frontiers in Plant Science

METHODS
published: 10 July 2018
doi: 10.3389/fpls.2018.00976

FNBtools: A Software to Identify Homozygous Lesions in Deletion Mutant Populations

Liang Sun, Yinbing Ge, Andrew Charles Bancroft, Xiaofei Cheng and Jiangqi Wen**

Noble Research Institute, Ardmore, OK, United States

Deletion mutagenesis such as fast neutron bombardment (FNB) has been widely used for forward and reverse genetics studies in functional genomics. Traditionally, the time-consuming map-based cloning is used to locate causal deletions in deletion mutants. In recent years, comparative genomic hybridization (CGH) has been used to speed up and scale up the lesion identification process in deletion mutants. However, limitations of low accuracy and sensitivity for small deletions in the CGH approach are apparent. With the next generation sequencing (NGS) becoming affordable for most users, NGS-based bioinformatics tools are more appealing. Although several deletion callers are available, these tools are not efficient in detecting small deletions. Population-scale deletion callers that can identify both small and large deletions are rare. We were motivated to create a population-scale deletion detection tool, called FNBtools, to identify homozygous causal deletions in mutant populations by using NGS data. FNBtools is a tool to call deletions at a population-scale and to achieve high accuracy at different levels of coverage. In addition, FNBtools can detect both small and large deletions with the ability to identify unique deletions in a mutant pool by filtering deletions that exist in a wild-type or control pool. Furthermore, FNBtools is also able to visualize all identified deletions in a genome-wide scope by using Circos. From simulated data analysis, FNBtools outperforms four existing popular deletion callers in detecting small deletions at different coverage levels. To test the usefulness of FNBtools in real biological applications, we used it to analyze a salt-tolerant mutant in *Medicago truncatula* and identified the unique deletion locus that is tightly linked with this trait. The causal deletion in the mutant was confirmed by PCR amplification, sequencing and genetic linkage analyses. FNBtools can be used for homozygous deletion identification in any species with reference genome sequences. FNBtools is publicly available at: https://github.com/noble-research-institute/fnbtools.

Keywords: deletion mutant, homozygous deletion, next generation sequencing, FNBtools, salt tolerance

OPEN ACCESS

Edited by:
Alfredo Pulvirenti, Università degli Studi di Catania, Italy

Reviewed by:
Gaurav Sablok, University of Helsinki, Finland
Prashanth N. Suravajhala, Birla Institute of Scientific Research, India

****Correspondence:***
Liang Sun
lsun@noble.org
Jiangqi Wen
jwen@noble.org

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: *01 May 2018*
Accepted: *15 June 2018*
Published: *10 July 2018*

Citation:
Sun L, Ge Y, Bancroft AC, Cheng X and Wen J (2018) FNBtools: A Software to Identify Homozygous Lesions in Deletion Mutant Populations. Front. Plant Sci. 9:976. doi: 10.3389/fpls.2018.00976

INTRODUCTION

Induced mutagenesis is a powerful approach in plant breeding and has been widely used for functional genomics. Commonly used mutagens for induced mutagenesis include physical (e.g., fast neutron bombardment [FNB]), chemical (e.g., ethyl methane sulfonate [EMS]) and biological (e.g., T-DNA and transposons) mutagens. For instance, T-DNA has been successfully used to generate large-scale mutant populations in the model plant species *Arabidopsis thaliana* and

Oryza sativa (Alonso et al., 2003; Sha et al., 2004). In legumes, *Tnt1* and *LORE1* insertion mutagenesis has also been successfully used to tag the *Medicago truncatula* and *Lotus japonicus* genomes, respectively (d'Erfurth et al., 2003; Malolepszy et al., 2016). Though T-DNA and transposon-based insertion mutagenesis have been widely used for reverse genetics because of their feasibility and convenience in identifying mutated genes (Alonso et al., 2003; Tadege et al., 2008), insertion mutagenesis only creates random mutations in single genes. In many genomes, tandemly repeated genes account for a considerable portion of the genome. These genes are intractable using insertion mutagenesis because their proximity on the chromosomes hinders the creation of higher order mutants. However, deletion mutagenesis achieved by irradiation or FNB can delete adjacent genes (Rogers et al., 2009). Compared to other methods, FNB mutagenesis is easy and effective, and does not require genetic transformation or tissue culture that are typical for T-DNA or *Tnt1* insertion mutagenesis. FNB mutants are non-transgenic and can be grown in fields without regulation.

Fast neutrons are high energy particles that induce a broad range of deletions (from a single base pair to thousands or even millions of base pairs) and other structural variations (SVs) (including combinations of inversions, deletions, substitutions and rearrangements) in cells (Rogers et al., 2009). For example, a deletion of ∼35 kb contains the causal gene *DNF4* that plays an essential role in nitrogen-fixing symbiosis in *M. truncatula* (Kim et al., 2015). FNB can also delete small genes in a genome that are relatively hard to tag by insertion mutagenesis (Rogers et al., 2009). In addition, FNB can be used to generate mutant populations with more complete genome coverage than is possible using other approaches (Li et al., 2001). Though comparative genomic hybridization (CGH) and De-TILLING methods identified deletions in FNB mutants (Carter, 2007; Rogers et al., 2009), both methods have low resolution and low accuracy, especially for small deletions. For these reasons, reverse genetic platforms for FNB have not been exploited extensively because of the complexity of mutations generated by FNB. Map-based cloning is a traditional forward genetics methodology to identify causal deletions in FNB mutants. However, this method is time-consuming and expensive. With the rapid development of next generation sequencing (NGS) technologies, whole genome sequencing has become more affordable, providing a new means of identifying causal deletions in FNB mutants.

To identify the causal mutation for a specific FNB mutant phenotype, researchers usually backcross the mutant line to wild-type to obtain a segregating F2 population. If the segregation ratio between wild-type and mutant plants is close to 3:1, this mutant phenotype is likely to be caused by a single recessive mutation. Sequencing pooled mutant DNA and wild-type DNA samples separately is a good strategy for identifying homozygous deletions, which are likely causal deletions.

To call large deletions, mapping-based deletion callers along with paired-end reads are the most commonly used techniques (O'Rourke et al., 2013). The sequence alignment map (SAM) files are successfully used to call small insertions and deletions in bioinformatics tools such as FreeBayes (Garrison and Marth, 2012) and Samtools (Li et al., 2009). Three types of signals from mapping files (e.g., SAM files) can also be used to capture informative reads for large deletions. These signals include: (1) Soft-clipped reads, which occur when one partial fragment of a single read is perfectly mapped to one genomic region and the other partial fragment is perfectly mapped to another nearby genomic region. Pindel (Chiang et al., 2009), Delly (Rausch et al., 2012), and Sprites (Zhang et al., 2016) are all examples of such tools that are currently available to call deletions with soft-clipped read information. (2) Discordant reads, which occur when one read of a pair is mapped to one genomic region and the other is mapped to a different nearby genomic region. Paired-end reads with discordant distance are captured and considered as insertions or deletions. Many tools adopt this signal to call deletions, for example, BreakDancer (Chen et al., 2009), VariationHunter (Hormozdiari et al., 2009; Hormozdiari et al., 2010), and PEMer (Korbel et al., 2009). (3) Read depth-based method, where high-coverage reads are mapped to the up/downstream regions of the deletion site, but fewer or no reads are mapped to the deletion region. Examples of read depth-based methods include SegSeq (Chiang et al., 2009), EWT (Yoon et al., 2009), and CNVnator (Abyzov et al., 2011). However, neither discordant read-based method nor read depth-based method is able to predict the exact positions of breakpoints.

Challenges of NGS Data in Deletion Analysis

Although several tools are available for deletion calling (Li et al., 2009; McKenna et al., 2010; Koboldt et al., 2012), challenges still exist. These challenges include, but are not limited to, (1) Small and large deletion detection. Small homozygous deletions can cause frame shift and/or introduce early stop codon, leading to disruption of gene function; whereas large deletions cover multiple genes, enabling functional characterization of tandem duplicated genes (Rogers et al., 2009). Despite several calling tools, such as Samtools (Li et al., 2009), GATK (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013), and VarScan (Koboldt et al., 2009, 2012), have been created in the past years. These tools are primarily designed for small deletion callings in human and some animal systems, where naturally occurring small deletions are common but induced large deletions are rare. Therefore, these tools cannot identify homozygous small and large deletions with a high accuracy. Additionally, none of these tools is able to filter out deletions in a control population. (2) Complexity of deletion identification. Given the complexity of identifying deletions, comparing deletions from multiple samples to achieve population-scale studies is still a challenge (Guan and Sung, 2016). (3) Annotation and visualization of deletions. Sometimes a long list of unique homozygous deletions is identified in a mutant sample. To narrow down candidate deletions from the list and focus on potential causal deletions, annotation of the listed candidate deletions facilitates the discovery of causal genes. Overall visualization of candidate deletions across the whole genome can also help generate hypotheses.

FNBtools as a Solution

To provide a solution to above-mentioned challenges, we employed BWA (Li and Durbin, 2009) and customized PERL and PYTHON scripts to extract and cluster all informative deletion reads from SAM files and created a homozygous deletion-calling tool, FNBtools. FNBtools aims to combine all three types of signals and effectively identify both small and large homozygous deletions in an FNB population. Not only can FNBtools simultaneously analyze multiple samples and identify homozygous deletions, but it can also filter deletions that exist in wild-type or control samples. Users can easily identify unique deletions in mutant samples that are likely to be causal deletions for phenotypes of interest. To better visualize identified deletions, we employed one of the most popular whole genome visualization tools, Circos (Krzywinski et al., 2009). FNBtools integrates the Circos library to generate high-resolution images that can be used for publication.

A Case Study: Deletion Detection in Salt-Tolerant Mutants in *M. truncatula*

In a joint effort between the Noble Research Institute and the John Innes Centre, an FNB population of 156,000 M2 plants in the *M. truncatula* Jemalong A17 background was generated[1]. In a forward genetic screen for the salt-tolerant phenotype, we isolated two mutants exhibiting enhanced salt tolerance. Backcross and segregation analysis indicated that these mutants are non-allelic, recessive and caused by single locus mutations. To identify the causal locus in each mutant, we sequenced the whole genomes of these mutants using Illumina NextSeq and used FNBtools to test the efficiency of the bioinformatics software.

MATERIALS AND METHODS

Paired-End Mapping

Because paired-end reads are more informative than single-end reads in detecting SV, FNBtools is designed for paired-end reads and is benchmarked by paired-end simulated data. FNBtools aligns all paired-end reads to the reference genome, similar to mapping-based structural variation tools such as Delly, Sprites, and Pindel. To perform alignment, FNBtools uses BWA MEM. BWA MEM produces SAM files, which are used to extract all informative reads.

Informative Reads Extraction

For small deletions, FNBtools simply extracts informative reads with small deletions from the SAM file produced by BWA MEM. Suppose that a read has a CIGAR string '100M5D45M' in the SAM file where 'M' and 'D' represent matching and deletion sequences, respectively. The number before the characters represents the number of base pairs involved. This type of small deletion read (5 bp in this example) is extracted and labeled as 'SMD.'

For large deletions, two signals are used to capture informative reads: discordant reads and soft-clipped reads. Paired-end sequence reads that have discordant distance (i.e., if the inner distance of normal paired-end reads is ∼200 bp, the inner distance of discordant reads might be 1,000 bp) larger than two times the DNA-Seq library fragment length with each pair's reads perfectly mapped to the reference genome are considered discordant reads. These reads are extracted and labeled as 'CRR'. For soft-clipped reads, one read is partially aligned to two different genomic positions nearby on the same chromosome. For example, a soft-clipped read has two CIGAR strings 'mMxS' and 'nSyM' (m and y are the number of base pairs mapped; n and x are the number of soft-clipped base pairs; M indicates matching and S indicates soft-clipping). If the difference between m and n is less than 10 bp, we treat this read as a soft-clipped read, and the read is labeled as 'CLR'. The SAM information of soft-clipped reads is modified to add deletion bars for large deletions and rewritten to a new SAM file. This new SAM file is sorted and indexed to a binary alignment/map (BAM) file by SAMtools (Li et al., 2009). The deletion bars can be visualized in IGV (Robinson et al., 2011; Thorvaldsdottir et al., 2013) by using the new index BAM file.

Informative Reads Clustering

All informative reads are represented by a sextuple (seq, chr, st, end, len, type) where seq, chr, st, end, len, and type indicate read id, chromosome, breakpoint position, deletion end position, deletion length, and deletion type (SMD, CLR, or CRR), respectively. They are clustered based on st, end and len. Different types of deletions use different window sizes. For SMD, CLR and CRR deletions, we use a window size of 5 bp, 50 bp and the library mean (500 bp by default), respectively. Informative reads are sorted in ascending order based on the chr and st, and are then clustered into small clusters. In each small cluster, there are three result sets for SMD, CLR and CRR reads separately. For example, three sets for SMD reads are represented by R_{st_smd}, R_{end_smd}, and R_{len_smd}, and similarly for CLR and CRR reads. Because the total number of each set for the same read type should be the same, we usedn_{smd}, n_{clr}, and n_{clr} to represent the total number of informative reads for SMD, CLR, and CRR deletions. The preliminary breakpoint position, end position and length of deletions are generated based on clustering of informative reads by using **Table 1**.

Gaps in SAM File

If homozygous deletions really exist in a specific genomic region, the read depth of this region should theoretically be zero. This information is very valuable and we used it to filter out heterozygous and false positive deletions. We use BEDTools (Quinlan and Hall, 2010) to capture all gapped regions for each sample by using BAM files (minimal MAPQ 30 to filter out low quality mappings) and compare them with the preliminary deletions from the above clustering step. If the start position of a preliminary deletion is within 3 bp (small deletion) or 20 bp (large deletion with soft clipped reads) or 50 bp (large deletion with only discordant reads) wiggle region of the gap region and the calculative gap length is at least 90% of the preliminary deletion

[1] https://medicago-mutant.noble.org/mutant/index.php

TABLE 1 | Strategy for determining breakpoint for informative read in one cluster.

Read type	Condition	Breakpoint position del_{st}	End position del_{end}	Deletion length del_{len}
SMD,CLR, CRR	$n_{smd} \geq n_{clr}$	$mode(R_{st\_smd})$	$mode(R_{end\_smd})$	$mode(R_{len\_smd})$
SMD,CLR, CRR	$n_{smd} < n_{clr}$	$mode(R_{st\_clr})$	$mode(R_{end\_clr})$	$mode(R_{len\_clr})$
CRR	$n_{smd} = 0 \cap n_{clr} = 0$	$median(R_{st\_crr})$	$median(R_{end\_crr})$	$median(R_{len\_crr})$

n_{smd}, R_{st_smd}, R_{end_smd}, and R_{len_smd} represent the total number, the breakpoint position, the deletion end position, and deletion length of small deletion reads; n_{clr}, R_{st_clr}, R_{end_clr}, and R_{len_clr} represent the total number, the breakpoint position, the deletion end position, and deletion length of soft clipped reads; n_{crr}, R_{st_crr}, R_{end_crr}, and R_{len_crr} represent the total number, the breakpoint position, the deletion end position, and deletion length of discordant reads; del_{st}, del_{end}, and del_{len} represent the breakpoint position, end position, and length of preliminary deletions.

TABLE 2 | Criteria for homozygous deletion identification.

Read type	Condition
Small deletion ($n_{smd}>0$)	$n_{smd} \geq n_{clr} \cap \mid gap_{st} - del_{st} \mid < 3$
Large deletion ($n_{clp} > 0$)	$n_{smd} < n_{clr} \cap \ \lvert gap_{st} - del_{set} \rvert \ < 20 \cap \frac{\sum gap_{len}}{del_{len}} \geq 0.9$
Large deletion ($n_{crr} > 0$, $n_{clp} = 0$ and $n_{smd} = 0$)	$n_{clr} = 0 \cap n_{smd} = 0 \cap \lvert gap_{st} - del_{st} \rvert < 50 \cap \frac{\sum gap_{len}}{del_{len}} \geq 0.9$

n_{smd} and n_{clr} represent the total number of small deletion reads and soft clipped reads respectively. del_{st} and del_{len} represent the breakpoint position and length of preliminary deletions. gap_{st} and $\sum gap_{len}$ represents the start position of gap and the calculative gap length in the deleted region. For large deletion, there is no clipped reads and small deletion reads but only discordant reads, thus the criteria were loosened to allow the difference of gap start position and deletion start position.

length, we report the deletion as a real homozygous deletion. The detailed criteria are listed in **Table 2**.

One potential problem in deletion analyses is sample contamination, which can be encountered when multiple samples are pooled. If a pooled sample is contaminated, the causal deletion may not be a homozygous deletion. FNBtools addresses this issue by providing a function to identify all deletions in the mutant pool and calculate deletion frequencies. Users can select those non-homozygous deletions with high frequencies for confirmation purposes if they believe their mutant pool may be contaminated.

The deletion frequency is determined via the following equation:

$$deletion\,frequency = \frac{N_{clr} + N_{smd} + N_{crr}}{N_{clr} + N_{smd} + N_{crr} + N_{der}} * 100$$

N_{clr} is the number of soft clipped reads spanning deletion region;

N_{smd} is the number of small deletions spanning the deletion region;

N_{crr} is the number of discordant reads spanning deletion region;

N_{der} is the number of reads in the deletion region. Theoretically, $N_{der} = 0$ if deletions are homozygous.

Unique Deletions Identification

In addition to identifying all homozygous deletions in the mutant population, FNBtools also can filter out deletions in wild-type (control) samples. FNBtools provides two options to filter out deletions from the control pool depending on the composition of the control pool. (1) Filter out homozygous deletions commonly existing in the control (wild-type) pool (see **Figure 1** for the pooling strategy A). In this case, the control pool contains both real wild-type individuals and heterozygous individuals from a segregating M2 or F2 progeny; therefore, only homozygous deletions that commonly exist in the mutant pool and the control pool are filtered out. (2) Filter out homozygous and heterozygous deletions commonly existing in the control pool (see **Figure 1** for pooling strategy B). In this case, only real wild-type individuals (no segregation of mutants is observed from selfed M2 or F2 wild-type-like individuals) in M3 or F3 generation are pooled as the control pool. Since the control pool is pure wild-type, there are no heterozygous deletions for the causal locus, both homozygous and heterozygous deletions commonly existing in the control pool and the mutant pool are filtered out. Only deletions that uniquely exist in the mutant population are reported by FNBtools. If users prefer to include all homozygous deletions regardless of uniqueness, FNBtools accepts a parameter to toggle what is reported in this regard. FNBtools currently uses fixed cut-off values for supporting reads $> = 3$ to reliably identify homozygous deletions.

Annotation and Visualization of Deletions

If deletions fall in gene regions (including 5′UTR and 3′UTR), exons and introns are annotated by gene IDs. These deletions and associated gene IDs can be visualized in Circos by FNBtools. In the Circos graph, there are three layers of visualization. The outermost layer shows deletion lengths smaller than 100 bp. The middle layer shows deletions with a length between 100 bp and 1 kb. The innermost layer shows deletion lengths greater than 1 kb.

Plant Materials and Sequencing

Wild-type *M. truncatula* (ecotype Jemalong A17) seeds were mutagenized by fast neutron irradiation at the 35 Gy dosage level. Approximately 8,600 M2 plants derived from 1,720 M1 lines were screened on half-strength Murashige-Skoog (1/2 MS) medium containing 1.3% NaCl and 0.5% Phytogel, resulting in the isolation of two salt-tolerant mutants, S1 and S2. To generate sequencing materials, we separately backcrossed the two mutants to wild-type A17. Resultant BC1F1 plants were self-pollinated and the seeds produced were grown to produce BC1F2 plants. Segregating F2 seeds were scarified with concentrated sulfuric acid for 8 min and

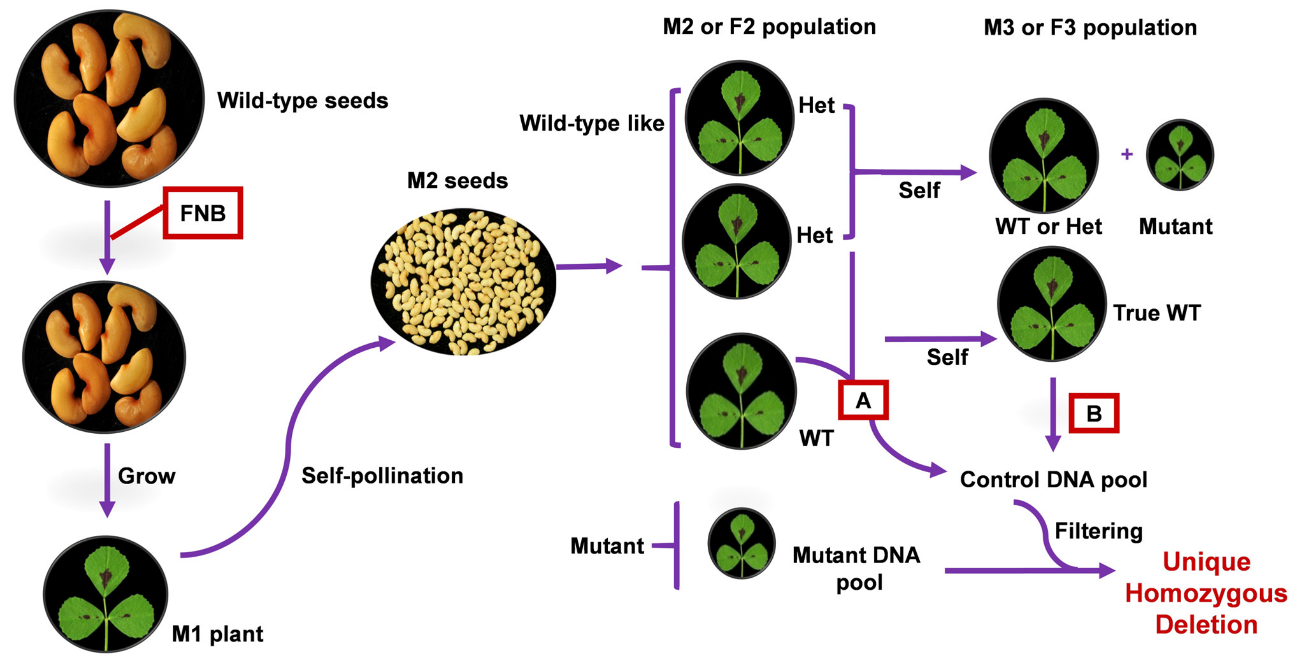

FIGURE 1 | Simplified schema for application of FNBtools to *M. truncatula* FNB population. Once a mutant is isolated from a segregating M2 progeny, it can be directly used as the mutant pool. Alternatively, the mutant can be backcrossed with wild-type to reduce background noise. In this scenario, segregated mutant plant from a F2 progeny can be used as the mutant pool. Pooling strategy A represents the control DNA pool from wild-type and heterozygous individuals in the segregating M2 or F2 progeny. With this control pool, only homozygous deletions commonly existing in the control pool and the mutant pool are filtered out from the mutant pool. Pooling strategy B represents the control DNA pool from true wild-type individuals in the M3 or F3 population. With this control pool, both heterozygous and homozygous deletions commonly existing in the control pool and the mutant pool are filtered out from the mutant pool.

washed thoroughly with tap water. The scarified seeds were further sterilized with 30% bleach for 10 min, followed by extensive rinsing with autoclaved ddH_2O and cold treatment at 4°C for 7 days on solidified 1/2 MS medium. Germinated seeds were first grown on regular 1/2 MS for 7 days in a growth chamber with a regime of 18 h light/25°C and 6 h dark/22°C photoperiod. At least 50 1-week-old F2 seedlings from each mutant were transferred onto solidified 1/2 MS medium containing 1.3% NaCl for 2 weeks. Surviving plants are salt-tolerant and considered mutants, while dead plants are wild-type or heterozygous. Because wild-type and heterozygous plants from the segregating progenies were dead during the salt selection, no materials were left to be the control. In this case, we used the mutant S1, which has the same parental background as mutant S2 but is not allelic to S2, as the control for FNBtools data processing and analysis. One trifoliate leaf from each surviving seedling was collected and frozen in liquid nitrogen for DNA isolation. Genomic DNA from 10 mutant S2 seedlings and 10 control S1 seedlings were isolated individually using the Dellaporta miniprep method (Weigel and Glazebrook, 2009) and pooled as the mutant and the control DNA samples for sequencing. The integrity of each DNA sample was visually examined on a 1% agarose gel. DNA concentration and purity were assessed using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, United States). DNA samples were submitted to the Genomics Core Facility at Noble Research Institute for 150 bp paired-end sequencing on an Illumina NextSeq 500 sequencing system. From the same segregating F2 progeny, 65 S2 mutant plants were sampled for genomic DNA isolation and subsequent genetic linkage analysis.

RESULTS AND DISCUSSION

Experimental Design

Our deletion mutant population design derives from MutMap (Abe et al., 2012). The principle of FNBtools is illustrated using FNB mutants in *M. truncatula* (**Figure 1**) because we have a large FNB mutant collection in this legume species. The approach is designed for mutagenized plants (M0) that have been used to generate M1 mutant lines, which have been selfed to generate an M2 population. Multiple mutant plants identified from a segregating M2 progeny can be directly used as the mutant pool. Alternatively, the mutant can be backcrossed with a wild-type plant to reduce background noise. Several mutant plants are then selected from the F2 progeny and pooled to be used as the mutant pool. This has the advantage of averaging-out non-causal mutations. To better pinpoint the causal homozygous deletions from a long list of identified deletions in a mutant pool (sometime the deletions can number in the thousands), a wild-type or control pool is used. We used two strategies for control sample pooling (**Figure 1**). In pooling strategy A, the control pool

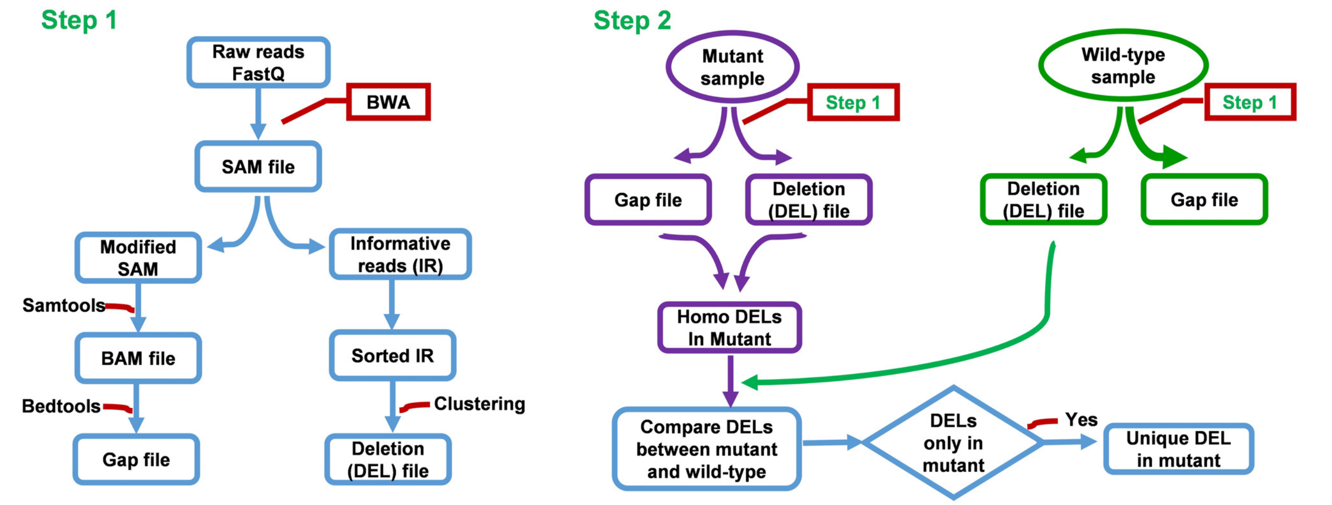

FIGURE 2 | Flow chart of FNBtools. Step 1 aligns NGS reads to the reference genome. Step 2 compares the deletions between the mutant sample and the wild-type or the control sample.

consists of individuals from a segregating M2 or F2 progeny that have a wild-type phenotype (containing a mixture of individuals that are homozygous wild-type and heterozygous mutant). In pooling strategy B, only true wild-type individuals, identified by progeny testing of the M3 or F3 generation, are pooled as the control. The pooled mutant and control DNA samples are sequenced separately using Illumina Hiseq or NextSeq. With the control pool from strategy A, only homozygous deletions commonly existing in the control pool and the mutant pool are filtered out from the mutant pool. With the control pool from strategy B, both heterozygous and homozygous deletions commonly existing in the control pool and the mutant pool are filtered out from the mutant pool.

FNBtools is able to accept multiple samples and align all reads to the reference genome. **Figure 2** provides additional details about our FNBtools in the Materials and Methods section. A flowchart summarizing the methodology is shown in Supplementary Figure S2.

Simulation Data

FNBtools was benchmarked on simulated data. The *M. truncatula* A17 genome (version 4.0 ∼400 Mb) (Tang et al., 2014) was used to generate random deletions using SVsim.[2] A total of 315 deletions with different deletion sizes were generated. The distribution of deletion sizes is shown in Supplementary Figure S1. Based on the mutated *M. truncatula* genome, 150 bp paired-end reads with 5x, 10x, 20x, and 40x coverage were generated using wgsim (Li et al., 2009), assuming a 0.5% sequencing error rate under each deletion size. To test the functionality of filtering heterozygous deletions in the wild-type sample, we also generated heterozygous 150 bp paired-end reads for the wild-type sample with 0.05, 0.1, 0.2, and 0.5 deletion frequencies at 5x, 10x, 20x, and 40x coverage, respectively.

Comparison With Similar Tools

We chose deletion callers such as Pindel, BreakDancer, Delly, and Sprites to compare with FNBtools. For a large deletion (i.e., $n_{smd} = 0$, n_{smd} represents the total number of small deletion reads) be counted as a successful detection, the breakpoints of deletion positions should be ±100 bp from the breakpoints of true deletions, and the deletion length should be 90% overlapping with the true deletions in simulated data. For small deletions ($n_{smd} > 0$), the breakpoints of each deletion position should be ±5 bp from the breakpoints of true deletions, and the overlapping rate should be 50%. Recall and precision values were measured together with calculating an accuracy score, F-score, described below:

$$\text{Precision} = \frac{TP}{TP + FP}\text{TP: true positive; FP: false positive}$$

$$\text{Recall} = \frac{TP}{TP + FN}\text{TP: true positive; FN: false negative}$$

$$\text{F-score} = 2^{*}\frac{precision^{*}Recall}{precision + Recall}$$

To evaluate how FNBtools, Pindel, BreakDancer, Delly, and Sprites perform at different deletion sizes, we visualized the performance of all five tools for comparison in **Figures 3**, **4A**. In terms of *F*-score, FNBtools outperforms almost all other tools for detecting homozygous deletions at different coverage levels. We found that FNBtools has a high *F*-score at almost every deletion size range, indicating that FNBtools performs very well even at low coverage. It is noteworthy that FNBtools has high precision and recall values at small deletions, which commonly

[2]https://github.com/GregoryFaust/SVsim

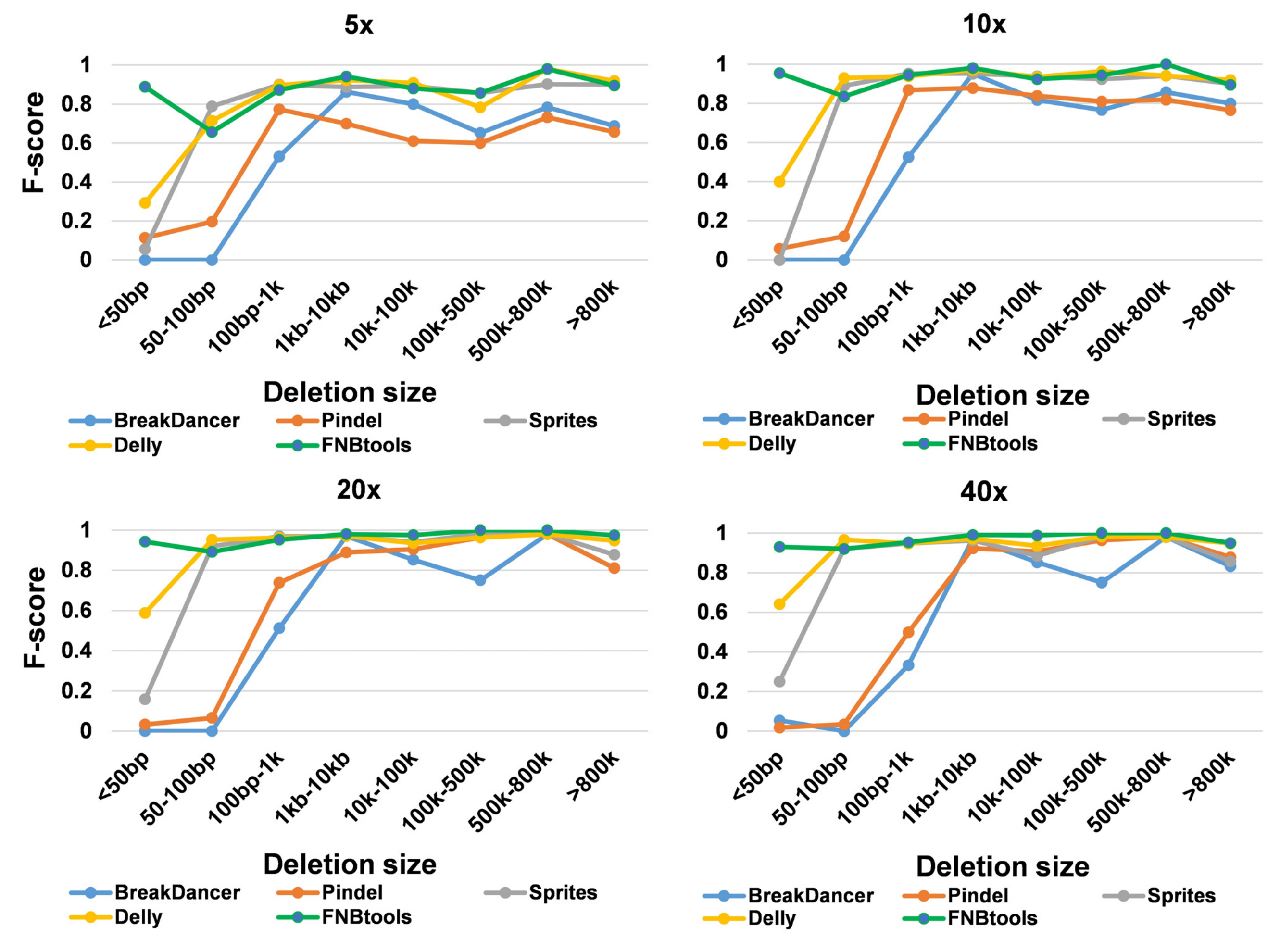

FIGURE 3 | Accuracy summary of five deletion callers with different coverage simulated data.

occur in FNB populations. In contrast, the other four tools have a limited ability to detect small deletions. Sequencing at lower coverage can greatly reduce the cost of NGS. Furthermore, from a comparison analysis using masked and unmasked *Medicago truncatula* reference genome, we found that FNBtools can also detect deletions in repetitive regions (Supplementary Figure S3).

FNBtools includes a function to filter deletions that exist in the wild-type (or control) population to identify unique deletions in the mutant population (see Materials and Methods for details). To validate the efficiency of this filtering function, we compared the results at different coverage levels and deletion frequencies. We found that the filtering efficiency increases as either the coverage level or deletion frequency increases. When the deletion frequency is 0.5 (50%) in the wild-type population, 97.2, 99.3, and 99.7% non-unique deletions can be filtered out at 10x, 20x, and 40x coverage levels, respectively (see **Figure 4B**).

Real Biological Data

To further evaluate the usefulness of FNBtools in real biological samples, we employed the tool in the identification of the causal deletion in a salt-tolerant FNB mutant. In most cases, wild-type (strategy B) or wild-type-like plants (strategy A) from a segregating progeny (**Figure 1**) are pooled and used as the control. However, in our case study, since wild-type and heterozygous plants from the segregating progenies were dead during the salt selection, no materials were left to be used as the control. In this case, we used pooled DNA from another mutant S1, which has the same parental background as the S2 mutant but is not allelic to S2, as the control for FNBtools data processing and analysis (see Materials and Methods for more details on sample pooling). The S1 and S2 pools were then sequenced using the Illumina NextSeq platform to generate paired-end reads. Approximately 131 million and 45 million raw reads from samples S1 and S2, respectively, were obtained from sequencing. After filtering out low quality reads, ∼112 million and 39 million clean reads for S1 and S2, respectively, were used for analysis. The coverage of the control and the mutant sample is 93x and 33x, respectively. These clean reads were mapped to the *M. truncatula* reference genome, and homozygous deletions were called using FNBtools. In total, 28,637 deletions were identified in sample

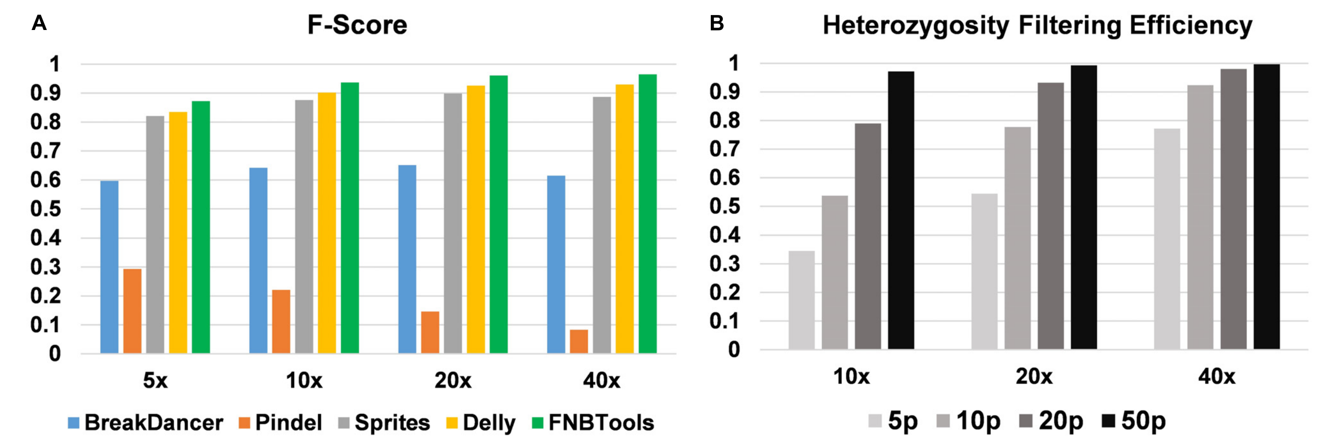

FIGURE 4 | (A) Performance comparison of homozygous deletion detection. **(B)** Heterozygosity filtering efficiency by FNBtools. 5p, 10p, 20p, and 50p represent 5, 10, 20, and 50% deletion frequencies in the wild-type sample.

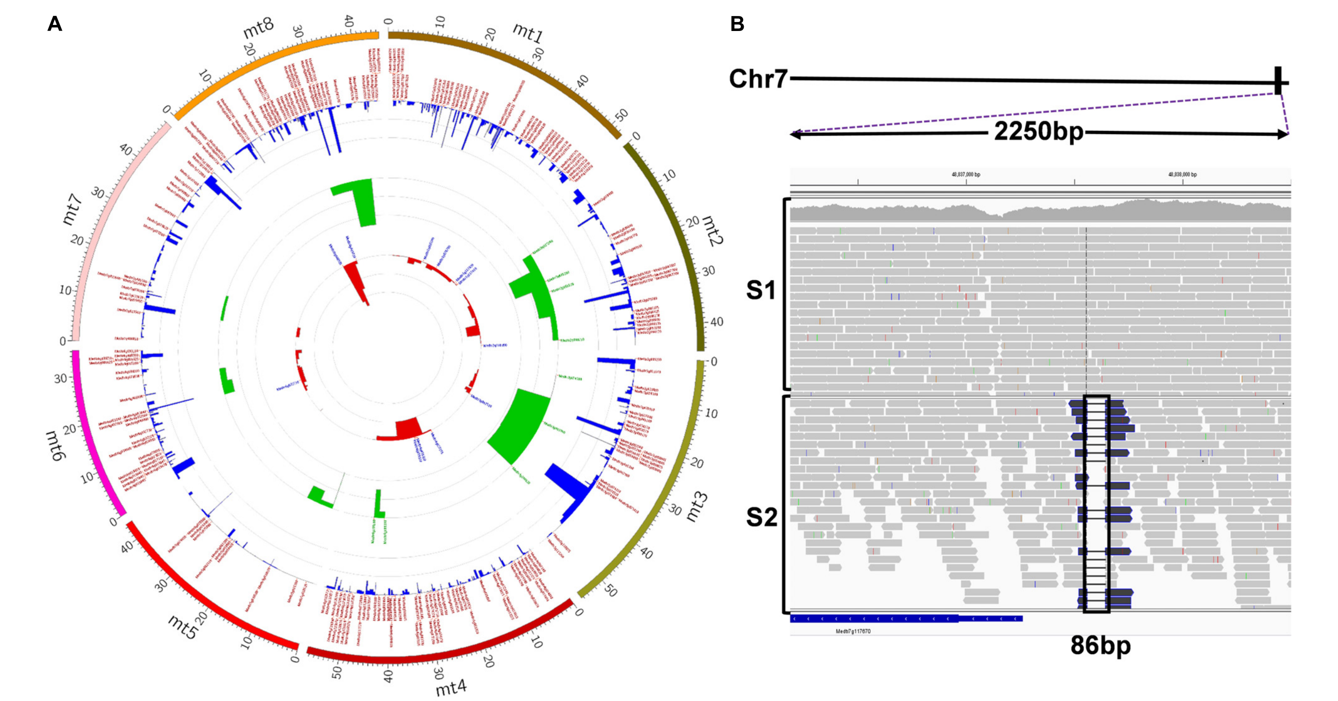

FIGURE 5 | (A) Circos visualization of identified deletions in S2. The inner-most circle (red) represents deletions > 1kb; the middle circle (green) represents deletions between 100 bp and 1 kb; the out-most circle (blue) represents deletions < = 100 bp. Each circle was annotated by gene IDs if deletions contain or are in genes (high-resolution image link: https://github.com/noble-research-institute/fnbtools/blob/master/example/Circos_S2_deletion.png). **(B)** Visualization of the causal deletion using IGV. The deletion is located at the bottom of chromosome 7 (solid black rectangular box). A 2250 base pair window from IGV, as indicated by the purple dotted lines, shows the reads around the deletion. The histogram plot illustrates the read depth with individual reads plotted below. There is no deletion in the control sample S1. Black box highlights the 86bp deletion region in the mutant sample S2.

S2 (Supplementary File S1). Out of these deletions, 5,373 are homozygous deletions that can be visualized in **Figure 5A** (high-resolution image link[3]). We also identified 2651 homozygous deletions in the control sample S1. Interestingly, we found 1,542 homozygous deletions that are present at the exact same locations with exact same deletion sizes in both control (S1) and mutant (S2) samples, indicating these deletions are systematically present in our materials. One explanation for this observation is that the starting materials we used for mutagenesis are different from the materials used for reference genome sequencing.

[3] https://github.com/noble-research-institute/fnbtools/blob/master/example/Circos_S2_deletion.png

To confirm the accuracy of FNBtools in determining homozygous deletions, we randomly selected 23 deletions with different read coverages for examination in the S1 sample. Supplementary Table S1 shows that when a deletion has four or more informative reads (either SMD or CLR), the deletion prediction by FNBtools is accurate with a 100% success rate. When a deletion has three reads, the prediction has a 74% accuracy. When the coverage of NGS data is very high, for example 93x coverage in our case study, we recommend to use more stringent cut-off values for supporting reads number.

From reciprocal genetic crosses, we knew that S1 and S2 are non-allelic mutant lines and the phenotype is caused by different causal genes/deletions. We filtered out all homozygous deletions from S2 that are either heterozygous or homozygous in S1. After filtering out deletions that commonly exist in S1 and S2, 12 unique homozygous deletions were identified in S2. All of these unique deletions have at least 12 informative reads, and the deletion sizes and positions were confirmed by PCR amplification and sequencing.

To find the causal deletion(s) in the mutant, we performed a genetic linkage analysis for these 12 deletions using a small population consisting of 65 mutant samples segregated from a backcrossed F2 progeny. Theoretically, based on traditional parametric linkage analysis, if a homozygous deletion is present in 25% of mutant samples, this deletion is recombining freely during meiosis and is not linked with the causal mutation (Morton, 1955). This number may fluctuate depending on the population size and the mutation location on chromosomes. If a deletion is the causal mutation, it should be present in all mutants. As shown in Supplementary Table S2, deletion 11 is present in all mutant plants, indicating strong linkage with the phenotype. All other deletions show free segregation patterns; thus, they are not linked with the phenotype. However, deletion 11 falls in the intergenic region between Medtr7g117670 and Medtr7g117675 (see **Figure 5B**). Though we successfully identified the linked locus, it will take more effort to pinpoint the causal gene. One approach is to identify mutants of these two genes from the *M. truncatula Tnt1* insertion mutant population (Tadege et al., 2008) and examine whether the mutants show a salt tolerance phenotype.

CONCLUSION

We have developed a software, FNBtools, which can identify both small and large homozygous deletions in FNB populations. FNBtools was developed by taking two types of reads (soft-clipped reads and discordant reads) into consideration. Using simulated data, FNBtools outperforms existing popular deletion callers, BreakDancer, Pindel, Delly, and Sprites, in detecting small deletions in all tested coverage levels. In a real biological case study using FNBtools, we successfully identified a locus linked with a phenotype in an FNB mutant from a *M. truncatula* mutant population. The linkage between the identified causal deletion and exhibited phenotype in the mutant was confirmed by PCR in a small segregating population. In plants, many deletion mutant populations have been generated by different groups, for example, *Arabidopsis thaliana* (Belfield et al., 2012), soybean (Bolon et al., 2011), peanut (Wang et al., 2015), common bean (O'Rourke et al., 2013), etc. Furthermore, the genome sequences of these plant species are also available. We speculate that FNBtools can be used for deletion detection in these mutant populations. With the ease of NGS sequencing, more plant species have been and will be sequenced. In fact, as long as genome sequences are available, FNBtools can be used for deletion identification in any species, such as microbes, worms and fruit flies. Application of FNBtools in crops will provide a quick and reliable tool in non-transgenic molecular breeding.

Our FNBtools is currently a reference genome-based tool. If deletions occur in the gap regions of the reference genome, for example, an entire scaffold is deleted, it is hard to identify this type of deletion.

AVAILABILITY

FNBtools is an open source program available in the GitHub repository (https://github.com/noble-research-institute/fnbtools).

AUTHOR CONTRIBUTIONS

LS and JW conceived the original research plans. LS, YG, and AB performed all bioinformatics analysis. XC performed most of the experiments. LS and JW supervised and complemented the writing.

FUNDING

This work was supported by the Computing Services Innovation Project of Noble Research Institute and in part by Forage Genetics International, Inc to JW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We would like to thank all committee members of Computing Services Innovation Project of Noble Research Institute for their help: Melanie Davis, Jody Beard, Perdeep Mehta, Mike Komp, Blake Michael, and Tora Williamsen-Berry. We thank Jeremy Murray at Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, for his critical reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00976/full#supplementary-material

REFERENCES

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30, 174–178. doi: 10.1038/nbt.2095

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301, 653–657. doi: 10.1126/science.1086391

Belfield, E. J., Gan, X. C., Mithani, A., Brown, C., Jiang, C. F., Franklin, K., et al. (2012). Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of *Arabidopsis thaliana*. *Genome Res.* 22, 1306–1315. doi: 10.1101/gr.131474.111

Bolon, Y. T., Haun, W. J., Xu, W. W., Grant, D., Stacey, M. G., Nelson, R. T., et al. (2011). Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiol.* 156, 240–253. doi: 10.1104/pp.110.170811

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16–S21. doi: 10.1038/ng2028

Chen, K., Wallis, J. W., Mclellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–676. doi: 10.1038/nmeth.1363

Chiang, D. Y., Getz, G., Jaffe, D. B., O'kelly, M. J. T., Zhao, X. J., Carter, S. L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103. doi: 10.1038/nmeth.1276

d'Erfurth, I., Cosson, V., Eschstruth, A., Lucas, H., Kondorosi, A., and Ratet, P. (2003). Efficient transposition of the Tnt1 tobacco retrotransposon in the model legume *Medicago truncatula*. *Plant J.* 34, 95–106. doi: 10.1046/j.1365-313X.2003.01701.x

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Garrison, E., and Marth, G. (2012). *Haplotype-Based Variant Detection from Short-Read Sequencing*. Available at: https://ui.adsabs.harvard.edu/#abs/2012arXiv1207.3907G [accessed 01 July, 2012].

Guan, P., and Sung, W. K. (2016). Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* 102, 36–49. doi: 10.1016/j.ymeth.2016.01.020

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 1270–1278. doi: 10.1101/gr.088633.108

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., et al. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26, i350–i357. doi: 10.1093/bioinformatics/btq216

Kim, M., Chen, Y., Xi, J., Waters, C., Chen, R., and Wang, D. (2015). An antimicrobial peptide essential for bacterial survival in the nitrogen-fixing symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15238–15243. doi: 10.1073/pnas.1500123112

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., Mclellan, M. D., Mardis, E. R., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285. doi: 10.1093/bioinformatics/btp373

Koboldt, D. C., Zhang, Q. Y., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., et al. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 10:R23. doi: 10.1186/gb-2009-10-2-r23

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

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X., et al. (2001). A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.* 27, 235–242. doi: 10.1002/cfg.148

Malolepszy, A., Mun, T., Sandal, N., Gupta, V., Dubin, M., Urbanski, D., et al. (2016). The LORE1 insertion mutant resource. *Plant J.* 88, 306–317. doi: 10.1111/tpj.13243

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

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7, 277–318.

O'Rourke, J. A., Iniguez, L. P., Bucciarelli, B., Roessler, J., Schmutz, J., Mcclean, P. E., et al. (2013). A re-sequencing based assessment of genomic heterogeneity and fast neutron-induced deletions in a common bean cultivar. *Front. Plant Sci.* 4:210. doi: 10.3389/fpls.2013.00210

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333–i339. doi: 10.1093/bioinformatics/bts378

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Rogers, C., Wen, J. Q., Chen, R. J., and Oldroyd, G. (2009). Deletion-based reverse genetics in *Medicago truncatula*. *Plant Physiol.* 151, 1077–1086. doi: 10.1104/pp.109.142919

Sha, Y., Li, S., Pei, Z., Luo, L., Tian, Y., and He, C. (2004). Generation and flanking sequence analysis of a rice T-DNA tagged population. *Theor. Appl. Genet.* 108, 306–314. doi: 10.1007/s00122-003-1423-9

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., et al. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. *Plant J.* 54, 335–347. doi: 10.1111/j.1365-313X.2008.03418.x

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* 15:312. doi: 10.1186/1471-2164-15-312

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Wang, J. S., Sui, J. M., Xie, Y. D., Guo, H. J., Qiao, L. X., Zhao, L. L., et al. (2015). Generation of peanut mutants by fast neutron irradiation combined with in vitro culture. *J. Radiat. Res.* 56, 437–445. doi: 10.1093/jrr/rru121

Weigel, D., and Glazebrook, J. (2009). Quick miniprep for plant DNA isolation. *Cold Spring Harb. Protoc.* 2009:pdb.prot5179. doi: 10.1101/pdb.prot5179

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Zhang, Z., Wang, J. X., Luo, J. W., Ding, X. J., Zhong, J. C., Wang, J., et al. (2016). Sprites: detection of deletions from sequencing data by re-aligning split reads. *Bioinformatics* 32, 1788–1796. doi: 10.1093/bioinformatics/btw053

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.