SOFTWARE NOTE

Applications
in Plant Sciences

InDelGT: An integrated pipeline for extracting indel genotypes for genetic mapping in a hybrid population using next-generation sequencing data

Zhiliang Pan ⓘ | **Jinpeng Zhang** ⓘ | **Shengjun Bai** ⓘ | **Zhiting Li** ⓘ | **Chunfa Tong** ⓘ

Co-Innovation Center for Sustainable Forestry in Southern China, College of Forestry, Nanjing Forestry University, Nanjing 210037, China

Correspondence

Chunfa Tong, College of Forestry, Nanjing Forestry University, 159 Longpan Road, Nanjing 210037, China.
Email: tongchf@njfu.edu.cn

Abstract

Premise: Although several software packages are available for genotyping insertion/deletion (indel) polymorphisms in genomes using next-generation sequencing data, simultaneously calling indel genotypes across many individuals for use in genetic mapping remains challenging.

Methods and Results: We present an integrated pipeline, InDelGT, for the extraction of indel genotypes from a segregating population such as backcross or F_2 lines, or from an F_1 cross between outbred species. The InDelGT algorithm is implemented in three steps: generating an indel catalog, calling indel genotypes, and analyzing indel segregation. We demonstrated the use of the pipeline with an example data set from an F_1 hybrid population of *Populus* and successfully constructed the two parental genetic linkage maps.

Conclusions: InDelGT is a practical tool that can quickly genotype a large number of indel markers within a population following Mendelian segregation. The InDelGT pipeline is freely available on GitHub (https://github.com/tongchf/InDelGT).

KEYWORDS

genetic mapping population, genotype calling, insertion/deletion polymorphism (indel), next-generation sequencing (NGS)

Insertion/deletion polymorphisms (indels) are a common and functionally important type of sequence variation (Albers et al., 2011). The insertion of mobile elements such as retrotransposons, slippage in simple sequence replication, or unequal crossovers during meiosis can all lead to indel generation (Britten et al., 2003). Consequently, indels are widely and densely distributed across the genome, with a distribution density second only to single-nucleotide polymorphisms (SNPs); indels are thus much more common than simple sequence repeats (SSRs). Indels can be multiallelic and codominant, potentially providing more genomic information than the typical biallelic SNPs (Lv et al., 2016). Indels in coding regions have a significant impact on the structure and function of the resulting protein, which in turn leads to phenotypic differences in the organism, meaning indels have made a significant contribution to the processes of speciation and phenotypic evolution (Tian et al., 2008). Indels have also become an effective marker system for genetic analysis due to their multiallelic nature, codominant inheritance, and wide genome coverage (Das et al., 2015).

With advances in next-generation sequencing (NGS), a large amount of DNA sequence data can be generated at low cost, facilitating the bioinformatic identification of genetic variants. A number of software programs have been developed for calling indel genotypes from sequence reads, including SOAPindel (Li et al., 2013), Pindel (Ye et al., 2009), VarScan (Koboldt et al., 2009), and GATK (McKenna et al., 2010). Although the analytical processes of these software programs are described in detail in the literature, each process involves multiple steps, with each step requiring the input of multiple parameters and commands. This means the analytical processes of the available

indel-calling programs are all relatively complicated and difficult to use. Moreover, there is interest in using indels for constructing genetic maps because they are useful in marker-assisted selection and quantitative trait locus mapping (Li et al., 2015; Song et al., 2015); however, no specific tools have been developed for extracting indel genotypes across many individuals in a hybrid population for use in genetic mapping.

Here, we report an integrated pipeline called InDelGT, which can genotype a large number of indel markers in a population following Mendelian segregation to construct genetic linkage maps. InDelGT can be used to perform indel genotyping of traditional backcross (BC) and F_2 populations from inbred lines, and we demonstrate its use here in an F_1 hybrid population derived by crossing two individuals from two outbred species. InDelGT can complete the entire genotyping process by simply entering a command line, and it can perform multicore or multithreaded parallel computing, which greatly accelerates the indel genotyping process.

METHODS AND RESULTS

Implementation of InDelGT

InDelGT was implemented with Perl to extract indel genotypes in a hybrid population. It can be run in three steps: (1) generating two parental indel catalogs, (2) calling indel genotypes across a hybrid population, and (3) analyzing indel segregation. A flow chart of the implementation is shown in Figure 1.

Generating two parental indel catalogs

In the first step, InDelGT generates parental indel catalogs with the Perl program *parent_genotyping.pl*. First, the paired-end reads of each parent are aligned to a reference sequence using BWA, generating a SAM format file (Li et al., 2009). The SAM files for the two parents are converted to BAM format files using the sorting and indexing process of SAMtools. Subsequently, BCFtools is used to generate BCF files from the BAM files, then to further convert them into VCF files. Finally, the indel genotype of each parent at each site is extracted from its VCF file, ensuring a read depth of at least 3 for an allele and a genotype quality greater than 30. The indel sites from the two parents are merged and saved in a text file (namely "parent.cls") as an indel catalog. With this catalog file, the genotypes of the female and male parents are called and saved in two files, "female_merge.gt" and "male_merge.gt", respectively.

Calling indel genotypes across a hybrid population

In the second step, the Perl program *progeny_genotyping.pl* is used to call the indel genotypes of all progeny with the catalog file generated above. The calling procedures are similar to the generation of the parental indel catalog. After generating the BCF files for each individual, the VCF files are generated according to the indel catalog by using BCFtools. Next, indel genotyping is performed according to the VCF or BCF files. A heterozygous indel genotype can be easily determined for a sample from the records in its VCF file; however, when calling a homozygous genotype at an indel site, we need to know the site start and end positions in the parental indel catalog. Only when every position from start to end is homozygous can we consider the indel genotype to be homozygous. Finally, the indel genotypes at all indel sites are saved for each sample. The file contains the indel site positions, allele nucleotides, and genotypes.

Analysis of indel segregation

In the last step, InDelGT uses the Perl program *segregation_ratio.pl* to identify the segregation types of indels, to count the number of individuals with each genotype, and to calculate a P value to test whether each indel follows Mendel's law of segregation. The indel sites that deviate from the Mendelian segregation ratio ($P < 0.01$) and have more than a certain percentage (e.g., 10%) of samples missing genotypes in the progeny are filtered out. The result files are saved in text files that vary for the different populations. For a BC or F_2 population, InDelGT will only produce one result file because there is only one type of segregation in the population. By contrast, for a hybrid population between two outbred species, up to seven result files will be produced, corresponding to the seven segregation types: $ab \times aa$, $aa \times ab$, $ab \times ab$, $ab \times cc$, $aa \times bc$, $ab \times ac$, and $ab \times cd$, where the first two letters of each type indicate the genotype of the female parent and the last two are the genotype of the male parent.

Use of InDelGT

InDelGT can run on Linux systems. To run InDelGT, users must download and install the necessary software packages, including BWA (Li and Durbin, 2009), SAMtools (Li et al., 2009), and BCFtools (Li et al., 2009). Moreover, a parameter setting file must be created, namely 'parameters.ini', which consists of three parts: folders, parameters, and data files (Figure 2). The first part, 'folders', gives the paths to software packages and data files, including BWA, SAMtools, BCFtools, and InDelGT itself, as well as the reference sequence file and FASTQ files. The second part, 'parameters', includes the minimum score of indel genotypes, the minimum score of the mapping quality of the reads, the percentage of missing genotypes allowed, the minimum P value allowed for testing the segregation ratio at an indel site, and the number of threads used for parallel computing. The third part, 'data files', shows the FASTA file of a reference sequence and the first and second read files of
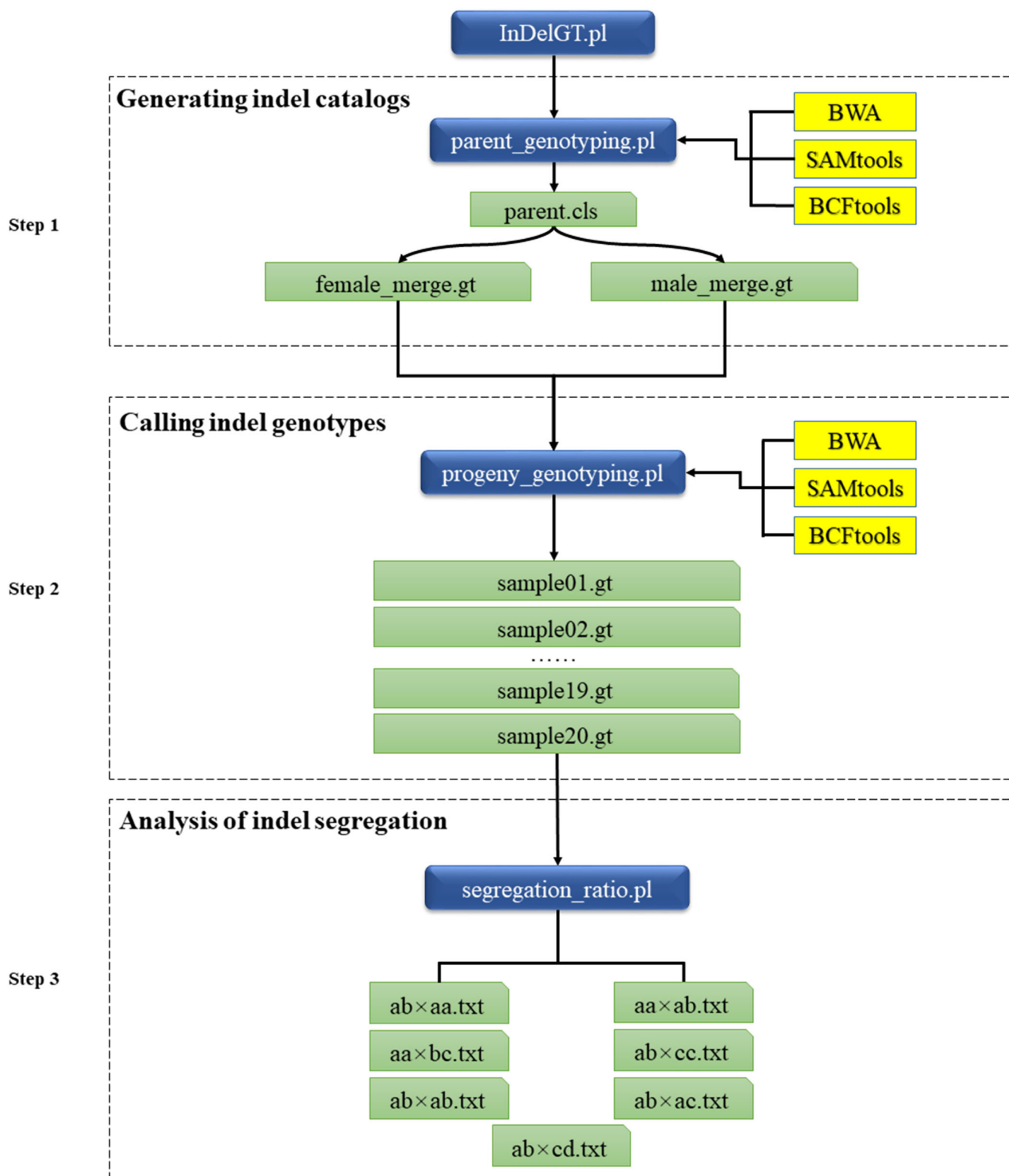
FIGURE 1 Schematic diagram of the InDelGT workflow. Three steps are required to run InDelGT: (1) generating a parental indel catalog, (2) calling indel genotypes across a hybrid population, and (3) analyzing indel segregation. The blue rounded rectangles represent the internal program written in Perl. The ordinary yellow rectangles indicate the external software packages used in InDelGT, including BWA, SAMtools, and BCFtools. The green cut-corner rectangles represent text files containing the intermediate and final results.

```
[folders]
BWA_FOLD:~/bwa-mem2
SAMTOOLS_FOLD:~/panzl/a/samtools-1.9
BCFTOOLS_FOLD:~/panzl/a/bcftools-1.9
INDELGT_FOLD:~/panzl/a/7/CI/InDelGT
DATAFOLD:~/panzl/a/7/CI
[parameters]
# the minimum score of InDel genotypes
GQ:30
# the minimum score of the mapping quality of reads
MAPQ:10
# the number of progeny
PROGENY_NUMBER:117
# percent of missing genotypes at each InDel
MISPCT:10
HETEROZYGOUS_DEPTH:3
HOMOZYGOUS_DEPTH:5
PVALUE:0.01
THREADS:20
[data files]
REFERENCE_FILE:Ptrichocarpa_533_v4.0.fa
# PARENT1 is the heterozygous parent in the BC1 population and PARENT2 is the
heterozygous parent in the BC2 population.
PARENT1: female_merge_1.fq      female_merge_2.fq
PARENT2: male_merge_1.fq      male_merge_2.fq
PROGENY1: 3-0_1.fq_filtered      3-0_2.fq_filtered
PROGENY2: 3-100_1.fq_filtered      3-100_2.fq_filtered
PROGENY3: 3-104_1.fq_filtered      3-104_2.fq_filtered
PROGENY4: 3-107_1.fq_filtered      3-107_2.fq_filtered
……
PROGENY117: 3-99_1.fq_ filtered      3-99_2.fq_filtered
```

FIGURE 2   The style of the parameter-setting file for launching InDelGT.

the two parents and all progeny in FASTQ format. After installing the external software packages and preparing the parameter file, users can run InDelGT with a simple command line 'perl InDelGT.pl -o' followed by an output.

InDelGT is freely available at https://github.com/tongchf/InDelGT. For users to quickly grasp the use of InDelGT, we have also made test data available online including a reference sequence file and the FASTQ files of two parents and 20 progeny.

Genotyping of indel markers in poplar: A case study using InDelGT

To demonstrate the function of InDelGT, we performed a real data analysis using two species in the genus *Populus* L. (poplar). Poplar has a small genome and high heterozygosity, making this an appropriate taxon to use for the evaluation of the computational performance of InDelGT. In previous studies, we performed restriction site–associated

TABLE 1 Mendelian segregation patterns and the numbers of indel markers identified.

Female genotype	Male genotype	Expected segregation ratio	Number of indels
ab	aa	1:1	1539
aa	ab	1:1	1262
ab	ab	1:2:1	3
ab	ac	1:1:1:1	1
Total			2805

DNA sequencing (RAD-seq) of the two parents and 117 progeny in an F$_1$ hybrid population derived from a cross between *P. deltoides* W. Bartram ex Marshall and *P. simonii* Carrière (Mousavi et al., 2016; Tong et al., 2016). The raw data are available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/Traces/sra) using the accession numbers listed in Appendix S1. After performing quality control on the data using the NGS QC toolkit (Patel and Jain, 2012), these RAD-seq data were used to call indel genotypes with InDelGT. We selected the *Populus trichocarpa* v4.0 (Tuskan et al., 2006) reference sequence to enable the calling of indel genotypes across all individuals, including the two parents. We set the genotype quality to be greater than 30, the coverage depth for each allele in a heterozygote as not less than 3, and the allele coverage depth in a homozygote as not less than 5. Additionally, we filtered out those indel sites for which more than 10% of the genotypes were missing data and with *P* values less than 0.01 in the data sets. According to the segregation types, the final genotype data were classified and saved in different text files. Consequently, a total of 50,532 indel sites were identified between the two parents, of which 2805 followed different Mendelian segregation ratios ($P > 0.01$) in the F$_1$ hybrid population. The indel genotype data included the segregation types $ab \times aa$, $aa \times ab$, $ab \times ab$, and $ab \times ac$, with totals of 1539, 1262, three, and one occurrence, respectively (Table 1).

To validate the indel genotype data used for genetic mapping, we constructed two parental genetic linkage maps using the two indel data sets from the segregation types $ab \times aa$ and $aa \times ab$. The results showed that the indels with the segregation type $ab \times aa$ were assigned into 19 linkage groups based on the maternal *P. deltoides* genome under a logarithm of odds (LOD) threshold ranging from six to 12 (Appendix S2), which was consistent with the karyotype of *Populus* ($2n = 38$) (Eckenwalder, 1996). In the same way, the indels with the segregation type $aa \times ab$ were divided into 20 linkage groups based on the paternal *P. simonii* genome under a LOD threshold ranging from six to 11 (Appendix S3). The maternal genetic linkage map contained 1125 indels spanning 3146.05 cM, with an average genetic distance between adjacent indels of 2.84 cM. The paternal genetic linkage map contained 1024 indels with a genetic distance of 3760.90 cM, and an average genetic distance of 3.75 cM.

TABLE 2 The primer sequences of the eight indel markers developed and used in this study.

| Marker name	Physical position	Indel size (bp)	Forward and reverse primers	Product length (bp)
P1	Chr01:45105144	8	F: CAGAATAGCAAGATGTGTAAGCGA	396
			R: GCAATCTTCTACGATCCTTTGACC	
P2	Chr01:7550584	8	F: ACTCCATTTCTATCCTTTGTAACT	317
			R: AATTCTAATGGTGGAAAGTGGAAAC	
P3	Chr03:14905134	11	F: TTGTCCTTTTATTGTCGTCTTGGT	280
			R: AACAACGCTACAACAAATCCCTCT	
P4	Chr12:11595547	20	F: GTTTTTCCAGTCTCTTTCTCCTCT	353
			R: AGCAAGTCCACATATAAGCAAGTT	
P5	Chr01:11673547	12	F: GAACCCTTGCCTAGACCTCC	264
			R: TGCTGTGTTTGCCCAACCTG	
P6	Chr01:48209479	12	F: AGACAGTTGAAGAGTTAAGGC	186
			R: GATTCTTGAGCTTGACCCCTT	
P7	Chr02:1340072	23	F: AAGATCCATAATGCACACGTCA	219
			R: TGGTCACTGGCTATAACTACT	
P8	Chr13:1129169	15	F: ATCTGCAAGAATAATGTCGGT	190
			R: GTTTACAGTTTCAGGAGTTTGCTT	
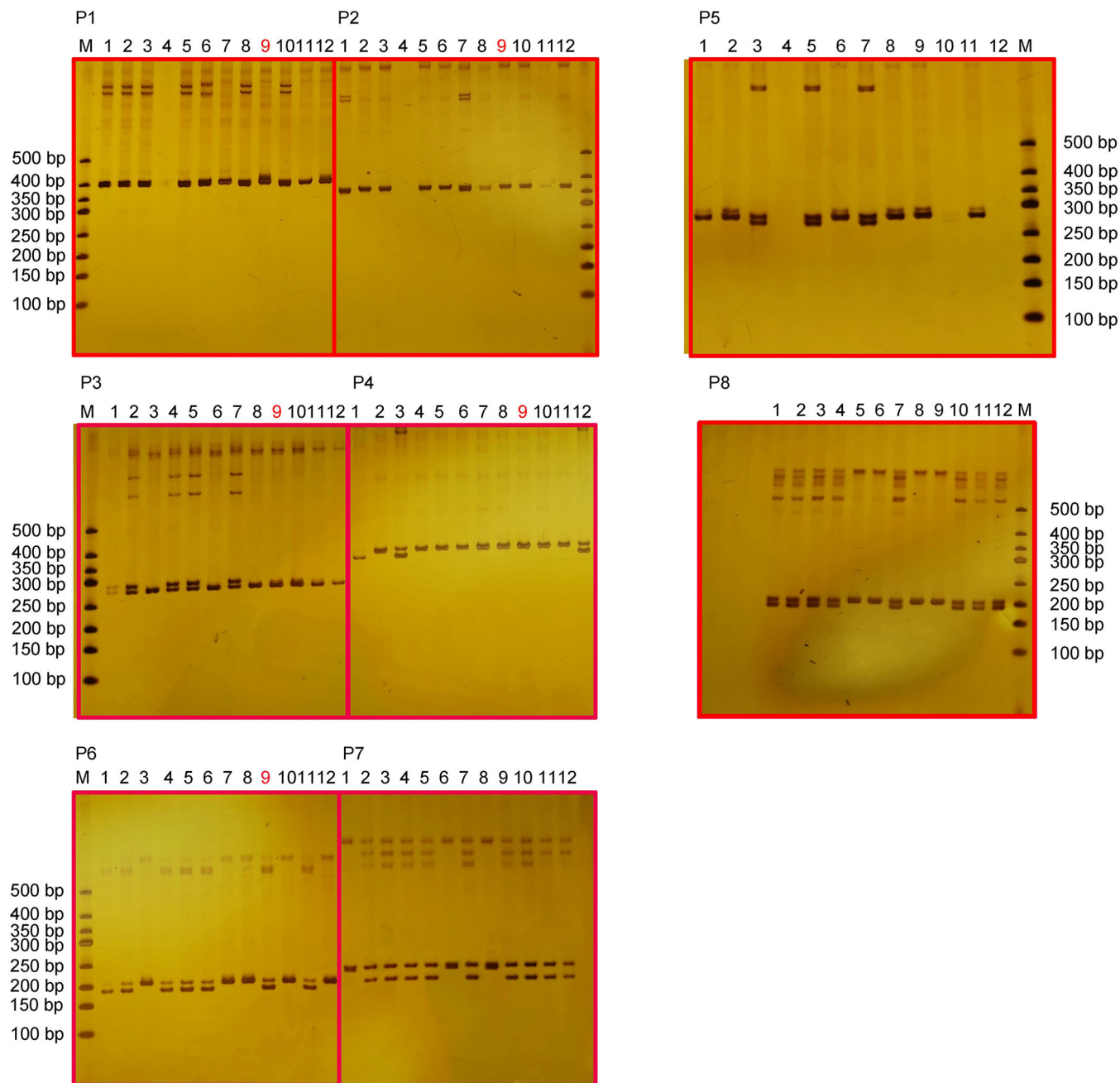
FIGURE 3 The PCR profiles of eight indels in 12 individuals. M represents the indel marker, and red numbers indicate that the electrophoretic bands of the individual are not consistent with its genotype called from InDelGT.

Experimental validation of indel genotypes

We randomly selected eight indels to verify the accuracy of our indel genotyping for 12 progeny using polyacrylamide gel electrophoresis (PAGE) technology (Table 2). First, the Oligo 7 program (https://www.oligo.net) was used to design specific primers based on the sequences flanking the indel sites. Genomic DNA was extracted from mature leaves of each sample using a Plant Genomic DNA Kit (Tiangen Biotech, Beijing, China), and the DNA fragments were amplified using PCR. The PCR was performed in a 10-µL

reaction volume containing 0.05 U of *Taq* DNA polymerase, 1 µL of 10× PCR buffer (Takara Bio, Kusatsu, Japan), 1 µL genomic DNA (30 ng/µL), 0.3 µL MgCl$_2$ (25 mM), 0.5 µL dNTPs (10 mM), 0.4 µL of each primer (10 µM), and 6.35 µL of double-distilled water. The PCR program began with denaturation at a temperature of 94°C for 3 min; followed by 30 cycles of 94°C for 30 s, 58°C for 30 s, and 72°C for 30 s; with a final extension at 72°C for 7 min. The PCR products were separated using electrophoresis on 8% denatured polyacrylamide gels run at 180 V for 1.5 h, and then detected using silver staining. Finally, we observed that

all but five genotypes were amplified, of which 86 (95%) genotypes were consistent with the results of the InDelGT genotyping (Figure 3). This experimental result showed that InDelGT has a high level of accuracy in determining indel genotypes.

DISCUSSION

In this paper, we report an integrated pipeline, InDelGT, for the extraction of indel genotypes in a hybrid population using next-generation DNA sequencing data. InDelGT can be used to implement the complex process of genotyping indels across multiple individuals in a population, such as traditional BC and F_2 populations, as well as F_1 hybrid populations between outbred species. The actual operation is very convenient, and the user can complete the entire calculation process with only one command line. In addition, InDelGT can not only extract indel genotype data but also perform Mendelian segregation ratio tests on the identified indel sites. To demonstrate the effectiveness of InDelGT, we performed indel genotyping on 117 progeny and two parents of an F_1 hybrid population derived from a cross between *P. deltoides* and *P. simonii*. The results showed that these indel markers could be divided into five segregation types. When the genetic linkage maps of the two parents were constructed using the two most abundant indel segregation types, these segregating markers were divided into 19 linkage groups or 20 linkage groups within a large LOD threshold range, which largely matched the karyotype of *Populus*.

The coverage depth of reads is an important factor affecting the accuracy of indel genotyping. A higher read coverage depth can reduce the error rate of detecting indels. Compared with small indels (1–5 bp), large indels (>5 bp) will complicate or distort the correct mapping of reads, so detecting large indels requires a higher coverage depth (Fang et al., 2014). The determination of heterozygotes requires a higher coverage depth than that of homozygotes (McKernan et al., 2009). Ajay et al. (2011) suggested that when a read alignment rate of 85% is reached, the average sequencing coverage depth needs to be more than 40×. Fang et al. (2014) believed that 60× is a more ideal average coverage depth for detecting indels to obtain more accurate indel genotypes. In the current study, the average genome coverage depth of reads for each progeny was approximately 20× (Appendix S1). Because the paired-end reads were from the restriction sites of *Eco*RI, they covered at most 20% of the poplar genome (Tong et al., 2016); therefore, the actual coverage depth may have reached as high as 100×, which is sufficient to satisfy the requirements of indel genotyping. On the other hand, it is worth emphasizing that InDelGT is specifically designed to extract indel genotypes across a mapping population. Unlike other software packages (Koboldt et al., 2009; Ye et al., 2009; McKenna et al., 2010; Li et al., 2013), the genotype data generated from InDelGT are filtered to follow the Mendelian segregation ratio, which could further improve the data accuracy overall. Our experimental validation demonstrated that InDelGT can generate high-quality indel genotype data for genetic linkage mapping.

CONCLUSIONS

InDelGT is a practical tool that can quickly genotype a large number of indel markers largely following Mendelian segregation in a hybrid population. While the example here focuses on an F_1 hybrid population from crosses between outbred species, InDelGT can also be used with traditional BC and F_2 populations. InDelGT is a useful tool for various studies in genetics, genomics, and marker-assisted breeding, especially in forest trees.

ORCID
Zhiliang Pan http://orcid.org/0000-0002-7973-0588
Jinpeng Zhang http://orcid.org/0000-0002-3007-1139
Shengjun Bai http://orcid.org/0000-0003-1343-8944
Zhiting Li http://orcid.org/0000-0002-0854-9538
Chunfa Tong http://orcid.org/0000-0001-9795-211X

REFERENCES
Ajay, S. S., S. C. Parker, H. O. Abaan, K. V. Fajardo, and E. H. Margulies. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome Research* 21: 1498–1505.

Albers, C. A., G. Lunter, D. G. MacArthur, G. McVean, H. W. Ouwehand, and R. Durbin. 2011. Dindel: Accurate indel calls from short-read data. *Genome Research* 21: 961–973.

Britten, R. J., L. Rowen, J. Williams, and R. A. Cameron. 2003. Majority of divergence between closely related DNA samples is due to indels. *Proceedings of the National Academy of Sciences, USA* 100: 4661–4665.

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

Eckenwalder, J. E. 1996. Systematics and evolution of *Populus*. *In* R. F. Stettler, H. D. Bradshaw, P. E. Heilman, and T. M. Hinckley [eds.],

Biology of *Populus*, and its implications for management and conservation, 7–32. NRC Research Press, Ottawa, Canada.

Fang, H., Y. Wu, G. Narzisi, J. O'Rawe, L. Barrón, J. Rosenbaum, M. Ronemus, et al. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine* 6: 89.

Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, et al. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25: 2283–2285.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Li, S., R. Li, H. Li, J. Lu, Y. Li, L. Bolund, M. H. Schierup, et al. 2013. SOAPindel: Efficient identification of indels from short paired reads. *Genome Research* 23: 195–200.

Li, W., J. Cheng, Z. Wu, C. Qin, S. Tan, X. Tang, J. Cui, et al. 2015. An InDel-based linkage map of hot pepper (*Capsicum annuum*). *Molecular Breeding* 35: 32.

Lv, Y., Y. Liu, and H. Zhao. 2016. mInDel: A high-throughput and efficient pipeline for genome-wide InDel marker development. *BMC Genomics* 17: 290.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

McKernan, K. J., H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* 19: 1527–1541.

Mousavi, M., C. Tong, F. Liu, S. Tao, J. Wu, H. Li, and J. Shi. 2016. De novo SNP discovery and genetic linkage mapping in poplar using restriction site associated DNA and whole-genome sequencing technologies. *BMC Genomics* 17: 656.

Patel, R. K., and M. Jain. 2012. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* 7: e30619.

Song, X., H. Wei, W. Cheng, S. Yang, Y. Zhao, X. Li, D. Luo, et al. 2015. Development of InDel markers for genetic mapping based on whole genome resequencing in soybean. *G3: Genes, Genomes, Genetics* 5: 2793–2799.

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

Tong, C., H. Li, Y. Wang, X. Li, J. Ou, D. Wang, H. Xu, et al. 2016. Construction of high-density linkage maps of *Populus deltoides* × *P. simonii* using restriction-site associated DNA sequencing. *PLoS ONE* 11(3): e0150692.

Tuskan, G., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. 2009. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865–2871.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. The RAD-seq data information for the two parents and 117 progeny in the hybrid population of *Populus deltoides* and *P. simonii*.

Appendix S2. The genetic map of linkage groups 1–19 for the maternal *Populus deltoides*. The length and name of each linkage group are presented above and below, respectively. Each indel is named by its position on the reference genome.

Appendix S3. The genetic map of linkage groups 1–20 for the paternal *Populus simonii*. The length and name of each linkage group are presented above and below, respectively. Each indel is named by its position on the reference genome.