

RESEARCH ARTICLE

Open Access



Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data

Jian Liu^{1,2*}, Jingtao Qu^{1†}, Cong Yang^{1†}, Dengguo Tang¹, Jingwei Li¹, Hai Lan^{1,2} and Tingzhao Rong^{1,2}

Abstract

Background: Insertions and deletions (indels) are the most abundant form of structural variation in all genomes. Indels have been increasingly recognized as an important source of molecular markers due to high-density occurrence, cost-effectiveness, and ease of genotyping. Coupled with developments in bioinformatics, next-generation sequencing (NGS) platforms enable the discovery of millions of indel polymorphisms by comparing the whole genome sequences of individuals within a species.

Results: A total of 1,973,746 unique indels were identified in 345 maize genomes, with an overall density of 958.79 indels/Mbp, and an average allele number of 2.76, ranging from 2 to 107. There were 264,214 indels with polymorphism information content (PIC) values greater than or equal to 0.5, accounting for 13.39 % of overall indels. Of these highly polymorphic indels, we designed primer pairs for 83,481 and 29,403 indels with major allele differences (i.e. the size difference between the most and second most frequent alleles) greater than or equal to 3 and 8 bp, respectively, based on the differing resolution capabilities of gel electrophoresis. The accuracy of our indel markers was experimentally validated, and among 100 indel markers, average accuracy was approximately 90 %. In addition, we also validated the polymorphism of the indel markers. Of 100 highly polymorphic indel markers, all had polymorphisms with average PIC values of 0.54.

Conclusions: The maize genome is rich in indel polymorphisms. Intriguingly, the level of polymorphism in genic regions of the maize genome was higher than that in intergenic regions. The polymorphic indel markers developed from this study may enhance the efficiency of genetic research and marker-assisted breeding in maize.

Keywords: Maize, Next-generation sequencing, Insertion and deletion, Indel, Molecular marker, Polymorphism

Background

Maize (*Zea mays* ssp. *mays*) is an incredibly important cereal crop grown widely throughout the world. Increased demand for maize owing to population growth and biofuel production, and the impacts of climate change on maize production will ratchet up the pressure for increased and more sustainable maize production. Since the 1980s, molecular markers have been widely used in maize genetics and breeding. High-density molecular

markers in applied and basic research are advantageous and necessary for map-based cloning and genome-wide association study.

In contrast to single nucleotide polymorphisms (SNPs), Insertions and deletions (indels) are the second most common type of polymorphism. In a previous study, Mills reported that indels are distributed throughout the human genome at an average density of one indel per 7.2 Kb [1]. Several studies have suggested numerous indels that may cause human disease [2]. One of the most common genetic diseases in humans, cystic fibrosis, is frequently caused by various indels in coding regions of the *CFTR* gene [3]. The genetic diseases of tuberous sclerosis, Rett syndrome, and hemophilia B are also caused by small indels [4–6]. Similarly, indels can alter the phenotype of plants (e.g. the maize domestication

* Correspondence: ljian79@hotmail.com

†Equal contributors

¹Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, China

²Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Ministry of Agriculture, Sichuan Agricultural University, Chengdu 611130, China

gene *teosinte branched*; the gene *Gn1a*, which increases the number of reproductive organs in rice; and the wheat stripe rust resistance gene *Yr36* [7–9]). Consequently, indels have been increasingly recognized as an important source of molecular markers, and indel markers have been successfully used for several genetic studies in crops [10].

The maize genome was completed in 2009 using traditional Sanger sequencing technology, [11]. However, the subsequent development of next-generation sequencing (NGS) technology has generated an enormous amount of short reads that science is scrambling to analyze. NGS technology has also drastically reduced the time and cost requirements of sequencing, which has enabled the re-sequencing of a large number of genomes. This has provided for the possibility of large-scale genetic variation surveys, where many individuals within a single species have been sequenced. Examples include 40 silkworm samples [12], 31 soybean samples [13], 900 sorghum samples [14], and 1800 rice samples [15–18]. As of 2012, more than 350 maize inbred line and landrace genomes had been completely re-sequenced and published [19–21]. Rapid bioinformatics developments have introduced various software programs designed to identify indels, including Dindel [22], VarScan [23], GATK [24], and SAMtool's mpileup [25].

The first step in most of these programs is to map reads directly to a reference sequence. Accurate indel calls from reads are challenging for a couple of reasons. First, reads covering indels are generally more difficult to map correctly to reference genomes, especially ones containing large indel events. Moreover, incorrect alignments at the nucleotide level lead to an incorrect placement of gaps in the alignments. Schuler proposed a PCR amplification computer simulation concept known as the electronic polymerase chain reaction (e-PCR), which has been used widely in various aspects of biology, including the chromosomal localization of DNA sequences, genomic sequencing, PCR primer design, and gene cloning [26]. Unlike traditional methods of sequence alignment, e-PCR programs search sequence databases using fragments similar in length to actual PCR primers that would target both ends of a sequence, instead of using a full-length sequence. Coupled with these developments in bioinformatics, NGS data enables the discovery of genome-wide indel polymorphisms by comparing the whole genome sequences of individuals within a species. Lai et al. re-sequenced a group of six elite maize inbred lines, and uncovered more than 30,000 polymorphic indels with sizes smaller than six bp [19]. Despite such progresses, information regarding indel polymorphism, size, and alleles is still inadequate, particularly considering the indel polymorphism levels observed in large populations.

We developed a set of highly polymorphic indel markers with large size differences and high-density occurrence using the NGS reads of 344 maize inbred lines and landraces along with one B73 maize reference genome [11], which was used as the template. Furthermore, we identified and analyzed these genome-wide indel polymorphisms among the populations employed using an e-PCR strategy, with the aim of enhancing the efficiency of maize genetic research and molecular marker assisted breeding.

Results

Identification and distribution of unique primers in the maize genome

We designed a total of 102,929,122 pairs of e-PCR primers using the maize B73 genome as a template. Of these, 93,492,302 pairs are located in intergenic regions, and 9,436,820 pairs are located in genic region, accounting for 90.83 % and 9.17 % of the total, respectively (Table 1). We mapped 11,807,240 of these primer pairs to unique genomic regions, 11.47 % of the total. Of these, 7,569,844 pairs are located in intergenic regions, and 4,237,396 pairs are located in genic regions, accounting for 64.11 % and 35.89 % of the unique primer pairs, respectively. Chromosome 1 contains the maximum number of unique primers (1,925,944), whereas chromosome 10 has the least (861,222). This implies that the number of primers located on a particular chromosome can be positively correlated with chromosome length. The average unique primer density is 5.73 per Kb of DNA. The highest density occurs on chromosome 1 (6.39 per Kb), and the lowest on chromosome 2 (3.81 per Kb). The density of unique primers in different genomic regions varies, and follows, in descending order: from 0.5 Kb upstream of to transcription start sites (TSS_up_0.5Kb), from transcription end sites to 0.5 Kb downstream of them (TES_down_0.5Kb), code determining sequences (CDSs), introns, 5'-untranslated regions (UTRs), 3'-UTRs, and intergenic regions.

Indel variation in the maize genome

The NGS sequencing data from 344 maize genomes consisted of 22,920,398,978 reads, with an average length of 92.78 bp, and an average sequencing depth of 3.02 \times . The sequencing depth of read qi410 was the lowest, at 0.07 \times , whereas W64A had the highest sequencing depth, that of 41.46 \times (Additional file: 1). Read 478 had the most e-PCR hits, that of 8,807,473, accounting for 74.59 % of the total unique primers, whereas qi410 had the fewest hits, 167,605, accounting for 1.42 % of unique primers. There were a sum total of 3,168,631.39 e-PCR hits, accounting for 26.84 % of the unique primers (Additional file: 2). We were able to locate 89.48 % of the unique primers (10,565,398) on over 20 genomes. A total of

Table 1 Distribution of e-PCR primers and polymorphic indels in different regions of the maize genome

Genome region	Total ^a	Unique Primer ^b			Indel ^c (PIC > 0)			High polymorphic indel (PIC ≥ 0.5)		
		Count	Density ^d (Kb)	Rate ^e (%)	Count	Density (Kb)	Rate ^f (%)	Count	Density (Kb)	Rate ^g (%)
TSS_up_0.5Kb	978,913	519935	26.27	53.11	135326	6.84	27.76	20419	1.03	4.19
5'-UTR	614,397	352736	11.62	57.41	88155	2.90	25.86	13294	0.44	3.90
3'-UTR	614,453	352784	11.61	57.41	88166	2.90	25.86	13297	0.44	3.90
CDS	2,204,900	1025151	16.29	46.49	77294	1.23	7.89	7272	0.12	0.74
Intron	5,081,799	2050762	11.99	40.36	389605	2.28	19.60	54527	0.32	2.74
TES_down_0.5Kb	967,784	488116	24.66	50.44	116096	5.87	25.17	16589	0.84	3.60
Intergenic	93,492,302	7569844	4.08	8.10	1206314	0.65	18.56	157841	0.08	2.43
Total ^h	102,929,122	11807240	5.74	11.47	1973746	0.96	18.68	264214	0.13	2.50

^aAll primers on the whole 10 chromosomes

^bPrimers were located in unique genomic region

^cPolymorphism primers were primers with length information in 20 or more than 20 genomes and PIC Value >0

^dDensity was calculated by number/Kb

^eThis rate was the percentage of unique primer against overall primer

^fThis rate was the percentage of unique primer with polymorphisms against unique primer

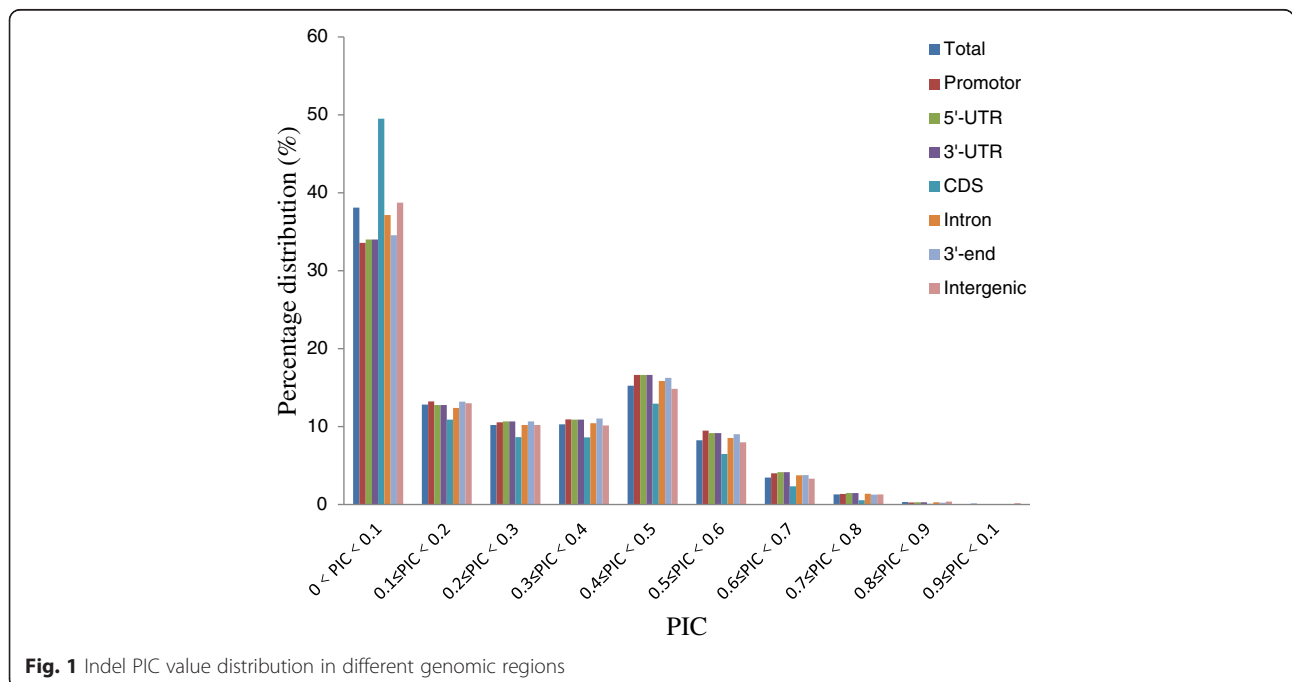
^gThis rate was the percentage of unique primer with polymorphisms greater than or equal to 0.5 against unique primer

^hThere were 102,929,122 primers in total, and the same primer might be divided into different regions and double counting due to the alternative splicing occurring in maize genome

1,973,746 indels were identified, a rate of 18.68 % of the unique primers. Chromosome 1 contains the maximum number of indels (330,526), whereas chromosome 10 has the minimum (141,538). The highest indel density occurs on chromosome 1, at 1.09 per Kb; whereas the lowest is on chromosome 2, at 0.62 per Kb; average indel density is 958.79 indels/Mb. The density of indels in various genomic regions follows, from highest to lowest: TSS_up_0.5Kb, TES_down_0.5Kb, 5'-UTR, 3'-UTR, intron, CDSs, and intergenic regions. Accordingly, 38.88 % of the indels are

located in genic regions, while 61.12 % are located in intergenic regions.

The rate of indel polymorphism varies by genomic region and is listed, from highest to lowest, as follows: TSS_up_0.5Kb, 5'-UTR, 3'-UTR, TES_down_0.5Kb, introns, intergenic regions, and CDSs. Indels in CDS regions not only have a low rate of polymorphism, but also have lower polymorphism information content (PIC) values; 49.52 % of indels within CDSs have PIC values lower than 0.1 (Fig. 1).



Frequency and distribution of indels in different genomic regions

The number of indels decreased as the PIC value increased, and the PIC value of 751,925 indels ranged from 0 to 0.1, accounting for 38.10 % of all indels. There were 264,214 indels with PIC values greater than or equal to 0.5, accounting for 13.39 % of all indels. The number of polymorphic allelic indels ranged from 2 to 107, with an average of 2.76; however, most of the polymorphic indels only had two or three alleles, 72.49 % and 15.63 %, respectively. Proportionally, the number of alleles is nearly the same in different genomic regions, although, CDS regions have a higher proportion of two allele polymorphisms than do other regions, accounting for 82.81 % of all two allele indel polymorphisms (Fig. 2). The size difference between the smallest and largest alleles varies from 1 to 211 bp in length, with the number decreasing as indel size increases. Indels of 1 bp in length account for 33 % of all indels, and indels of 2 bp length account for 14.08 %, while sizes smaller than 11 bp account for 84.53 %. Other than within CDS regions, the proportion of indels in other regions all reduces with an increase in the difference in length between polymorphic indel alleles (Fig. 3).

Primer design for highly polymorphic indel markers with large major allele differences

Primer pairs for 83,481 unique indels, all with PIC values greater than or equal to 0.5, and with major allele differences greater than or equal to 3 bp, were designed from the maize B73 genome. These indel loci primers were designed to generate PCR products with lengths of

60–100 bp, sizes that polyacrylamide gels can resolve (Additional file: 3). Primer pairs were also designed for 29,403 unique indels, with PIC values greater than or equal to 0.5, and with major allele differences greater than or equal to 8 bp. These indel loci primers were designed with PCR product lengths of 150–300 bp, sizes that agarose gels can resolve (Additional file: 4). The exact positions of these indel markers in the maize genome, as well as the primer sequences, amplicon lengths, PIC values, major allele differences, number of alleles, and number of e-PCR products in the maize genomes evaluated are presented in Additional files: 3 and 4. This data should prove useful in furthering maize genetic research by facilitating primer design in sequences with indels.

Experimental validation of indel accuracy and polymorphism

Indel accuracy was experimentally validated between the maize 1212 genome and the B73, Mo17, and Zheng58 genomes. With genomic DNA from 1212, B73, Mo17, and Zheng58 as templates, 100 indel loci were PCR-amplified. Of the 100 indel loci, 98 were readily amplified, and 89 of these indel loci were polymorphic between the 1212 and B73 genomes, with an accuracy of 90.82 %. Indel accuracy between 1212 and Mo17 was 90.90 %, while that between 1212 and Zheng58 was 89.80 % (Fig. 4).

Indel polymorphisms were also experimentally validated. We selected another 100 indel loci for PCR amplification. The PIC value of these indel loci in 345 maize genomes ranged from 0.50 to 0.80, with an average PIC value of 0.55, and the allele number ranged

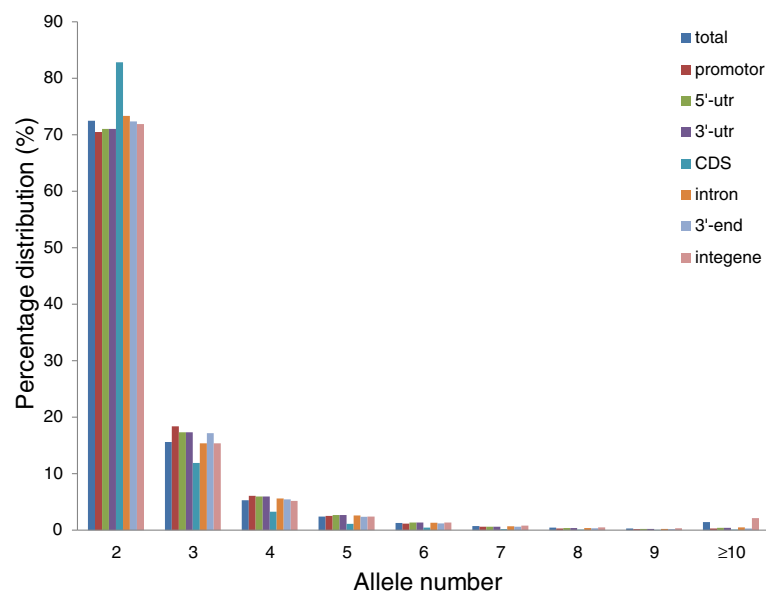


Fig. 2 Indel allele number distribution for different genomic regions

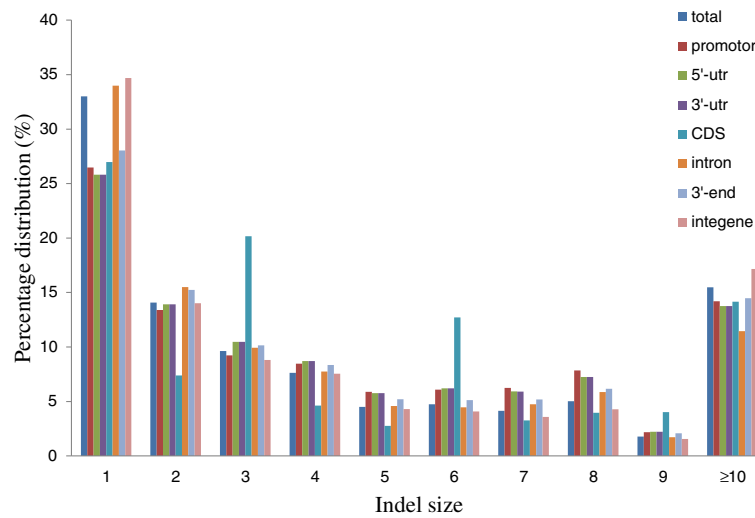


Fig. 3 Indel size distribution in different genomic regions

from 2 to 13, with an average of 3.50. The PIC value of the indel loci in 20 maize inbred lines ranged from 0.19 to 0.74, with an average of 0.54, and the allele number ranged from 2 to 5, with an average of 2.83 (Fig. 5, Additional file: 5).

Discussion

Next-generation sequencing technology can produce a huge quantity of DNA sequence data, which becomes a powerful tool for the discovery of high-density molecular markers. A large variety of indel identification software with the main objectives of optimal performance, sensitivity, and specificity is rapidly becoming available [27]. The e-PCR strategy we used identifies indels by aligning flanking sequences to indels rather than mapping complete reads to reference sequences containing indels. This can effectively reduce the influence of indels on alignment. The results of this study show that indels can be efficiently and accurately identified using e-PCR *in silico* assays, which also save considerable time and laboratory costs over using traditional *in vivo/vitro* approaches.

The development of high-density molecular markers significantly increases the efficiency of map-based cloning and marker-assisted selection. In this study high-density indel markers, widely distributed across the

maize genome, were developed at an average density of one indel per 0.96 Kb, significantly higher than the 0.01 per Kb density developed by Lai et al. [19] in their maize study, and the 0.14 per Kb density used by Mills et al. in their human genome study [1]. Our indel loci reside largely within intergenic regions (1,206,314, 61.12 % of the total), compared with genic regions (767,432, 38.88 %) (Table 1). Indel markers within genes are genic or functional markers [28]. Functional markers are superior to random DNA markers, such as restriction fragment length polymorphism (RFLP), simple sequence repeat (SSR), and amplified fragment length polymorphism (AFLP), because functional markers are completely linked with trait locus alleles [29]. Polymorphism among homologous indels is the basis for developing indel markers. Our experimental results show that the average PIC value and allele number of *in vitro* experimentally developed indel primers are both lower than the *in silico* primers used in e-PCR. This is primarily caused by the difference in resolution between agarose/polyacrylamide gels and e-PCR. We also note that the use of short amplicons reduces the opportunity for the formation of secondary structure, and minimizes length-dependent differential amplification. PCR products were effectively separated and easily scored after 30 min of electrophoresis on 6 %

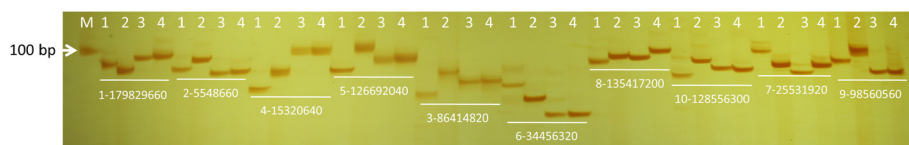


Fig. 4 Indel accuracy experimental validation. PCR products from lines 1 through 4 (1212, B73, Mo17, and Zheng58); The numbers below the horizontal line indicate the chromosome and locus of the indel. M: Marker DL2000

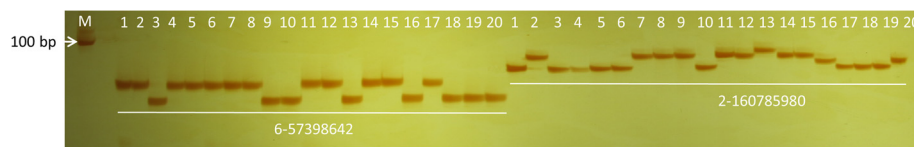


Fig. 5 Indel polymorphisms experimental validation. PCR products from lines 1 to 20 are Y0921, JD7275, RP1282, LH8012, 5220-2, 2054, 9HT1736, F19, Sn811, QA356, SC17931, 9HT1736, 9LB050, JD7275, liao1478, mian04185, SCML103, H1277, Qi31912, and Y1035; The numbers below the horizontal line indicate the chromosome and locus of the indel. M: Marker DL2000

polyacrylamide gels owing to our use of short amplicons and indel to amplicon size ratios larger than 3 %.

The average rate of indel polymorphism is 18.68 % across the maize genome, with TSS_up_0.5Kb regions having the highest polymorphism rate, whereas the CDS regions have the lowest. Intriguingly, polymorphism levels in genic regions of the maize genome were higher than that in intergenic regions. This may be caused by differences in complexity between genic and intergenic regions, as intergenic regions in the maize genome are very rich in repeat sequences. This makes sequence analysis very difficult. The majority of indels in CDS regions had lengths divisible by three, which is a direct result of selection against frameshift mutations (Fig. 3).

The accuracy and identification of unique loci are critically important for developing molecular markers. The predominant error associated with Illumina NGS platforms are substitution errors, which have relatively little effect on indel identification [30]. PCR-based experimental validation shows that indel markers have an accuracy of approximately 90 %. 97.05 % of 20 maize inbred lines contained PCR products for those primers we synthesized for polymorphism validation, which also showed that these indel markers were highly effective. Those PCR products inconsistent with e-PCR results may occur due to non-specific amplification. An increase in read depth per locus can be used to improve the accuracy of indel identification. Moreover, alignment parameters can be set to reduce non-specific amplification by increasing the number of mismatches in the identification of unique loci in the genome, and accuracy can be improved by decreasing the number of mismatches in the analysis of variation between populations. Gel electrophoresis results show that the actual PCR products of indel markers have fewer non-specific bands, which suggests that e-PCR can be a powerful tool for reducing non-specific amplification.

Conclusions

Here we report a large-scale analysis of genome-wide indel polymorphisms among maize populations, including inbred lines from different stages of breeding history and landraces. Most of these populations are the parents of the commercial hybrid and key lines in today's global

germplasm pool. The indel markers developed in this study provide a simple and efficient tool for any laboratory focusing on map-based gene cloning and molecular marker assisted breeding in maize.

Methods

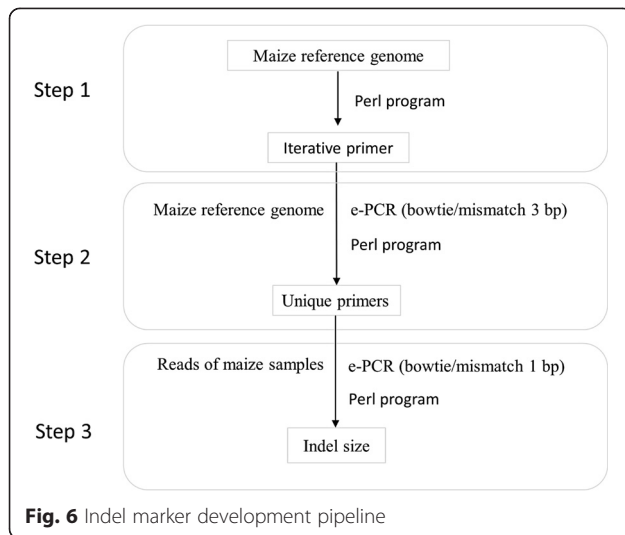
Maize genome sequence sources and plant materials

The genome sequences for B73 (Release ZmB73_Ref-Gen_v2) and Mo17 (454 pyrosequencing data) were downloaded from <http://ftp.maizesequence.org/release-5b/assembly/> and <http://www.phytozome.net/maize.php>, respectively. The 5'-UTR, CDS, 3'-UTR, exon, intron, and intergenic regions were determined based on database annotation (ZmB73_5b_FGS, <http://ftp.maizesequence.org/release-5b/filtered-set/>). Our TSS_up_0.5 Kb regions are defined as those sequences from 0.5 kb upstream of to transcription start sites, and TES_down_0.5Kb regions are defined as those sequences from transcription end sites to 0.5 kb downstream of them. The re-sequencing data of 344 maize samples were downloaded from National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov/sra?term=SRA049859> and <http://www.ncbi.nlm.nih.gov/sra?term=SRA051245>). We determined the genotype of the Chinese waxy maize inbred line 1212, at approximately 10-fold coverage, using Illumina HiSeq 2000 genome sequencing technology.

Twenty-four maize inbred lines were used to test for indel primer accuracy and to validate indel polymorphism levels. This included three elite maize inbred lines (B73, Mo17, and Zheng58), one Chinese waxy maize inbred line (1212), and twenty new maize inbred lines (Y0921, JD7275, RP1282, LH8012, 5220-2, 2054, 9HT1736, F19, Sn811, QA356, SC17931, 9HT1736, 9LB050, JD7275, liao1478, mian04185, SCML103, H1277, Qi31912, and Y1035).

e-PCR primer design and unique locus primer identification

The development of indel markers consists of three steps (Fig. 6). The first step is e-PCR primer design. Using the maize B73 reference genome as a template, 20 bp sequences extracted via a Perl script were used as upstream primers for e-PCR, and after intervals of 20 bp, the reverse complements of another 20-bp sequence were used as



downstream primers. The next pair of primers was designed the same way, but 20 bp away from the beginning of the previously designed primer, so that the primers covered the entire genome (step 1 in Fig. 6; Fig. 7). The second step removes redundant/ambiguous primers. All primer sequences were then aligned to the maize B73 reference genome using Bowtie software, an ultrafast and memory efficient tool for aligning short DNA sequences to large sequence databases [31]. Default parameter values were used, except for the number of allowable mismatches between primer and genomic sequences (we used $n=3$). Those primers that mapped to multiple positions were eliminated (step 2 in Fig. 6).

Indel variation among maize populations

The third step is to align all of the valid primer sequences identified in the previous step to short sample reads, and to estimate the length of the resultant e-PCR products (Fig. 6, step 2). The NGSQC toolkit (v.2.3.3) was first used to filter all of the raw data for high-quality reads using a quality score of 20 and

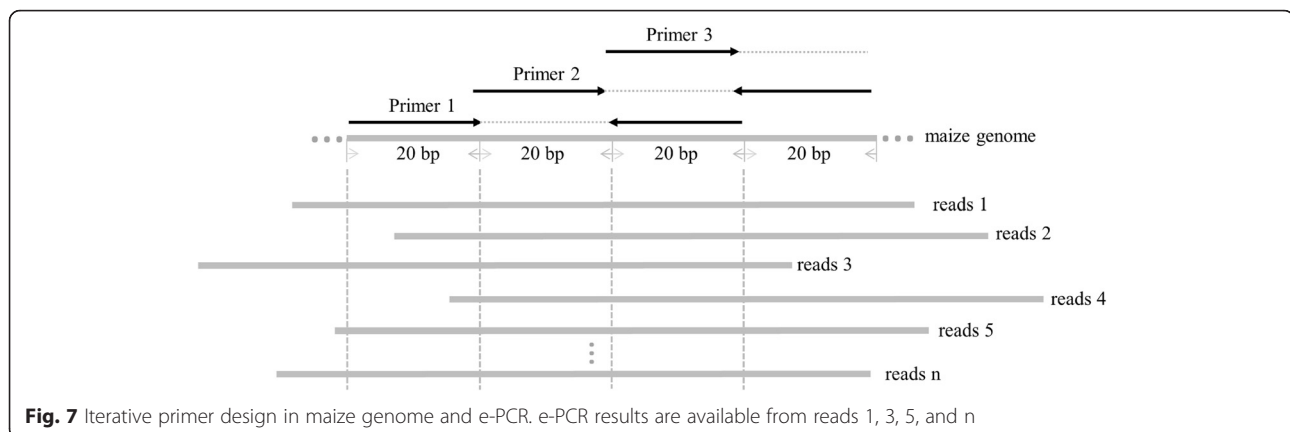
above for our cut-off [32]. Unique primer sequences were then aligned against filtered reads from 344 maize sample template sequences via Bowtie, using default parameters, except only allowing one mismatch (step 3 in Fig. 6; Fig. 7). The amplicon length of each e-PCR primer was extracted using a Perl script, and the most frequent length was selected when there were data of various lengths. The allelic diversity of each indel locus that had length information results from ≥ 20 genomes was assessed using the PIC value, which was defined as $PIC_i = 1 - \sum_{j=1}^n p_{ij}^2$, where p_{ij} is the frequency of the j th pattern for the i th marker [33].

PCR primer design

Unique indel loci were selected for PCR-based primer design. Sequences of 100 bp, including a 20-bp variation region and two 40-bp flanking sequences on each side of the locus were used to automatically design primers with Primer3 [34]. The following parameters were employed: a primer length range from 20 nt to 28 nt, with a 23-nt optimum; a thermal melting temperature (T_m) of 60 °C to 65 °C, with an optimum temperature of 63 °C, and primer pairs must have similar T_m values; a GC content of around 50 %, ranging from 30 to 70 %; and an expected product size of 60 to 90 bp ending with G- or C-rich region at the 3'-end.

Experimental validation

One hundred pairs of primers evenly distributed on maize chromosomes, all of which had 3–10-bp size differences simultaneously in the genomes of maize inbred lines between 1212 and B73, Mo17, and Zheng58, were selected randomly. Then, DNA from 1212, B73, Mo17, and Zheng58 were used as the template for PCR amplification to validate the accuracy of the indel primer design. One hundred pairs of primers (with PIC values ≥ 0.45 and major allele differences ≥ 3 bp) were selected, and DNA from twenty new maize inbred lines was used as the template for indel polymorphism validation.



Genomic DNA was extracted from 2-week-old seedlings using a modified CTAB (cetyltrimethylammonium bromide) DNA extraction protocol [35]. PCR was performed in a reaction mixture of 15 μ L, containing 50 ng of total genomic DNA as a template, 1.5 μ L 10 \times buffer (Mg²⁺), 2.0 μ L dNTP (2.5 mM), 100 nM of each primer, 2 U Taq polymerase, and ddH₂O. A C1000 thermal cycler (Bio-Rad, Inc., Hercules, CA) was used for the amplification with the following protocol: an initial denaturation for 3 min at 95 °C, 35 cycles of denaturation for 30 s at 95 °C, annealing for 90 s at 55 °C, with an extension for 90 s at 72 °C; and a final extension for 10 min at 72 °C. PCR products were electrophoresed on a 6.0 % polyacrylamide gel. The PIC value for each marker was calculated using the formula previously described.

Additional files

Additional file 1: List of inbred lines sequenced. (XLSX 39 kb)

Additional file 2: Genome differences between varieties and B73. (XLSX 37 kb)

Additional file 3: Polymorphic indel markers (3 bp). The indel markers had PIC values ≥ 0.5 , and major allele differences ≥ 3 bp. (XLSX 17940 kb)

Additional file 4: Polymorphic indel markers (8 bp). The indel markers had PIC values ≥ 0.5 , and major allele differences ≥ 8 bp. (XLSX 6664 kb)

Additional file 5: Indel marker experimental validation. (XLSX 16 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JL conceived the project. QJT performed the bioinformatics analyses. JWL, DGT, HL, and CY tested the indel markers. JL, TZR, and JWL prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge Haopeng Yu and Bin Wei for technical help. Financial supported was provided by the National Basic Research Program of China (the "973" project, 2014CB138203) and National Natural Science Foundation of China (31101161).

Received: 5 November 2014 Accepted: 24 July 2015

Published online: 13 August 2015

References

- Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 2006;16(9):1182–90.
- Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science.* 1994;266(5182):66–71.
- Collins FS, Drumm ML, Cole JL, Lockwood WK, Woude GV, Iannuzzi MC. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science.* 1987;235(4792):1046–9.
- Roberts PS, Chung J, Jozwiak S, Dabora SL, Franz DN, Thiele EA, et al. SNP identification, haplotype analysis, and parental origin of mutations in TSC2. *Hum Genet.* 2002;111(1):96–101.
- Trappe R, Laccione F, Cobilanschi J, Meins M, Huppke P, Hanefeld F, et al. MECP2 mutations in sporadic cases of Rett syndrome are almost exclusively of paternal origin. *Am J Hum Genet.* 2001;68(5):1093–101.
- Ketterling RP, Vielhaber EL, Lind TJ, Thorland EC, Sommer SS. The rates and patterns of deletions in the human factor IX gene. *Am J Hum Genet.* 1994;54(2):201.
- Clark RM, Wagler TN, Quijada P, Doebley J. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat Genet.* 2006;38(5):594–7.
- Ashikari M, Sakakibara H, Lin S, Yamamoto T, Takashi T, Nishimura A, et al. Cytokinin oxidase regulates rice grain production. *Science.* 2005;309(5735):741–5.
- Fu D, Uauy C, Distelfeld A, Blechl A, Epstein L, Chen X, et al. A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science.* 2009;323(5919):1357–60.
- Wu D-H, Wu H-P, Wang C-S, Tseng H-Y, Hwu K-K. Genome-wide InDel marker system for application in rice breeding and mapping studies. *Euphytica.* 2013;192(1):131–43.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science.* 2009;326(5956):1112–5.
- Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science.* 2009;326(5951):433–6.
- Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet.* 2010;42(12):1053–9.
- Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci.* 2013;110(2):453–8.
- Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet.* 2010;42(11):961–7.
- Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet.* 2012;44(1):32–9.
- Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol.* 2012;30(11):105–11.
- Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature.* 2012;490(7421):497–501.
- Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet.* 2010;42(11):1027–30.
- Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, et al. Genome-wide genetic changes during modern breeding of maize. *Nat Genet.* 2012;44(7):812–5.
- Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet.* 2012;44(7):803–7.
- Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res.* 2011;21(6):961–73.
- Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics.* 2009;25(17):2283–5.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.
- Schuler GD. Sequence mapping by electronic PCR. *Genome Res.* 1997;7(5):541–50.
- Neuman JA, Isakov O, Shomron N. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform.* 2013;14(1):46–55.
- Varshney RK, Mahendar T, Aggarwal RK, Börner A. Genic molecular markers in plants: development and applications. *Genomics-assisted crop improvement.* the Netherlands: Springer; 2007. p. 13–29.
- Andersen JR, Lübberstedt T. Functional markers in plants. *Trends Plant Sci.* 2003;8(11):554–60.
- Yang X, Chockalingam SP, Aluru S. A survey of error-correction methods for next-generation sequencing. *Brief Bioinform.* 2013;14(1):56–66.

31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.
32. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One.* 2012;7(2):e30619.
33. Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME. Optimizing parental selection for genetic linkage maps. *Genome.* 1993;36(1):181–6.
34. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.* 2000;132(3):365–86.
35. Allen G, Flores-Vergara M, Krasynanski S, Kumar S, Thompson W. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat Protoc.* 2006;1(5):2320–5.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

