frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 28 June 2023
DOI 10.3389/fpls.2023.1216505

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Yaping Yuan,
Jilin University, China
Kun Li,
Chinese Academy of Agricultural Sciences (CAAS), China

*CORRESPONDENCE
Han Zhao
zhaohan@jaas.ac.cn
Jiuran Zhao
maizezhao@126.com
Fengge Wang
fenngewangmaize@126.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 04 May 2023
ACCEPTED 12 June 2023
PUBLISHED 28 June 2023

CITATION
Liu Z, Zhao Y, Zhang Y, Xu L, Zhou L, Yang W, Zhao H, Zhao J and Wang F (2023) Development of Omni InDel and supporting database for maize. *Front. Plant Sci.* 14:1216505. doi: 10.3389/fpls.2023.1216505

Development of Omni InDel and supporting database for maize

Zhihao Liu[1,2†], Yikun Zhao[1†], Yunlong Zhang[1†], Liwen Xu[1], Ling Zhou[3], Weiguang Yang[2], Han Zhao[3*], Jiuran Zhao[1*] and Fengge Wang[1*]

[1]Key Laboratory of Crop DNA Fingerprinting Innovation and Utilization (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Beijing Academy of Agricultural and Forest Sciences (BAAFS), Beijing, China, [2]College of Agriculture, Jilin Agricultural University, Changchun, China, [3]Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing, Jiangsu, China

Insertions–deletions (InDels) are the second most abundant molecular marker in the genome and have been widely used in molecular biology research along with simple sequence repeats (SSR) and single-nucleotide polymorphisms (SNP). However, InDel variant mining and marker development usually focuses on a single type of dimorphic InDel, which does not reflect the overall InDel diversity across the genome. Here, we developed Omni InDels for maize, soybean, and rice based on sequencing data and genome assembly that included InDel variants with base lengths from 1 bp to several Mb, and we conducted a detailed classification of Omni InDels. Moreover, we screened a set of InDels that are easily detected and typed (Perfect InDels) from the Omni InDels, verified the site authenticity using 3,587 germplasm resources from 11 groups, and analyzed the germplasm resources. Furthermore, we developed a Multi-InDel set based on the Omni InDels; each Multi-InDel contains multiple InDels, which greatly increases site polymorphism, they can be detected in multiple platforms such as fluorescent capillary electrophoresis and sequencing. Finally, we developed an online database website to make Omni InDels easy to use and share and developed a visual browsing function called "Variant viewer" for all Omni InDel sites to better display the variant distribution.

KEYWORDS

Omni InDel, maize, InDel, crops, database

Introduction

Insertions–deletions (InDels) rank second only to single-nucleotide polymorphisms (SNPs) as the most prevalent form of genetic variation in plants. Increasing numbers of studies have shown that InDels, which have the advantages of high polymorphism, co-dominance, high density, high reliability, and ease of genotyping, are the main source of plant genetic structural variation and are widely distributed in plant genomes (Das et al., 2015; Jain et al., 2019; Sun et al., 2019; Cui et al., 2021; Loewenthal et al., 2021; Pan et al., 2022). In practical applications, InDels are also favored by breeders compared to SNPs and

simple sequence repeats (SSRs) because InDel detection merely necessitates straightforward techniques such as gel-based size separation or polyacrylamide gel electrophoresis, common procedures in genetics and breeding laboratories. (Păcurar et al., 2012; Liu et al., 2013; Wu et al., 2013; Moghaddam et al., 2014; Yang et al., 2014; Li et al., 2015; Zhou et al., 2015; Feng et al., 2020; Pan et al., 2021). Moreover, InDels are readily compatible with multi-platform detection techniques, including electrophoresis, chip arrays, and KASP tools that have been widely developed and applied for species diagnosis, marker-assisted selection breeding, evolutionary studies, genetic linkage map construction, and ancestral sequence reconstruction (Ashkenazy et al., 2012; Liu et al., 2013; Yamaki et al., 2013; Li et al., 2015; Song et al., 2015; Vialle et al., 2018; Jain et al., 2019; Pan et al., 2022; Seo et al., 2022). In addition, InDels are an important aspect of molecular evolutionary research because these variations shape genes and genomes, and some studies have shown their effects on proteins, including potential functional gains/losses in organisms (Lovett, 2004; Cartwright, 2009; Tian et al., 2008; Terakami et al., 2012; Rockah-Shmuel et al., 2013; Lin et al., 2017; Ramakrishna et al., 2018). However, in current crop research studies, InDel variant mining and marker development is usually focused on a single type of dimorphic InDel, which does not reflect the overall InDel diversity across the genome (Zhao et al., 2017; Wang et al., 2018; Paudel et al., 2019; Wang et al., 2020; Fei et al., 2021; Hechanova et al., 2021; Liu et al., 2022; Seki, 2022). Only small, dimorphic InDels have been extensively examined in whole plant genomes; InDels of various lengths and types, which have important value and application potential in genetic research, have not been fully elucidated and systematically analyzed.

Despite maize, rice, and soybean being significant model plants in fields of genetic breeding, gene function, evolutionary genetics, and plant diversity research (Carter et al., 2004; Hirsch et al., 2014; Li et al., 2014; Schatz et al., 2014; Wing et al., 2018; Li et al., 2020; Liu et al., 2020; Hufford et al., 2021; Qin et al., 2021), several factors have inhibited the comprehensive development of InDel markers within these species. First, mining InDel loci of different lengths has traditionally relied on the examination of a single reference genome, which leads to genomic bias and the inevitable loss of some InDel loci. Furthermore, these mining strategies based on single reference genomes often use second-generation short reads for mapping, but many important InDels associated with agronomic traits are too large to be easily detected with Illumina short reads (Liu et al., 2020; Matar and Melzer, 2021). Therefore, the identification of medium (3–50 bp) and large (≥50 bp) InDels requires the use of a multi-genome collinearity approach rather than a single-genome mining strategy. Second, almost all research related to InDel mining and marker development has focused on dimorphic InDels, but an InDel itself is polymorphic (Li et al., 2019; Bennett et al., 2020). Third, Multi-InDel analysis, which examines the number of useful InDel polymorphisms while retaining the advantages of SNPs and SSRs, has been applied in medicinal research but not yet in plants (Sun et al., 2019; Qu et al., 2020). Fourth, Perfect InDels, which exist naturally in both animals and plants, are single-copy, mainly dimorphic, easily detected and typed, and characterized by a high minor allele frequency (MAF), conserved flanks, and a length of 3–10 bp; thus, they are compatible with multiple detection platforms. Perfect InDels can be used to track certain key traits and identify cultivars but have not been mined in model crop systems. Furthermore, no studies have been performed to identify and analyze InDel loci with insertions or deletions in the form of repeating units or to uncover InDel loci embedded in other types of variants. In summary, the types, and lengths of InDel loci in crop plants have not been systematically obtained and sorted. Overcoming the above-mentioned problems associated with InDel variation mining and application to model crops such as maize, rice, and soybean would provide an important foundation for breeding and domestication research on other crop species.

To comprehensively capture and analyze the intricacies of InDel variations in crops, our initial step involved formulating a strategy based on Omni InDel. In this approach, variation mining in the studied crop species was carried out using multiple assembled genomes within the species; simultaneously, all collected InDel variants within the genome were integrated into a thorough variation map of InDels comprising dimorphic, polymorphic InDels; Perfect InDels; and Multi-InDels that range from 1 bp to several Mb. We also classified types of SSRs and non-SSR InDels. To enrich our InDel variation map and provide plant researchers with richer variation information, we then collected information on variable InDel loci in maize, soybean, and rice and constructed Omni InDel databases for these species. We also established a multi-species Omni InDel database website for data sharing and dissemination. Finally, we selected several loci for chip verification and, using maize as an example, identified 3,587 germplasm resources to verify the effectiveness of the developed markers.

Materials and methods

Small and medium Omni InDel identification

To identify small Omni InDels (SOIs, 1~2bp) and medium Omni InDels (MOIs, 3~50bp), the illumina paired-end reads of each maize species were mapped to genome assembly B73 with BWA (v. 0.7.17-r1198-dirty) (Li, 2013) with default settings. The bam files obtained with BWA were sorted by SAMtools (v. 1.9) (Danecek et al., 2021). The variants were called using GATK (v. 4.2) (McKenna et al., 2010) with default settings, and the InDels were extracted on the basis of the GATK results.

Large and huge Omni InDel identification with MUMMER

To identify large Omni InDels (LOIs, 50–1000 bp) and huge Omni InDels (HOIs, >1000 bp), we aligned the 11 maize inbred line assemblies to the B73 reference genome based on Nucmer (v. 4.0.0beta2) (Marçais et al., 2018) with the parameters "—mum -g 1000 -c 90 -l 40." Then, the alignment files were filtered to generate 1-to-1 mapping by delta-filter with the parameters "-m -i 90 -l 100."

The Nucmer output was analyzed using SyRI (https://github.com/schneebergerlab/syri) (Goel et al., 2019) with default parameters to identify variation. According to the sequence variation definitions in SyRI outputs, we extracted the InDel variation from the raw results.

Germplasm genetic pattern analysis

A phylogenetic tree was constructed using medium Omni InDels (MOIs) by unweighted pair group method with arithmetic mean (UPGMA) in the R (v. 3.5.1) package "phangorn" (Schliep, 2011). The data used to construct the tree were from a genetic distance matrix calculated by the "poppr" (Kamvar et al., 2014) package in R.

Principal components analysis (PCA) of InDel loci from 3,558 *Zea mays* inbred lines was also conducted in PLINK v1.90b6.21 64-bit (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007) with the same dataset. The PCA results further confirmed and supported the clusters classified by the phylogenetic tree.

Development of the Omni InDel web-based database

The Omni InDel online platform was built using the Next.js framework. Next.js is a lightweight React.js server-side rendering application framework that connects web pages to application programming interfaces through a multifunctional router. This platform integrates powerful JavaScript tools and makes it easier to build and deploy the released versions through the Webpack package. We standardized the code format using the Prettier tool, used the Eslint tool to eliminate obvious errors from the code, and used Babel to maintain compatibility between normal JS and ES6. On the front end, we used Ant Design, which is a front-end user interface component based on the React.js.

Results

Omni InDel statistics of crops

Omni InDel sites encompass insertion-deletion variations that range from 1 bp to 9.4 Mb. Specifically, InDel sites spanning 1 bp to 2 bp are classified as small Omni InDel (SOIs), those with lengths from 3 bp to 50 bp were considered MOIs, those with lengths from 50 bp to 1000 bp were considered LOIs, and those with lengths >1000 bp were considered HOIs.

In order to construct Omni InDel datasets for SOI and MOI sites less than 50 bp in length, next-generation paired-end sequencing (average 4×) on 438 maize inbred lines commonly used in breeding at home and abroad was performed, followed by InDel calling using GATK with B73 as the reference genome. After filtering and screening, 17,820,869 InDels were obtained. These sites were predominantly distributed in the chromosome arms and less so in the centromeric region (Figure 1A). Comparison with the annotated gene positions of the B73 genome revealed that 15.49% of the sites were distributed in the gene region, and the others were distributed in the intergenic region. Insertions and deletions accounted for 37.92% and 57.73% of the sites, respectively. In addition, 4.35% of the sites showed characteristics of both insertions and deletions (Figure 1B). These InDel sites could not be traditionally classified; therefore, these InDel named as Mixed-Omni InDel. Among the non-mixed sites, more than 60% were within 1–2 bp; additionally, 6.08% had more than two alleles, and the rest were biallelic sites (Figure 1C).

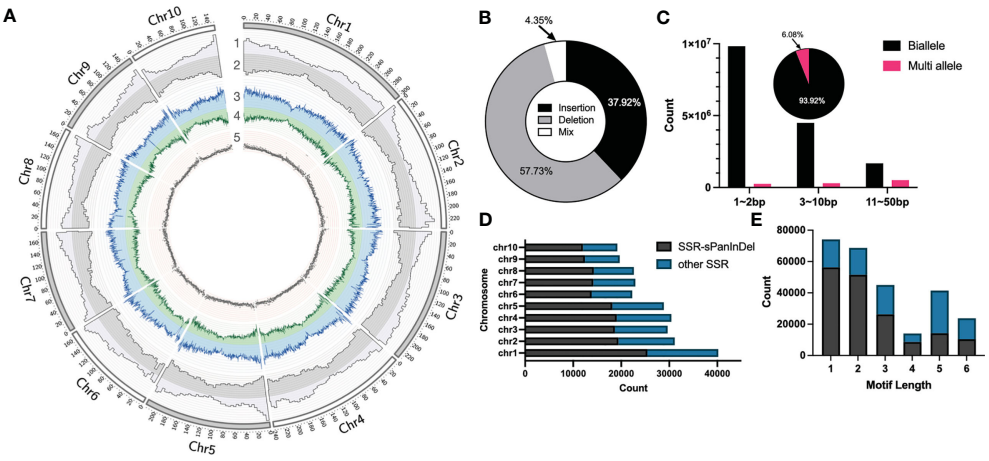

FIGURE 1
Statistical overview of small Omni InDels (SOIs) and medium Omni InDels (MOIs). **(A)** Circos (Krzywinski et al., 2009) display of SOI and MOI density distribution. 1, Gene density of B73. 2, Distributions of SOI and MOI. 3, Deletion density. 4, Insertion density. 5, Distribution of SSR-InDels in chromosomes, the window size for density calculation is 5 million bp. **(B)** The proportion of insertion, deletion and insertion-deletion mixed type sites in the MOI and SOI sets. **(C)** Counts of biallele and multi-allele InDels of different sizes (1–2 bp, 3–10 bp, 11–50 bp), and the pie is the ratio of biallele and multi-allele InDels. **(D)** Numbers of SSRs in each maize chromosome and the ratio between Omni InDel SSRs and other SSRs. **(E)** Numbers of SSRs with different motif lengths and the ratio between Omni InDel SSRs and other SSRs.

The polymorphisms of SSR molecular markers manifest as different numbers of motifs in different plant varieties, and sequence scanning and identification of motifs of an assembled genome is generally used when mining such sites. Utilizing Misa, a total of 267,041 SSR sites was identified within the B73 reference genome (motif repeat numbers: 1-10, 2-6, 3-5, 4-4, 5-3, and 6-3), of which a minimal number were SSR sites with a motif length of 4. It was found that a total of 167,590 SSR sites existed in the form of motifs in the sites corresponding to Omni InDels, which accounted for 62.76% of all SSR sites, when the location and motif type of the SSR sites in SOIs and MOIs (but not LOIs and HOIs) were compared. It was found that some of these Omni InDel sites with two or more alleles exhibited tandem repeat motifs with different numbers of repetitions; they presented as double polymorphisms of InDel and SSR sites, which we propose could be used as Omni InDel sites with different allele types (Figures 1D, E). Therefore, sites with both SSR and InDel characteristics were referred to as SSR-Omni InDel sites.

Traditional mutation mining based on next-generation sequencing technology can only obtain small InDels, whereas pan-genomic analysis has been recently become popular, in which large structural variants, such as presence/absence variations (large and huge InDels) can be obtained by collinearity between genomes. For LOI and HOI sites with the lengths of ≥50 bp, collinear comparison was performed on 11 portions of maize genomes assembled with third-generation sequencing data and the B73 reference genome and a total of 92,196 LOI sites and 81,898 HOI sites in the 11 maize inbred lines were obtained (Figures 2B, C). Most of these sites were evenly distributed on chromosomes, although some sites were concentrated in the chromosome arms, and a few were distributed in the centromere region (Figure 2D). The greatest number of sites was present in the EP1 variety (LOIs and HOIs: 20,387), and the smallest number of sites was present in the RP125 variety (9,103). Insertions and deletions accounted for 49.26% and 50.74% of InDels, respectively (Figure 2A). Among the 11 maize inbred lines, the proportion of deletions to insertions was between 0.91 and 1.6, with P1566673 accounting for the largest proportion (deletions: 8,819/insertions: 5,479).

Omni InDels are variants that exist in a variety of species, but their markers have not been completely developed. With the continuous improvement of third-generation sequencing technology and bioinformatics tools, complete Omni InDel markers for a variety of organisms have been developed to meet the needs of complex analyses in plant sciences, crop breeding, forensic science, human genetics, evolutionary science, and animal breeding-related fields. To fully elucidate the role of Omni InDels in different crops and enrich the site pool of Omni InDels, InDels from rice (Yan et al., 2020; Shang et al., 2022) and soybean (Liu et al., 2020) were collected in addition to maize and built an Omni InDel database for each species. (Http: https://omni-indel.plantdna.site)

Development and validation of perfect InDels and germplasm analysis

Mutation mining is very common in crop research, but developing mutation sites into usable molecular markers requires mining and experimental verification. To establish the reliability and applicability of our developed Omni InDel database, especially with respect to molecular marker genotyping and further analysis, we proceeded to filter the MOIs and constructed a Perfect InDel set composed of single-copy, mainly dimorphic, easily genotyped InDels that are characterized by a high MAF (≥0.05) and a mutation length of 3–10 bp. Thus, the Perfect InDels were compatible with multiple detection platforms, such as polyacryamide gel electrophoresis (PAGE), fluorescent capillary

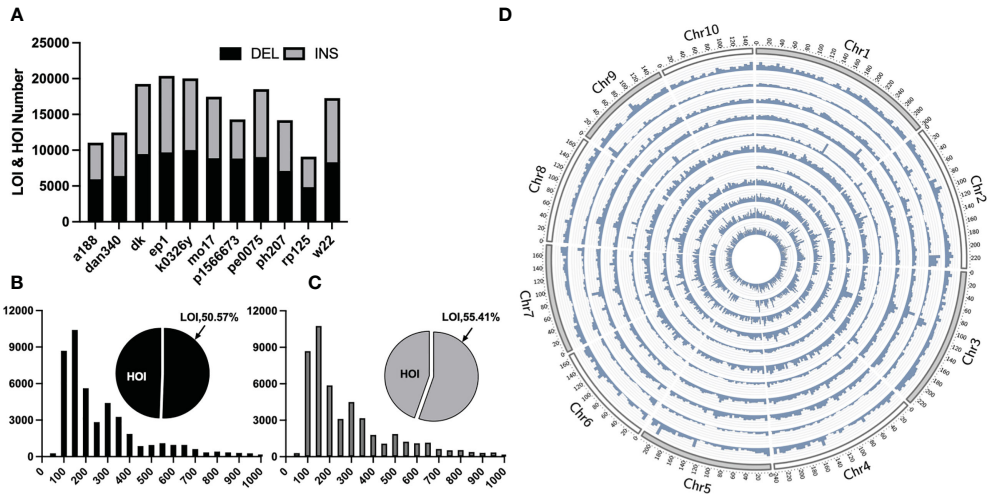

FIGURE 2
Statistical overview of large Omni InDels (LOIs) and huge Omni InDels (HOIs). **(A)** Numbers of LOIs and HOIs of 11 maize inbred lines and the gray and black blocks represent the proportions of DEL and INS respectively within LOIs and HOIs. **(B)** Proportion and distribution of different lengths of the deletion-type LOIs. **(C)** Proportion and distribution of different length of the insertion-type LOIs. **(D)** Density distribution of LOIs and HOIs in 11 maize inbred lines; from the outer ring to the inner ring: A188, Dan340, DK, EP1, K0326Y, Mo17, P1566673, PE0075, PH207, RP125, W22, the window size for density calculation is 5 million bp.

electrophoresis (CE), Kompetitive Allele-Specific PCR (KASP) and sequencing. We first extracted the upstream and downstream 30-bp sequences of the sites with the above characteristics, performed BLAST comparison in the genome, and screened out 4,246,656 sites with at least one flank being a single copy upstream or downstream; then, the sequences were filtered under the condition of MAF ≥0.05. Finally, 1,212,345 sites were obtained, and these sites formed the set of Perfect InDels.

To validate the maize Perfect InDel sites, 4,538 InDels that covered 10 maize chromosomes in the Perfect InDel set were arbitrarily selected. The selected InDels varied in length from 3 bp to 10 bp; 3,558 maize inbred lines at home and abroad were collected and genotyped using the InDels selected above using the chip platform. The results demonstrated that the selected Omni InDel sites facilitated satisfactory genotyping across all maize inbred lines, and the average number of missing sites per inbred line was 87, the smallest number of missing sites of inbred line samples was only 7, and the average missing rate of samples was 1.93%.

In order to classify the germplasm groups among these 3,558 maize inbred lines, phylogenetic tree was built (Figure 3A) for these lines based on the genetic distance between lines constructed with the selected InDels mentioned above using maximum likelihood. Based on the genetic distance between the samples, the 3,558 maize inbred lines were divided into 11 groups. Of these 11 groups, the Reid group had the most inbred lines (676), followed by improved Reid (527), and the total proportion of the two groups reached 33.7%. Among them, some maize inbred lines in the HZS-improved line group had similar genetic distances, which may be related to the frequent improvement of some elite inbred lines in the HZS-improved line group by breeders, resulting in more HZS-improved line group with very similar genetic backgrounds.

Then, PCA was performed to test the reliability of the phylogenetic results. According to the PCA results, PC1, PC2, and PC3 represented 21.84%, 14.11%, and 12.66% of the total variation, respectively (Figures 3B–D). There was a long genetic distance between the HZS-improved line group, Reid group, and X group. The genetic distance between the Reid group and the improved Reid group was small because the improved Reid group was developed from the Reid group.

Development of multi-InDels with Omni InDels

The concept of Multi-InDels was first introduced in the field of human research by Qu et al., 2020. Based on this concept, to expand the application of Omni InDels in crops, a set of Multi-InDels was developed that can be applied in crop research based on Omni InDels that met the following conditions: (1) mutation length between 3–10 bp, namely select sites in the MOI set; the missing rate is <50% and the MAF is >0.1; (2) each Multi-InDel contains 3 or more MOI sites; (3) the combination of variation length of InDels in the Multi-InDel namely the length of each possible theoretical haplotype of Multi-InDel should be different to facilitate detection of each possible allele on the detection platform; (4) each site in the Multi-InDel should be dimorphic; (5) to ensure that PCR can be conducted and that the sites are compatible with fluorescent capillary electrophoresis platforms as well as sequencing platforms, the maximum length of each Multi-InDel is within 200 bp; (6) the distance between two adjacent Multi-InDel sites is more than 200 bp.

Given these six core parameters, Multi-InDels offer the advantages of high polymorphism and can be detected using both sequencing and fluorescence platforms, thus presenting significant practical value.

A total of 441,572 candidate sites, each with a MAF of at least 0.1, were pinpointed when we examined the maize MOI locus database. Next, using the sliding window approach, 22,774

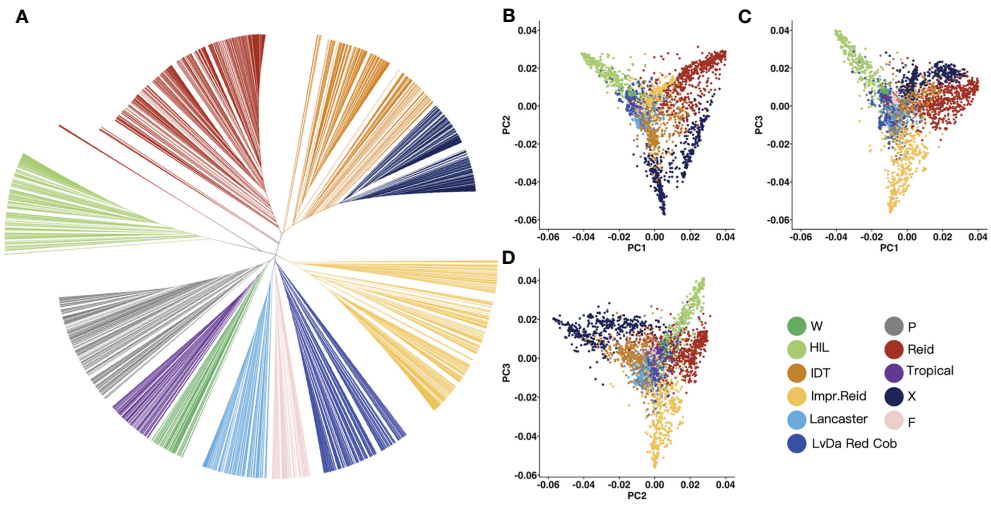

FIGURE 3
Phylogenetic tree and PCA result for 3,558 maize inbred lines. **(A)** UPGMA tree constructed based on the genetic distance matrix between maize inbred lines, 3,558 samples were genetically clustered into the following 11 groups: tropical, flint inbred lines (F), waxy (W), LvDa Red Cob, HZS-improved line (HIL), Improved Reid (Impr.Reid), Reid, Lancaster, P(P78599 line), Iodent (IDT), and X. **(B–D)** Scatter plots of the first principal component (PC1) versus PC2 and PC1 versus PC3 and PC2 versus PC3.

candidate Multi-InDels were selected, which included a total of 81,976 MOIs. The maximum distance between MOIs within the candidate Multi-InDels did not exceed 200 bp.

However, some haplotypes of these candidates had the same PCR length, which made genotyping impossible. The length of haploid genotypes that may appear in the candidate Multi-InDels was filtered to address this. Finally, 6,432 sites were obtained with different lengths for all theoretical haplotypes. 20 sites with two adjacent Multi-InDels spaced within 200 bp of each other was excluded, and eventually revealed 6,312 Multi-InDel sites (Figure 4A). These sites were distributed across chromosomes, and they contained up to 5 MOIs at most which accounting for the proportion 0.2% and 3 MOIs at least which accounting for 90.6% (Figure 4B).

Establishment of an Omni InDel network database

To enhance the accessibility of Omni InDel resources, we developed a web-based database (https://omni-indel.plantdna.site) that contains all Omni InDel information for maize, rice, and soybean (Figure 5). Researchers can effortlessly access, browse, and download information on variable loci of interest from the Omni InDel database.

The Omni InDel online platform provides two core functional modules: "Browse" and "Variant viewer." Among these, the "Browse" module, which utilizes MySQL for database storage, provides key information including "location details," "Omni classification," and "mutation specifics" of all types of Omni InDels; this facilitates seamless switching between maize, rice, and soybean. Additionally, all the data can be downloaded. The "Variant viewer" function is a variation display of newly developed online interactive genome structure. The HTML5 Canvas function was used to draw the distribution map of variants on chromosomes and added interactive functions such as scrolling and clicking. By clicking on different blocks, users can shift the visualization area, resulting in a clear representation of chromosomal variations. As a browser feature, Canvas allows for superior customization and browser rendering efficiency in genome visualization compared to other tools.

Using maize as an example, the web database also facilitates variation visualization as a part of its browsing functionality, which allows researchers to browse the whole genome density distribution of variation at different scales and easily choose a single variation from the millions recorded. Four different types of Omni InDel are displayed in the window and become interactive points. After the users click, detailed variation information of each site is displayed.

Discussion

This study presents the Omni InDel database for multiple crops (maize, soybean, and rice). It offers a rich reservoir of insertion and deletion molecular variation information to support research related to molecular-assisted breeding, population structure genetic analysis, and differentiation of breeding varieties. The Omni InDel database can serve as a critical resource for researchers working on the breeding of new crop cultivars. This database contains InDels for which the length varies from 1 bp to several Mbp, which could allow researchers to flexibly choose the molecular variation that meets their needs. Moreover, InDel has an incomparable advantage over SNP in terms of the length of the variation type. Firstly, in the detection of molecular markers, SNP is mainly detected by methods such as sequencing, KASP, etc., while InDel can be detected by simple gel electrophoresis. Secondly, InDel may cause more severe mutations than SNP, because some relatively large InDels may even cause chromosomal structural changes.

More than 60% of SSR loci were found in the Omni InDel database. These are dual-characteristic variable loci, and some of them showed polymorphism in the database. Hence, when mining SSR molecular markers, it's crucial to filter the assembled genome by motif and employ the locus information in the Omni InDel

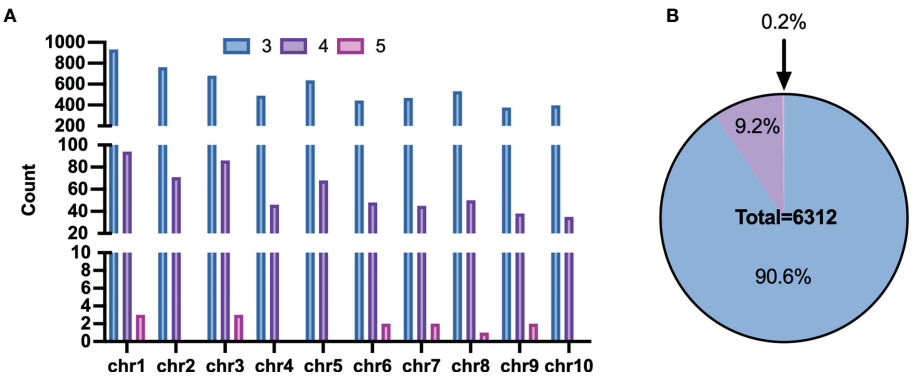

FIGURE 4
Distribution of Multi-InDels. **(A)** Count of Multi-InDels containing 3-5 InDels in each chromosome. **(B)** Proportion of Multi InDels containing 3–5 InDels.

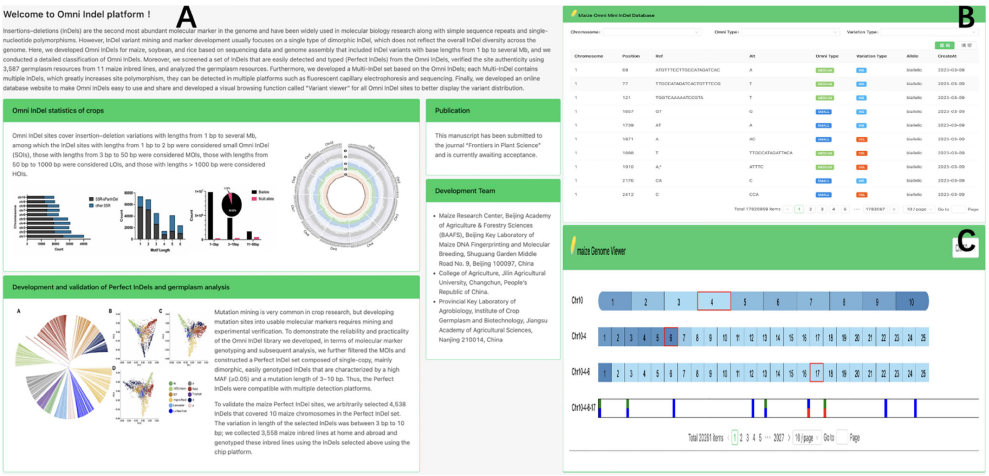

FIGURE 5
Display Omni InDel network database. **(A)** The homepage of Omni InDel website **(B)** Data browser of Omni InDel. **(C)** Screenshot of the"Variant Viewer".

database for dual verification of the mined SSR loci. This process aids in the identification of dual-characteristic variable loci. This can effectively reveal many reliable SSR loci, and we can also eliminate InDel loci with SSR characteristics when necessary.

To extend the application scope of Omni InDels, we developed Multi-InDels based on Omni InDels. The polymorphism of Multi-InDels that contained 3–5 InDels within 200 bp greatly increased compared with that of single InDels, with 8 theoretical haplotypes for Multi-InDels containing 3 InDels and 32 theoretical haplotypes for Multi-InDel containing 5 InDels. The polymorphism of Multi-InDels is not less than that of SSRs. Moreover, because SSR-InDels were removed from Multi-InDels, the chance of stutter in the amplification process of this locus was greatly reduced. Moreover, DNA fingerprinting reveals that SSRs exhibit polymorphism through the insertion or deletion of several base motifs. The size difference in the amplification fragment between different SSR haplotypes corresponds to a multiple of the number of consecutive motifs, thereby adding complexity to the fingerprinting process. All possible amplification lengths of Multi-InDels are known and the amplification products will not show stutter, which makes fingerprinting simpler and clearer. This is the first Multi-InDel database on crops since the release of Multi-InDel database in human research in 2020. This database should be verified and applied in crop-related research.

In addition to the development of this Omni InDel database for multiple crops using maize as an example, we verified some MOIs that were arbitrarily chosen from the whole database to genotype 3,558 domestically and foreign-collected maize inbred lines on an array-based platform. Then, we carried out germplasm analysis using these genotype fingerprints, and the results supported Omni InDels as an effective resource for germplasm analysis. Furthermore, those markers developed from the database (such as Perfect InDels) can also be compatible with multiple detection platforms include array based and KASP which are high throughput platform. These characteristics make them high throughput, high compatible marker (HCM).

A web-based Omni InDel database was developed to facilitate the sharing and dissemination of Omni InDel data, making it readily accessible for browsing on this website.

Data availability statement

The datasets presented in this study can be found in onlinerepositories. The names of the repository/repositories and accessionnumber(s) can be found below: NCBI BioProject accessionnumber: PRJNA976755.

Author contributions

FW and HZ designed the experiments and managed the project. JZ managed the project. ZL performed the data analysis and wrote the manuscript. YKZ and WY reviewed and edited the manuscript. YLZ performed the data curation and validation. LX and LZ provided the resource data. All authors contributed to the article and approved the submitted version.

Funding

This research supposed by Beijing Scholars Program (BSP041), Beijing Academy of Agricultural and Forestry Sciences (Grant Nos. KJCX20230301).

Conflict of interest

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.

Publisher's note

References

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., et al. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40 (Web Server issue), W580–W584. doi: 10.1093/nar/gks498

Bennett, E. P., Petersen, B. L., Johansen, I. E., Niu, Y., Yang, Z., Chamberlain, C. A., et al. (2020). INDEL detection, the 'Achilles heel' of precise genome editing: a survey of methods for accurate profiling of gene editing induced indels. *Nucleic Acids Res.* 48 (21), 11958–11981. doi: 10.1093/nar/gkaa975

Carter, T. E., Nelson, R., Sneller, C. H., and Cui, Z. (2004). Genetic diversity in soybean. in *Soybeans: Improvement, Production, and Uses* Eds. R. M. Shibles, J. E. Harper, R. F. Wilson and R. C. Shoemaker. doi: 10.2134/agronmonogr16.3ed.c8

Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Mol. Biol. Evol.* 26 (2), 473–480. doi: 10.1093/molbev/msn275

Cui, J., Peng, J., Cheng, J., and Hu, K. (2021). Development and validation of genome-wide InDel markers with high levels of polymorphism in bitter gourd (Momordica charantia). *BMC Genomics* 22 (1), 190. doi: 10.1186/s12864-021-07499-0

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10 (2), giab008. doi: 10.1093/gigascience/giab008

Das, S., Upadhyaya, H. D., Srivastava, R., Bajaj, D., Gowda, C. L. L., Sharma, S., et al. (2015). Genome-wide insertion-deletion (InDel) marker discovery and genotyping for genomics-assisted breeding applications in chickpea. *DNA Res.* 22 (5), 377–386. doi: 10.1093/dnares/dsv020

Fei, Q., Ding, H., and Yan, X. (2021). Development of InDel markers and establishment of a specific molecular marker of the new strain (SW-81) in pyropia haitanensis. *Algal Res.* 60), 102480. doi: 10.1016/j.algal.2021.102480

Feng, J. J., Zhu, H. Y., Zhang, M., Zhang, X. X., Guo, L. P., Qi, T. X., et al. (2020). Development and utilization of an InDel marker linked to the fertility restorer genes of CMS-D8 and CMS-D2 in cotton. *Mol. Biol. Rep.* 47 (2), 1275–1282. doi: 10.1007/s11033-019-05240-5

Goel, M., Sun, H., Jiao, W., and Schneeberger, K.. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20, 277. doi: 10.1186/s13059-019-1911-0

Hechanova, S. L., Bhattarai, K., Simon, E. V., Clave, G., Karunarathne, P., Ahn, E.-K., et al. (2021). Development of a genome-wide InDel marker set for allele discrimination between rice (Oryza sativa) and the other seven AA-genome oryza species. *Sci. Rep.* 11, 8962. doi: 10.1038/s41598-021-88533-9

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell.* 26 (1), 121–135. doi: 10.1105/tpc.113.119982

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., et al. (2021). *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373 (6555), 655–662. doi: 10.1126/science.abg5289

Jain, A., Roorkiwal, M., Kale, S., Garg, V., Yadala, R., and Varshney, R. K. (2019). InDel markers: an extended marker resource for molecular breeding in chickpea. *PLoS One* 14 (3), e0213999. doi: 10.1371/journal.pone.0213999

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an r package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281. doi: 10.7717/peerj.281

Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (arXiv:1303.3997). *arXiv*. doi: 10.48550/arXiv.1303.3997

Li, W., Cheng, J., Wu, Z., Qin, C., Tan, S., Tang, X., et al. (2015). An InDel-based linkage map of hot pepper (Capsicum annuum). *Mol. Breed* 35, 32. doi: 10.1007/s11032-015-0219-3

Li, W., Liu, D., Tang, S., Li, D., Han, R., Tian, Y., et al. (2019). A multiallelic indel in the promoter region of the cyclin-dependent kinase inhibitor 3 gene is significantly associated with body weight and carcass traits in chickens. *Poult Sci.* 98 (2), 556–565. doi: 10.3382/ps/pey404

Li, M. W., Wang, Z., Jiang, B., Kaga, A., Wong, F. L., Zhang, G., et al. (2020). Impacts of genomic research on soybean improvement in East Asia. *Theor. Appl. Genet.* 133, 1655–1678. doi: 10.1007/s00122-019-03462-6

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32 (10), 1045–1052. doi: 10.1038/nbt.2979

Lin, M. X., Whitmire, S., Chen, J., Farrel, A., Shi, X. H., Guo, J. T., et al. (2017). Efects of short indels on protein structure and function in human genomes. *Sci. Rep.* 7, 9313. doi: 10.1038/s41598-017-09287-x

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182 (1), 162–176.e13. doi: 10.1016/j.cell.2020.05.023

Liu, X., Liu, Y., Yang, S., Wang, J., Lu, X., Wei, X., et al. (2022). Development and characterization of 29 InDel markers from the mangrove kandelia obovata genome using a resequencing dataset. *Conserv. Genet. Resour* 14, 263–266. doi: 10.1007/s12686-022-01272-5

Liu, B., Wang, Y., Zhai, W., Deng, J., Wang, H., Cui, Y., et al. (2013). Development of InDel markers for brassica rapa based on whole-genome resequencing. *Theor. Appl. Genet.* 126 (1), 231–239. doi: 10.1007/s00122-012-1976-6

Loewenthal, G., Rapoport, D., Avram, O., Moshe, A., Wygoda, E., Itzkovitch, A., et al. (2021). A probabilistic model for indel evolution: differentiating insertions from deletions. *Mol. Biol. Evol.* 38 (12), 5769–5781. doi: 10.1093/molbev/msab266

Lovett, S. T. (2004). Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* . 52, 1243–1253. doi: 10.1111/j.1365-2958.2004.04076.x

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14 (1), e1005944. doi: 10.1371/journal.pcbi.1005944

Matar, S., and Melzer, S. (2021). A 598-bp InDel variation in the promoter region of Bna.SOC1.A05 is predominantly present in winter type rapeseeds. *Front. Plant Science.* 12. doi: 10.3389/fpls.2021.640163

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Moghaddam, S. M., Song, Q., Mamidi, S., Schmutz, J., Lee, R., Cregan, P., et al. (2014). Developing market class specific InDel markers from next generation sequence data in phaseolus vulgaris l. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00185

Păcurar, D. I., Păcurar, M. L., Street, N., Bussell, J. D., Pop, T. I., Gutierrez, L., et al. (2012). A collection of INDEL markers for map-based cloning in seven arabidopsis accessions. *J. Exp. Bot.* 63, 2491–2501. doi: 10.1093/jxb/err422

Pan, G., Li, Z., Huang, S., Tao, J., Shi, Y., Chen, A., et al. (2021). Genome-wide development of insertion-deletion (InDel) markers for cannabis and its uses in genetic structure analysis of Chinese germplasm and sex-linked marker identification. *BMC Genomics* 22 (1), 595. doi: 10.1186/s12864-021-07883-w

Pan, Z., Li, Z., Zhang, J., Zhao, W., and Tong, C. (2022). Investigation of genome-wide InDel distribution and segregation in populus with restriction-site associated DNA sequencing data. *Trop. Plant Biol.* 15, 171–180. doi: 10.1007/s12042-022-09312-y

Paudel, L., Clevenger, J., and McGregor, C. (2019). Refining of the egusi locus in watermelon using KASP assays. *Sci. Hortic.* 257, 108665. doi: 10.1016/j.scienta.2019.108665

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81 (3), 559–75. doi: 10.1086/519795

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184 (13), 3542–3558.e16. doi: 10.1016/j.cell.2021.04.046

Qu, S., Lv, M., Xue, J., Zhu, J., Wang, L., Jian, H., et al. (2020). Multi-indel: a microhaplotype marker can be typed using capillary electrophoresis platforms. *Front. Genet.* 11. doi: 10.3389/fgene.2020.567082

Ramakrishna, G., Kaur, P., Nigam, D., Chaduvula, P. K., Yadav, S., Talukdar, A., et al. (2018). Genome-wide identifcation and characterization of InDels and SNPs in glycine max and glycine soja for contrasting seed permeability traits. *BMC Plant Biol.* 18, 141. doi: 10.1186/s12870-018-1341-2

Rockah-Shmuel, L., Toth-Petroczy, A., Sela, A., Wurtzel, O., Sorek, R., and Tawfik, D. S. (2013). Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. *PLoS Genet.* 9, e1003882. doi: 10.1371/journal.pgen.1003882

Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., et al. (2014). Whole genome *de novo* assemblies of three divergent strains of rice, oryza sativa, document novel gene space of aus and indica. *Genome Biol.* 15, 506. doi: 10.1186/s13059-014-0506-z

Schliep, K. (2011). "phangorn: phylogenetic analysis in R". *Bioinformatics* 27 (4), 592–593. doi: 10.1093/bioinformatics/btq706

Seki, K. (2022). Detection of candidate gene LsACOS5 and development of InDel marker for male sterility by ddRAD-seq and resequencing analysis in lettuce. *Sci. Rep.* 12, 7370. doi: 10.1038/s41598-022-11244-2

Seo, J., Dhungana, S., Kang, B., Baek, I., Sung, J., Ko, J., et al. (2022). Development and validation of SNP and InDel markers for pod-shattering tolerance in soybean. *Int. J. Mol. Sci.* 23 (4), 2382. doi: 10.3390/ijms23042382

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., et al. (2022). A super pan-genomic landscape of rice. *Cell Res.* 32 (10), 878–896. doi: 10.1038/s41422-022-00685-z

Song, X., Wei, H., Cheng, W., Yang, S., Zhao, Y., Li, X., et al. (2015). Development of INDEL markers for genetic mapping based on whole genome resequencing in soybean. *G3* 5, 2793–279912. doi: 10.1534/g3.115.022780

Sun, K., Yun, L., Zhang, C., Shao, C., Gao, T., Zhao, Z., et al. (2019). Evaluation of 12 multi-InDel markers for forensic ancestry prediction in Asian populations. *Forensic Sci. International: Genet.* 43, 102155. doi: 10.1016/j.fsigen.2019.102155

Terakami, S., Matsumura, Y., Kurita, K., Kanamori, H., Katayose, Y., Yamamoto, T., et al. (2012). Complete sequence of the chloroplast genome from pear (Pyrus pyrifolia): genome structure and comparative analysis. *Tree Genet. Genomes* 8, 841–854. doi: 10.1007/s11295-012-0469-8

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455 (7209), 105–108. doi: 10.1038/nature07175

Vialle, R. A., Tamuri, A. U., Goldman, N., and Thorne, J. (2018). Alignment modulates ancestral sequence reconstruction accuracy. *Mol. Biol. Evol.* 35 (7), 1783–1797. doi: 10.1093/molbev/msy055

Wang, S., Li, Z., Guo, X., Fang, Y., Xiang, J., and Jin, W. (2018). Comparative analysis of microsatellite, SNP, and InDel markers in four rhododendron species based on RNA-seq. *Breed. Sci.* 68, 536–544. doi: 10.1270/jsbbs.18092

Wang, D., Yang, T., Liu, R., Li, N., Wang, X., Sarker, A., et al. (2020). RNA-Seq analysis and development of SSR and KASP markers in lentil (Lens culinaris medikus subsp. culinaris). *Crop J.* 8, 953–965. doi: 10.1016/j.cj.2020.04.007

Wing, R. A., Purugganan, M. D., and Zhang, Q. (2018). The rice genome revolution: from an ancient grain to green super rice. *Nat. Rev. Genet.* 19, 505–517. doi: 10.1038/s41576-018-0024-z

Wu, D. H., Wu, H. P., Wang, C. S., Tseng, H. Y., and Hwu, K. K. (2013). Genome-wide InDel marker system for application in rice breeding and mapping studies. *Euphytica* 192, 131–143. doi: 10.1007/s10681-013-0925-z

Yamaki, S., Ohyanagi, H., Yamasaki, M., Eiguchi, M., Miyabayashi, T., Kubo, T., et al. (2013). Development of INDEL markers to discriminate all genome types rapidly in the genus oryza. *Breed Sci.* 63 (3), 246–254. doi: 10.1270/jsbbs.63.246

Yan, J., Zou, D., Li, C., Zhang, Z., Song, S., and Wang, X. (2020). SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genomics Proteomics Bioinf.* 18 (2), 173–185. doi: 10.1016/j.gpb.2020.03.002

Yang, J., Wang, Y., Shen, H., and Yang, W. (2014). In silico identification and experimental validation of insertion-deletion polymorphisms in tomato genome. *DNA Res.* 21, 429–438. doi: 10.1093/dnares/dsu008

Zhao, S., Li, A., Li, C., Xia, H., Zhao, C., Zhang, Y., et al. (2017). Development and application of KASP marker for HighnThroughput detection of AhFAD2 mutation in peanut. *Electron. J. Biotechnol.* 25, 9–12. doi: 10.1016/j.ejbt.2016.10.010

Zhou, G., Zhang, Q., Tan, C., Zhang, X., and Li, C. (2015). Development of genome-wide InDel markers and their integration with SSR, DArT and SNP markers in single barley map. *BMC Genomics* 16, 804. doi: 10.1186/s12864-015-2027-x