-Kemp et al., 2015).

Several different possible dosages are available for a specific SNP in polyploids, which should also be taken into account when selecting SNPs. This is specifically an issue for autopolyploids with a high ploidy, such as sugarcane. The estimation of SNP dosages during SNP mining stage has been reported in sugarcane. In a study aiming at SNP discovery in sugarcane, the different dosage levels of SNPs were called by using UnifiedGenotyper in the Genome Analysis ToolKit (GATK) (Song et al., 2016). Since gene dosage and allelic configuration are unknown, there is a paucity of statistical approaches (Luo et al., 2001; Baker et al., 2010) for linkage analysis using all the markers with different dosages. Therefore, single dose markers (SDMs) (Wu et al., 1992) have become the primary marker choice for linkage analysis in polyploids like sugarcane, whose segregation pattern follows that of a diploid (1:1 and 3:1 in F1 populations or 3:1 in selfing populations) (Pastina et al., 2012; Vukosavljev et al., 2016; Balsalobre et al., 2017). This makes many available statistical methods for diploids applicable for polyploids (Baker et al., 2010). The proportion of SDMs in the markers generated from high throughput NGS data can be high. In a study mining sequence variations among sugarcane accessions, the percentage of single dose SNPs ranged from 38.3 to 62.3% with an averaging of 49.6% (Yang et al., 2017). Another research group developing a 345K sugarcane SNP array specifically included single dose SNPs as much as possible (Aitken et al., 2016).

Above are the general and technical considerations in SNP selection process for Illumina and Affymetrix platforms. In certain cases, if there are multiple sources of SNPs called from different sequence approaches, then SNP source should be taken into account. It was reported in polyploid species that SNP validation rates of SNP array were associated with the SNP identification approaches (Hulse-Kemp et al., 2015). SNPs derived from gene-enriched sequencing had a higher marker validation rate (88%), such as RNA-seq and gene-enrichment restriction libraries from five *G. hirsutum* lines, than SNPs from genomic re-sequencing (50%) from 12 *G. hirsutum* lines (including the five lines from RNA-seq). Similar validation rates were obtained using RNA-seq in *Eucalyptus grandis* (83%) (Novaes et al., 2008) and *Brassica napus* (87%) (Barbazuk et al., 2007). Therefore, RNA-seq and gene-enrichment are common strategies to identify SNPs for SNP array development in polyploids including peanut (Hulse-Kemp et al., 2015), potato (Hamilton et al., 2011), wheat (Cavanagh et al., 2013; Wang et al., 2014), and sugarcane (Aitken et al., 2016).

SNP ARRAY DEVELOPMENT AND APPLICATIONS IN POLYPLOID SPECIES

Although complicated, several recent studies have reported the development and application of SNP array in polyploid crops (**Table 1**) including tetraploid, hexaploid, octoploid, and even dodecaploid species.

Tetraploid

The recent progress in SNP array development in tetraploid species was made in the allotetraploid peanut (*Arachis hypogaea*; AABB-type genome; $2n = 4x = 40$; ∼2.7 Gb genome size) (Bertioli et al., 2016). A total of 163.8K SNPs were identified from 30 represented tetraploid cultivars and 11 wild diploid accessions by DNA resequencing and RNA-seq methods, of which 58K SNPs were selected for developing a peanut SNP array (Pandey et al., 2017). The array was further utilized to investigate the genetic architecture of 300 diverse accessions, which showed 44,424 (73.3%) polymorphic SNPs on this array (Pandey et al., 2017). Meanwhile, the validation of the 58K SNP array among 384 diverse cultivars revealed 54,564 (93.7%) polymorphic SNPs between diploid species, 47,116 (81.0%) polymorphic SNPs between cultivars and interspecific hybrids, and 15,897 (27.3%) polymorphic SNPs within *A. hypogaea* accessions (Clevenger et al., 2017).

Potato (*Solanum tuberosum* L.; $2n = 4x = 48$; ∼844 Mb genome size) is a highly heterozygous autotetraploid, and is the most important non-grain food crop in the world (Consortium, 2011). Three SNP arrays were developed for this crop, 96 BeadXpress SNP Array (Hamilton et al., 2011), 8K Infinium Potato Array (Felcher et al., 2012), and 20K Infinium SNP Array (Vos et al., 2015). Through transcriptome sequencing and Sanger EST sequencing, a total of 69,011 high confidence SNPs were identified by Hamilton et al. (2011), of which 96 SNPs were selected for the 96 BeadXpress SNP array development. This study revealed distinct relationships among different potato market classes. In addition, another 8,303 SNPs were selected from the same SNP set for 8K Infinium Potato Array development for linkage mapping (Felcher et al., 2012). By using previously identified SNPs from two studies (Hamilton et al., 2011; Uitdewilligen et al., 2013), the 20K SNP array was designed, which was performed on 569 potato accessions to study drift and selection effects that influenced the genetic components of European potato. The array was also deployed to evaluate and estimate linkage disequilibrium (LD) decay in another study (Vos et al., 2017). In addition, this 20K SNP array was used to evaluate the double-reduction (DR) landscape in 237 individuals, which showed the phenomenon that the rate of DR increased with the distance from the centromeres (Bourke et al., 2015). Cotton (*Gossypium hirsutum* L.; $2n = 4x = 52$; ∼2.5 Gb genome size) is widely cultivated, providing over 95% of cotton production in the world (Zhang et al., 2008). A cotton 63K SNP array was developed based on identified SNPs from 13 different data sets, thus covering a diversity range of SNP sources, containing 45K intra-specific SNPs and 18K inter-specific SNPs. This array was applied successfully to distinguish differences between *G.hirustum* and other Gossypium species, between wild and cultivated genotypes, and among cultivars (Hinze et al., 2017). It was also used to produce two high-density genetic maps (Hulse-Kemp et al., 2015), and to identify 160 quantitative trait loci (QTLs) related to 16 agronomic traits by GWAS among 503 *G.hirustum* accessions (Huang et al., 2017).

Alfalfa (*Medicago sativa* L.; $2n = 4x = 32$) is the main forage legume crop in the world (Li et al., 2014b). It plays important ecological roles in livestock farming systems by stabilizing soil and increasing soil fertility through symbiotic biological nitrogen fixation (Li et al., 2014b; De Vega et al., 2015; O'Rourke et al., 2015). A total of 900K SNPs were identified between 27 diverse alfalfa genotypes by RNA-seq (Li et al., 2012), of which 9K SNPs were used to develop an Illumina alfalfa SNP array (Li et al., 2014a). To validate this SNP array, 280 alfalfa genotypes were assayed with an 81% SNP marker polymorphic rate, and results showed clear population structure, analyzed genetic diversity of sub-populations, and evaluated the LD across all genotypes (Li et al., 2014a).

Oilseed rape (*Brassica napus* L., AACC type-genome; $2n = 4x = 38$; ∼845 Mb genome size) is an essential economical oilseed crop, which can be used for extracting oil from seed and edible vegetable (Chalhoub et al., 2014; Clarke et al., 2016). A 60K Brassica Infinium array was designed by Clarke et al. (2016). The sources of SNPs used on this array mainly included previously identified SNPs (Bus et al., 2012; Dalton-Morgan et al., 2014; Cheung et al. unpublished) and newly called SNPs in this study originating from published sequence data in Harper et al. (2012), Clarke et al. (2013) and unpublished genomic and transcriptome sequence data in Clarke et al. To validate this 60K array, 327 and 432 diverse genotypes were independently genotyped at two laboratories, which obtained ∼60% of genome specific markers in diverse *B. napus* genotypes, 26.5 and 29.7K scorable SNP markers in *B. oleracea* and *B. rapa* respectively, and a map of *B. napus* with 46% (21,766 SNPs) mapped markers in one of DH population was constructed (Clarke et al., 2016). This 60K *Brassica* SNP array has been widely used for oilseed rape, such as identification of three QTLs associated with Sclerotinia stem rot resistance (Wu et al., 2016), 79 QTLs related with seed quality, flowering time, and root morphology traits (Wang et al., 2017), 117 genomic regions involved in selective sweeps (Zou et al., 2018), etc. In addition, a user guide was reported for this Brassica 60K SNP array (Mason et al., 2017).

Hexaploid

Wheat (*Triticum aestivnm* L.; AABBDD type-genome; $2n = 6x = 42$) is one of the world's most important cereal crops and has the most progresses in SNP array development among hexaploid crop species (Consortium, 2014). A 9K iSelect SNP array was used to assess genetic variations in coding regions of 2,994 hexaploid wheat accessions (Cavanagh et al., 2013), and a high density SNP map was constructed. Subsequently, a 90K SNP iSelect array was developed and used to assess genetic variations in allohexaploid and allotetraploid wheat populations (Wang et al., 2014). With the 90K SNP iSelect array, researchers working on wheat were able to detect QTLs conveying leaf rust resistance (Gao L. et al., 2016), identify QTLs associated with physiological and agronomic traits (Gao F. et al., 2016; Zou

et al., 2016), perform phylogenetic analysis (Turuspekov et al., 2015), and detect candidate loci involved in domestication and improvement (Gao et al., 2017). Recently, Winfield et al. (2016) used Exom-seq to identify 921K putative SNPs from 43 bread wheat accessions. A total of 820K SNPs were included for SNP array, which were then validated on 475 accessions with an average call rate of 98.4%. This array was also used to map three populations and characterize wheat accessions and relatives. Furthermore, to update this 820K SNP array, with consideration of higher level polymorphic and more evenly distributed SNPs, 35K SNPs were selected to develop a commercial high-density Axiom array. This 35K SNP array was used to assess a diverse panel of 1,843 samples, which constructed genetic maps and characterized novel genetic diversity among those samples (Allen et al., 2017).

Cultivated hexaploid oat (*Avena sativa* L.; AACCDD type-genome; $2n = 6x = 42$) is one of the important cereal crops in the world (Andon and Anderson, 2008). The first oat SNP array contained 3,072 SNPs, which was applied to build a physically anchored consensus map of oat with 985 mapped SNPs (Oliver et al., 2013). To expand this SNP array, various bioinformatic pipelines were applied to discover SNPs from multiple available DNA sequence sources, including eight alternate methods, and a new SNP calling method was also assessed by Tinker et al. (2014). Finally, a 6K oat BeadChip was designed, which produced 86.6% success rate by validating in 1,100 samples (Tinker et al., 2014). Recently, this 6K SNP array was applied to genotype 138 oat accessions for mapping QTL associated with frost tolerance using GWAS (Tumino et al., 2016).

Octoploid and Dodecaploid

Cultivated strawberry [*Fragaria* × *ananassa* (Duch.); AABBCCDD genome; $2n = 8x = 56$, ∼698 Mb genome size] is an allo-octoploid crop species (Hirakawa et al., 2014), while *Fragaria* ($2n = 2x = 14$, genome size of 240 Mb) is a diploid, called woodland strawberry (Shulaev et al., 2011). A total of 36 million unique variants were identified as SNP resources among 19 octoploid and six diploid strawberry accessions, from which a 90K SNP array (ISTRAW90) was designed (Bassil et al., 2015). The ISTRAW90 array was validated by genotyping 384 octoploid strawberry samples, and 12,609 SNPs had the highest quality, which exhibited relatively low success rate (13.3%). However, it is still a useful high-throughput tool to construct high density linkage maps (Mahoney et al., 2016), to distinguish cultivars [application of Poly High Resolution *(PHR)* SNPs] (Jung et al., 2017), and to perform genomic selection (Gezan et al., 2017) in octoploid strawberry.

As a highly heterozygous species, commercial sugarcane (*Saccharum* complex, $2n = 12x = 100\sim120$, ∼10 Gb genome size), is autopolyploid mostly derived from the interspecific cross between auto-octoploid *S. officinarum* ($2n = 8x = 80$) and autoploid *S. spontaneum* ($2n = 4x–16x = 32\sim128$) (Roach, 1989; D'hont et al., 1998) followed by backcrosses with *S. officinarum*. Sugarcane is not only highly heterozygous, but also exhibits chromosome number variation, mixed ploidy, as well as aneuploidy. To identify large numbers of SNPs for development of sugarcane SNP array, 16 sugarcane clones were used for deep sequencing with target on gene-rich regions, which identified 4.5M SNPs (Aitken et al., 2016). Based on the combination of different allele dosage levels, the 4.5M SNPs were clustered into three classes and evaluated by Affymetrix. Finally, a total of 345K SNPs were selected with high quality score from all classes and with wide distribution across sugarcane contigs. Subsequently, a 345K sugarcane SNP array was developed and used for genotyping 367 sugarcane clones, resulting in 48,802 (14.1%) validated polymorphic markers for further analysis and 11,443 (3.3%) highest quality markers. These 48,802 polymorphic markers have been included in a 50K cost-effective SNP array to genotype over 2,000 sugarcane clones in Australia (Aitken et al., 2016).

SNP ARRAY GENOTYPE CALLING AND DATA ANALYSIS PIPELINES

After the SNP array assay, calling SNP genotypes from the assay is a critical next step. For polyploid species, the genotyping calling is complicated. Rather than having three possible genotypes (homozygote with reference allele, heterozygote, and homozygote with alternative allele) at a SNP locus in diploid species, polyploid species usually have more than three possible genotypes. Theoretically, the number of genotypes can be up to five in tetraploids, seven in hexaploids, nine in octoploids, and 13 in dodecaploid. So far, there are two software, fitTetra and ClusterCall, written for tetraploids, which can call five genotypes. Another software, SuperMASSA, was written for all ploidies (so far only successfully reported in sugarcane). Most other polyploid crops were using genotype calling software accompanying Affymetrix or Illumina platform. However, the disadvantage of the software is their inability to identify >3 clusters for Affymetrix or >5 clusters for Illumina platform. The GenomeStudio software from Illumina is able to provide five clusters. As an example, the markers from the 20K Infinium SNP Array (Vos et al., 2015) were first automatically scored using fitTetra, after which the rejected markers were further manually scored using GenomeStudio. A total of 843 markers were recovered from the 1,832 rejected markers with a 46% recovery rate (Vos et al., 2015). Consequently, the GenomeStudio may be impractical to use for large amounts of markers, as it requires manual adjustment of the cluster boundaries for each marker.

There are different scenarios for genotyping calling for autopolyploids and allopolyploids. The difficulty of SNP calling in allopolyploids mainly comes from the complexity of genome architecture. Ideally, the allelic SNP variation should be derived for each sub-genome, which is why sub-genome specific SNPs are desired (Kaur et al., 2012). With sub-genome specific SNPs, the genotype calling in allopolyploids becomes no different with diploids. For example, in an Affymetrix Axiom SNP array of strawberry and an Illumina Infinium SNP array of cotton, the sub-genome specific SNPs would produce three distinct clusters, behaving like a co-dominant marker in diploids with two homozygous genotypes and one heterozygous genotype (Bassil et al., 2015; Hulse-Kemp et al., 2015). These types of SNPs are distinct from the homoeologous SNPs that are polymorphic

between sub-genomes within a sample. In practice, the sub-genome specific SNPs can be difficult to obtain for allopolyploid species with highly identical or closely related sub-genomes. For allopolyploids, due to the presence of homoeologous sequences on different sub-genomes, the SNP probes could hybridize to sequences not only on the target sub-genome, but also to the other sub-genomes or paralogs. With increasing ploidies, the fluorescent signal from a specific allele would be hard to separate from the signals of remaining alleles. Moreover, a mutation in any of the homoeologous copies may lead to the failure of probe hybridization, resulting in complex cluster types. Consequently, some attrition can be generated other than the genome-specific SNPs. Take a study applying the 90K SNP array in wheat as an example, the genotype calling revealed 35,684 (44%) assays showing three clusters, corresponding to genome-specific SNPs, 25,199 (31%) showing monomorphic clusters, and 20,704 (25%) showing complex clustering patterns due to this attrition (Wang et al., 2014).

For autopolyploids, the major complication is distinguishing between different allele dosages. However, this becomes more difficult as the ploidy increases. As SDMs only have two or three clusters, they can be easily and clearly distinguished in genotype calling for both Affymetrix and Illumina platforms. More details on SDMs are discussed in section SNP Selection for SNP Array Design and Genotype Calling in Polyploid Species above.

General Genotype Calling Based on Affymetrix Axiom and Illumina Infinium Platforms

After obtaining raw probe intensities of all the SNPs for each sample in cell intensity files (CEL) from Affymetrix Axiom SNP array, there are several main steps for genotype calling (**Figure 1A**) by Best Practices Genotyping Workflow on this platform, starting from the normalization of raw intensities and ending at SNP clustering and filtering (Liu et al., 2017). More technical and computational details are available from Affymetrix: (http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf). From the report of Affymetrix, SNPs are sorted into six quality classes based on their performance of clustering, three of which report accurate genotypes and are recommended for further validation (PHR, NMH-no minor homozygote, and OTV-off target variant). The remaining three classes are not considered for further processing, because MHR (mono high resolution) cluster is monomorphic, and a simple three-cluster genotype model are not able to generate complex intensity over three clusters correctly (CRBT-call rate below threshold, and Other) (Bassil et al., 2015). Additionally, it is easy to miscall the heterozygous as homozygous especially for single-dose markers in high polyploid species when the data quality is low (Lu et al., 2013; Li et al., 2014a). This SNP calling pipeline can limit the efficiency of genotype calls for polyploid species because it is not able to call multiple allele dosages.

For Illumina Infinium genotyping data analysis, a flowchart (**Figure 1B**) was summarized according to GenomeStudio® Polyploid Genotyping Module v2.0 Software Guide (https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2-0/genomestudio-polyploid-genotyping-module-user-guide-1000000012407-00.pdf) and Infinium Genotyping Data Analysis (https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf). The steps are generally similar as that of Affymetrix except for some platform-specific parameters. The SNP genotypes are called by using GenomeStudio software based on intensity data file and SNP information file. Take allotetraploid cotton as an example (Hulse-Kemp et al., 2015), six distinct SNP classes are generated after SNP genotype calling, of which two classes of SNPs show only one genotype and thus are monomorphic SNPs. The remaining four classes of SNPs show three genotypes or clusters, which are clustered based on their GenTrain score (the SNP cluster quality score), thus are polymorphic. The SNPs from the third class, representing three genotypes (AA, AB, BB), are genome-specific markers and are recommended SNPs with reliable genotypes. The fourth class SNPs are also genome-specific markers, but assaying two homoeologous loci, one monomorphic and the other polymorphic (AAAA, AAAB, AABB). Both the fourth and fifth class SNPs require manual adjustment, which could be a challenge. The sixth class SNPs are usually thought to be failed.

Available Tools for Genotype Calling Based on Array Data in Polyploid Species

Fortunately, three open access tools have been developed and used for genotype calling in polyploid crop species based on array data, including the R package fitTetra (Voorrips et al., 2011), superMASSA (Serang et al., 2012) (http://statgen.esalq.usp.br/SuperMASSA/), and ClusterCall (Carley et al., 2017). The R package fitTetra uses an automated method based on fitting a mixture of normal distributions. It can handle mixed ploidy with diploids and tetraploids simultaneously, where the reference diploids show the two homozygote extremes (Voorrips et al., 2011). It contains three functions, CodomMarker, fitTetra, and saveMarkerModels (Voorrips et al., 2011). The last two are written exclusively for tetraploid species, while CodomMarker function can possibly be used for other ploidies depending on the data quality (Voorrips et al., 2011; Vos et al., 2015). SuperMASSA is a web-based software with window interface (python scripts also available upon request), which implements an algorithm to find the exact maximum a posteriori (MAP) genotype configuration using a Bayesian model (Serang et al., 2012). It can also handle mixed ploidy for a diverse panel, though the ploidy of two parents should be the same for the F1 model (Garcia et al., 2013). So far, superMASSA is still the only software supporting all ploidies in genotype calling (Serang et al., 2012; Garcia et al., 2013; Balsalobre et al., 2017). Recently, Carley et al. (2017) designed an R package, ClusterCall, an automated method to convert signal intensity into different allele dosages for tetraploid genotypes which are called based on hierarchical clustering among multiple F1 populations. This process is implemented independently to each marker in two phases, training and prediction (Carley et al., 2017; Endelman et al., 2017). However, ClusterCall was designed specifically for autotetraploids.

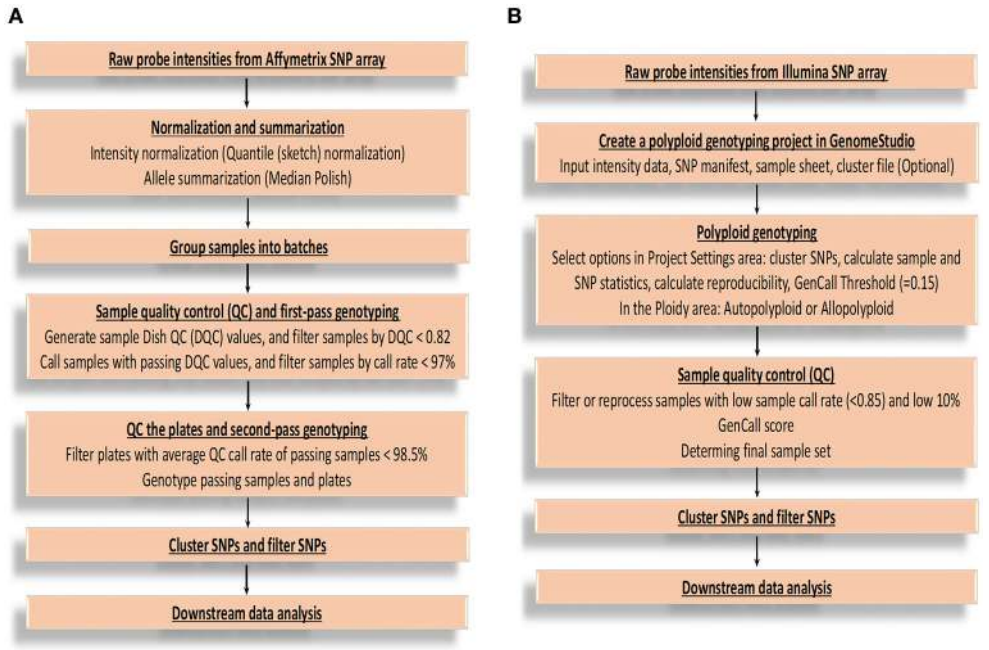

FIGURE 1 | Flowcharts of genotype calling for Affymetrix and Illumina platforms. **(A)** The flowchart of executing genotype calling mainly based on Best Practices Genotyping Workflow on the Affymetrix Axiom SNP array platform. **(B)** The flowchart of executing genotype calling based on Illumina Infinium SNP array platform. GenCall score: a quality metric calculated for each genotype (data point), and ranges from 0 to 1; 10% GenCall score: 10th percentile GenCall score for all samples.

Different genotype calling programs may perform differently on calling genotypes, because various models are implemented for each program. Given the difference of data formatting requirements from superMASSA (x_1, y_1) and ClusterCall (requirement of F1 population data sets), comparison of these three software programs was impractical. Therefore, to evaluate the impact of different genotype calling software on genotype calling in tetraploid as an example, the results comparing the software between fitTetra and ClusterCall were completed by Carley et al. (2017) and the results are summarized below. In addition, since superMASSA has been only successfully used in autopolyploid sugarcane, the comparison of its performance and accuracy with other tools is not available. Thus, we performed genotype calling by using the published data (intensities) from Voorrips et al. (2011) as input data for both software, superMASSA and fitTetra, and compared their results in section Comparison of Genotype Calling between fitTetra And SuperMASSA with Potato Illumina GoldenGate™ Assay Results.

Comparison of Genotype Calling between fitTetra and ClusterCall with Potato Infinium 8303 SNP Array Results

SNP Array Data

Three potato F1 populations, comprised of 160, 191, and 162 progeny respectively, were genotyped with the potato Infinium 8,303 SNP array (Hamilton et al., 2011; Felcher et al., 2012). Genotype calling of these three SNP array data sets was performed by using R packages ClusterCall and fitTetra separately. For fitTetra, the saveMarkerModels function was selected, and three parameters were adjusted (p.threshold = 0.85, peak.threshold = 1, and sd.threshold = 0.1) (Carley et al., 2017). For ClusterCall, to obtain the maximum number of markers with $\geq$ 0.95 concordance ratio across the three potato populations, default parameter values were used (Carley et al., 2017).

Results of Comparison

Across the three potato F1 population genotype calling results, the number of markers with over 95% concordance (the proportion of samples whose genotype was consistent with the prevalent genotype in that cluster) scored in at least one F1 population was higher when using ClusterCall (5,729 or 94.6% of the total markers) than using fitTetra (5,325, or 82.5% of its total). By increasing the threshold to one (1) concordance (perfect concordance), the number of concordant markers with ClusterCall decreased to 4,217 compared to 3,478 with fitTetra. Therefore, ClusterCall called genotypes with much higher concordance rate, thus is more accurate than fitTetra. ClusterCall used F1 populations with a large number of progeny as training data sets, which could increase the reliability of inferring genotypes based on chi-squared segregation test. In addition, ClusterCall showed higher accuracy (94.6 vs. 82.5%) and less computation time (9.5 min vs. 7.3 h) than fitTetra. However, fitTetra is dependent on fitting eight settled models and selecting the best fit by using constraints on parameters, such as the means, mixing ratios of the distributions, and Bayesian Information Criterion (BIC). Therefore, ClusterCall can

improve the quality of genotype calling by using large F1 training populations, while this has no advantage for fitTetra (Voorrips et al., 2011; Vos et al., 2015; Carley et al., 2017). However, without training data of the F1 population, fitTetra would be more widely applied than ClusterCall, specifically to call genotypes from a population of large and different families or lines, as the call rate of fitTetra should improve with enough genotypes sitting in each distribution.

Comparison of Genotype Calling between fitTetra and SuperMASSA with Potato Illumina GoldenGate™ Assay Results

SNP Array Data

The data set from an Illumina GoldenGate™ assay (Voorrips et al., 2011), comprised of 384 SNPs and used to genotype 224 tetraploid potato individuals, was analyzed. Three settings of fitTetra were adjusted according to Vos et al. (2015) (p.threshold = 0.95, peak.threshold = 0.99, and call.threshold = 0.60). After data filtering, out of the 86,016 data points (genotypes × individuals), 64,168 reached above criteria, and 69 of the 384 SNPs (18.0%) were rejected. Simultaneously, the same dataset from array assay was input into superMASSA software following the recommended MAP (Serang et al., 2012). The same model for genotype distribution (Hardy-Weinberg) was chosen for both softwares (Voorrips et al., 2011), with ploidyset to four.

Results of Comparison

Using 64,168 data points from the output that survived filtering for both tools, a total of 52,364 (81.6%) common data points (a sample assigned into same cluster with same marker by both tools) were obtained. When comparing the proportion of common genotypes with respect to the total genotypes of each individual, it was noteworthy that both software programs performed better on calling the two homozygous genotypes than calling remaining heterozygous genotypes, according to the concordance rate (e.g., common homozygous/total homozygous called from a software). To be specific, the average concordance rates in calling two homozygous genotypes were 93.9% in fitTetra and 90.5% in superMASSA. However, the software became more diverged on calling three remaining heterozygous genotypes (average concordance rate 74.1% in fitTetra vs. 76.1% in superMASSA).

The Hardy-Weinberg equilibrium (HWE) model was used to assign the genotypes in both fitTetra and superMASSA, which revealed high concordance rate of genotype calls (>80% for total calls). The difference of genotype calls could be due to the usage of two different parameters to evaluate the models: smallest BIC value for fitTetra, and MAP for superMASSA. Furthermore, the common situations were described in these two software that assuming equal distance for clusters at fixed positions, which may result in assignment of samples to improper clusters (misclassification) (Grandke et al., 2016). The comparisons of the automated genotyping calling softwares are based on minimal data. To conclude on which software performs better, further validation would be needed. Checking the segregation ratio of markers in a mapping population is an important way to validate the marker genotype calling. For example, SDM are expected to segregate in 1:1 ratio in a bi-parental population. However, for high dosage markers, to have a high resolution of the segregation ratio, the population size should be big enough for this purpose. The other way could be using the NGS methods to sequence the SNP regions to a high depth and using the ratio of each haplotype allele read to determine the dosage of each SNP allele. Alternatively, the concordance genotypes can be considered as more reliable than discordance genotypes.

CONCLUSION AND SUGGESTIONS

Although challenges exist for SNP discovery in polyploid species, progress has been made with the improved and increased number of available analysis tools (Clevenger et al., 2015; Song et al., 2016). As a high-throughput genotyping assay, SNP array technology is rightfully gaining popularity for SNP genotyping due to its flexibility, relative cost-efficiency, and automatic genotype calling, as discussed in this review.

SNP selection for SNP array design is a critical step for successful SNP array application. SNPs identified from gene enriched sequences are preferable for SNP array development (Hulse-Kemp et al., 2015; Clevenger et al., 2017). In regards to techniques, the first thing to consider for SNPs to be included in the array is whether the SNPs are suitable for probe design based on the requirements of selected platform, such as SNP depth, SNP types, SNP frequency, additional variations within probe sequence of target SNPs, and Affymetrix P-convent value or Illumina ADT value. The second consideration will be the dosage of SNPs. For polyploids such as sugarcane, given the lack of software for high dosage SNP marker analyses, SDMs could be preferred for SNP array design, as they can be treated as markers in diploids for genotype calling. Similarly, genome-specific SNPs are preferred for allopolyploids. Markers from elite cultivars, accessions and related species can be preferentially selected due to wider genetic backgrounds. To reduce the influence of ascertainment bias, the third optional consideration can be SNP distribution according to different homologous chromosomes and functions (may focus on genic regions) if the reference genome is available and fully annotated, and adding SNP resources from wild species or related species. In fact, before developing a large SNP array, a small scale of SNP array to validate some of SNPs is viable.

To call the genotypes (or dosage of each SNP locus) based on the SNP array results in polyploids, three additional open access tools are available beside the native callers from the platforms. In the only comparison of ClusterCall and fitTetra to date, ClusterCall showed higher accuracy and less computation time than fitTetra for genotype calling in an auto-tetraploid species (Carley et al., 2017). However, without training data of a F1 population, fitTetra will be more widely applied than ClusterCall, specifically for calling genotypes from a population with different families or lines. Therefore, a hybrid approach to combine ClusterCall training set calls with fitTetra prediction set calls can be applied (Carley et al., 2017). The performance of fitTetra and superMASSA were both better in calling homozygous genotypes

than calling heterozygous genotypes. Currently, superMASSA is the only open-access tool that has been successfully utilized in determining ploidies and genotype calling based on SNP array data for highly polyploid species like sugarcane (Garcia et al., 2013; Costa et al., 2016; Balsalobre et al., 2017). To analyze SNP array data for highly polyploid species (ploidy >4), superMASSA software (Carley et al., 2017) with adjusted settings (default maybe not suggested for all data sets) can be used to call all the possible genotypes. An updated version of fitTetra, fitPoly, seemed to be available soon (Van Geest et al., 2017), which can be explored for calling genotypes with multiple dosages. Otherwise, the Affymetrix and Illumina genotype calling platforms are still the main choice for genotype calling of SDM or genome-specific SNPs. Hopefully, new technologies and tools will be available to analyze different allele dosages of SNP array data for highly polyploid crops such as sugarcane in the near future.

AUTHOR CONTRIBUTIONS

JW guided the intention and outline of the review. QY prepared the manuscript draft. XY, ZP, LX, and JW critically revised and re-wrote parts of the manuscript.

ACKNOWLEDGMENTS

We would like to thank Dante Leventini and Erik Hanson for editing the manuscript. This work was supported by the Scientific Research Foundation of Graduate School at the Fujian Agriculture and Forestry University, Chinese Government Scholarship (CSC No. 201608350089), Florida Sugarcane League, and USDA National Institute of Food and Agriculture, Hatch Project 1011664.

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001

Aitken, K., Farmer, A., Berkman, P., Muller, C., Wei, X., Demano, E., et al. (2016). Generation of a 345K sugarcane SNP chip. *Proc. Aust. Soc. Sugar Cane Technol.* 29, 1165–1172.

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi: 10.1093/molbev/msq148

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., et al. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* 15, 390–401. doi: 10.1111/pbi.12635

Allwright, M. R., and Taylor, G. (2016). Molecular breeding for improved second generation bioenergy crops. *Trends Plant Sci.* 21, 43–54. doi: 10.1016/j.tplants.2015.10.002

Andon, M. B., and Anderson, J. W. (2008). State of the art reviews: the oatmeal-cholesterol connection: 10 years later. *Am. J. Lifestyle Med.* 2, 51–57. doi: 10.1177/1559827607309130.

Baker, P., Jackson, P., and Aitken, K. (2010). Bayesian estimation of marker dosage in sugarcane and other autopolyploids. *Theor. Appl. Genet.* 120, 1653–1672. doi: 10.1007/s00122-010-1283-z

Balsalobre, T. W. A., Da Silva Pereira, G., Margarido, G. R. A., Gazaffi, R., Barreto, F. Z., Anoni, C. O., et al. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics* 18:72. doi: 10.1186/s12864-016-3383-x

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., and Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *Plant J.* 51, 910–918. doi: 10.1111/j.1365-313X.2007.03193.x

Bassil, N. V., Davis, T. M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., et al. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16:1. doi: 10.1186/s12864-015-1310-1

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denance, C., Theron, A., et al. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant J.* 86, 62–74. doi: 10.1111/tpj.13145

Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., et al. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* × *domestica* Borkh). *PLoS ONE* 9:e110377. doi: 10.1371/journal.pone.0110377

Borrill, P., Adamski, N., and Uauy, C. (2015). Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol.* 208, 1008–1022. doi: 10.1111/nph.13533

Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics* 201, 853–863. doi: 10.1534/genetics.115.181008

Bus, A., Hecht, J., Huettel, B., Reinhardt, R., and Stich, B. (2012). High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics* 13:281. doi: 10.1186/1471-2164-13-281

Carley, C. A. S., Coombs, J. J., Douches, D. S., Bethke, P. C., Palta, J. P., Novy, R. G., et al. (2017). Automated tetraploid genotype calling by hierarchical clustering. *Theor. Appl. Genet.* 130, 717–726. doi: 10.1007/s00122-016-2845-5

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., et al. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE* 7:e31745. doi: 10.1371/journal.pone.0031745

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., et al. (2014). A high-density SNP genotyping array for rice biology and molecular breeding. *Mol. Plant* 7, 541–553. doi: 10.1093/mp/sst135

Chen, Z. J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. *Nat. Rev. Genet.* 14, 471. doi: 10.1038/nrg3503

Chen, Z. J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* 28, 240–252. doi: 10.1002/bies.20374

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* 129, 1887–1899. doi: 10.1007/s00122-016-2746-7

Clarke, W. E., Parkin, I. A., Gajardo, H. A., Gerhardt, D. J., Higgins, E., Sidebottom, C., et al. (2013). Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L. *PLoS ONE* 8:e81992. doi: 10.1371/journal.pone.0081992

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., and Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Mol. Plant* 8, 831–846. doi: 10.1016/j.molp.2015.02.002

Clevenger, J., Chu, Y., Chavarro, C., Agarwal, G., Bertioli, D. J., Leal-Bertioli, S. C., et al. (2017). Genome-wide SNP genotyping resolves signatures of selection and tetrasomic recombination in peanut. *Mol. Plant* 10, 309–322. doi: 10.1016/j.molp.2016.11.015

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Consortium, I. W. G. S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788. doi: 10.1126/science.1251788

Consortium, P. G. S. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Costa, E. A., Anoni, C. O., Mancini, M. C., Santos, F. R. C., Marconi, T. G., Gazaffi, R., et al. (2016). QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. *Euphytica* 211, 1–16. doi: 10.1007/s10681-016-1746-7

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., et al. (2014). A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Funct. Integr. Genomics* 14, 643–655. doi: 10.1007/s10142-014-0391-2

Deschamps, S., Llaca, V., and May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology* 1, 460–483. doi: 10.3390/biology1030460

Deulvot, C., Charrel, H., Marty, A., Jacquin, F., Donnadieu, C., Lejeune-Henaut, I., et al. (2010). Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics* 11:468. doi: 10.1186/1471-2164-11-468

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, A., et al. (2015). Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci. Rep.* 5:17394. doi: 10.1038/srep17394

D'hont, A., Ison, D., Alix, K., Roux, C., and Glaszmann, J. C. (1998). Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* 41, 221–225. doi: 10.1139/g98-023

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B., Carley, C. A. S., Douches, D. S., Coombs, J. J., Bizimungu, B., De Jong, W. S., et al. (2017). Pedigree reconstruction with genome-wide markers in potato. *Am. J. Potato Res.* 94, 184–190. doi: 10.1007/s12230-016-9556-y

Ergül, A., Marasali, B., and Agaoglu, Y. (2015). Molecular discrimination and identification of some Turkish grape cultivars (*Vitis vinifera* L.) by RAPD markers. *VITIS J. Grapevine Res.* 41, 159.

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., et al. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE* 7:e36347. doi: 10.1371/journal.pone.0036347

Ganal, M. W., Durstewitz, G., Polley, A., Berard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* 6:e28334. doi: 10.1371/journal.pone.0028334

Gao, F., Liu, J., Yang, L., Wu, X., Xiao, Y., Xia, X., et al. (2016). Genome-wide linkage mapping of QTL for physiological traits in a Chinese wheat population using the 90K SNP array. *Euphytica* 209, 789–804. doi: 10.1007/s10681-016-1682-6

Gao, L., Turner, M. K., Chao, S., Kolmer, J., and Anderson, J. A. (2016). Genome wide association study of seedling and adult plant leaf rust resistance in elite spring wheat breeding lines. *PLoS ONE* 11:e0148671. doi: 10.1371/journal.pone.0148671

Gao, L., Zhao, G., Huang, D., and Jia, J. (2017). Candidate loci involved in domestication and improvement detected by a published 90K wheat SNP array. *Sci. Rep.* 7:44530. doi: 10.1038/srep44530

Garcia, A. A., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* 3:3399. doi: 10.1038/srep03399

García-Pereira, M. J., Carvajal-Rodriguez, A., Whelan, S., Caballero, A., and Quesada, H. (2014). Impact of deep coalescence and recombination on the estimation of phylogenetic relationships among species using AFLP markers. *Mol. Phylogenet. Evol.* 76, 102–109. doi: 10.1016/j.ympev.2014.03.001

Garvin, M. R., Saitoh, K., and Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Mol. Ecol. Resour.* 10, 915–934. doi: 10.1111/j.1755-0998.2010.02891.x

Gezan, S. A., Osorio, L. F., Verma, S., and Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Hortic Res.* 4, 16070. doi: 10.1038/hortres.2016.70

Grandke, F., Singh, P., Heuven, H. C., De Haan, J. R., and Metzler, D. (2016). Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum. *BMC Genomics* 17:672. doi: 10.1186/s12864-016-2926-5

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., et al. (2011). Demographic history and rare allele sharing among human populations. *?Proc. Natl. Acad. Sci. U.S.A.* 108, 11983–11988. doi: 10.1073/pnas.1019276108

Guerra, M. (2008). Chromosome numbers in plant cytotaxonomy: concepts and implications. *Cytogenet. Genome Res.* 120, 339–350. doi: 10.1159/000121083

Gupta, P., and Priyadarshan, P. (1982). Triticale: present status and future prospects. *Adv. Genet.* 21, 255–345. doi: 10.1016/S0065-2660(08)60300-4

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., et al. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics* 12:302. doi: 10.1186/1471-2164-12-302

Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., et al. (2012). Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nat. Biotechnol.* 30, 798–802. doi: 10.1038/nbt.2302

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. doi: 10.3389/fpls.2014.00484

Heslot, N., Rutkoski, J., Poland, J., Jannink, J. L., and Sorrells, M. E. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS ONE* 8:e74612. doi: 10.1371/journal.pone.0074612

Hilu, K. W. (1993). Polyploidy and the evolution of domesticated plants. *Am. J. Bot.* 80, 1494–1499. doi: 10.2307/2445679

Hinze, L. L., Hulse-Kemp, A. M., Wilson, I. W., Zhu, Q. H., Llewellyn, D. J., Taylor, J. M., et al. (2017). Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. *BMC Plant Biol.* 17:37. doi: 10.1186/s12870-017-0981-y

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., et al. (2014). Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species. *DNA Res.* 21, 169–181. doi: 10.1093/dnares/dst049

Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., et al. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* 15, 1374–1386. doi: 10.1111/pbi.12722

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63K SNP array for cotton and high-density mapping of intra-and inter-specific populations of *Gossypium* spp. *G3* 5, 1187–1209. doi: 10.1534/g3.115.018416

Jung, H.-J., Veerappan, K., Natarajan, S., Jeong, N., Hwang, I., Nagano, S., et al. (2017). A system for distinguishing octoploid strawberry cultivars using high-throughput SNP genotyping. *Trop. Plant Biol.* 10, 68–76. doi: 10.1007/s12042-017-9185-8

Kaur, S., Francki, M. G., and Forster, J. W. (2012). Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnol. J.* 10, 125–138. doi: 10.1111/j.1467-7652.2011.00644.x

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 37, 4181–4193. doi: 10.1093/nar/gkp552

Le, S. Q., and Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 21, 952–960. doi: 10.1101/gr.113084.110

Li, X., Acharya, A., Farmer, A. D., Crow, J. A., Bharti, A. K., Kramer, R. S., et al. (2012). Prevalence of single nucleotide polymorphism among 27 diverse alfalfa genotypes as assessed by transcriptome sequencing. *BMC Genomics* 13:568. doi: 10.1186/1471-2164-13-568

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., et al. (2014a). Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS ONE* 9:e84329. doi: 10.1371/journal.pone.0084329

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014b). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3* 4, 1971–1979. doi: 10.1534/g3.114.012245

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., and Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 13(Suppl. 8):S8. doi: 10.1186/1471-2164-13-S8-S8

Liu, S., Sun, L., Li, Y., Sun, F., Jiang, Y., Zhang, Y., et al. (2014). Development of the catfish 250K SNP array for genome-wide association studies. *BMC Res. Notes* 7:135. doi: 10.1186/1756-0500-7-135

Liu, S., Zeng, Q., Wang, X., and Liu, Z. (2017). "SNP array development, genotyping, data analysis, and applications," in *Bioinformatics in Aquaculture: Principles and Methods,* ed Z. Liu (Chichester: John Wiley & Sons, Ltd.), 308–337.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9:e1003215. doi: 10.1371/journal.pgen.1003215

Luo, Z., Hackett, C., Bradshaw, J., Mcnicol, J., and Milbourne, D. (2001). Construction of a genetic linkage map in tetraploid species using molecular markers. *Genetics* 157, 1369–1385.

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* 110, 99–104. doi: 10.1038/hdy.2012.79

Mahoney, L., Sargent, D., Wood, D., Ward, J., Bassil, N., Handcock, J., et al. (2016). A high-density linkage map of the ancestral diploid strawberry, *Fragaria iinumae*, constructed with single nucleotide polymorphism markers from the IStraw90 array and genotyping by sequencing. *Plant Genome* 9. doi: 10.3835/plantgenome2015.08.0071

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 345:1250092. doi: 10.1126/science.1250092

Martino, A., Mancuso, T., and Rossi, A. M. (2010). Application of high-resolution melting to large-scale, high-throughput SNP genotyping: a comparison with the TaqMan method. *J. Biomol. Screen.* 15, 623–629. doi: 10.1177/1087057110365900

Mason, A. S., Higgins, E. E., Snowdon, R. J., Batley, J., Stein, A., Werner, C., et al. (2017). A user guide to the*Brassica*60K Illumina Infinium^{TM} SNP genotyping array. *Theor. Appl. Genet.* 130, 621–633. doi: 10.1007/s00122-016-2849-1

McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., Mcnally, K. L., Fitzgerald, M., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nat. Commun.* 7:10532. doi: 10.1038/ncomms10532

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J. Jr., Grattapaglia, D., Sederoff, R. R., et al. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9:312. doi: 10.1186/1471-2164-9-312

O'Rourke, J. A., Fu, F., Bucciarelli, B., Yang, S. S., Samac, D. A., Lamb, J. F., et al. (2015). The *Medicago sativa* gene index 1.2: a web-accessible gene expression atlas for investigating expression differences between *Medicago sativa* subspecies. *BMC Genomics* 16:502. doi: 10.1186/s12864-015-1718-7

Oliver, R. E., Lazo, G. R., Lutz, J. D., Rubenfield, M. J., Tinker, N. A., Anderson, J. M., et al. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. *BMC Genomics* 12:77. doi: 10.1186/1471-2164-12-77

Oliver, R. E., Tinker, N. A., Lazo, G. R., Chao, S., Jellen, E. N., Carson, M. L., et al. (2013). SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. *PLoS ONE* 8:e58068. doi: 10.1371/journal.pone.0058068

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462. doi: 10.1016/j.cell.2007.10.022

Pandey, M. K., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., et al. (2017). Development and evaluation of a high density genotyping 'Axiom_*Arachis*' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* 7:40577. doi: 10.1038/srep40577

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Pastina, M. M., Malosetti, M., Gazaffi, R., Mollinari, M., Margarido, G. R., Oliveira, K. M., et al. (2012). A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theor. Appl. Genet.* 124, 835–849. doi: 10.1007/s00122-011-1748-8

Peng, Z., Fan, W., Wang, L., Paudel, D., Leventini, D., Tillman, B. L., et al. (2017). Target enrichment sequencing in cultivated peanut (*Arachis hypogaea* L.) using probes designed from transcript sequences. *Mol. Genet. Genomics.* 292, 955–965. doi: 10.1007/s00438-017-1327-z

Peng, Z., Gallo, M., Tillman, B. L., Rowland, D., and Wang, J. (2016). Molecular marker development from transcript sequences and germplasm evaluation for cultivated peanut (*Arachis hypogaea* L.). *Mol. Genet. Genomics* 291, 363–381. doi: 10.1007/s00438-015-1115-6

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and 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:e37135. doi: 10.1371/journal.pone.0037135

Poland, J. A., Brown, P. J., Sorrells, M. E., and 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:e32253. doi: 10.1371/journal.pone.0032253

Qi, H., Song, K., Li, C., Wang, W., Li, B., Li, L., et al. (2017). Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*). *PLoS ONE* 12:e0174007. doi: 10.1371/journal.pone.0174007

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., et al. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* 10, 1047–1064. doi: 10.1016/j.molp.2017.06.008

Rieseberg, L. H., and Willis, J. H. (2007). Plant speciation. *Science* 317, 910–914. doi: 10.1126/science.1137729

Roach, B. T. (1989). Origin and improvement of the genetic base of sugarcane. *Proc. Aust. Soc. Sugar Cane Technol.* 11, 34–47.

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. doi: 10.1111/pbi.12645

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* 11(Suppl. 1), 1–8. doi: 10.1111/j.1755-0998.2010.02979.x

Serang, O., Mollinari, M., and Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS ONE* 7:e30906. doi: 10.1371/journal.pone.0030906

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Soltis, P. S., Marchant, D. B., Van de Peer, Y., and Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125. doi: 10.1016/j.gde.2015.11.003

Song, J., Yang, X., Resende, M. F. Jr., Neves, L. G., Todd, J., Zhang, J., et al. (2016). Natural allelic variations in highly polyploidy *Saccharum* complex. *Front. Plant Sci.* 7:804. doi: 10.3389/fpls.2016.00804

Stupar, R. M., Bhaskar, P. B., Yandell, B. S., Rensink, W. A., Hart, A. L., Ouyang, S., et al. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics* 176, 2055–2067. doi: 10.1534/genetics.107.074286

Tayalé, A., and Parisod, C. (2013). Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenet. Genome Res.* 140, 79–96. doi: 10.1159/000351318

Thiel, T., Kota, R., Grosse, I., Stein, N., and Graner, A. (2004). SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. *Nucleic Acids Res.* 32:e5. doi: 10.1093/nar/gnh006

Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed. Biotechnol.* 2, 195–212. doi: 10.9787/PBB.2014.2.3.195

Tinker, N. A., Chao, S., Lazo, G. R., Oliver, R. E., Huang, Y.-F., Poland, J. A., et al. (2014). A SNP genotyping array for hexaploid oat. *Plant Genome* 7. doi: 10.3835/plantgenome2014.03.0010

Tinker, N. A., Kilian, A., Wight, C. P., Heller-Uszynska, K., Wenzl, P., et al. (2009). New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics* 10:39. doi: 10.1186/1471-2164-10-39

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., and Belzile, F. (2017). Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics* 18:5. doi: 10.1186/s12859-016-1431-9

Tumino, G., Voorrips, R. E., Rizza, F., Badeck, F. W., Morcia, C., Ghizzoni, R., et al. (2016). Population structure and genome-wide association analysis for frost tolerance in oat using continuous SNP array signal intensity ratios. *Theor. Appl. Genet.* 129, 1711–1724. doi: 10.1007/s00122-016-2734-y

Turuspekov, Y., Plieske, J., Ganal, M., Akhunov, E., and Abugalieva, S. (2015). Phylogenetic analysis of wheat cultivars in Kazakhstan based on the wheat 90 K single nucleotide polymorphism array. *Plant Genet. Resour.* 15, 29–35. doi: 10.1017/S1479262115000325

Uitdewilligen, J. G., Wolters, A.-M. A., Bjorn, B., Borm, T. J., Visser, R. G., and Van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* 8:e62355. doi: 10.1371/journal.pone.0062355

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 K SNP genotyping array. *BMC Genomics* 15:823. doi: 10.1186/1471-2164-15-823

Van Geest, G., Voorrips, R. E., Esselink, D., Post, A., Visser, R. G., and Arens, P. (2017). Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183k SNP array. *BMC Genomics* 18:585. doi: 10.1186/s12864-017-4003-0

Voorrips, R. E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12:172. doi: 10.1186/1471-2105-12-172

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., Van Eck, H. J., and Van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130, 123–135. doi: 10.1007/s00122-016-2798-8

Vos, P. G., Uitdewilligen, J. G., Voorrips, R. E., Visser, R. G., and Van Eck, H. J. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* 128, 2387–2401. doi: 10.1007/s00122-015-2593-y

Vukosavljev, M., Arens, P., Voorrips, R. E., van 't Westende, W. P., Esselink, G. D., Bourke, P. M., et al. (2016). High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68K rose SNP array. *Hortic. Res.* 3, 16052. doi: 10.1038/hortres.2016.52

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep.* 6:20728. doi: 10.1038/srep20728

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, X., Long, Y., Wang, N., Zou, J., Ding, G., Broadley, M. R., et al. (2017). Breeding histories and selection criteria for oilseed rape in Europe and China identified by genome wide pedigree dissection. *Sci. Rep.* 7:1916. doi: 10.1038/s41598-017-02188-z

Watanabe, K. (2015). Potato genetics, genomics, and applications. *Breed. Sci.* 65, 53–68. doi: 10.1270/jsbbs.65.53

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* 14, 1195–1206. doi: 10.1111/pbi.12485

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Wu, J., Zhao, Q., Liu, S., Shahid, M., Lan, L., Cai, G., et al. (2016). Genome-wide association study identifies new loci for resistance to Sclerotinia stem rot in *Brassica napus*. *Front. Plant Sci.* 7:1418. doi: 10.3389/fpls.2016.01418

Wu, K., Burnquist, W., Sorrells, M., Tew, T., Moore, P., and Tanksley, S. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theor. Appl. Genet.* 83, 294–300. doi: 10.1007/BF00224274

Würschum, T., Langer, S. M., Longin, C. F., Korzun, V., Akhunov, E., Ebmeyer, E., et al. (2013). Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. *Theor. Appl. Genet.* 126, 1477–1486. doi: 10.1007/s00122-013-2065-1

Yang, X., Song, J., You, Q., Paudel, D. R., Zhang, J., and Wang, J. (2017). Mining sequence variations in representative polyploid sugarcane germplasm accessions. *BMC Genomics* 18:594. doi: 10.1186/s12864-017-3980-3

Yu, H., Xie, W., Li, J., Zhou, F., and Zhang, Q. (2014). A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol. J.* 12, 28–37. doi: 10.1111/pbi.12113

Zhang, H. B., Li, Y., Wang, B., and Chee, P. W. (2008). Recent advances in cotton genomics. *Int. J. Plant Genomics* 2008:742304. doi: 10.1155/2008/742304

Zou, J., Hu, D., Mason, A. S., Shen, X., Wang, X., Wang, N., et al. (2018). Genetic changes in a novel breeding population of *Brassica napus* synthesized from hundreds of crosses between *B. rapa and B. carinata. Plant Biotechnol. J.* 16, 507–519. doi: 10.1111/pbi.12791

Zou, J., Semagn, K., Iqbal, M., N'diaye, A., Chen, H., Asif, M., et al. (2016). Mapping QTLs controlling agronomic traits in the 'Attila'×'CDC Go'spring wheat population under organic management using 90K SNP array. *Crop. Sci.* 57, 365–377. doi: 10.2135/cropsci2016.06.0459

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.