High throughput genotyping of structural variations in a complex plant genome using an original Affymetrix® Axiom® array

Clement Mabire, Duarte Jorge, Aude Darracq, Ali Pirani, Hélène Rimbert, Delphine Madur, Valerie Combes, Clémentine Vitte, Sébastien Praud, Nathalie Rivière, et al.

High throughput genotyping of structural variations in a complex plant genome using an original Affymetrix® Axiom® array

Mabire Clément[2*], Duarte Jorge[1*], Aude Darracq[2], Ali Pirani[3], Hélène Rimbert[1,4], Delphine Madur[2], Valérie Combes[2], Clémentine Vitte[2], Sébastien Praud[1], Nathalie Rivière[1], Johann Joets[2], Jean-Philippe Pichon[1], Stéphane D. Nicolas[2]

*Two authors contributed equally to work

Authors Affiliation:

1 Biogemma - Centre de Recherche de Chappes, CS 90126, Chappes 63720, France

2 GQE – Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

3 Thermo Fisher Scientific - 3450 Central Expy, Santa Clara, CA, 95051, USA

4 Present adress : GDEC, INRA, Université Clermont Auvergne, 63000 Clermont-Ferrand, France

Corresponding authors: stephane.nicolas@inra.fr

Abstract

Background

Insertions/deletions (indels), and more specifically presence/absence variations (PAVs) are pervasive in maize and have strong functional and phenotypic effect by removing or modifying drastically genes. Genotyping of such variants on large panels remains poorly addressed, while necessary for approaches such as association mapping or genomic selection.

Results

We have developed a new high throughput and cost-effective tool to genotype indel. We first identified 141,000 indels by aligning reads from the B73 line against the genome of three temperate maize inbred lines (F2, PH207, and C103) and reciprocally. Next, we designed an Affymetrix® Axiom® array to target these indels with a combination of probes selected at breakpoint sites (13%) and/or within the indel sequence either at polymorphic (25%) or non-polymorphic sites (63%) sites. The final array design is constituted of 662,772 probes and targets 105,927 indels including PAVs, ranging from 35bp to 129kbp. After Affymetrix® quality control, we successfully genotyped 89,393 indels (84%) on 445 maize DNA samples with 479,027 probes (72%). A principal coordinate analysis on dissimilarity estimated from a subset of 57,824 indels on 362 inbred lines is consistent with the structure obtained using 50K SNP arrays.

Conclusions

We efficiently genotyped thousands of small to large indels on a large number of individuals using a new Affymetrix® Axiom® array. This powerful tool opens the way to studying the contribution of indels to trait variation and heterosis in maize. The approach is easily extendable to other species and should contribute to decipher the biological impact of indels at a larger scale.

Keywords

Present Absent Variation, Copy Number Variation, Structural Variation, genotyping, array, Zea mays, Genome assembly, Breakpoint, Chromosomal rearrangements

Introduction

In the past decade, there has been growing evidence that structural variations (SVs) are pervasive within plant genomes (Anderson et al., 2014; Beló et al., 2010; Cao et al., 2011; Liu et al., 2015; Owens et al., 2018; Saintenac et al., 2011; Saxena et al., 2014; Springer et al., 2009; Swanson-Wagner et al., 2010). Insertion/deletions (indels) are one class of SVs of particular interest since they lead to the presence or absence of, sometimes large, genomic regions at a given locus, among individuals from the same species. The content of these indels can either be present elsewhere in the genome, but can also be completely absent from the genome, in which case they are referred to as presence/absence variants (PAVs). Some indels carry entire genes or affect gene regulatory elements, and are thus likely to have a functional and phenotyping impact (Chia et al., 2012; Hirsch et al., 2014; Lu et al., 2015; Mace et al., 2013; Saxena et al., 2014). Hundreds to thousands of SVs, including PAVs and copy number variations (CNVs), have been discovered in several plant species, including wheat (Montenegro et al., 2017), rice (Zhao et al., 2018), *Arabidopsis thaliana* (Lu et al., 2012), Potato (Hardigan et al., 2016), pigeon peas (Varshney et al., 2017), and Sorghum (Shen et al., 2015). These results support the idea that one single reference genome cannot properly represent the complete gene set of a given species. There has been an increasing interest for building new individual genomes in complement to the reference genome, in order to better describe the genetic diversity within a plant species (Appels et al., 2018; Cao et al., 2011; Darracq et al., 2018; Hirsch et al., 2016; Jiao et al., 2017; Pinosio et al., 2016; Sun et al., 2018; Varshney et al., 2017; Zhou et al., 2017).

In maize, BAC sequence comparison first revealed that gene and transposable element content greatly vary between inbred lines (Fu and Dooner, (2002); Brunner et al. (2005)). Whole genome sequencing of the B73 inbred line then provided the opportunity to explore the extent of SVs across the entire maize genome (Schnable et al., 2009) by designing Comparative Genomic Hybridization (CGH) technology (Pinkel et al., 1998). Several CGH studies found multiple CNVs between the B73 reference genome and other maize inbred lines or teosintes (Beló et al., 2010; Springer et al., 2009; Swanson-Wagner et al., 2010). These studies demonstrated the large extent of SVs among maize inbred lines, including presence/absence variations of low copy sequences such as genes. This was well illustrated by the discovery of a large 2 Mbp presence/absence region between Mo17 and B73 carrying several genes (Beló et al., 2010; Hirsch et al., 2016; Springer et al., 2009; Swanson-Wagner et al., 2010). However, CGH array technology shows several major drawbacks since (i) it does not allow the discovery of sequences that are not present in the reference genome used for designing probes of the arrays, (ii) it has a limited resolution which does not allow detection of indels smaller than 1kb, and (iii) it is costly and labor-intensive, and therefore not adapted for genotyping several hundreds of individuals.

Methods based on SNP array experiments have been developed to detect CNVs and were shown to be more cost effective and with higher throughput but to reduce breakpoint resolution than CGH arrays (Cooper et al., 2008; Dellinger et al., 2010 Wang et al., 2017). Didion et al. (2012) identified atypical patterns of reduced hybridization intensities that were highly reproducible, so called “off-target variants” (OTVs). OTV patterns could originate either from the absence of the sequence due to a PAV polymorphism, or to a single nucleotide polymorphism within the probe sequence, thus preventing the correct hybridization of the DNA sample. For instance, 45,974 OTVs were discovered in a maize population using the 600K Affymetrix® Axiom® SNP array (Unterseer et al., 2014). While these approaches proved to be useful, there is a strong risk of false positive detection of PAVs using OTV patterns, mainly because these arrays were not designed to target PAVs. In order to reduce this risk of false positive detection of PAVs and more largely CNVs, several methods based either on segmentation or Hidden Markov Chain have been developed to use variation of fluorescent intensity signal of contiguous probes along the genome (Hupe et al., 2004; Olshen et al., 2004; Picard et al., 2007, 2005, Marioni et al., 2006; Stjernqvist et al., 2007). These kind of approaches have been used on 600K Affymetrix® Axiom® SNP array to detect CNV and to explore the contribution of CNV to phenotypic variation (Lyra et al., 2018).

With the emergence of massive parallel sequencing, new methods have been developed to detect structural variations based on the alignment of resequencing reads onto a high quality reference genome sequence. Among these, three have been mainly used (Alkan et al., 2011): (i) the "read-depth" (RD) method which can only detect copy number variations, (ii) the "read-pair" (RP) method which can detect deletions as well as small insertions (up to the size of the insert), (iii) the "split-read" (SR) method which can also detect deletions and small insertions (up to the size of a read). Chia et al. (2012) used the RD approach to identify CNVs among 104 maize lines and performed association studies for several traits. However, the RD method does not allow the identification of novel sequences and is error prone, especially regarding the size of the discovered CNVs which greatly depends on the size of the sliding window used. The RP method has been implemented in many computational tools like BreakDancer (Chen et al., 2009) and has been widely used. Although it has proven to be highly efficient to detect deletions (Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005), this approach suffers from two limitations: it does not allow precise detection of breakpoints and the size of the insertions which can be detected is directly limited by the library insert size. The SR method, which was first implemented in Pindel (Ye et al., 2009), has the advantage of defining breakpoints at a single base resolution, but again the size of the detectable inserted sequence is limited.

The "assembly" (AS) method is able to detect all types of SVs of any size, but is also the most cost and computation-intensive. It is the only method able to detect large insertions with precise breakpoint definition. However, the assembly of large and complex genomes such as maize remains very expensive and computationally intensive despite recent progress in this area (Darracq et al., 2018; Hirsch et al., 2016; Jiao et al., 2017). There has been in the past some attempts to reduce this complexity by reducing the number of sequences to assemble. For instance, Lai et al., (2010) identified 104 deletions and 570 insertions among 6 maize inbred lines by assembling genomic regions from reads that did not map on the B73 reference genome. The sequences assembled by this approach were enriched in erroneous reads or reads coming from external contamination and they were too short to be anchored to the reference genome B73. Hirsch et al. (2014) identified several putatively expressed genes that were not present within B73 reference genome by assembling and comparing the transcriptome of hundreds of inbred lines. This new approach was limited to the transcribed part of the genome and suffered from a high level of false positives. More recently, Lu et al., 2015 used genotyping by sequencing approaches on 14,129 inbred lines to identify 1.1 million short and unique sequences (GBS tags) that (i) did not align on the B73 reference genome or were aligned but outside of a 10Mbp windows around their mapped position, or (ii) were mapped at the same location by joint linkage mapping in NAM populations using co-segregation with SNP and logistic regression between indel and SNP in an association panel. The main drawback of this approach is the high percentage of missing data due to the low depth of sequencing which requires imputation before being able to make genetic analysis. Recent whole genome sequence assembly of PH207 (Hirsch et al., 2016), and F2 (Darracq et al., 2018) have allowed the identification of thousands of large indel and PAV sequences. For instance, 2,500 genes were found either present or absent in PH207 and B73 genomes and 10,735 PAV sequences larger than 1kb were discovered between F2 and B73, including 417 novel genes in F2. These discovery approaches have been limited to a few individuals due to sequencing costs and computational challenges, so they have not been adapted for characterization of SVs on large maize panels. Darracq et al. (2018) developed an interesting approach for the genotyping of PAVs from mapping of low depth (5-20X) resequencing datasets. This method is based on the comparison of reads aligning to the region found in F2 and in the line of interest. While this method is potentially adapted to genotype PAVs on any set of line with low resequencing data, it has been so far used for PAV genotyping on a low (<30) number of maize lines. Moreover, it is restricted to the analysis of PAVs, and is not adapted for genotyping other types of SVs.

To our knowledge, no high-throughput genotyping approach has been developed for genotyping large numbers of indels, including PAVs, on a large sets of individuals. In this study, we present an approach which is both (i) comprehensive, as it includes the discovery and localization of deletions as well as insertions regarding the B73 reference genome at the base pair level and (ii) high throughput, as it allows to genotype thousands of indels on hundreds of individuals. Our strategy takes advantage of next generation sequencing (NGS) technologies and recent advances in assembly of complex genomes. It also benefits from the high efficiency of SNP arrays like the high-throughput Affymetrix® Axiom® technology. In this paper, we detail how we discovered thousands of small to large indels, including PAVs, from three maize inbred lines (F2, PH207 and C103) as compared to the B73 reference genome. We then describe how we designed and selected 600,000 probes to create a new Maize Affymetrix® Axiom® array to genotype these indels. Finally, we describe how we successfully used this array to genotype an association panel of 362 maize inbred lines.

Results

Indel and PAV discovery

To design a comprehensive indel genotyping array, we first needed to discover a set of indels which would be representative of the maize temperate germplasm. We already had access to sequence data for the European flint line F2 and we benefited from a first set of 42,330 F2-specific sequences larger than 150pb, and totaling 16Mb. This dataset was constituted from the *de novo* assembly of F2 paired-end that failed (at least for one read of the pair) to align onto the B73 AGPv2 sequence and which were totally devoid of coverage by B73 reads (“Reference guided assembly” in Figure S2 so called “no map” approach). We also took advantage of the work done by Darracq et al., 2018 to add another 10,044 F2-insertions (size >1 kb, total size of 88Mb) with less than 70% of their length covered by B73 reads.

To complement these two datasets of F2/B73 deletions and insertions, we generated Illumina® paired-end and mate-pair sequences from two other key founders of temperate maize breeding programs: PH207 and C103. We then used these F2, PH207 and C103 sequence data to detect, not only PAVs this time, but all indels, at base pair resolution between these three lines and B73. This methodology allowed us to have access both to their sequences and their breakpoints allowing to genotype such indels in several individuals (See material and methods for more details).

We first aligned F2, PH207 and C103 sequences against the B73 reference genome sequence in order to detect deletions. Here, the term “deletion” does not reflect any underlying biological process of DNA excision, but refers to a sequence of at least 100bp present in the B73 genome at one locus and absent in another line at the same locus. Deletions were detected for the three lines simultaneously using the “genotyping” option of Pindel (Ye et al., 2009), generating a set of 26,368 non-redundant deletions with precise identification of their breakpoints (Figure S1 A). The number of deletions found for each line was similar, respectively 12,165, 11,922 and 13,432 for F2, PH207 and C103. 67% of the deletions found were unique to one line, 24% were shared by two lines and 9% by three lines. These results confirm the good complementarity of the lines chosen in this study.

Next, we generated a draft genome assembly for each of these lines, which were used as template for alignment of B73 reads to detect insertions regarding B73 reference genomes. As for deletions, here the term “insertion” does not reflect any underlying biological process of DNA integration, but defines a sequence larger than 100bp that is present in one line at a given locus, and absent from B73 at the same locus. These three draft assemblies cover less than one third of the expected maize genome size but include a large portion of low copy sequences, including genes, as shown by BUSCO results (Table 1). Detection of insertions was processed this time separately for each inbred line, and generated 28,221 insertions for F2, 27,904 insertions for C103 and 26,795 insertions for PH207, with their precise breakpoints. The number of insertions is similar between lines, but significantly greater than this obtained for deletions. Among these insertions, 26,691 cases could be uniquely anchored at base pair resolution onto the B73 reference genome sequence (Figure S1b). Again, a majority of insertions were unique to one line (72%) confirming the complementarity of the material chosen.

Finally, the results from the different approaches were merged into a non-redundant set of 141,325 indel sequences (see material and methods) comprising 52,175 deletions and 89,150 insertions. These regions were then used for the design of genotyping probes.

Design of the genotyping array

Genotyping strategy

Large indels can be efficiently genotyped with a SNP array using a combination of two types of probes: (i) "external" probes which target breakpoints using the two flanking sequences of a given indel (BP probes) and (ii) "internal" probes which target presence/absence regions (PARs) within the internal sequence of indels on polymorphic (OTV probes) or monomorphic sites (MONO probes). We define PARs as small portions of DNA sequence of at least 35bp that were observed present or absent at the genome level when comparing two individuals. They are thus suitable for the design of presence/absence genotyping probes. Ideally, each indel should be called by two BP probes on either side and by multiple internal probes regularly distributed along the internal sequence of the indel (Figure 1 A). However, in practice, this combination of different probes is not always possible. For instance, precise breakpoints were not described for all PAVs from our "No-map" approach and Darracq et al., 2018), and PARs for internal probes were not always found in our indels.

Probe design

On one hand, BP probes, which should behave like classical SNP probes where one allele corresponds to the presence and the other to the absence of the indel. They are useful to explore the conservation of the localization of large insertion/deletion events across multiple individuals, even when no internal probe can be designed due to the absence of PARs (Figure S6). Among the 141,325 selected variants, 86,406 indels (22,420 deletions and 63,986 insertions as compared to the B73 reference genome sequence) had breakpoints defined at base pair resolution and were suitable for BP probe design. Four different breakpoint types were identified according to the presence of micro-homology and/or shorter non homologous sequence (Muñoz-Amatriaín et al., 2013) in place of a complete deleted sequence (Figure S3): (type I) 3,397 cases with sharp breakpoints; (type II) 45,987 cases with a micro-homology sequence (8.6 bp on average and no more than 237 bp) which was present in one copy in the reference sequence and duplicated at both extremities of the novel inserted sequence; (type III) 36,893 cases harboring insertion of a short non-homologous fragment (42.2 bp on average and up to 892 bp) in place of a large deleted sequence; and (type IV) 156 cases with a combination of type II and type III breakpoints. Following Affymetrix® recommendations, 19,010 indels with type II breakpoints having a micro-homology sequence longer than 5bp were excluded from the design process. In the end, 67,396 indels, representing 48% of all available indel variants, were submitted to the Affymetrix® design pipeline. Two probes, one on forward (FW) and one on reverse (REV) strand, were designed for each breakpoint. These probes were classified as *not possible* (18%), *not recommended* (33%), *neutral* (15%) and *recommended* (35%) by this automated pipeline (see Methods for details), leaving 33,430 indels (51%) that could be targeted by at least one *recommended* probe.

On the other hand, internal probes, which should behave like an “off-target” variants (Didion et al., 2012) where the hybridization of the probe stands for the presence and the absence of hybridization for the absence of the indel, are useful to explore the genetic diversity within indel sequences (Figure 1 D). They will also be particularly interesting to target indels for which no breakpoint could be identified (such as PAVs from the “no map” approach).

For the design of OTV probes, we benefited from the availability of SNPs which had been previously identified from the alignment of resequencing data from a core collection of 25 temperate maize inbred lines against the B73-F2 maize pan-genome from Darracq et al. (2018). As a consequence, OTV probes have only been designed for deletions positioned on B73 reference genome and F2 insertions coming from Darracq et al. (2018). Among these, the context sequences of 436,162 SNPs, corresponding to 21,390 indels, were extracted and submitted to Affymetrix® design pipeline. Again, two probes, one on forward (FW) and one on reverse (REV) strand, were designed for each SNP. Finally, a total of 872,324 OTV probes could be designed and scored as *not possible* (0.05%), *not recommended* (71%), *neutral* (14%) and *recommended* (16%), leaving 17,589 indels (82%) which could be targeted by at least one *recommended* probe.

For the design of BP and OTV probes we could rely on Affymetrix® design pipeline to identify probes localized in PARs and thus suitable for the Affymetrix® Axiom® technology. For the design of MONO probes, we first had to identify such PARs within 141,325 indels cumulating 133Mbp of sequence. We used sequence masking methods to exclude repeats based on similarity to known maize repeats or on occurrence of 17-mers found within the sequencing datasets we had for B73, F2, PH207 and C103 (more details in methods). By doing so, we identified 122,972 PARs, representing a cumulated size of 27Mbp, corresponding to 20.3% of the initial size and allowing the possibility to design MONO probes for 79,987 indels (56.5%). These PAR sequences were successfully used for the design of 25,735,797 MONO probes, among which 59% were scored as *recommended* and allowed to target 62,875 indels (79%).

With this combined approach, we designed a total of 26,715,361 probes targeting 117,756 indels, which represent a cumulated length of 250 Mbp including 27 Mbp of PARs (Table 2). Among these indels, 97,748 (83%) can only be targeted with either internal or external probes, but not both (Figure 3 A). These results support our overall strategy which includes the discovery of indels with precise breakpoints in a preliminary step, and the use of complementary internal/external probes for the genotyping of large indels.

Array design

We used the Affymetrix® recommendations to select the 700,000 probes to be included in the final array, plus some other criteria depending on the probe type. Nevertheless, because of their added value, we decided to keep all BP probes as soon as they had less than 3 hits on the B73 reference genome sequence. This first selection consumed 84,994 probes targeting 53,456 indels, among which 70% could only be targeted by BP probes. Concerning OTV and MONO probes, we first selected *neutral* and *recommended* probes having no hit at all (for insertions), and only one hit (for deletions), against the B73 reference genome sequence. We then considered their density with the objective to maximize the number of indels that could be surveyed, as well as to have an even distribution of probes along targeted indel sequences (see Methods for more details). We then performed a second selection among *not recommended* OTV and MONO probes for 4,541 indels that were still not targeted. After filtering some duplicated probes, we built a final array design containing 662,772 probes targeting 105,927 indels that represent a cumulated length of 232 Mbp, including 25.9 Mbp of PARs.

Description of the array content

The final array design allows to genotype indels with various sizes, ranging from 37 bp to 129.7 kbp, with a median of 501 bp (Figure S4). They are covered by 1 to 482 probes with a median of 3 probes per indel (Figure S5). The number of probes does not always reflect the length of the indels, as the proportion of PARs within indels is highly variable. Indeed, while 8,040 indels (ranging from 37 bp to 2,409 bp with a median of 163 bp) were completely covered by PARs and could thus be considered as a proper PAVs, 34,372 indels (ranging from 101 to 129,700 bp with a median of 320 bp) were not covered by any PAR at all (Figure 2). In fact, the number of internal probes were more strongly correlated to the size of the PARs (r2 = 0.79) rather than to the size of the indels (r2 = 0.16) (Figure S6).

As expected, the probe selection process did not impact the overall distribution of probe types among targeted indels as 35% of them can exclusively be genotyped by BP probes, whereas 50% can only be genotyped thanks to the use of internal probes, among which 73% are only targeted by the use of the original MONO probes (Figure 3b). Indeed, a large number of indels did not contain PARs and cannot be genotyped with 35bp internal probes but only with BP probes whereas some others indels contains PARs but have not BP due to Indel discovery approach (“No map”).

Among the 43,117 indels that could be anchored onto the B73 reference genome sequence and which were included in the array design, 13,737 were located inside a gene, 57 close to a gene (less than 1 kb away), 1,311 inside a pseudogene and 2,212 inside a transposable element. From the localization of these indels, evaluated indels and probe density across each chromosome. We observed a higher density in chromosome arms than in peri-centromeric regions (Figure S7). We also identified clusters of indels with large specific sequence at the beginning of chromosome 6 (10-20Mbp) or at the end of chromosome 5 (~190Mbp).

Assessing array quality by genotyping 105,927 indels on 480 maize DNA samples

Indel calling using dedicated Affymetrix® pipelines

We genotyped 480 maize DNA samples including 440 inbred lines, 24 highly recombinant inbred lines and 16 F1 hybrids. Dedicated Affymetrix® pipelines were implemented for each of the probe types to call genotype of the indels based on fluorescent intensity and contrast variation of the probes. It included two algorithms already developed by Affymetrix® (Didion et al., 2012) for BP and OTV probes (Figure S8 A et B) and a third one which was newly developed for the calling of presence/absence alleles using MONO probes (Figure 4). 35 DNA samples including all F1 hybrids, did not pass Affymetrix® quality control due to their low call rate (<0.9) and were eliminated. Out of 662,772 probes, 479,027probes representing 89,393 indels (84%) passed Affymetrix® quality control and were called on 445 DNA samples. Respectively 55%, 59% and 81% of BP, OTV and MONO probes were converted into recommended markers after clustering by Affymetrix® pipelines (Table S1, S2, and S3). Thanks to the 3 probe types and redundancy, 84% of indels could be called with an average of 5.4 probes per indel.

To evaluate the genotyping capacity of the probes, we first compared the clustering of inbred lines expected for three probe types (BP, OTV, and MONO) with the observed clustering of inbred lines based on fluorescence intensity and contrast of 445 inbred lines genotyped with the array. For BP probes, we expected at least two clusters corresponding to the individuals homozygous either for presence ("AA" or "BB") or absence ("OO"). A third cluster could be observed when individuals were heterozygous individuals for presence/absence ("OA" or "OB" hemizygous) (Figure 1 C). For OTV probes, we expected at least 3 different clusters: two cluster corresponding to the individuals homozygous for allele A or B of SNP ("AA", "BB"), and a third "off-target" cluster for the individuals homozygous for absence ("OO"). A fourth cluster could be observed when some individuals were heterozygous at the within-indel SNPs (AB). For MONO probes, we expected only two clusters corresponding to the individuals for which the sequence was present ("AA" or "BB") or absent ("OO ", "AA" or "BB") (Figure 1 C). The observed clustering by the three dedicated pipelines was consistent with the expected clustering for 43% of BP, 83% of OTV and 63% of MONO probes (Table 3).

We observed also some unexpected clustering. For 57% of BP probes, we observed an additional off-target cluster (OTV in Table 3). This indicates that some BP probes did not hybridize properly in some inbred lines, which can either be due to the presence of polymorphism within flanking sequences of the targeted indels or to the existence of more complex rearrangements removing the breakpoints. To explore these two hypotheses, we took advantage of the availability of forward (FW) and reverse (REV) probes for 12,150 indels to determine whether the clustering between FW and REV BP probes from the same indel was similar or different. While 12% of these indels had their FW and REV BP probes classified identically either as OTV, 35% had their FW and REV probes classified differently (one as BP and the other as OTV).

Regarding MONO probes, 25% displayed additional cluster(s) when sequence were present suggesting the presence of a single nucleotide polymorphisms at this position. Among these, we were able to distinguish two types of clustering (Table 3). 4.7% of MONO probes exhibited a clustering similar to those observed for OTV probes suggesting that these MONO probes revealed really by chance a single nucleotide polymorphisms. In contrast, 20.4% of MONO probes displayed an unexpected clustering pattern for inbred lines with the presence of a heterozygous cluster but absence of a second homozygous cluster for SNP (Figure S12 B). In the end, 2.8% of MONO probes displayed an additional heterozygous cluster for SNP when sequence is present but no "off target" cluster corresponding to individuals for which sequence are absent (Figure S12 D)

For 18% of OTV (Figure S12 A) and 8.3% of MONO probes, clustering displayed no "off target" cluster for absence suggesting no presence/absence polymorphism at this position (Table 3). Note that some BP were also classified as monomorphic for presence/absence but were filtered out by the BP pipeline (MonoHighResolution in Table S1).

Finally, 422,369 probes were able to call both presence and absence alleles, which allowed us to successfully genotype a total of 86,648 indels (82% of indels targeted by the array) on 445 inbred lines.

Evaluation of genotyping reproducibility and quality

Consistency of genotyping among the four inbred lines used for indel discovery

We used the 479,027 probes passing Affymetrix® quality controls to evaluate the quality of Presence/Absence genotyping by comparing the genotyping results from our array with those generated from sequencing data from the 4 lines used for the discovery of indels (B73, F2, PH207, and C103). Respectively 97.5%, 92.7% and 90.3% of the BP, OTV and MONO probes predicted a genotyping result consistent with this obtained with BLAST. We observed a strong asymmetry for concordance rate depending on whether we expect the locus to be present or absent from sequencing data (94.9% vs 86.2% for allele present and absent, respectively). Interestingly, we observed no asymmetry for BP probes and a strong asymmetry for OTV and MONO probes for concordance rate (Table 4). The four inbred lines showed very similar concordance rates, F2 being the most concordant (97.9%). The median consistency rate of probes within indels remained relatively high and stable, around 90%, independently of the number of probes per indel (Figure S9).

Consistency among probes from the same indel

To estimate the consistency of different probes for typing a given indel, we analyzed genotyping results for 48,486 indels genotyped with at least two probes in a collection of 24 temperate inbred lines. Among these 24 lines, there are the four lines used to discover PAVs and the twenty used to discover SNPs within specific regions of Indels (Darracq et al., 2018). For each indel and each inbred line, we calculated the frequency of presence call over all probes. Frequencies of 1 (presence) and 0 (absence) indicated that all probes displayed consistent genotyping for the corresponding inbred line. Overall, 78% of these indels genotyping displayed an average allelic frequency for the presence allele of 1 or 0 meaning that all probes had a consistent genotyping results for calling the allele at both present and absent states, respectively (Figure 5). A total of 12,308 indels (25%) displayed only two states across the 24 inbred lines, corresponding to the presence or the absence of the sequence, while for 75% at least one inbred line had at least one inconsistent probe conducting to the presence of more than two haplotypes across 24 inbred lines. Some contradictory calls were repeatedly found across the 24 samples (Figure S10), thus suggesting that some between-probe inconsistencies could have biological origins rather than being calling errors.

To investigate the consistency between the forward (FW) and reverse (REV) BP probes, we compared the genotyping results of 8,116 indels having both FW and REV BP probes called on our 24 inbred lines. 33% of these indels have a consistent calling between their FW and REVs probes for all inbred lines. The proportion of indels displaying an inconsistent calling between the FW and REV probes for 24 lines varied according to the breakpoint type and their classification (Figure S11). We observed also more similar calling when both FW and REV probes had similar classification (BP-BP or OTV-OTV) than when they had different classification status (BP-OTV) (Figure S11 A). Altogether, these results suggest that some calling inconsistencies could come from polymorphisms in the flanking sequence while some other could be due to local rearrangements in the lines under genotyping as compared to the lines used for INDELs discovery.

Assessing array quality to call highly hemizygous individuals using BP

In order to evaluate our ability to identify individuals displaying hemizygous genotype (heterozygous for presence / absence of the sequence), we rescued for BP probes the genotyping of DNA samples for 12 F1 hybrid eliminated by Affymetrix® quality control due to their low call rate. This low call rate came mainly from inability of current Affymetrix® algorithms to identify hemizygous cluster for OTV and MONO probes and therefore to assign a genotype to hemizygous individuals. As a consequence, it strongly increase missing data for F1 hybrids only for OTV and MONO probes. We selected 20,370 BP probes classified as expected by the design (Table 3) to compare them with those expected from their 9 parental lines. 89% of observed homozygous alleles were consistent with expected genotyping results of F1 hybrids and 94% of observed hemizygous alleles were consistent with expected genotyping results.

Reproducibility

We evaluated the reproducibility of genotyping by comparing the genotyping results of 13 different inbred lines that were replicated in the experiment (Table S4). Note that these are not perfect biological replicates as they represent the same variety but come either from different seed lots or from different accessions. These replicates exhibited a genotyping difference varying from 0.6% to 5.2%. This is similar to the amount of inconsistencies obtained on the same material using a 50K SNP array (Ganal et al., 2011) suggesting that indel genotyping inconsistencies for replicates come mostly from seed lot divergences rather than genotyping errors (Table S4).

Application: Diversity analysis of 362 maize inbred lines panel

In order to evaluate the interest of this new array to analyze the contribution of indels to the genetic diversity, we analyzed 57,824 polymorphic indels among a subset of 362 out 445 inbred lines, representing a large genetic diversity and previously studied (Bouchet et al., 2013; Camus-Kulandaivelu, 2005). To give same weight to each indel in the diversity analysis, we selected one single probe per indel based on the probe genotyping quality (see methods).

We first used these indels to calculate the genetic distance between inbred lines and to perform Principal Coordinate Analysis (PCoA) (Figure 6 A). To compare our indel-based results to this of previously characterized SNPs, we displayed on this PCoA the genetic structuration of these 362 inbred lines as obtained from the Panzea 50K SNP array (Bouchet et al., 2013). The first axis showed good discrimination of European Flint from Corn Belt Dent and Stiff Stalk lines, while the second axis discriminated European Flint and Northern Flint lines. Overall, the clustering of individuals based on genetic distance estimated with indels (1-IBS) by PCoA was consistent with the genetic structuration obtained from SNPs. We observed that B73 and F2, that were used to discover the majority of indels, deviated from other inbred lines. We thus performed a second PCoA excluding B73 and F2 (Figure 6 B). The two PCoAs gave similar patterns.

Discussion

1. An original high throughput approach for genotyping indels

The comparison of whole genome sequence assemblies is in theory the best approach to identify precisely and exhaustively structural variations between two individuals (Darracq et al., 2018; Hirsch et al., 2016; Jiao et al., 2017, Sun et al., 2018). But even though great progress has been made recently in this area, whole genome assembly is still too costly, time consuming and computationally intensive to be applied to hundreds of individual considering the complexity of maize genome (Darracq et al., 2018; Gabur et al., 2018). Other whole genome approaches based on sequencing and alignment of reads, and using "read-depth", "read-pair" and "split-read" identification methods (Chen et al., 2009; Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005; Ye et al., 2009) were mostly limited to the identification of deletions (i.e. sequences absent compared to a reference genome). Liu et al., (2015) partially addressed the lack of insertions (i.e. novel sequences compared to a reference genome) in previous studies by the identification 1,973,746 indels. Although, among these a majority were very small (85% smaller than 11bp) and the use of PCR markers to genotype them was time-demanding, labor-intensive and costly at large scale level. In this paper we describe an original approach combining the accuracy of the detection of insertions and deletions using high coverage sequence data and multiple reference genome assemblies, along with the high-throughput and accuracy of SNP arrays. We further show that using this approach, we were able to design and use an innovative array which allowed for the first time to genotype accurately thousands of small to large insertion/deletion variants, including PAVs, on hundreds of maize individuals. We used different methods to compile 52,175 deletions and 89,150 insertions between three newly sequenced maize inbred lines (F2, PH207 and C103) and the maize B73 AGPv2 reference genome, among which 75% were included in our array. Contrary to older studies, we did not focus solely on PAVs, but we also included in our array many insertion and deletion events, even if they contained non-unique sequences, by targeting their breakpoints.

By designing probes directly on indel breakpoints for both insertions and deletions , our approach overcomes some of the limitations of CGH or SNP array based studies. To our knowledge none of the previous studies which have used an array technology for genotyping indels have specifically targeted such a high number of insertion/deletion breakpoints. Unterseer et al., (2014) genotyped 6,759 small deletions which were discovered by aligning reads of 30 inbred lines against B73 genome but it included no insertions. However, CGH and SNP arrays did not usually design probes to target breakpoints and detected indels by analyzing the variation of fluorescent intensity signals of ordered probes (Cooper et al., 2008; Dellinger et al., 2010 Wang et al., 2017). As a consequence, these technologies targeted exclusively low copy regions of the genome excluding indels containing repeats such as TEs as soon as their breakpoints were not included in design (Beló et al., 2010; Lyra et al., 2018; Springer et al., 2009). This is a strong drawback for maize and many other crops since a large part of their sequence is composed of transposable elements (Feschotte et al., 2002; Schnable et al., 2009) that may be highly variable between individuals (Liu et al., 2015; Morgante et al., 2007; Sun et al., 2018) and may impact phenotypes (Ducrocq et al., 2008; Salvi et al., 2007, 2002). Using BP probes allow to target Present/Absent Variation whose sequence were unique and not present elsewhere in the genome as well transposable elements whose their internal sequence could be present/absent at one locus but present elsewhere in the genome. Another advantage to genotype breakpoints is that we are almost certain to genotype the same mutational event across all individuals of the population because it is highly unlikely that two independent mutational events can lead to the same breakpoint. On the contrary, when we detected classically indels using CGH or SNP array, it is much harder to identify common indels among a population of individuals as we don't know precisely the breakpoint at base pair level. Genotyping breakpoint is also very cheap since only one or two probes by indel are required. Indel size is therefore no longer a limitation for genotyping using breakpoints in the contrary to SNP and CGH arrays which have limited resolution when they used fluorescent intensity variation (Alkan et al., 2011). The genotyping of breakpoints by sequencing is possible with a tool like Pindel (Ye et al., 2009) which has a genotyping mode, but at a much greater cost and with lower call rate compared to the use of an SNP array. Finally, breakpoint probes are codominant markers and allow to accurately genotype hemizygous individuals (Heterozygous for presence/absence) since their genotyping are based on fluorescent contrast rather than fluorescent intensity variation which are known to be more noisy as for MONO and OTV probes (Alkan et al., 2011).

Although the use of BP probes is clearly the simplest way to genotype indels using an SNP array, breakpoints are not always available (no maps approach discovery) or "designable" with 35bp probes, for instance when sequences of microhomology at breakpoint site were larger than 5bp. In order to genotype the 52,471 indels without breakpoints and explore the genetic diversity within indels, we also designed 577,778 internal probes both on monomorphic and polymorphic sites on PARs for both insertions and deletions. To genotype PARs in indel internal sequences using SNPs, we took advantage of the already available Affymetrix® algorithms to call Off-Target Variants (OTVs) which can detect variation of fluorescent intensity signals for a single probe (Didion et al., 2012) (Figure 1 C). This approach was used by Unterseer et al. (2014) who was able to detect 45,974 OTVs on a set of maize inbred lines using a 600K SNP array. Nevertheless, the array was designed in a classical way to target SNPs and there was no prior evidence that the probes called as OTVs would belong to real indels like in our approach. Additionally, detecting SNP in insertion required to assemble a pangenome combining common and specific sequence from different individuals in order to retrieve SNP by aligning reads from sequenced lines. In our case, the sole use of OTV probes would have conducted to the elimination of a lot of indels since 87,372 indels including 74,648 insertions had no known SNPs within their internal sequence. In order to avoid this ascertainment bias due to prior knowledge of the presence of SNPs, we designed 414,500 MONO probes on putative monomorphic sites within PARs of indel sequences. It permitted to genotype 38,134 supplementary indels that could be targeted neither by OTV or BP probes. This new type of probes required the development of a new algorithm in order to cluster individuals according to their fluorescent intensity variation only, to be able to assign a genotype to each individual (Figure 4). A limitation of current Affymetrix® algorithms to genotype indels using OTV and MONO probes is that they are currently unable to genotype hemizygous individuals. While it was not a strong issue for maize inbred lines (or individuals from autogamous species) that are mostly homozygous, it is a strong issue for individuals from allogamous species that are highly heterozygous. By improving the current Affymetrix® algorithms, it should be possible to identify hemizygous cluster according to fluorescence intensity for OTV and MONO probes. We observed indeed some clusters that seem badly interpreted as heterozygote for SNP although they correspond more probably to hemizygous individuals for OTV and MONO probes (Figure S12B, see below for more detailed discussion). Alternatively, other algorithms/software based on fluorescent intensity variation of either a single probe or several ordered probes exists and could be used to detect copy number variation and therefore hemizygote in individuals (Hupe et al., 2004; Marioni et al., 2006; Olshen et al., 2004; Picard et al., 2007, 2005; Stjernqvist et al., 2007).

2. Reliability of genotyping / calling results

Our approach provides a reliable and reproducible genotyping strategy for indels since (i) 91.5% of alleles called from probes are consistent with expected genotype from the resequencing data available for the 3 lines (F2, PH207, C103), (ii) 78% of indels genotyping had internal calls totally consistent between each other exhibiting either absence or presence for an inbred line, and (iii) the genotyping results were highly reproducible (94.8-99.4%) between biological replicates.

We observed a higher inconsistency between observed and expected calls for genotype “absent” than for genotype “present” with MONO and OTV probes, but not with BP probes (Table 4). This asymmetry between present and absent for consistency suggests a greater number of false positives in absent than present. We found that 20,574 indels were in fact totally monomorphic and present across all lines suggesting they represented false-positive indels coming certainly from regions which were not assembled in our draft genomes. Indeed, the probes targeting sequence regions present in one line but not assembled in their draft genome assembly, were falsely expected absent but they correctly hybridized with DNA and were called “present” on the array. This explains why the number of false positives was higher for B73, as all B73 absence genotypes were defined in comparison to draft assemblies, whereas for the other 3 lines absence genotypes were defined in comparison with the gold standard B73 genome sequence. The fact that we obtained a better result on OTV probes coming from F2 can be explained because we used only SNPs discovered on the B73-F2 pan-genome and not on other genomes. On the contrary, the fact that BP probes had similar consistencies for genotype “absent” and “present” could be explained because the BP probes were designed exclusively on B73 reference genome whatever we genotype insertions or deletions. One possible improvement to our approach to reduce the number of false-positive absences would be to not only align B73 reads onto each draft genome assembly but to align reads from each sequenced genomes on each other and against itself. This would have several benefits: (i) it would allow to discover even more indels and of better quality since each putative deletion discovered in one sample could potentially benefit from supporting reads from another sample, (ii) this would also simplify the identification of indels common to more than on genotype, and last but not least (iii) it would help to identify and eliminate false-positive deletions by the alignment of each sample on its own draft assembly.

Nevertheless, the use of incomplete draft genomes does not explain all discrepancies between expected and observed genotypes. First, these genotyping errors could also be due to a wrong clustering leading to assign incorrectly genotype “present” instead of “absent” for a subset of individuals. It was well exemplified by some MONO probes classified as SNP although the clustering pattern looks like a MONO clustering with a strong difference of fluorescence intensity between two clusters. It suggests strongly for the cluster displaying the lowest fluorescent intensity a wrong assignment of homozygous genotype for one of two SNP alleles (presence of sequence) instead of the assignment of the homozygous absence of the sequence (Figure S12 C). Similarly, the more detailed inspection of the clustering of MONO probes displaying unexpected cluster pattern (Table 4, figure S12 D) and OTV probes classified as SNP (Table 4, figure S12 A) suggest a wrong assignment of genotype for the cluster displaying the lowest fluorescent intensity since the clustering looks like MONO and OTV clustering. Second, the genome divergence within probe sequence for some inbred lines could conduct to group those individuals in an OTV cluster and therefore lead this time to the assignment of an absent allele even though the sequence is present for these lines.

Surprisingly, 4.7% of MONO probes displayed a classical OTV clustering suggesting that an unknown SNP was targeted by these probes by chance. These 15,690 new OTVs are very interesting since they were discovered by chance on a large set of 445 inbred lines. We could therefore expect that these OTV have no ascertainment bias which can be very useful for analyzing genetic diversity within indels carrying PARs regions. On the contrary, 20.4% of MONO probes displayed an unexpected clustering with one off-target cluster corresponding to absence of the sequence, one cluster corresponding to heterozygous inbred lines for SNP but only one homozygous cluster (Unexpected MONO 1 in table 4). Considering these “unexpected MONO 1” as true SNP would conduct to a density of SNP (1 SNP every 5 bp) which are not compatible with level of diversity observed in maize in different previous studies (Brandenburg et al., 2017; Gore et al., 2009). Deeper investigation of these MONO probes clustering showed for some probes that the cluster of heterozygous inbred lines displayed intermediate position for both intensity and contrast between two clusters homozygous for presence and absence of the sequence, respectively (Figure S12 B). It suggested strongly that these clusters of inbred lines assigned as heterozygous were in fact inbred lines carried only one copy of the sequence (hemizygous genotype). An alternative hypothesis to explain this unexpected pattern is the presence of divergent duplicated sequence leading to the presence of an artefactual heterozygous cluster for SNP corresponding to the presence of two paralogous sequences rather than one copy. This result suggests therefore that there is probably room to improve Affymetrix® algorithms in order to better identify additional clusters corresponding to the presence of hemizygous individuals for both MONO and OTV probes and therefore improve the quality of the genotyping of indels when using a SNP array.

These potential clustering errors as well as the bad design of some probes previously mentioned can explain that only 27% of indels displayed consistent genotype for presence/absence between all probes from same indels across the 24 inbred lines. Interestingly, some indels showed reproducible inconsistent genotypes for presence/absence across their probes in several inbred lines (Figure S10). It suggested that this pattern could not be due to random errors but could have instead a biological origin with possibly rearrangements having occurred several times within the same genomic region in some inbred lines. Following this hypothesis, Gu et al. (2008) observed two different types of rearrangement which could explain our observations: (i) rearrangements with an unique breakpoint in population and therefore common size between individuals conducting to two haplotypes in a population (ii) rearrangement with non-unique breakpoints scattered in a genomic region which conducted to several haplotypes. This hypothesis is also supported in our experiment by the 56% of BP probes classified as OTVs indicating that FW or/and REV flanking sequence did not well hybridize in all lines.

The development of a statistical approach to merge either *a posteriori* the calling results of independent clustering of individual probes or *a priori* the fluorescent intensity signal of successive probes within a indel could be interesting in order to improve the robustness of indel genotyping. This would have the advantage to limit the effect of genotyping errors due to a bad clustering and to reduce the noise in fluorescent intensity signals. It would also help to identify true different haplotypes representative of the complexity of a region in a population.

Finally, 72% of probes were converted into markers, a number which is comparable to other maize Affymetrix® Axiom® SNP arrays in comparison to 74.9% in Unterseer et al., (2014). Out of these, only 88% were really polymorphic for presence/absence. This conversion rate is not so bad considering that Affymetrix® Axiom® array analysis pipelines, which have been optimized for the detection of bi-allelic SNPs, are more sensitive to variations in fluorescent contrast (x-axis) compared to variations in fluorescent intensity (y-axis) which are known to be more noisy (Alkan et al., 2011; Didion et al., 2012). Moreover, we did not always followed Affymetrix® recommendations as we didn't filtered out probes with a bad design score.

To conclude, we developed a high-throughput and cost-effective indel genotyping array based on the indels discovered by sequencing on four inbred lines. It could be highly valuable to use more lines for the initial indel discovery step since our four inbred lines do not well represent the whole genomic diversity of maize, notably tropical lines. As a consequence, it could lead to ascertainment bias by reinforcing the differentiation of inbred lines genetically close to the four inbred lines used to discover indels (Clark et al., 2005; Ganal et al., 2011; Gouesnard et al., 2017) as we observed in our diversity analysis for lines close to B73 and F2. Several new individual maize genome assemblies are now available in the public domain and more and more could become available in the future. Our approach could easily be applied to these new genome assemblies to discover new indels on a larger set of inbred lines representative of maize diversity with the aim to design a new indel array. Although our arrays were not yet designed to genotype duplications and inversions, our approach could be easily extended to genotype breakpoints of inversions but required further development of pipeline for genotyping duplication using internal probes.

Material and Methods

Indel and PAV discovery

Three maize inbred lines, which are key founders of maize breeding program and originated from three different heterotic groups, have been selected for depth sequencing and indel discovery: the European Flint line F2 and two American dent lines, PH207 (Iodent) and C103 (Lancaster). DNA for genotyping were extracted from leaves following a NaBisulfite method modified from Tai and Tanksley (1990) and Dellaporta et al. (1983). For each inbred line, paired-end and mate-pair whole genome shotgun libraries were sequenced on Illumina® HiSeq 2000 platforms (Table S6). A data set of B73 paired-end reads (35x) was downloaded from the Sequence Read Archive (accession SRR404240).

For deletion discovery step, F2, PH207 and C103 paired-end reads were aligned against B73 AGPv2 genome sequence using novoalign version 3.01.01 (http://www.novocraft.com) (default parameters). Samtools (Li et al., 2009) version 0.1.18 was used to coordinate sort and retain reads with a mapping quality of at least Q30. Duplicated reads were eliminated using MarkDuplicate from the picardtools suite (http://broadinstitute.github.io/picard) version 1.48. Pindel (Ye et al., 2009) version 0.2.5a2 was ran in parallel on each chromosome to perform multi-genotype calling of deletions. Raw formatted results were converted to VCF (Variant Calling Format) standard format using the script Pindel2vcf. BreakDancer (Chen et al., 2009) was used in complement to Pindel but only for F2. Deletions shorter than 100bp were discarded. Deletions spanning a B73 assembly gap or located in regions prone to mis-assemblies such as telomeric, knob and centromeric regions, were also excluded from further analysis using IntersectBed BEDTools (Quinlan and Hall, 2010) version 2.16.1.

For whole genome sequence reconstruction of F2, PH207 and C103 inbred lines, paired-end and mate-pair reads were used together and assembled using ALLPATHs-LG (Gnerre et al., 2011) version R41008. For F2, the script CacheToAllPathsInputs.pl was used to cache the data to use for assembly: 100% of the non-overlapping 230bp insert paired end data set, 100% of the overlapping 170bp insert paired end data set, 30% of the non-overlapping 370bp insert paired end data set, and 100% of the 2.4kb insert mate pair data set. Indeed, only overlapping paired end reads are used by ALLPATHs-LG for building contigs, but the supplementary non-overlapping paired end reads for F2 was used for error correction. RunAllPathsLG was then run for all three genotypes using these optional parameters. For each assembly, the coverage of the gene space was evaluated using BUSCO (Waterhouse et al., 2018) version 3.0.2 using genome mode and maize species (-m geno -sp maize).

B73 paired-end reads were successively aligned to ALLPATHs-LG F2, PH207 and C103 genome sequence assemblies. The same tools and parameters used to call deletions against B73 genome were applied to detect B73 deletions against F2, PH207 and C103 genome sequences. For commodity, these B73 deletions will be reciprocally called insertions of F2, PH207 and C103 compared to B73 reference. Again, only insertions smaller than 100bp were discarded, but not the ones spanning assembly gaps as they were real assembly gaps (with approximate size inferred from paired reads average distance) and not “unsized” gaps like in B73 genome. When possible, insertions were anchored onto B73 AGPv2 genome sequence using a dedicated pipeline combining Megablast version 2.2.19 (Altschul et al., 1990) and Age version 0.4 (https://github.com/abyzovlab/AGE). Again, insertions that could be anchored on B73 reference and were overlapping regions prone to mis-assemblies such as telomeric, knob and centromeric regions, were also excluded from further analysis using IntersectBed.

F2 specific sequences coming either from the no-map approach (Figure S2) or from the work of Darracq et al. (2018) were included as such, without any further filtering.

The multiple references and approaches used during the indel discovery step led to a set of indels with various levels of redundancy. Some “intra-tool” redundancy was found (*eg.* multiple calls found by one tool within the same genotype at highly polymorphic loci). These “ambiguous” calls were systematically identified using the Bedtools suite version 2.16.1 (Quinlan and Hall, 2010) and eliminated. Moreover, for F2 deletions, some “inter-approach” redundancy was also expected and eliminated using intersectBed utility also from the Bedtools suite. When redundancy was found, Pindel calls were preferred to BreakDancer ones because they had precise breakpoints and contained also the calls for PH207 and C103. The same filter was applied to all insertions that could be anchored to the B73 genome sequence. Furthermore, for non-anchored indels, in order to avoid too much redundancy in internal genotyping probes design, RepeatMasker (http://www.repeatmasker.org) was used to mask redundant regions by similarity using an iterative approach. First, “ALLPATHs-LG assembly” F2 insertions were masked with “ABySS assembly” F2 insertions (at least 95% of identity) to generate a non-redundant set of F2 insertions. Then C103 insertions were masked with F2 insertions (at 90% of identity), PH207 insertions were masked with C103 and F2 insertions (90%), and finally F2 No-Map specific sequences were masked with PH207, C103 and F2 insertions (90%).

Design of Affymetrix® Axiom® array

Preparation of sequences for probes for design

To identify presence/absence regions (PARs) within indel sequences more suitable for the design of “off-target” probes, we used the genometools Tallymer utility (Gremme et al., 2013) version 1.5.6 to create two indexes for B73, F2, PH207 and C103: one from their genome assemblies (17-mers with a minimal occurrence of 1) and one from a 5x genome equivalent subset of their raw sequenced data (17-mers with a minimal occurrence of 5). Then B73 genome was iteratively annotated with the script tallymer2gff3.plx (options used: -k 17 -min 35 -occ 1|5 depending on the index) to identify regions not covered by F2, PH207 and C103 kmers. Reciprocally, the two F2 draft genomes, PH207 and C103 ALLPATHs-LG draft genomes were ran through the same procedure to identify regions not covered by B73 kmers. The gff files generated by this process were then used in combination with gff files of repeats annotated with RepeatMasker to define PARs of a minimum size of 35bp for each type of indel and each draft genome.

BP preparation

Breakpoints could be targeted by probes (Figure 1 A) providing that the nucleotide flanking the breakpoint at the beginning of the deleted sequence were different from the nucleotide right after the end of deleted sequence (and reciprocally on the reverse strand). Type I and type III breakpoints without micro-homology sequence can be submitted to Affymetrix®’ straightforward design procedure whereas type II breakpoints have to go through an iterative design process, shifting the sequence by one base on each attempt until reaching a discriminative position. This iterative process stops after 5bp.

Probes scoring

All potential probes were evaluated in an in-silico analysis to predict their microarray performance. A p-convert value, which arises from a random forest model intended to predict the probability that the SNP will convert on the array, was determined for all probes. The model considers factors including probe sequence, binding energies, and the expected degree of non-specific binding and hybridization to multiple genomic regions. This degree of non-specific binding is estimated calculating 16-mer hit counts, which is the number of times all 16 bp sequences in the 30 bp flanking region from either side of the SNP have a matched sequence in the genome. These scores were generated both for forward and reverse probes. A probeset is recommended if p-convert>=0.6 and there are no expected polymorphisms in the flanking region. A probeset is neutral if p-convert>=0.4, the number of expected polymorphisms in the flanking region is less than 3, and the polymorphisms are further than 21 bp of the variant of interest. Probesets not falling into these two categories are scored as *not recommended*. Probesets that cannot be designed are scored as *not possible*.

Probes selection

Concerning OTV and MONO probes, we applied three successive filtering steps. First, we selected only probes classified as recommended and neutral based their scoring, with no more than one hit on B73 reference genome for deletion probes and no hit at all for insertion probes were selected. After this step, 204,213 OTV probes and 18,884,827 MONO probes remained. Secondly, only probes with more than 70% in PARs were kept. An additional filtering step was implemented specifically for MONO probes to optimize probes distribution along the targeted PARs. To optimize probes distribution along the targeted PARs, these ones were cut in 75bp windows using windowmaker (Bedtools) and the MONO probe with the highest p-convert value was selected for each window. In case there were indels with less than 4 MONO probes selected using 75bp windows, these probes were eliminated and a second iteration was made using this time 50bp windows, followed by a last iteration with 25bp windows. This gave at this point a total of 616,286 probes including BP and OTV probes targeting 108,703 indels (90% of indel selected for design).

We completed the design by rescuing 6,219 OTV and 3,441 MONO probes from indels or PARs still not targeted by any probes, bringing the total number of probes selected to 625,946 to target 109,292 indel. At the last step, duplicated probeset were removed based on their sequence by Affymetrix® during the chip design procedure, leaving 662,772 probeset (105,927 indels) corresponding to 1,404,570 different probes to be arrayed on the array.

Genotyping of 105k indels on 480 maize DNA samples

Vegetal Material for genotyping

662,772 probes selected in the array were used to genotype 480 diverse DNA samples including 440 inbred lines, 24 highly recombinant inbred lines and 16 F1 hybrids. Both F1 hybrids (obtained by crossing inbred lines) and their parental inbred lines were genotyped on the array but seed lots used to produce F1 hybrids and those used to extract DNA for genotyping were different. Among these 480 DNAs, 13 inbred lines were genotyped using two different DNAs from two different seedlots and was used to evaluate the reproducibility of the genotyping (Table S4). DNA samples of one F1 hybrid were also genotyped 6 times.

DNA for genotyping were extracted from leaves following a NaBisulfite method modified from Tai and Tanksley (1990) and Dellaporta et al. (1983).

Variant calling using Affymetrix® algorithm

DNA samples from 480 individuals were hybridized to array using the Affymetrix® system. The genotyping, sample QC, and marker filtering was performed according to the Axiom® Best Practice genotyping analysis workflow. Genotype calls and classifications were generated from the hybridization signals in the form of CEL files using the Affymetrix® Power Tools (APT) and the SNPolisher package for R according to the Axiom® Genotyping Solution Data Analysis Guide.

The APT results were then post-processed using SNPolisher, which is an R package specifically designed by Affymetrix®. Markers metrics were generated using the *Ps_Metrics* function. The markers QC metrics were used to classify probesets into 14 categories (Figure S13) using the *Ps_Classification* and *Ps_Classification_Supplemental* functions with all default setting for diploid, except for an empirically determined, more stringent heterozygous variance filter (AB.varY.Z.cut=2.6). Example of clusters from each classification were visualized using the *Ps_Visualization* function (Figure S13).

Each type of probe had a dedicated algorithm (Figure 4 and Figure S8) to call genotyping according to expected behavior from the probe design. Variant were preferentially selected as recommended if they were exhibiting stable category assignments with clearly separated clusters. Each variant was ranked into a category (Figure S13) at each step of the algorithm.

Algorithms used to convert BP and OTV were similar, as BP and OTV behaved like classical SNP. For initial genotype calling, a priori cluster position were used since no information about expected position was available. A first analysis was performed according to Affymetrix® recommendations. Secondly, level of inbreeding was taking into account for a posteriori cluster definition because of the high amount of inbred lines in the panel. This parameter took values from 0 for fully heterozygous to 16 for completely homozygous samples. For OTV and BP algorithms, an inbred penalty of 4 (lower penalty for inbred species) was applied to try to re-labelled probes that fall into categories: CallRateBelowThreshold (CRBT), HomHomResolution (HHR), NoMinorHom (NMH), Other and UnexpectedHeterozygosity after the first cluster analysis. Markers that were classified as OTV may also be considered recommended after *OTV_caller* function has been used to re-label the genotype calls. The SNPolisher *OTV_Caller* function performed post-processing analysis to identify miscalled AB clustering and identify which samples should be in the OTV cluster and which samples should remain in the AA, AB, or BB clusters. Samples in the OTV cluster were re-labelled as OTV. Finally, the recommended markers list is created by combining the list of markers that are classified into the recommended categories (PolyHighResolution (PHR), MonoHighResolution (MHR), and OTV).

BP and OTV probes that exhibited only two clusters (AA or BB and OTV) should fall into monomorphic classification and classify as not recommended. A new MONO algorithm were developed (Figure 4) because we expected for this probes fluorescence pattern no polymorphism in the present sequence (Figure 1 C). Contrary to BP and OTV algorithm, *OTV_caller* was used before inbred penalty for MONO probes analysis. To classify monomorphic sequence genotyping, the *OTV_Caller* function was called and as we expected monomorphic genotyping, only MHR and NMH were considered as recommended. Other monomorphic probes are then analyzed with an inbred penalty of 16 (highest level) to re-labelled probes considering maximum level of heterozygosity. Finally, a new function called *Hom2OTV* was used to classified probes exhibiting two homozygous clusters but with a different position in the Y axis (high and low position). This function tried to decide if the difference of contrast represent actually one homozygous and one OTV cluster as we expect (respectively presence and absence of the corresponding probe sequences).There are no parameters in this function. The lower intensity homozygous cluster is recalled as OTV.

Evaluation of genotyping quality

We compared the genotyping for 479,027 probes from indel array with expected genotyping from resequencing of 4 inbred lines used to discover indels: B73, F2, PH207 and C103. Expected genotyping was built from alignment of probes sequences on reference genome B73 and de novo assembly of 3 inbred lines (F2, PH207 and C103) with Blast software. Sequences were considered present in lines when the probes were aligned with less than 5% of mismatch and absent when not.

Genotyping consistency for B73, F2, PH207 and C103 was calculated between expected and observed genotyping for “presence” and “absence” (Table 4). For this purpose, Affymetrix® genotyping was converted into two genotypes, present and absent and hemizygote from BP were considered as missing data. Consistency of Presence/Absence genotypes between resequencing and array genotyping was analyzed for four individuals (B73, F2, PH207, C103) according to probe types (BP, OTV, MONO): Number of similar genotypes between observed and expected/number of genotype observed. Note that the seed lot used for B73 and F2 genotyping is different from this used for indel discovery, while it is the same one for inbred lines PH207 and C103.

In order to evaluate the consistency of probes genotyping within indels (Figure 5), we used 24 inbred lines including 20 inbred lines from a core collection (Darracq et al., 2018) and the 4 inbred lines used for indel discovery. From 479,027 probes, we selected 294,650 polymorphic probes and totally consistent between sequencing and array genotyping in order to limit the genotyping errors due either to array or sequencing. These probes allowed us to genotyped 72,555 indels. We selected 48,486 polymorphic indels that are genotyped with at least two probes (corresponding to 270,581 probes), and calculated the frequency of presence allele for each indel and inbred lines.

To evaluate quality of genotyping for hybrids, we predicted the genotype of hybrids based on the genotyping of 2 parental lines for 20,370 BPs probes without OTV cluster. This expected genotype for hybrids was then compared with the observed genotyping from array of the corresponding hybrid. With following formula (Number of similar alleles (homozygous or hemizygous) between expected and observed)/(number of expected alleles (homozygous or hemizygous)).

To evaluate the reproducibility of the 479,027 probes of the array (Table S4), we compared genotyping of 13 duplicated inbred lines (A554, A632, A654, B73, C103, CO255, D105, EP1, F2, F252, KUI3, Oh43, and W117) originated from different seed sources. The genotyping of these 13 duplicated lines were also compared using 43,982 SNPs from the Illumina 50K SNP array.

Diversity analysis

We performed diversity analysis on 362 inbred lines from an association panel representing a wide range of diversity (Bouchet et al., 2013; Camus-Kulandaivelu, 2005) using genotyping from our indels Affymetrix® Axiom®. We compared these results with diversity analysis performed on same lines using genotyping of Illumina 50K SNP array (Ganal et al., 2011). Genotyping of indels were treated as bi-allelic 0/2 for “present” and “absent” respectively.

To perform diversity analysis, we first selected 237,629 probes among the 479,027 probes for which (i) the clustering observed were consistent with expected one (Table 3) and (ii) for which genotyping produced by our array for 4 lines used for discovered indels were totally consistent with genotyping based on the alignment of probes on genome assemblies using BLAST software. We filtered out 219,068 probes based on their genotyping quality (missing data rate below 20%, heterozygous rate below 15% and minor allele frequency above 5%). In the end, we selected a single probe by indels that are the best considering both genotyping and Affymetrix® quality leading to a set of 57,824 probes genotyping 57,824 indel to analyze diversity in 362 inbred lines.

We estimated two kinship matrices between 362 lines using “identity by state” estimators (IBS) based on 57,824 indels (Figure 6). Kinship matrices were estimated with the “ibd” function in R package GenABEL (Aulchenko et al., 2007). Genetic structuration were estimated using only 28,143 panzea SNPs using admixture software (Alexander et al., 2009). We selected Admixture results corresponding to five genetic groups (Q=5) since it corresponded to the number of genetics group defined in previous studies using panzea SNP from Illumina 50K (Bouchet et al., 2013). Lines were assigned to one genetic group providing that the probability of assignment to the groups were superior to 0.6 whereas lines below this threshold were considered “admixed”. In order to compare genetic structuration based on indels and SNP, we performed Principal Coordinate Analysis (PcoA) on genetic distance between lines with (362 lines) and without F2 and B73 (360 lines) based on their dissimilarity (1-IBS) using Indels. Each lines were plotted on two first plan of PcoA and colored according to assignment to 5 genetics groups (Figure 6).

Data Access

The array content is available at https://doi.org/10.15454/DWB4UT

Acknowledgements

This work was supported by the project CNV-MAIZE (ANR-10-GENM-003) and the project Investement for the future AMAIZING ANR-10-BTBR-01 (ANR-PIA AMAIZING) and France Agrimer. PhD student C. Mabire is jointly funded by the program CNV4sel in the framework of metaprogram Selgen and by the Plant Biology and Breeding department of the French National Institute for Agricultural Research (INRA). We are very grateful to Patrick Schnable and Cheng-Ting "Eddy" Yeh to provide a subset of Presence/Absent Variants coming from their RNAseq and Sequence capture approach and for his helpful discussion. We are also very grateful to Alain Charcosset for his helpful discussion and comments on the manuscript.

Disclosure declaration

Ali Pirani is an employee of Affymetrix®.

Authors' contributions

SDN designed and supervised the study and conducted CNVMaize project

CM, JD and SDN drafted the manuscript, CV and JJ corrected the manuscript;

NR, SDN, JPP and SP conceived the array, AP, SDN, JJ and JD designed the array;

AP develop calling Affymetrix® pipelines and did the call of indel;

JPP, JJ and CV contributed to the sequencing;

JD, AD, HR and JJ performed the indel discovery, JD and JJ build genome assemblies, JJ and AD discovered SNP within indels;

CM evaluated the quality of genotyping and conducted genetic diversity analysis;

DM and VC did DNA extraction and prepared the samples for arrays genotyping;

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109

Alkan, C., Coe, B.P., Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376. https://doi.org/10.1038/nrg2958

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990 Basic Local Alignment Search Tool 8.

Anderson, J.E., Kantar, M.B., Kono, T.Y., Fu, F., Stec, A.O., Song, Q., Cregan, P.B., Specht, J.E., Diers, B.W., Cannon, S.B., et al., 2014. A Roadmap for Functional Structural Variants in the Soybean Genome. G3amp58 GenesGenomesGenetics 4, 1307–1318. https://doi.org/10.1534/g3.114.011551

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Stein, N., Choulet, F., Distelfeld, A., et al., 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191. https://doi.org/10.1126/science.aar7191

Aulchenko, Y.S., Ripke, S., Isaacs, A., van Duijn, C.M., 2007. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Beló, A., Beatty, M.K., Hondred, D., Fengler, K.A., Li, B., Rafalski, A., 2010. Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor. Appl. Genet. 120, 355–367. https://doi.org/10.1007/s00122-009-1128-9

Bouchet, S., Servin, B., Bertin, P., Madur, D., Combes, V., Dumas, F., Brunel, D., Laborde, J., Charcosset, A., Nicolas, S., 2013. Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8) locus. PLoS One 8, e71377.

Brandenburg, J.-T., Mary-Huard, T., Rigaill, G., Hearne, S.J., Corti, H., Joets, J., Vitte, C., Charcosset, A., Nicolas, S.D., Tenaillon, M.I., 2017. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLOS Genet. 13, e1006666. https://doi.org/10.1371/journal.pgen.1006666

Brunner, S., 2005. Evolution of DNA Sequence Nonhomologies among Maize Inbreds. PLANT CELL ONLINE 17, 343–360. https://doi.org/10.1105/tpc.104.025627

Camus-Kulandaivelu, L., Veyrieras, J.B., Madur, D., Combes, V., Fourmann, M., Barraud, S., Dubreuil, P., Gouesnard, B., Manicacci D., Charcosset A., 2005. Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172, 2449–2463. https://doi.org/10.1534/genetics.105.048603

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al., 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43, 956–963. https://doi.org/10.1038/ng.911

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681. https://doi.org/10.1038/nmeth.1363

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., et al., 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803–807. https://doi.org/10.1038/ng.2313

Clark, A.G., 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15, 1496–1502. https://doi.org/10.1101/gr.4107905

Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., Nickerson, D.A., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. 40, 1199–1203. https://doi.org/10.1038/ng.236

Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J.-P., Mary-Huard, T., Chevalier, C., Bérard, A., Le Paslier, M.-C., Rogowsky, P., et al., 2018. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19. https://doi.org/10.1186/s12864-018-4490-7

Dellaporta, S.L., Wood, J., Hicks, J.B., 1983. A plant DNA minipreparation: Version II. Plant Mol. Biol. Report. 1, 19–21. https://doi.org/10.1007/BF02712670

Dellinger, A.E., Saw, S.-M., Goh, L.K., Seielstad, M., Young, T.L., Li, Y.-J., 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 38, e105–e105. https://doi.org/10.1093/nar/gkq040

Didion, J.P., Yang, H., Sheppard, K., Fu, C.-P., McMillan, L., de Villena, F.P.-M., Churchill, G.A., 2012. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics 13, 34. https://doi.org/10.1186/1471-2164-13-34

Ducrocq, S., Madur, D., Veyrieras, J.-B., Camus-Kulandaivelu, L., Kloiber-Maitz, M., Presterl, T., Ouzunova, M., Manicacci, D., Charcosset, A., 2008. Key Impact of Vgt1 on Flowering Time Adaptation in Maize: Evidence From Association Mapping and Ecogeographical Information. Genetics 178, 2433–2437. https://doi.org/10.1534/genetics.107.084830

Feschotte, C., Jiang, N., Wessler, S.R., 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. 3, 329–341. https://doi.org/10.1038/nrg793

Fu, H., Dooner, H.K., 2002. Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. 99, 9573–9578.

Gabur, I., Chawla, H.S., Snowdon, R.J., Parkin, I.A.P., 2018. Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. https://doi.org/10.1007/s00122-018-3233-0

Ganal, M.W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E.S., Charcosset, A., Clarke, J.D., Graner, E.-M., Hansen, M., Joets, J., et al., 2011. A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. PLoS ONE 6, e28334. https://doi.org/10.1371/journal.pone.0028334

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., et al., 2009. A First-Generation Haplotype Map of Maize. Science 326, 1115–1117. https://doi.org/10.1126/science.1177837

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P., Moreno-Gonzalez, J., Madur, D., Combes, V., Tollon-Cordet, et al., 2017. Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. Theor. Appl. Genet. 130, 2165–2189. https://doi.org/10.1007/s00122-017-2949-6

Gremme, G., Steinbiss, S., Kurtz, S., 2013. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 645–656. https://doi.org/10.1109/TCBB.2013.68

Gu, W., Zhang, F., Lupski, J.R., 2008. Mechanisms for human genomic rearrangements. PathoGenetics 1, 4. https://doi.org/10.1186/1755-8417-1-4

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., et al., 2016. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum tuberosum*. Plant Cell 28, 388–405. https://doi.org/10.1105/tpc.15.00538

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Penagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., et al., 2014. Insights into the Maize Pan-Genome and Pan-Transcriptome. Plant Cell 26, 121–135. https://doi.org/10.1105/tpc.113.119982

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al., 2016. Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell 28, 2700–2714. https://doi.org/10.1105/tpc.16.00353

Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., Barillot, E., 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20, 3413–3422. https://doi.org/10.1093/bioinformatics/bth418

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., et al., 2017. Improved maize reference genome with single-molecule technologies. Nature. https://doi.org/10.1038/nature22971

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al., 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64. https://doi.org/10.1038/nature06862

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al., 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318, 420. https://doi.org/10.1126/science.1149504

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. 42, 1027.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, J., Qu, J., Yang, C., Tang, D., Li, J., Lan, H., Rong, T., 2015. Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data. BMC Genomics 16. https://doi.org/10.1186/s12864-015-1797-5

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Yu, Li, Yongxiang, Semagn, K., Zhang, X., et al., 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. Nat. Commun. 6. https://doi.org/10.1038/ncomms7914

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T., Ma, H., 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. 22, 508–518. https://doi.org/10.1101/gr.127522.111

Lyra, D.H., Galli, G., Alves, F.C., Granato, Í.S.C., Vidotti, M.S., Bandeira e Sousa, M., Morosini, J.S., Crossa, J., Fritsche-Neto, R., 2018. Modeling copy number variation in the genomic prediction of maize hybrids. Theor. Appl. Genet. https://doi.org/10.1007/s00122-018-3215-2

Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X., et al., 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat. Commun. 4. https://doi.org/10.1038/ncomms3320

Marioni, J.C., Thorne, N.P., Tavare, S., 2006. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22, 1144–1146. https://doi.org/10.1093/bioinformatics/btl089

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.-K.K., Visendi, P., Lai, K., Doležel, J., Batley, J., Edwards, D., 2017. The pangenome of hexaploid bread wheat. Plant J. 90, 1007–1013. https://doi.org/10.1111/tpj.13515

Morgante, M., Depaoli, E., Radovic, S., 2007. Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 10, 149–155. https://doi.org/10.1016/j.pbi.2007.02.001

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T., Richmond, T.A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., et al., 2013. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. 14. https://doi.org/10.1186/gb-2013-14-6-r58

Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572. https://doi.org/10.1093/biostatistics/kxh008

Owens, G.L., Baute, G.J., Hubner, S., Rieseberg, L.H., 2018. Genomic sequence and copy number evolution during hybrid crop development in sunflowers. Evol. Appl. https://doi.org/10.1111/eva.12603

Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.-J., 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 14.

Picard, F., Robin, S., Lebarbier, E., Daudin, J.-J., 2007. A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63, 758–766. https://doi.org/10.1111/j.1541-0420.2006.00729.x

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al., Albertson, D.G., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20, 207–211. https://doi.org/10.1038/2524

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M.C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., Morgante, M., 2016. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. Mol. Biol. Evol. 33, 2706–2719. https://doi.org/10.1093/molbev/msw161

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

Saintenac, C., Jiang, D., Akhunov, E.D., 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12, R88.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K.A., Meeley, R., Ananiev, E.V., Svitashev, S., Bruggemann, E., et al., 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. 104, 11376–11381. https://doi.org/10.1073/pnas.0704145104

Salvi, S., Tuberosa, R., Chiapparino, E., Maccaferri, M., Veillet, S., van Beuningen, L., Isaac, P., Edwards, K., Phillips, R.L., 2002 Toward positional cloning of Vgt1, a QTL controlling the transition from the vegetative to the reproductive phase in maize 13.

Saxena, R.K., Edwards, D., Varshney, R.K., 2014. Structural variations in plant genomes. Brief. Funct. Genomics 13, 296–307. https://doi.org/10.1093/bfgp/elu016

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al., 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326, 1112–1115. https://doi.org/10.1126/science.1178534

Shen, X., Liu, Z.-Q., Mocoeur, A., Xia, Y., Jing, H.-C., 2015. PAV markers in Sorghum bicolour: genome pattern, affected genes and pathways, and genetic linkage map construction. Theor. Appl. Genet. 128, 623–637. https://doi.org/10.1007/s00122-015-2458-4

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., et al., 2009. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet. 5, e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stjernqvist, S., Rydén, T., Sköld, M., Staaf, J., 2007. Continuous-index hidden Markov modelling of array CGH copy number data. Bioinformatics 23, 1006–1014. https://doi.org/10.1093/bioinformatics/btm059

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, Haiming, Zhao, Hainan, Song, W., Zhang, M., et al., 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. 50, 1289–1295. https://doi.org/10.1038/s41588-018-0182-0

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., Springer, N.M., 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20, 1689–1699. https://doi.org/10.1101/gr.109165.110

Tai, T.H., Tanksley, S.D., 1990. A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. Plant Mol. Biol. Report. 8, 297–303. https://doi.org/10.1007/BF02668766

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al., 2005. Fine-scale structural variation of the human genome. Nat. Genet. 37, 727–732. https://doi.org/10.1038/ng1562

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R., Pausch, H., et al., 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15, 823. https://doi.org/10.1186/1471-2164-15-823

Unterseer, S., Seidel, M.A., Bauer, E., Haberer, G., Hochholdinger, F., Opitz, N., Marcon, C., Baruch, K., Spannagl, M., Mayer, K.F., 2017. European Flint reference sequences complement the maize pan-genome. bioRxiv 103747.

Varshney, R.K., Saxena, R.K., Upadhyaya, H.D., Khan, A.W., Yu, Y., Kim, C., Rathore, A., Kim, D., Kim, J., An, S., et al., 2017. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. Nat. Genet. 49, 1082–1088. https://doi.org/10.1038/ng.3872

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. https://doi.org/10.1093/molbev/msx319

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z., 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. https://doi.org/10.1093/bioinformatics/btp394

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al., 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. https://doi.org/10.1038/s41588-018-0041-z

Zhou, P., Silverstein, K.A.T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A.D., Steele, K.P., Stupar, R.M., Miller, J.R., et al., 2017. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3654-1

List of Table and Figure

List of figure

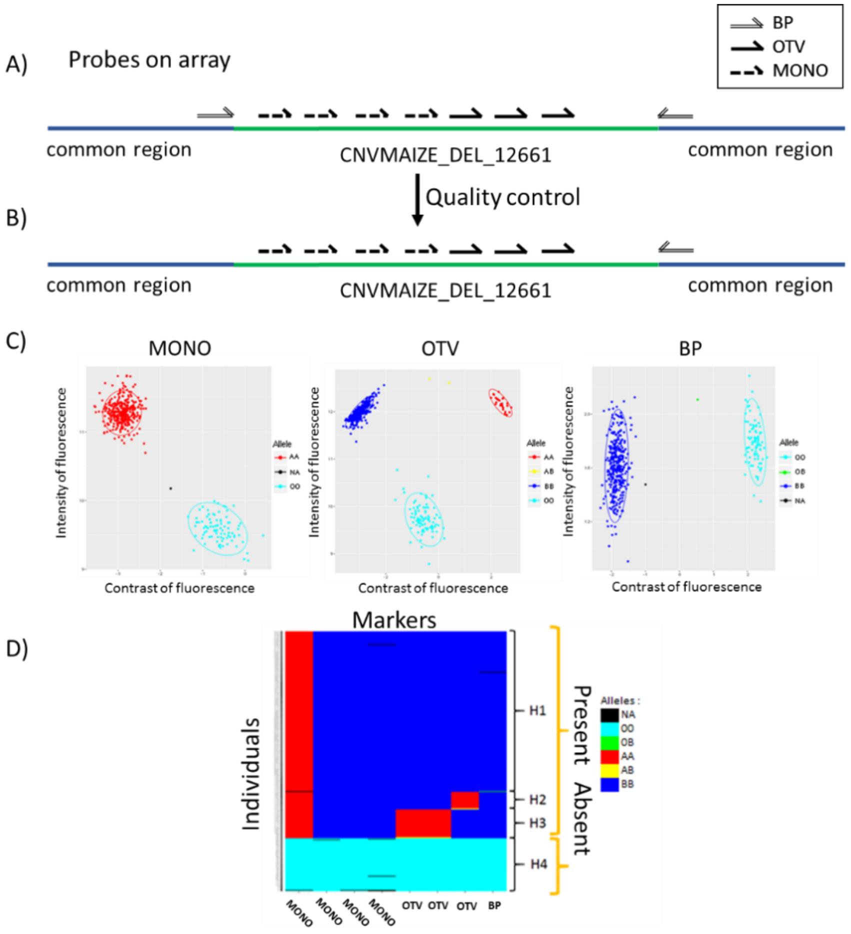


Figure 1: Genotyping of indel CNVMAIZE_DEL_12661 using three probe types on 445 individuals. A) Schematic distribution of the 9 probes along the sequence of indel CNVMAIZE_DEL_12661 (green line) and the bordering sequence common between all individuals (blue line) genotyped by the array. Double, dotted, and full arrows represented the probes designing on the forward and reverse flanking sequences of the breakpoint sites (BP), at not polymorphic (MONO) and polymorphic sites (OTV) within internal sequence of indel. B) Schematic distribution of the 8 probes passing Affymetrix® quality control and called by Affymetrix® pipeline C) Clustering produced by Affymetrix® algorithm for an OTV, MONO and BP probe from indel based on both fluorescence contrast (X axis) and intensity (Y axis) of the 445 inbred lines. Red, blue and yellow dots indicated the presence of the sequence (genotype "present") either homozygous for allele A (AA), or allele B (BB) or heterozygous (AB), respectively. Cyan and green indicated that the sequence were absent in the individual (OO), or only in one copy of the sequence, e.g hemizygous for presence/absence (OB or OA). Black dots indicated individuals for which no genotype could be assigned (Missing data) D) Haplotypes displayed by the genotyping using 8 probes (column) on the 445 inbred lines (row). Colors corresponded to the genotype of individuals produced by clustering in C)

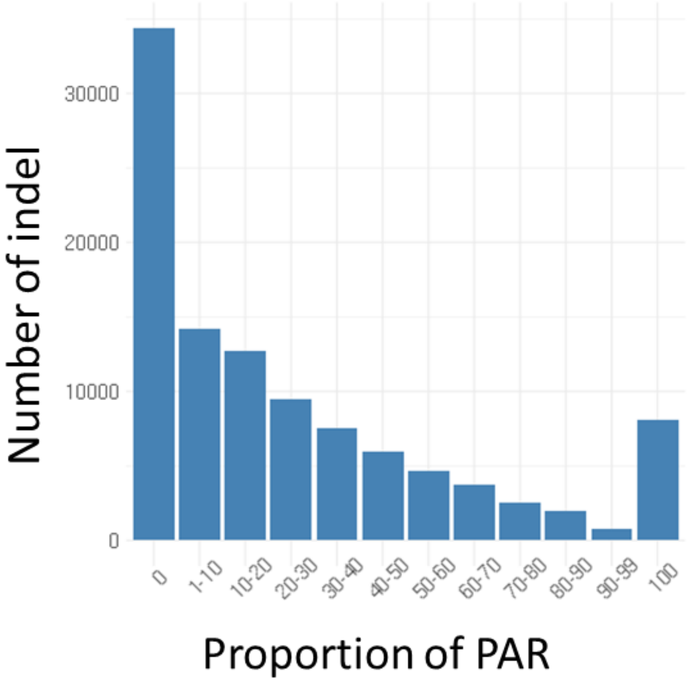
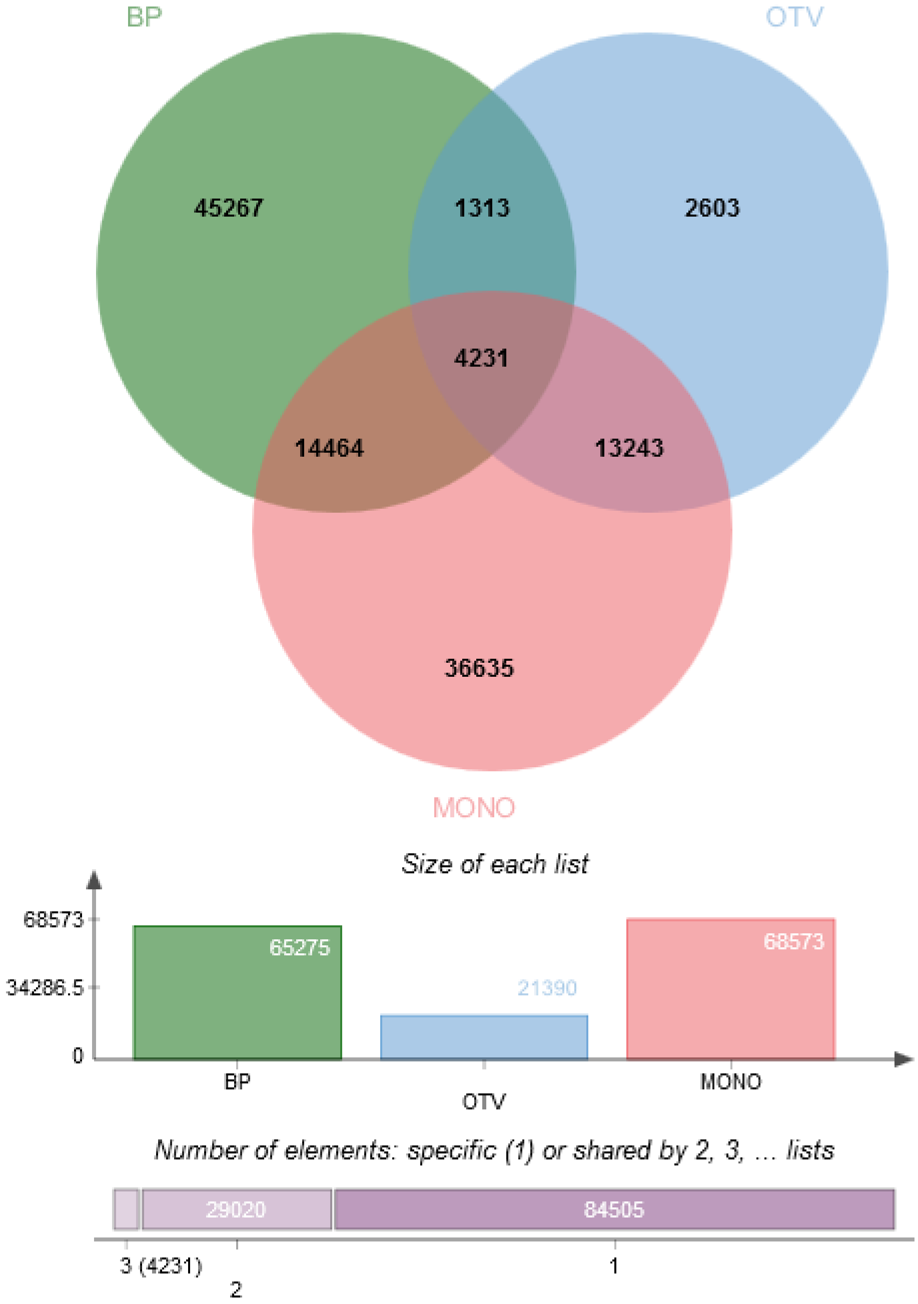


Figure 2: Distribution of the number of indels genotyped by the array according to the proportion of presence/absence regions (Specific fraction) identified in their internal sequence.

A)


BP
OTV
45267
1313
2603
4231
14464
13243
36635
MONO
Size of each list
68573
34286.5
0
65275
21390
68573
BP
OTV
MONO
Number of elements: specific (1) or shared by 2, 3, ... lists
29020
84505
3 (4231)
2
1

B)

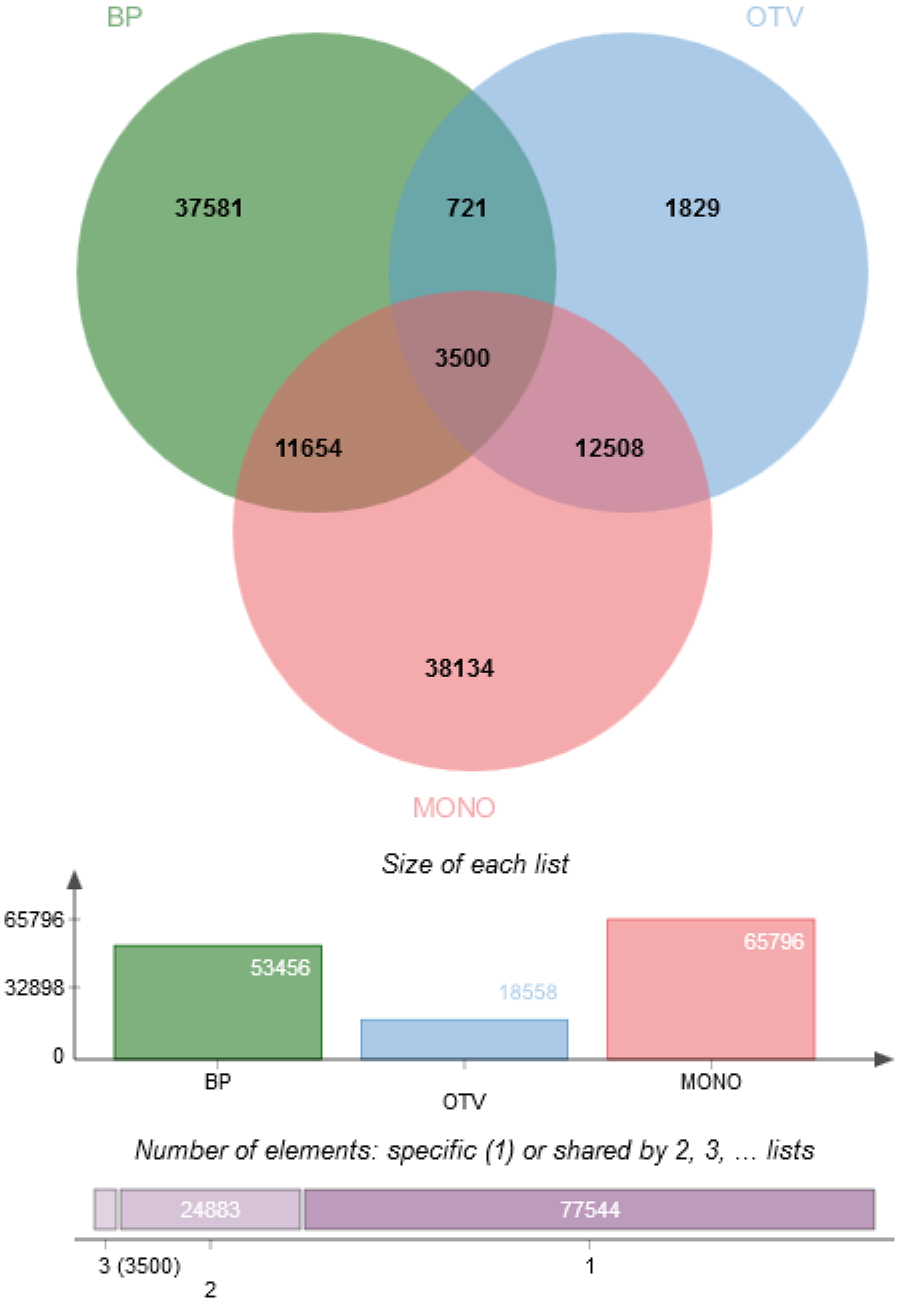


Figure 3: Number of indels that could be targeted by each type of probes designed (A) and selected to be included in the final array (B).

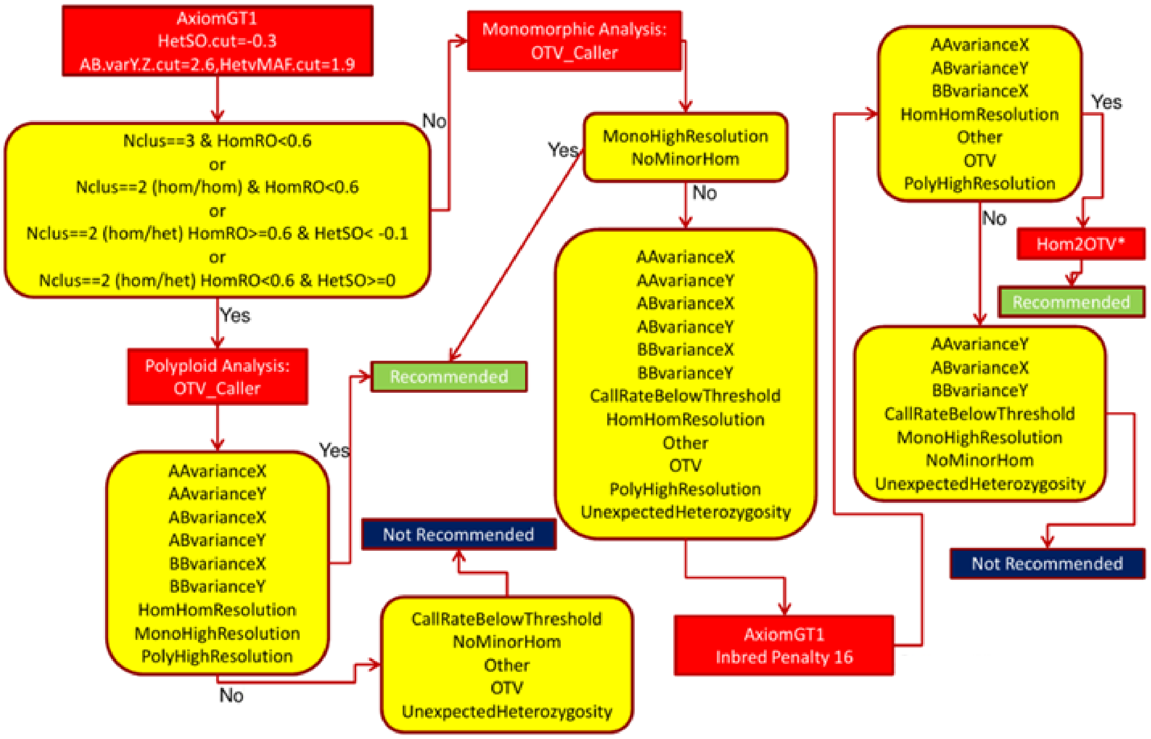


Figure 4: Dedicated Affymetrix® pipeline used for calling indel polymorphisms from the fluorescent intensity variation of MONO probes. Each probe was classified into different categories according to the number of cluster, the call rate and quality metrics of the clustering based on the position, variance and separation of different cluster. In order to retrieve the best clustering for each probes, successive step of clustering using different clustering algorithms (Red square, Axiom GT1, OTV caller, Hom2OTV) or/and with different parameters. According to their classification at each step (yellow square) and threshold used for quality metrics, probes could be classified as recommended (green square), not recommended (blue square) or to be submitted to another step. A new pipeline and an algorithm (Hom2OTV) have to be specifically developed for calling indel genotype of MONO probes since we expected only 2 clusters (absence / presence) that varied exclusively for fluorescent intensity rather than for fluorescent intensity ratio between two labelled nucleotides. At the end, all probes were classified into 14 categories either as recommended or not recommended depending on threshold.

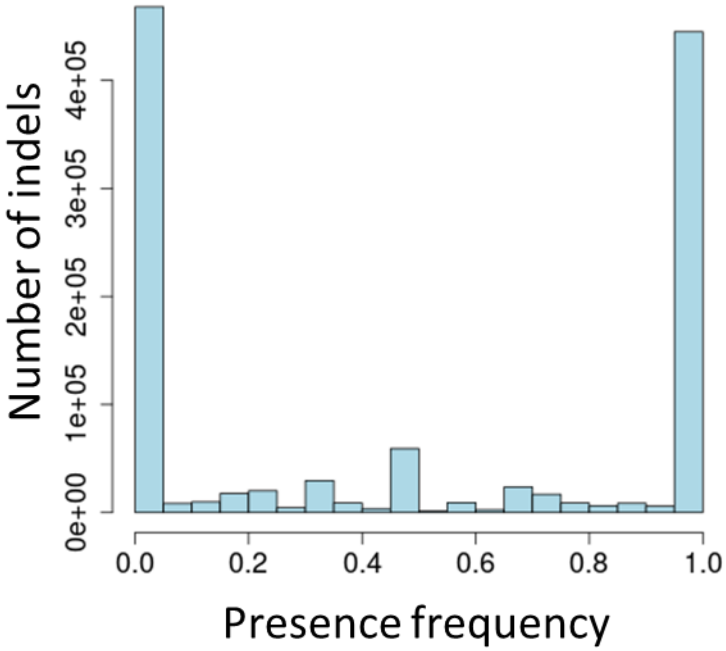


Figure 5: Distribution of the average allelic frequencies of the presence across different probes within 48,486 indels with at least two probes genotyped for 24 inbred lines.

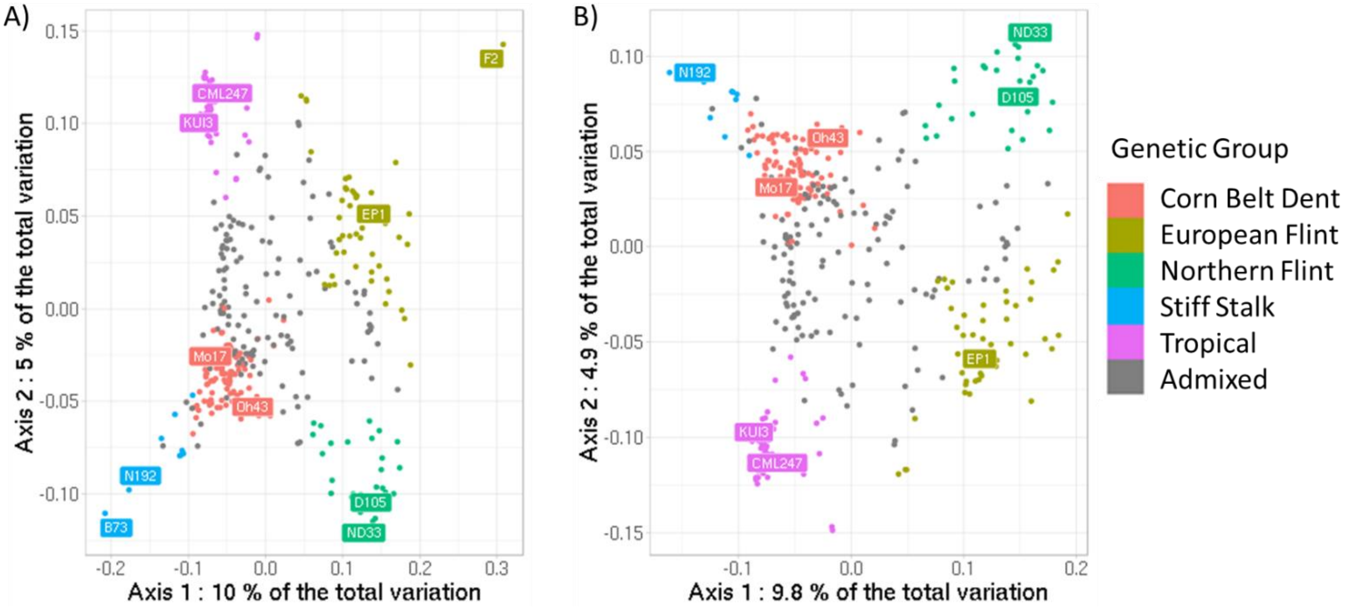


Figure 6: Principal coordinate analysis on the genetic distance (1-IBS) between inbred lines from an association panel estimated by 57,824 indels. A) 362 maize inbred lines were represented B) 360 maize inbred lines were represented excluding B73 and F2 that are used for discovering indels. Colors represented the assignation of the inbred lines to the 5 genetic groups defined by admixture using Panzea SNPs from 50K Illumina array when the probability of assignation to a group (membership) were superior to 60%. Inbred lines that are not assigned to a group (membership<60%) were considered admixed. Common name of two maize accessions typical of each genetic groups were indicated.

List of Table

Table 1: F2, PH207 and C103 de novo assembly metrics. For each assembled genome are detailed: the number of scaffold sequences which were assembled, the length of the shortest scaffold, the length of the longest scaffold, the average size, the N50 of the assembly, the total number of bases included in the assembly, the percentage of Ns present in the assembly and finally the BUSCO statistics including the percentage of complete (C), fragmented (F) and missing (M) BUSCO genes from a total of 1440 BUSCO groups searched for maize.

Maize line	Number of scaffolds	Min size	Max size	Average size	N50	Total (Mb)	% of Ns	Complete BUSCOs (C)	Fragmented BUSCOs (F)	Missing BUSCOs (M)
F2	76563	892	112956	16900	14042	646.3	9.48%	89.3%	4.9%	5.8%
PH207	81688	884	2024489	29557	16860	797.5	8.90%	91.8%	2.7%	5.5%
C103	84990	886	120582	19305	16146	793	8.21%	90.6%	4.2%	5.2%

Table 2: Number of probes before and after selection for array design and passing the Affymetrix® quality control. At each step, are detailed the number (and percentage) of each probe type and the corresponding number (and percentage) of targeted indels. Note that a same indel could be genotyped by several probe types which conducted to a sum of percentage superior to 1 in indel columns.

	Before selection		**On array**		**Called by Affymetrix® pipeline**	
	Probes	**indel**	**Probes**	**indel**	**Probes**	**indel**
BP_Type1	6,648 (0.02%)	3,324 (2.82%)	4,691 (0.71%)	2,751 (2.6%)	2,092 (0.44%)	1,482 (1.66%)
BP_Type2	51,770 (0.2%)	25,885 (21.98%)	38,790 (5.85%)	22,662 (21.39%)	20,540 (4.29%)	14,407 (16.12%)
BP_Type3	71,820 (0.27%)	35,910 (30.5%)	41,272 (6.23%)	27,897 (26.34%)	23,631 (4.93%)	18,485 (20.68%)
BP_Type4	312 (0.001%)	156 (0.13%)	241 (0.04%)	146 (0.14%)	119 (0.02%)	93 (0.1%)
OTV	872,324 (3.26%)	21,390 (18.16%)	163,278 (24.64%)	18,558 (17.52%)	96,867 (20.22%)	15,064 (16.85%)
MONO	25,735,797 (96.25%)	68,573 (58.23%)	414,500 (62.54%)	65,796 (62.11%)	335,778 (70.1%)	63,597 (71.14%)
ALL	26,738,671	117,756	662,772	105,927	479,027	89,393

Table 3: Comparison between the clustering expected for BP, MONO and OTV probe type and the clustering produced by Affymetrix® pipelines based on the fluorescent intensity and contrast of 445 inbred lines for 479,027 probes. Clustering example: typical example of clustering based on the fluorescent intensity (y-axis) and contrast (x-axis). Colors indicate the assignation of the individuals to different clusters identified by pipeline. Description: Brief characteristic of each classification based on the clustering of individuals (homoz.= homozygote, het=heterozygous, OT= off-target)

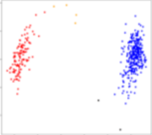
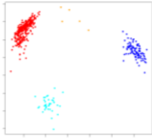
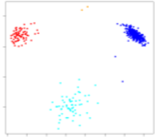
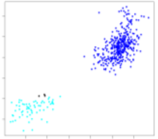
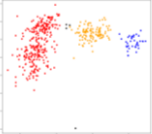
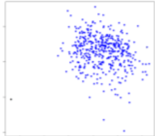
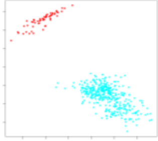
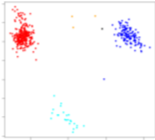
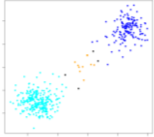
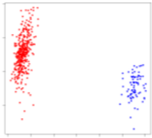
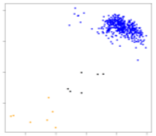
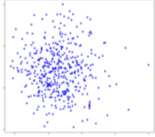
		Classification based on the clustering produced by Affymetrix® pipelines and genotyping assignment					
Probe types		**BP**	**OTV**				
BP	**Nbr (%)**	20,370 (43.9%)	26,012 (56.1%)				
	Clustering example						
	Description	Two homoz. clusters	Two homoz. and one OT clusters				
		OTV	**MONO**	**SNP**	**monomorphic**		
OTV	**Nbr (%)**	78,799 (81.3%)	502 (0.5%)	17,562 (18.1%)	4 (0.0%)		
	Clustering example						
	Description	Two homoz. and one OT clusters.	One homoz. and one OT clusters.	Two homoz. clusters.	One cluster		
		MONO	**OTV**	**Unexpected MONO 1**	**SNP**	**Unexpected MONO 2**	**monomorphic**
MONO	**Nbr (%)**	212,434 (63,3%)	15,690 (4,7%)	68,562 (20,4%)	1,981 (0.6%)	9,525 (2.8%)	27,586 (8.29%)
	Clustering example						
		One homoz. and one OT clusters	Two homoz. and one OT clusters.	One homoz., one OT and one het. clusters.	Two homoz. clusters.	One homoz. and one het. clusters.	One cluster

Table 4: Consistency rate between expected and observed genotype for 4 individuals used to discover indel, according to the three type of probes and the two different genotype expected: presence (P) or absence (A) of the sequence.

Probe types	Expected genotyping	B73	F2	C103	PH207	All individuals
BP	**A**	0.96	0.93	0.94	0.94	0.94
	P	0.96	0.95	0.95	0.95	0.95
OTV	**A**	0.85	0.89	0.80	0.78	0.83
	P	0.93	0.97	0.96	0.96	0.96
MONO	**A**	0.77	0.81	0.82	0.81	0.80
	P	0.90	0.98	0.94	0.94	0.95
All probe types	**A**	0.80	0.85	0.83	0.82	0.82
	P	0.92	0.97	0.94	0.94	0.95
	A & P	**0.85**	**0.94**	**0.89**	**0.89**	**0.89**