Theoretical and Applied Genetics
https://doi.org/10.1007/s00122-021-03965-1

ORIGINAL ARTICLE

Genome-wide analysis of deletions in maize population reveals abundant genetic diversity and functional impact

Xiao Zhang[1,2,3] · Yonghui Zhu[4] · Karl A. G. Kremling[3] · M. Cinta Romay[3] · Robert Bukowski[5] · Qi Sun[5] · Shibin Gao[1,2] · Edward S. Buckler[3,6] · Fei Lu[3,7,8,9]

Received: 4 June 2021 / Accepted: 30 September 2021

Abstract

Key message **Two read depth methods were jointly used in next-generation sequencing data to identify deletions in maize population. GWAS by deletions were analyzed for gene expression pattern and classical traits, respectively.**

Abstract Many studies have confirmed that structural variation (SV) is pervasive throughout the maize genome. Deletion is one type of SV that may impact gene expression and cause phenotypic changes in quantitative traits. In this study, two read count approaches were used to analyze the deletions in the whole-genome sequencing data of 270 maize inbred lines. A total of 19,754 deletion windows overlapped 12,751 genes, which were unevenly distributed across the genome. The deletions explained population structure well and correlated with genomic features. The deletion proportion of genes was determined to be negatively correlated with its expression. The detection of gene expression quantitative trait loci (eQTL) indicated that local eQTL were fewer but had larger effects than distant ones. The common associated genes were related to basic metabolic processes, whereas unique associated genes with eQTL played a role in the stress or stimulus responses in multiple tissues. Compared with the eQTL detected by SNPs derived from the same sequencing data, 89.4% of the associated genes could be detected by both markers. The effect of top eQTL detected by SNPs was usually larger than that detected by deletions for the same gene. A genome-wide association study (GWAS) on flowering time and plant height illustrated that only a few loci could be consistently captured by SNPs, suggesting that combining deletion and SNP for GWAS was an excellent strategy to dissect trait architecture. Our findings will provide insights into characteristic and biological function of genome-wide deletions in maize.

Introduction

With the improvement of sequencing technologies, many studies have indicated that a single reference genome is insufficient in terms of capturing all genomic diversity. The introduction of the pan-genome concept has allowed us to discover new gene sets and partition them as core or dispensable, according to whether they contribute to essential functions or adaption and diversity (Golicz et al. 2016; Tranchant-Dubreuil et al. 2018). For 20 years (Kim and Misra 2007), most studies have focused on single nucleotide polymorphisms (SNPs) as the main source of genetic differences because of their high throughput and easy detection. However, many large structural variations are found throughout the genome and are regarded as a cornerstone of the pan-genome (Golicz et al. 2016). Compared with SNP, SV is longer variation (> 50 bp) can be divided into deletion, insertion, copy number variation (CNV), inversion, and translocation (Chiang et al. 2017; Yuan et al. 2021). As many plants have large and complex genomes that contain an abundance of repetitive sequences, the identification of SVs is challenging and low-sensitivity when using short-read sequencing (Sedlazeck et al. 2018), indicating that SV calling needs filters for robust results. With the development of DNA sequencing technology, long-read sequencing revealed extensive SV detection among diverse individuals within crop species, which may assist crop genomics and improvement (Della Coletta et al. 2021). To build "near-complete" plant genomes (Michael and VanBuren 2020),

Communicated by Thomas Lubberstedt.

Xiao Zhang and Yonghui Zhu contributed equally to this work.

✉ Xiao Zhang
hunterzap@163.com

✉ Fei Lu
flu@genetics.ac.cn

Extended author information available on the last page of the article

Published online: 18 October 2021

pan-genome studies in soybean and rice have constructed several graph-based genomes to demonstrate the hidden genomic variations (Liu et al. 2020b; Qin et al. 2021). These studies indicated that SVs were pervasive throughout the genome and play an important role in gene expression, environmental adaption, domestication, and trait architecture. Early studies demonstrated that SVs were pervasive but not evenly distributed in the maize genome by multiple strategies such as array comparative genomic hybridization (aCGH) (Springer et al. 2009; Swansonwagner et al. 2010) and genotyping by sequencing tags (Lu et al. 2015). Currently, many assembled genomes have been released with the exception for B73, which provide references for SV identification. A recent study even reported the de novo assembly of 26 parents of NAM (nested association mapping) population (Hufford et al. 2021). The genome assembly of 26 NAM parents demonstrated that SVs were more common on chromosome arms with the highest recombination rate. GWAS using SVs detected a significant SV associated with northern leaf blight on chromosome 10, which cannot be identified by SNPs. The assembly of a tropical small-kernel line indicated the 22% of SVs cannot be detected via SNP approach, and some of them can impact expression and trait performance (Yang et al. 2019). The SV analysis of Mo17 indicated that more than 20% predicted genes contained either large-effect mutations or SVs (Sun et al. 2018). The alignment of the A188 and B73 genome sequences identified extensive SVs, including duplication and copy number increases, for carotenoid accumulation (Lin et al. 2021). Other than de novo assembly, SVs can be also detected by read mapping, including paired read (PR), read depth (RD), and split read (SR). As the assembly of high-quality reference genomes is challenging and relatively expensive, the comparison of the segment alignment between genomes is limited, while read mapping strategies are more common, but need rigorous design. In humans, several studies have already applied multiple RD, SR, and sequencing assembly methods to identify SVs with high genetic diversity from hundreds to two thousand human genomes (Consortium 2012; Mills et al. 2011; Sudmant et al. 2015). In plants, a CNV map was constructed using more than 1000 Arabidopsis accessions: the CNV was found to overlap with 18.3% of protein-coding genes, and these genes were enriched in evolution, stress, and defense. CNV markers can accurately explain population structure and migration patterns. The dosage effect of genes triggered by CNV impacts transcript and protein level (Zmienko et al. 2020).

As a type of genomic imbalance, deletions can modulate phenotypes by altering the transcriptome in different feedback loops. A deletion can affect the expression of a single gene through several mechanisms: (1) gene duplication or deletion altering gene dosage in a fraction of dosage-sensitive genes (Gamazon and Stranger 2015); (2) the formation of novel transcripts through the disruption of the structure in genes partially overlapped with deletions; (3) "position effects," by the alteration of distance from *cis*- or *trans*-regulators or missing regulatory elements, unmasking recessive functional polymorphisms (Feuk et al. 2006); (4) long-distance *trans*-regulation generated by the modification of the position of genes or regulatory elements within the nucleus and/or chromosome territory of a genomic region (Cremer et al. 2006; Fraser and Bickmore 2007; Reymond et al. 2007). The identification of expression quantitative trait loci (eQTL) has been identified as a good strategy to explore how genomic variants impact gene regulation.

According to the relative distance between the polymorphism and target gene, eQTL can be considered as "local eQTL" and "distant eQTL": eQTL located within the gene sequence or near the encoded transcript are "local eQTL"; meanwhile, eQTL mapped elsewhere in the genome are determined as "distant eQTL." Further, eQTL can be categorized as either *trans*- or *cis*-depending on whether the regulatory mechanism is allele-independent or allele-dependent (Albert and Kruglyak 2015; Fan et al. 2020). *cis*-eQTL are known to mainly have a larger effect than *trans*-eQTL in global transcript profiling, as *cis*-eQTL directly influenced the *cis* sequence polymorphisms (Hansen et al. 2008; Liu et al. 2017a). In maize, using recombination inbred lines and GWAS panel, eQTL mapping was applied to link expression and phenotypic variation, including root, leaf, kernel development, and oil-related traits (Holloway et al. 2011; Kremling et al. 2019; Liu et al. 2017a; Pang et al. 2019). With the development of "omics," many studies were able to integrate eQTL mapping with GWAS in order to identify candidate genes in multiple dimensions. For example, by integrating GWAS, eQTL, and quantitative trait transcript analysis, 137 putative kernel length-related genes were identified in total, including 43 previously reported QTL regions (Pang et al. 2019). In another study combining GWAS, eQTL mapping, and trait correlation to dissect leaf development in maize, 25 prioritized candidate genes were identified and were enriched in specific functional categories (Miculan et al. 2021). However, unlike SNPs that are often bi-allelic, SVs are deemed multi-allelic both in length and copy number (Handsaker et al. 2015). A few studies have employed eQTL mapping using SVs in maize, even for the SVs identified by short-read sequencing, which is extremely challenging because of the complexity of the genome (Schnable et al. 2009).

As the SV-gene pairs exhibit subtle and significant gene regulation, the regulation of these genes can form a network that influences the complex quantitative trait variation (Alonge et al. 2020). In humans, the role of SVs has been linked to many severe diseases (Gonzalez et al. 2005; Helbig et al. 2009; Nuytemans et al. 2009; Prasad et al. 2012; Yang et al. 2013a). In plants, numerous examples have

demonstrated that SV has a potential role in biotic (Cook et al. 2012; Dolatabadian et al. 2017; Liu et al. 2017b; Zuo et al. 2014) and abiotic resistance (Gabur et al. 2019; Maron et al. 2013), domestication (Lye and Purugganan 2019; Studer et al. 2011), kernel development (Liu et al. 2019, 2015), and heterosis (Lai et al. 2010; Springer et al. 2009). Moreover, SVs are shown to function in environmental adaption, including stay-green traits (Qian et al. 2016) and flowering time (Díaz et al. 2012; Huang et al. 2018). A PanSV genome of 100 tomato lines revealed that SVs changed gene dosage and expression levels that resulted in differences of fruit flavor, size, and production (Alonge et al. 2020). As SVs can potentially explain "missing" heritability, they are considered as an important complement to GWAS (Manolio et al. 2009) and genomic prediction models. Currently, the third version of the maize haplotype map has developed, extending the number of inbred lines from 27 (Gore et al. 2009) to 1218 (Bukowski et al. 2017), and providing an abundance of short-read sequencing data for the analysis of genetic diversity. However, although many studies have already investigated the gene expression and phenotypic effect of SVs from comparative genomics between the assembled genomes in maize, it is still valuable to perform eQTL mapping and GWAS for expression analysis to understand the underlying functional effect using short-read sequencing data from large-scale population.

In this study, we applied two methods, based on the depth of sequencing reads, in order to detect deletions in the resequencing data generated from 270 diverse maize inbred lines (Bukowski et al. 2017; Flint-Garcia et al. 2005) (Table S1). Moreover, eQTL of gene expression in different tissues (Kremling et al. 2018) and the loci associated with the flowering time and plant height were also analyzed using GWAS. The detection efficiency of GWAS for expression and phenotypes was also compared between deletions and SNPs. Our study will thus provide information on developing high-efficiency methods for application in plants with complex genomes and to uncover the distribution and functional impact of deletion in the maize genome.

Materials and methods

Materials, tissues, and sequencing

Paired-end sequencing data were produced for a total of 270 maize inbred lines that could be divided into non-stiff stalk (NSS), stiff stalk (SS), tropical and subtropical (TS), popcorn, and sweetcorn lines (Bukowski et al. 2017; Flint-Garcia et al. 2005). The sequencing depth ranged from 0.004 to 40.904× and from 0.001 to 20.025× under $q > 30$ filtering via mapping quality. The duplicates were marked using MarkDuplicates from Picard (version 2.5.0) (Institute 2019). The raw sequencing reads were aligned to B73_RefV3 reference in BWA (version 0.7.13) (Li and Durbin 2009) and then transformed and merged to BAM files for each taxon in SAMtools (version 1.3.1) (Li et al. 2009). The details of sample collection and sequencing for the RNA-seq data were introduced in Kremling et al. (Kremling et al. 2018).

Estimation of a suitable window size for the read-depth approach

The "suitable window size" in this study means the minimum size of the window that ensures there are enough reads. To ensure statistical power, we set a threshold of at least 30 reads for sufficient power for analysis. Based on the threshold for the read count, we have calculated a suitable window size for each taxon using the formula below:

$$\text{Suitable window size} = \frac{\text{Expected read depth} \times \text{Read length}}{\text{Sequencing depth}}$$

where "suitable window size" means the suitable window size estimated for each taxon; "expected read depth" is the minimum accepted threshold in the number of reads (30); "read length" represents the read length; and "sequencing depth" means the sequencing depth for each taxon. As the sequencing depth of each taxon is different and a consistent standard is needed to calculate the read count in all inbred lines, the largest suitable window size in 262 lines (2900 bp) was set as a consistent standard in this panel. The window size was set to 3000 bp for convenience.

PAV calls based on multiple reference genomes

PAV locations and sequences were identified using a sliding window method against three released reference genomes, that is, Mo17, W22, and CML247. This method was slightly modified based on the identification of Mo17 PAVs by Sun et al. (Sun et al. 2018). We then divided the B73 genome into 150-bp overlapping windows with a 1-bp step size. The sliding window size was set to 150 bp because it is the expected read length, and the step size was set to 1 bp for higher resolution. We then aligned those sequences to the B73 genome and the other three genomes using BWA (version 0.7.13) with the options of "-w < bandwidth > < aligned genome fasta > < divided genome fasta > -M." When the gap proportion cutoff and the coverage cutoff were set to 0.25, the sequences of windows that could be aligned to the B73 reference genome but could not be aligned to the other genomes were considered B73-specific sequences and a PAV interval of the corresponding genome. The overlapping windows were then merged into suitable windows estimated above, and the PAV proportion for each window was calculated.

Deletion detection via HMMCopy and dynamic window methods

For the HMMCopy method, the whole genome was split using the suitable window size described above. Read count, GC content, and mappability in these windows were estimated using HMMcopy_utils (Ha et al. 2012). A 150 mer was set to calculate the mappability file for each base pair, and the parameter E was set to 0.9 to assure the precision of copy numbers for smaller windows. After removing "NA" in the copy number estimated by HMMCopy, the deletion windows with negative copy numbers were retained and combined with the PAV windows in which the proportion of PAVs identified by three reference genomes above was greater than 0.9. The copy number threshold for this population was determined by the boundary of the distribution located in the negative axis.

According to the mappability file described above, the suitable window size for the dynamic window method was set as the standard for the unique mapped base pairs in each dynamic window. The read numbers for these windows were then counted via BEDTools (version 2.26.0) (Quinlan and Hall 2010), whereas the GC percentage was calculated using a customized script. According to McConnell et al. (McConnell et al. 2013) and Knouse et al. (Knouse et al. 2016), the read counts were normalized by the genome-wide median read count with a similar GC percentage (1% interval) to eliminate the GC content bias. Following the same protocol as the HMMCopy method, dynamic deletion windows with negative copy numbers were kept and combined with the PAV proportion estimated by the three genomes. Then, the copy number cutoff of this population for this method was determined by the normal distribution boundary located only on the negative axis.

To eliminate the false-positive deletions called for each method, for each inbred line, deletions from both methods with the same type and a reciprocal overlap larger than 50% of their size were identified as the same deletion and used for further analysis.

Transferring deletion windows to a bi-allelic genotype for analysis of population structure and linkage disequilibrium (LD) decay distance

The whole genome was split into the suitable window size described above. For each taxon, we used the allele "A" if the window was identified as a deletion window or "T" if the window could not be identified. After filtering the markers with minor allele frequency (MAF) > 0.05, 50,000 markers were randomly selected across the genome and imported to Tassel 5 (5.2.57) (Bradbury et al. 2007) for principal component analysis (PCA) analysis.

LD decay was also analyzed via Tassel 5 by setting the window size to 100. The LD decay distance was defined by a cutoff of $r^2 = 0.1$. We then divided the population into NSS, SS, TS, and mixed subgroups, and the same steps were repeated for each group.

Deletion proportion and genomic features

Based on the "Zea_mays.AGPv3.26.gff3" file, we extracted repeats and gene density into 1-Mb window size across the whole genome. The recombination, genomic evolutionary rate profiling (GERP) score, and sorting intolerant from tolerant (SIFT) were collected from the NAM population (Rodgersmelnick et al. 2015). Pearson's correlation analysis was performed pairwise by setting the P value cutoff to 0.01.

Deletion proportion and gene expression

For each tissue and stage, the genes with no expression in this population were removed. The raw gene expression abundance was Box-Cox-transformed, and the Z-score was then calculated. The gene expression rank was estimated in accordance with the Z-score of each taxon. For each expressed gene in this population for each tissue, the Z-score data were separated into two groups based on whether the deletions overlapped or not. The Wilcoxon test was used for comparison between these two groups. Moreover, simple linear regression was applied between the mean deletion proportion and the corresponding rank.

eQTL mapping by deletions and SNPs

To eliminate the effects of hidden determinants on gene expression, 25 hidden factors were estimated using PEER (Stegle et al. 2010) for each tissue. Combined with five multidimensional scaling (MDS) covariates, 30 covariates were imported into the Tassel 5 "EqtlAssociationPlugin" for eQTL mapping of Box-Cox-transformed expression values for each gene. The significance level was set to 0.05 and corrected by Bonferroni test. To filter the repetitively false positives caused by LD, we grouped significant sites that were separated by < 150,000 bp. The lead eQTL was defined as the eQTL with the most significant site and the largest R^2 in each group. The LD analysis of each pair of sites was performed for the lead eQTL of each gene, and the R^2 cutoff for LD was set to 0.1. If two lead eQTL were in LD, only the site with more significant and larger effect was kept. An eQTL was considered "local" if the eQTL was found within 50 kb of the target gene. The remaining eQTL were considered "distant." The eQTL mapping process for SNPs was similar except that the cutoff for the interval size was reduced to 5000 bp.

Distant eQTL hotspots were defined as the genomic regions that were enriched in eQTL influencing the expression of genes. For their discovery, we applied a permutation approach. First, we performed a sliding window analysis with 0.9 Mb windows and 3000 bp steps. In each permutation, distant eQTL of tissues and total distant eQTL were randomly assigned to windows in the whole genome. eQTL were counted using the sliding window. For each sliding window, the 95th percentile of the eQTL number of all permutations was recorded as the eQTL cutoff. To reduce false positives, only the maximum of cutoffs in all sliding windows was set as a threshold to extract distant eQTL hotspots. The overlapped sliding windows were then merged, and the target genes regulated by distant eQTL in each hotspot were extracted. Transcription factors information was extracted from PlantTFDB database (Jin et al. 2016) (http://planttfdb.cbi.pku.edu.cn/).

Gene ontology (GO) enrichment and KEGG analysis

Gene names were extracted using the AgriGO tool (Tian et al. 2017) (http://systemsbiology.cau.edu.cn/agriGOv2/index.php). The option "Fisher Test" was selected as the test method, and the significance level was set to 0.05 in 'Yekutieli (FDR under dependency)' in multi-test adjustment, and "Plant GO slim" was chosen as gene ontology type. We also extracted gene names and transformed them into "Entrez Gene" ID using the MaizeGDB gene center (https://www.maizegdb.org/gene_center/gene). The "Entrez Gene" IDs were imported in KOBAS 3.0 (Xie et al. 2011) (http://kobas.cbi.pku.edu.cn/), and default parameters were selected for KEGG analysis. The KEGG pathways in which the *P* value was less than 0.05 were considered as significantly enriched pathways.

GWAS of common traits using deletion alleles

Best linear unbiased predictions for 11 traits of 254 overlapped inbred lines were collected from Peiffer et al. (2014). The data included days to anthesis (DTA), days to silking (DTS), anthesis-silking interval (ASI), growing degree days to anthesis (GDD-DTA), growing degree days to silking (GDD-DTS), growing degree days to anthesis-silking interval (GDD-ASI), plant height (PH), ear height (EH), PH minus EH (PH-EH), EH divided by PH (EHdivPH), and PH divided by days to anthesis (PHdivDTR). In total, 239,484 deletion alleles with MAF > 0.05 were used as alleles for GWAS. The GWAS analysis was performed by FarmCPU with the first three PCs as covariates. The significance level was set to 0.01 and adjusted by Bonferroni test.

Results

Identification of deletion windows in the population using multiple reference genomes

Two different methods were applied to identify deletions based on read counts. The analysis flows are shown in Fig. 1a. First, we estimated PAVs from multiple genomes to evaluate the copy number threshold of the population. The B73 sequence was partitioned into 150-bp segments with one base pair step size and aligned to the currently assembled genomes of Mo17, CML247, and W22 (https://www.maizegdb.org/genome). The segments that were present in B73 but absent in others were extracted and combined into long fragments as PAV. As per our results, it showed that the total size of PAV in CML247 was the largest (37.02 Mb), whereas Mo17 contained the smallest length (34.58 Mb), a difference of approximately 2 Mb (Fig. 1b, Table S2). In contrast, PAVs that were present in CML247, Mo17, and W22 but absent in B73 reference genome were compared (Fig. 1c, Table S3). The results showed that the PAV length in CML247 (37.02 Mb) was the largest, whereas that in W22 (29.17 Mb) was the smallest, a difference of approximately 8 Mb. The PAV segments from two datasets were concentrated to less than 1000 bp.

After filtering the mapping quality to values larger than 30, the window size was set to 3000 bp for statistical power. Eight lines were then removed for the following analysis because of insufficient depth based on the suitable window size formula (refer to Methods) (Fig. 1d, Table S1). After correction for GC and mappability, the copy number of the corrected read counts for each window was calculated using the R packages HMMCopy (Ha et al. 2012) and dynamic window method, respectively. As expect, RD was closely correlated with the average read count of windows (Figs. S1a, b), but showed little correlation with the average copy number per window in both methods (Figs. S1c, d). For each method, only windows with a PAV proportion of more than 90% were collected to estimate the copy number threshold. The distribution of copy number was bimodal, which indicated that the copy numbers of many PAV windows were larger than zero. Because deletion can result in read count loss, the right boundary of the negative copy number distribution was set as the copy number threshold. To ensure reliability and avoid false positives, we set − 1.70 and − 1.85 as the thresholds of deletion for HMMCopy and dynamic window method, respectively, in this study (Figs. 1e, f). Finally, each deletion window was further verified by at least 50% reciprocal overlap between the two methods. To verify the accuracy of deletions identified in this study, the

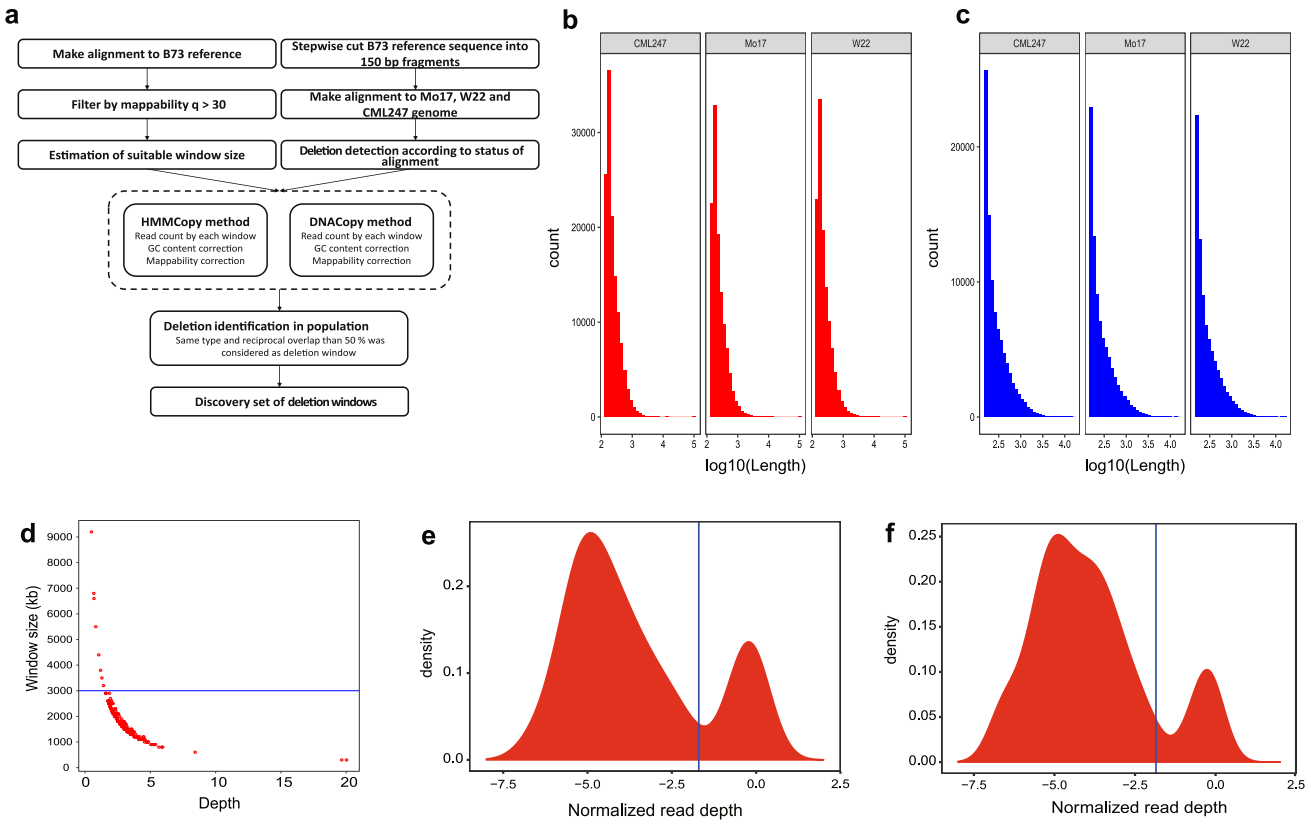

Fig. 1 The route map of deletion detection and the estimation of deletion window size. **a.** Route map of deletion discovery using next-generation sequencing data of the maize population. **b.** Length distribution of PAVs exists in B73 but not in CML247, Mo17, and W22. **c.** Length distribution of PAVs exists in CML247, Mo17, and W22 but was not in B73. **d.** Correlation between sequencing depth and estimated window size. The blue line was the window size (3000 bp) used in this study. **e.** The distribution of normalized read depth calculated by the HMMCopy package. The blue line was the threshold of normalized read depth for the HMMCopy method. **f.** The distribution of normalized read depth calculated by the dynamic window method. The blue guideline was the threshold of normalized read depth for the dynamic window method

Benchmarking Universal Single-Copy Orthologs (BUSCOs) were introduced as single-copy benchmarks. When comparing the deletion frequency between BUSCO overlapping windows and non-overlapping windows, the deletion frequency of non-overlapping windows was significantly larger than the overlapping windows (0.11 vs 0.22, Wilcoxon test $P < 0.01$). The distribution of frequency also indicated BUSCO overlapping windows enriched in the lower frequency interval compared with non-overlapping windows (Fig. S2).

Deletions showed uneven distribution across the genome and the population

The deletion windows were compared in the genome and the population. In total, 318,269 deletion windows were identified, accounting for 46.35% of the total windows across the genome. The deletions were distributed throughout the maize genome, but rarely located in centromere regions, except for chromosomes 2 and 9 (Fig. 2a, 4). Among them, 45 deletion windows showed a high frequency (> 95%), which were considered as deletion hotspot windows, suggesting that many common deletions were found in different inbred lines (Table S4). In total, 19,754 deletion windows were found to have overlapped with 12,751 genes, and most deletion windows were enriched in the coding for amino acids sequences (CDS) and introns (Fig. 2b). After merging the continuous deletion windows, numerous large deletion regions were identified across the genome. The largest deletion region, which contained 46 continuous deletion windows, was located on chromosome 1 (Table S5). The existence of the deletion hotspot windows and large deletion regions illustrated that deletions were not evenly distributed in the maize genome. The mean frequency of all deletion windows for each inbred line was close to 0.10 in this population (Fig. S3). Unsurprisingly, B73 contained the fewest deletion windows, whereas B57 contained the largest number of deletion windows in the population. To dissect deletion distribution among different heterotic

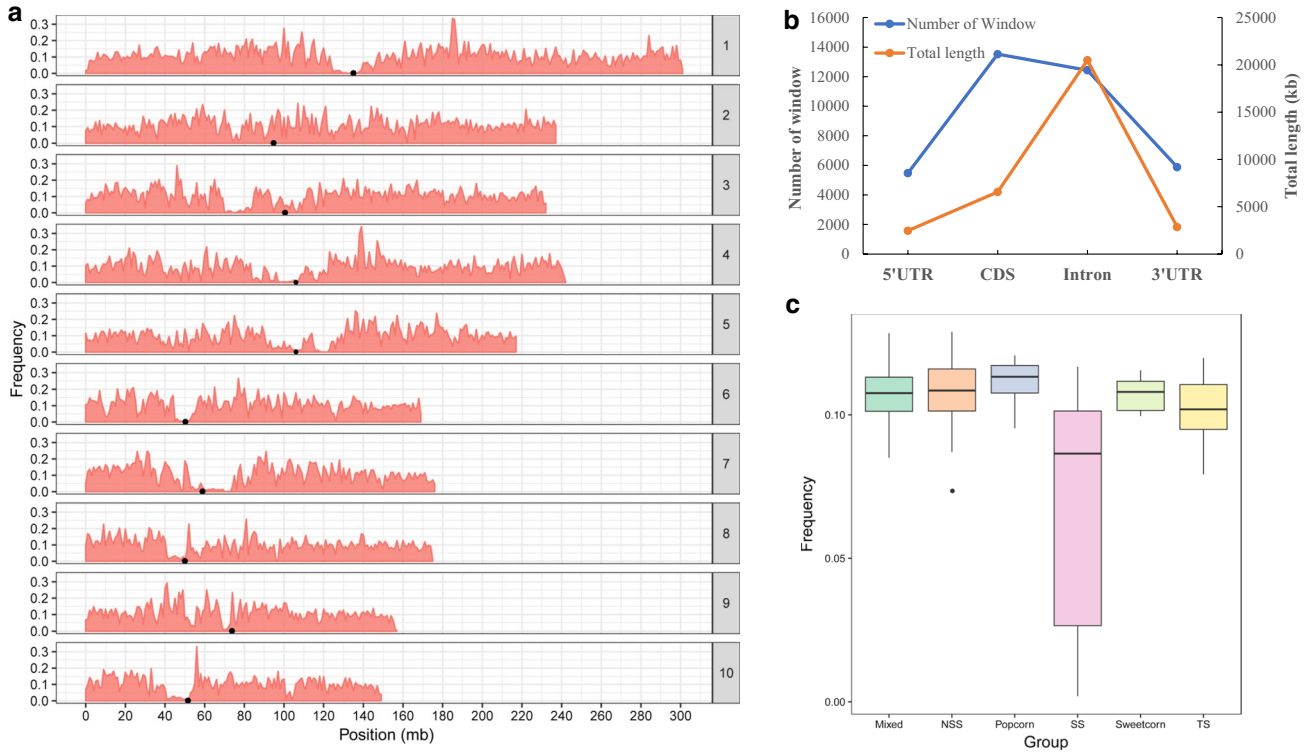

Fig. 2 Deletion distribution for the whole genome, gene regions, and different subgroups. **a.** Deletion distribution throughout the whole genome, black circles are centromeres. **b.** Number and total length of overlapped deletion windows in gene regions. **c.** Box plot of deletion frequency for each subgroup. Subgroups were clustered by SSRs according to the previous study (Flint-Garcia et al. 2005)

groups, the inbred lines were clustered into six groups as described in a previous study (Flint-Garcia et al. 2005). The deletion frequency was significantly different among groups ($P < 2e\text{-}16$, one-way ANOVA) (Table S6). The average deletion frequency represents the average deletion frequency of inbred lines belong to the corresponding subgroup. The average deletion frequency in the stiff-stalk (SS) group, where B73 belongs, was the lowest, whereas that in the popcorn group was noted to be the highest (Fig. 2c, Table S7). This was mainly due to the bias of the reference genome and identity by descent similarity within SS relative to the NSS group.

Across the inbred lines, the site frequency spectrum was compared between deletions and SNPs called from the same raw data (Fig. S4a). Relative to B73, the highest frequency of SNP and deletion allele was 0–0.20, but the proportion of SNPs was much higher than that of deletions in low-level allele frequency. To better understand the relationship between SNP and deletion allele frequency, the average allele frequency of each polymorphism in the deletion window was calculated (Fig. S4b). However, the correlation was not strong ($P < 2.2e\text{-}16$, $r = -0.044$, Pearson's correlation) between deletion allele frequency and SNP frequency across the whole genome.

Deletions can explain population structure and were significantly correlated with genomic features

Population structure is often affected by selection, which can guide breeding progress. In GWAS, the population structure is also imported as a covariance. To explore whether deletions could explain the population structure, the deletion windows were transformed into two-allele genotypes (see Methods). After filtering alleles with a MAF of < 0.05, 50,000 deletion windows were randomly selected for PCA. Six window numbers from 5000 to 100,000 were simulated to investigate the effect of window number for PCA (Fig. S5); the results showed little differences among the variant window numbers. SS, NSS, and TS groups showed clear divisions in the population. The inbred lines with mixed genetic backgrounds were mainly located between NSS and TS. Popcorn and sweetcorn were not grouped independently, but were within the NSS group, which was consistent with a previous study (Flint-Garcia et al. 2005). In contrast, SS group was dispersed, and some inbred lines were close to the NSS group (Fig. 3a). PCA analysis was also performed with 50,000 randomly selected SNPs extracted from the same sequencing data (Fig. 3b). After a simulation of different

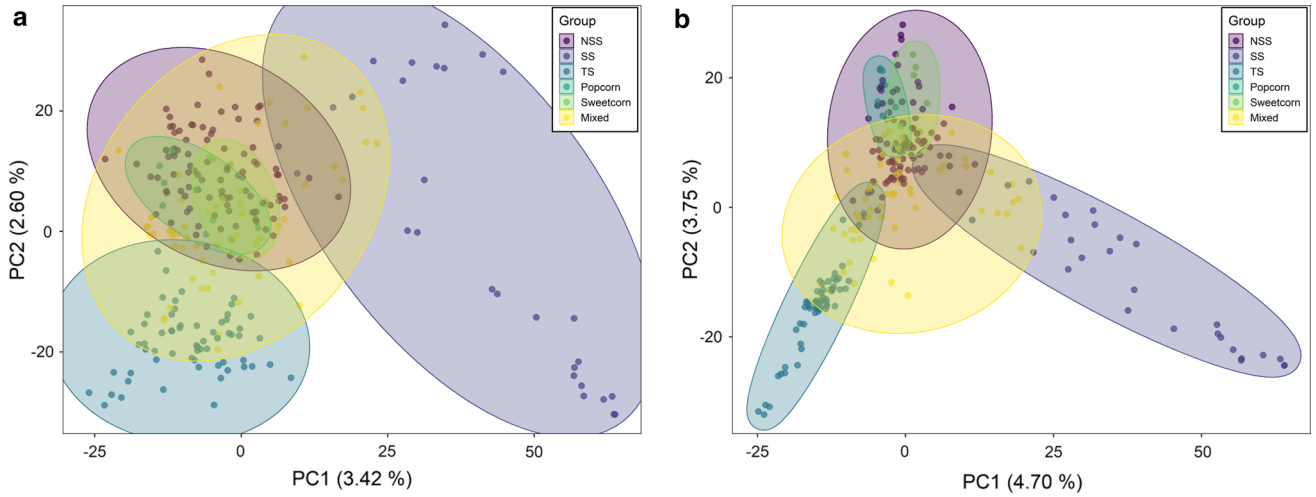

Fig. 3 Distribution of inbred lines between SNPs and deletions in PCA analysis. **a.** Distribution of inbred lines in PCA analysis of deletions. **b.** Distribution of inbred lines in PCA analysis of SNPs

randomly selected SNPs for PCA analysis, the results were similar to deletion that showed little difference in PCA cluster (Fig. S6). In conclusion, PCA plots of deletion and SNP showed the same group clusters, and even similar discrete ranges of each subgroup, although the PCA plot of SNPs were slightly concentrated.

LD is an essential factor affecting the accuracy and confidence intervals of association analysis. At a cutoff of $r^2 = 0.1$ for each chromosome, the average distance was 102 kb. The distance was greatest between chromosomes 4 and 5, that is, at 126 kb (Fig. S7a). Compared with the LD decay distance estimated by SNPs, which was usually less than 10 kb (Lu et al. 2011; Yan et al. 2009), the distance calculated using deletions was larger, probably because of the lower density of deletion markers across the genome. A previous study concluded that the LD decay distance is closely related to the genetic background of the population (Zhang et al. 2016). Therefore, we separated individuals into SS, NSS, and TS groups to evaluate the LD decay in different germplasm groups. Similar to deletion frequency, the SS group exhibited the furthest distance (294 kb) (Figs. S7b, c). In contrast, the TS group had the shortest distance (84 kb) (Fig. S7d), which was consistent with the results estimated using SNPs in previous studies (Lu et al. 2011; Yan et al. 2009; Zhang et al. 2016). We further analyzed whether SNPs within the deletion windows showed LDs with located deletion. After filtering SNPs for MAF > 0.05, heterozygosity < 0.2, and missing rate < 0.2, only 37,098 deletion windows accounting for only 32.8%, showed LD with inside SNPs among 113,040 deletion windows contained SNPs by setting the LD cutoff to $r^2 = 0.1$. This result indicated that the SNPs within deletion windows cannot replace deletion alleles for association analysis (Fig. 5a).

To understand whether deletions were affected by genomic features, the average deletion proportion in 1-Mb windows was calculated. The distribution trends of six elements, including deletion frequency, recombination, repeat, gene density, GERP score (Rodgersmelnick et al. 2015), and centromere position, were integrated for Pearson's correlation analysis (Fig. S8). The results indicated that deletion frequency was significantly ($P < 0.05$) correlated with most features, including repeats, gene density, centromere position, and sites of GERP > 0. However, no strong correlation was detected between deletion frequency and other features, showing a low correlation coefficient (Fig. S8). The distribution of each feature also showed a similar trend, with deletions often present in the regions at both ends of the chromosomes and became gradually sparser closer to the centromere (Fig. 4), which might be due to the poor assembly and low recombination rates around the centromeres.

The proportion of deletion in a gene impacts its expression

A previous study confirmed that deletions may affect gene expression by dosage and transcript regulation (Gamazon and Stranger 2015). To investigate the overall relationship between deletion and gene expression, we compared the gene expression of the deletion overlap in seven different tissues (GRoot, GShoot, Kern, L3Base, L3Tip, LMAD, and LMAN) (Kremling et al. 2018). Normalized expression levels were used to rank the genes, and the mean deletion

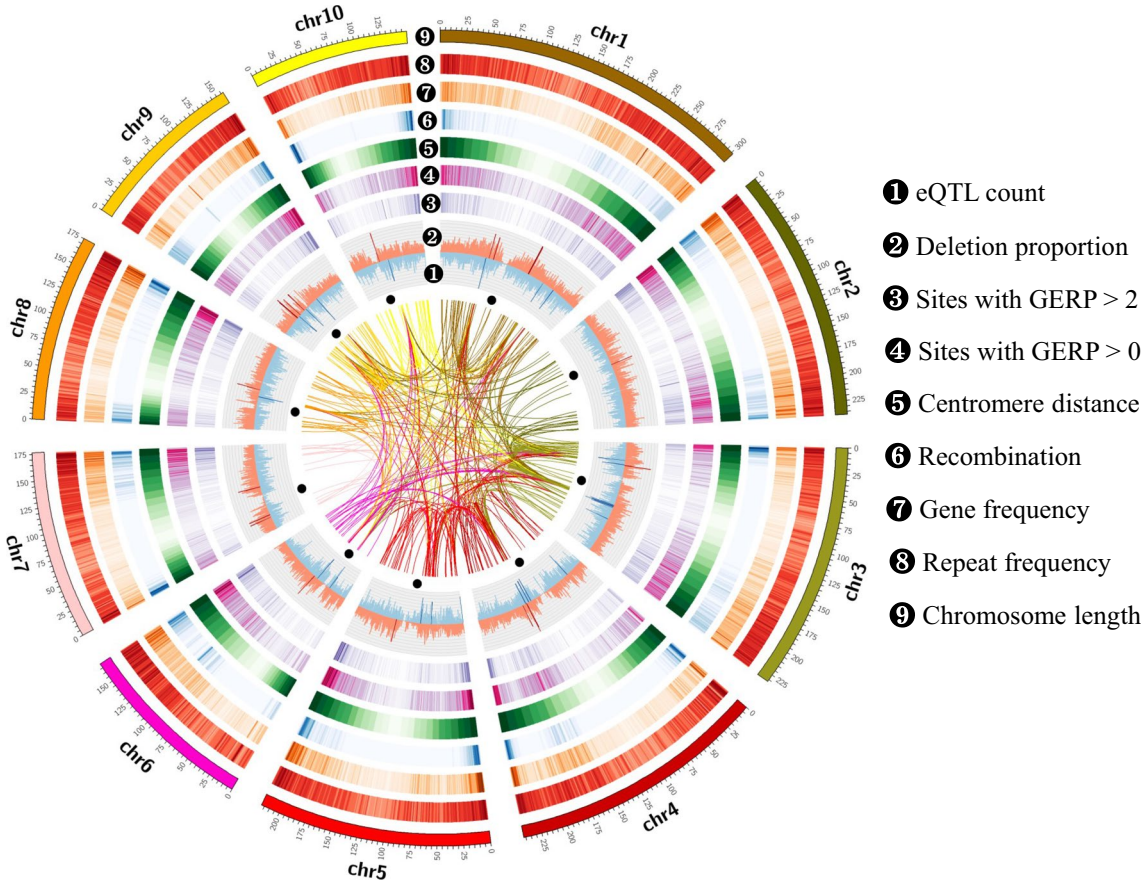

Fig. 4 Circular plot shows the distribution of deletion frequency and other genomic features. The genomic features from outer to inner layer are chromosome length scale plate of the genome, repeat frequency in 1-Mb windows, gene frequency in 1-Mb windows, average recombination in 1-Mb windows, relative centromere distance in 1-Mb windows, site frequency with GERP > 0 in 1-Mb windows, site frequency with GERP > 2 in 1-Mb windows, average deletion proportion in 1-Mb windows (dark red bars mean the windows with deletion proportion larger than 95% of windows), eQTL count in 1-Mb windows (dark blue bars mean windows with eQTL count larger than 95% of windows), the black dots represent centromere position in each chromosome, the inside lines link the eQTL position and associated genes (only eQTL and associated genes located in different chromosomes are displayed)

proportion for each rank was calculated. The deletion proportion was slightly positive correlated with rank in all tissues and stages, indicating that a high proportion of deletion resulted in lower expression values (Figs. S9a–g).

For each gene, inbred lines were divided into two groups according to the presence of absence of deletion overlap or not, and the expression abundance was then compared via Wilcoxon test. After Bonferroni correction ($P < 0.01$), only 697 genes showed a significant difference between the two groups, 181 of which were all-expressed genes, accounting for less than 4% genes of the total gene set in maize (Table 1, Table S8). In multiple tissues, only a small proportion of genes are consistently influenced by deletions, especially in deletion overlapping genes expressed in all inbred lines. This implies that deletions could influence gene expression through a complex regulatory mechanism during different development processes.

Table 1 Number of genes affected by deletion overlapping in expression

Tissue and stage	Detected overlapped genes	All-expressed genes
GRoot	315	81
GShoot	426	97
Kern	383	92
L3Base	380	87
L3Tip	331	92
LMAD	298	66
LMAN	397	92
Unique tissue[a]	201	76
All tissues[b]	149	6

[a]Deletion overlapping can impact gene expression in only one tissue;
[b]Deletion overlapping can impact gene expression in all tissues

GWAS to detect large-scale eQTL caused by deletions

Although deletion can influence gene expression, it is more important to understand which deletion windows can regulate the expression of specific genes. For eQTL mapping, the expression values of each gene were Box-Cox-transformed to fit the normal distribution. However, in Box-Cox transformation, the zero values were adjusted via the addition of a small random value beneath the minimum detection threshold. Thus, only the genes expressed in all individuals of the population for eQTL mapping. Here, we used GWAS for large-scale eQTL detection in seven tissues using deletion alleles. To avoid false positives, the two-step method, according to Fu et al. (Fu et al. 2013), was applied. For each gene, raw GWAS results were filtered using Bonferroni-corrected P thresholds ($P < 0.05$). Then, when accounting for deletion intervals and LD, the deletion windows with lower association or in LD were removed. In total, 36,237 eQTL were identified for 25,669 genes in 7 different tissues. The list included 10,641 unique genes (Table S9), which meant that there were 3.41 eQTL for each gene. Among the genes with eQTL, 71.5% of genes had only one eQTL, 20.1% of genes had two eQTL, and 8.4% of genes had three or more eQTL. When considering the eQTL frequency of each deletion window, 24 deletion windows were enriched with more than ten eQTL across all chromosomes (Fig. 4, Table S10). The transformed positions between the gene and eQTL were plotted to show the exhibition of relative distance. A strong enrichment was observed along the diagonal, implying that majority of eQTL were located around the genes (Fig. 5c, Fig. S10). However, only a small number of eQTL were detected in the gene region (0.3%), and more eQTL were located in the different chromosomes relative to its regulated genes. These results indicated that deletions in promoter or distal regulation played a more important role in gene expression regulation than that in the gene region (Fig. 5b, Table S9). The eQTL distribution peaked at 5–20 kb, descending smoothly until 100 kb away from the target genes (Fig. 5d). We divided eQTL by the cutoff of 50 kb from the regulatory genes into "local" and "distant" eQTL, and found that distant eQTL were much more frequent than local eQTL, although local eQTL showed significantly larger effects (R^2, Wilcoxon test, $P < 0.01$) (Fig. 5e). This implied that although more *trans*-eQTL may broadly play a critical role in gene expression, *cis*-eQTL have more significant effects on a single gene.

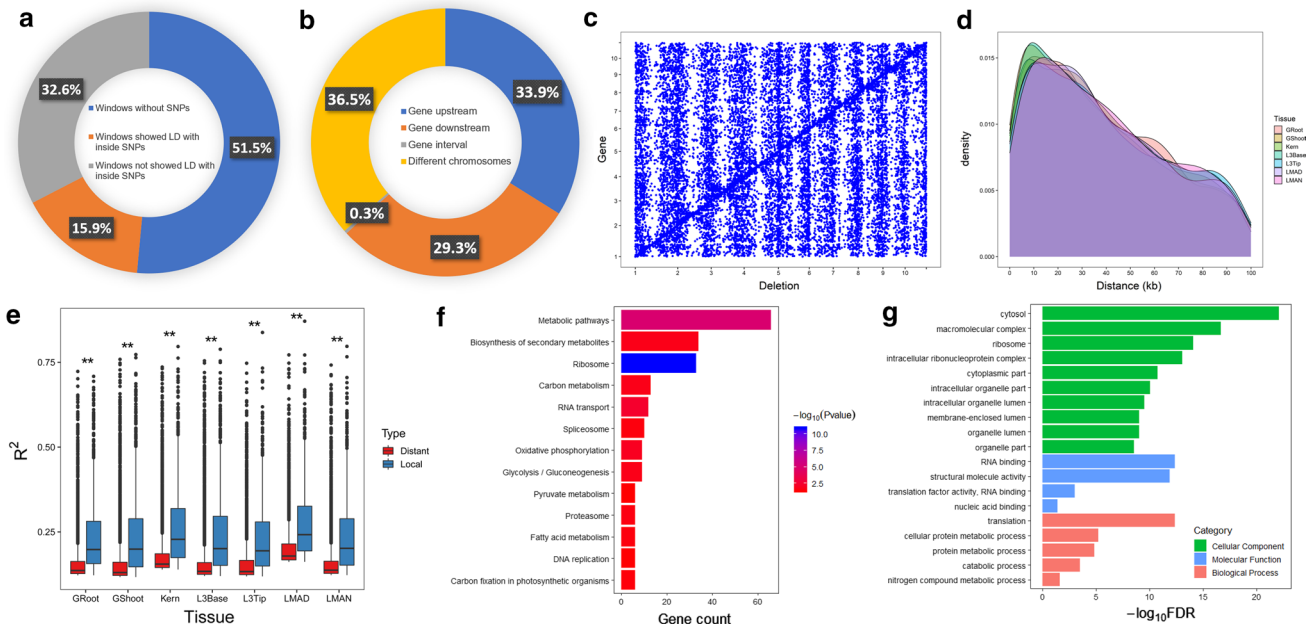

Fig. 5 Summary of eQTL detection in different tissues. **a.** The ratio of deletion windows without SNPs, deletion windows showed LD with SNPs inside the windows, and deletion windows did not show LD with SNPs inside the windows. **b.** The ratio of eQTL located in upstream of the associated gene, eQTL located in downstream of the associated gene, eQTL located within the associated gene region, and eQTL located in the different chromosomes relative to the associated gene. **c.** Genome-wide relative position between deletion window and the associated gene. **d.** Distribution of distances between eQTL and the associated genes in seven tissues. Only eQTL and the associated genes located in the same chromosome were included. **e.** eQTL effects (R^2) between distant eQTL and local eQTL in different tissues using deletions. ** means very significant difference at $P < 0.01$. **f.** KEGG pathway analysis of the genes located in the distant eQTL hotspots. **g.** GO analysis of the commonly associated genes in all tissues

As distant eQTL may be contained core pathways regulating a series of genes, we further studied the hotspots of distant eQTL in the whole genome. The "total" was calculated by merging the distant eQTL of all tissues, and applying a sliding window permutation to estimate the cutoff of the hotspots. Interestingly, hotspots were enriched on chromosomes 1, 3, 4, and 5, and some were consistently detected in different tissues (Fig. S11). Overall, 124 hotspots and 1,619 regulated genes were identified. The target genes regulated by hotspots were enriched in a specific metabolic pathway, including the biosynthesis of secondary metabolites and ribosomes (corrected $P < 0.05$), as shown by KEGG results (Fig. 5f, Table S11).

To determine whether eQTL can be affected by multiple tissues, the eQTL frequency was calculated by multiple tissues, and 301 genes were associated with eQTL in all seven tissues (Table S12). The GO analysis illustrated that these genes were enriched in some essential functions, such as translation, cellular protein metabolic processes, and protein metabolic processes (Fig. 5g). In contrast, total genes and tissue-specific genes were usually enriched on stimulus responses, such as response to stimulus and response to stress (Fig. S12). In addition, the GO analysis of genes specific to maize kernel revealed enrichment of macromolecule metabolic processes and protein metabolic processes, most likely associated with kernel maturity (Fig. S13). Three pairs of relevant tissues in the same stages (GRoot and GShoot, L3Base and L3Tip, LMAD and LMAN) were integrated to analyze eQTL distributions (Fig. S14). The results showed under 50% of the total genes was associated with deletions, and among these, less than half of genes were further regulated by the same eQTL between each pair of tissues.

eQTL identified by deletions cannot be fully captured by SNPs

To compare the effect of eQTL detected by deletions (deletion-eQTL) and SNPs (SNP-eQTL), we also detected eQTL from genome-wide SNPs extracted from the same raw data. The results were analyzed using a similar pipeline as deletions, but the window interval was reduced to 5,000 bp. Because there were more SNPs than deletion alleles, a total of 421,680 eQTL were identified for 42,328 gene/tissue combinations, including 13,180 unique genes (32.00 eQTL for each gene) (Table S13, Fig. S15). By setting a relative distance of 50 kb from the target gene, all eQTL were grouped into "local" or "distant" eQTL. Similar to deletions, there were more distant eQTL than local eQTL, but local eQTL tended to have larger effects (Wilcoxon test, $P < 0.01$) (Fig. S16).

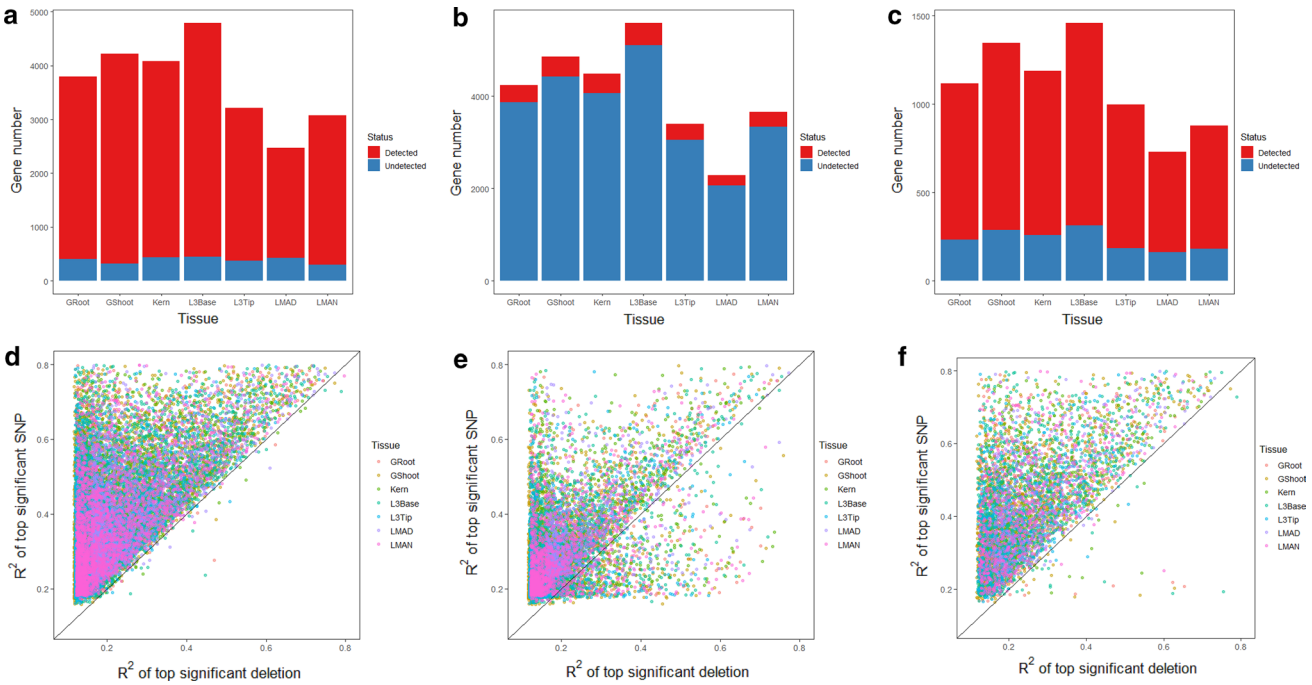

Fig. 6 eQTL detection between deletions and SNPs in seven tissues. **a.** Genes with eQTL consistently detected by deletions and SNPs in different tissues. **b.** Genes with distant eQTL consistently detected using deletions and SNPs in different tissues. **c.** Genes with local eQTL consistently detected using deletions and SNPs in different tissues. **d.** QQ plot of effect (R^2) between the top significant SNP and the top significant deletion of total eQTL for the same associated gene. **e.** QQ plot of effect (R^2) between the top significant SNP and the top significant deletion of distant eQTL for the same gene. **f.** QQ plot of effect (R^2) between the top significant SNP and the top significant deletion of local eQTL for the same gene

To explore the eQTL detection efficiency by SNPs, the eQTL detected by deletions and SNPs were combined for further analysis. The results showed that eQTL were also detected by SNPs in 22,959 out of 25,669 genes, accounting for 89.4% of the total genes (Fig. 6a). To better understand the relationship between eQTL identified in deletions and SNPs, the distance between the SNP-eQTL and deletion-eQTL was estimated for the same gene. With a threshold of 50 kb, 79.00% of the local deletion-eQTL could be consistently detected, whereas only 9.11% of distant deletion-eQTL could be captured by SNP-eQTL (Fig. 6b, c). A total of 67 deletion-eQTL associated with 60 unique genes were well identified by SNPs, with the SNP-eQTL was located within the deletion windows (Table S14).

Next, we compared the effect (R^2) between deletion-eQTL and SNP-eQTL. Owing to the existence of multiple deletion-eQTL and SNP-eQTL for a single gene, only the top deletion-eQTL and top SNP-eQTL with the largest R^2 were compared. The results showed that SNP-eQTL always showed a significantly (Student's t-test, $P < 0.01$) larger effect compared to deletion-eQTL (Fig. S17). QQ plots of paired comparisons between the top deletion-eQTL and top SNP-eQTL illustrated that only a few top deletion-eQTL have larger effects than the top SNP-eQTL. Among them, more distant top deletion-eQTL exhibited larger effects than top SNP-eQTL compared with local deletion-eQTL (Fig. 6d, e, f).

GWAS of deletion alleles in common traits

As this population has been widely used for GWAS, there is sufficient phenotypic data for analysis. In our study, two classical traits, that is, flowering time and PH, were used for GWAS. After Bonferroni correction for a $P < 0.01$, 44 loci were significantly associated with six different traits (Fig. S18, Fig. S19). Among them, the seven deletion alleles were associated with multiple traits of flowering time (Table S15). The association of the deletion window Chr. 8: 130,311,001–130,314,000 with four traits (GDD_DTA, GDD_DTS, DTA, and DTS) close to *Vgt1*, was validated in several previous studies (Li et al. 2016; Salvi et al. 2007). In addition, compared with the eQTL detection results, many significant deletion windows were also associated with genes in expression. The deletion window of Chr. 8: 144,354,001–144,357,000 was associated with the expression of GRMZM2G159053 in four different tissues, indicating that these genes may be located in the same regulatory pathway and jointly regulate flowering time together.

Similarly, we performed GWAS on the same phenotypic data using SNPs derived from the same sequencing dataset (Fig. S20, Fig. S21). After a Bonferroni-corrected $P < 0.01$, a total of 73 SNPs across all chromosomes were significantly associated with nine traits (Table S16), including 6 SNPs associated with all traits related to flowering time (ASI, DTA, and DTS). Compared with Li et al. (Li et al. 2016), four genes, that is, GRMZM2G101852, GRMZM2G137387 (*mads20*), GRMZM2G169927, and GRMZM2G127121 (*zcn16*), consistently located within 1 Mb of 220 flowering time candidate genes. Compared with the results of deletions, only a few markers were consistently detected in flowering time and PH traits (Table 2, Fig. S22), which indicated that combining deletions and SNPs from the same dataset for GWAS was an excellent strategy to dissect trait architecture.

Discussion

The joint application of multiple read-depth methods is a good strategy for calling deletions based on a reference genome with resequencing data in a large population owing to its high calculation efficiency and lack of interruption by complex genomic variation. In this study, to reduce the false-positive rate of deletion detection, we adopted a strict standard; that is, deletions called by both methods were to be reciprocally overlapped by more than 50%. In addition, we also corrected the read count bias for GC content and the mappability of the whole genome. PAVs identified from three currently released reference genomes were used as a control to determine the deletion threshold. Although this method could not confirm the exact position of deletions, it

Table 2 Associated loci consistently detected by deletion and SNP alleles

Deletion interval	Associated trait	*P* value	Effect	SNP position	Associated trait	*P* value	Effect
3: 243,810,001–243,813,000	DTA	1.64E-7	1.17	1: 243,758,149	EHdivPH	4.70E-18	0.029
3: 243,810,001–243,813,000	DTS	2.45E-9	1.38	1: 243,758,149	EHdivPH	4.70E-18	0.029
3: 18,849,001–18,852,000	GDD_DTA	2.85E-10	37.02	3: 18,948,262	EHdivPH	3.80E-19	0.030
8: 35,223,001–35,226,000	DTS	1.67E-7	1.47	8: 34,540,548	DTA	1.60E-14	21.66
8: 35,223,001–35,226,000	DTS	1.67E-7	1.47	8: 34,540,548	DTS	5.49E-13	19.38
10: 88,347,001–88,350,000	GDD_DTS	9.43E-9	−24.04	10: 90,166,649	GDD_DTS	1.08E-11	−119.72
10: 88,347,001–88,350,000	GDD_DTS	9.43E-9	−24.04	10: 88,856,148	PH	8.38E-11	7.81
10: 122,877,001–122,880,000	EH	1.09E-7	4.11	10: 123,641,720	EHdivPH	1.21E-10	−0.012

was an efficient solution to the high false positives caused by shallow depth in the split-read or pair-end methods. Therefore, this method is considered to be suitable for the identification of deletions in shallow sequencing data, but can also be applied to long-read sequencing data. Owing to the great variance among the different maize inbred lines (Schnable et al. 2009), with the development of the maize pan-genome, more reference genomes will be used for alignment, which will expand variance calling in the future.

In this study, the deletions were distributed throughout the whole genome, as shown in Fig. 2a. There was a low deletion frequency in most of the centromere regions, except on chromosomes 2 and 9. As the deletion hotspots are only a 3000-bp window, no deletion hotspots were located in centromeres (Table S4, Fig. 4). The same conclusion was drawn: that centromeres were located near low diversity regions of CNV/PAV except chromosome 9 between B73 and Mo17 (Springer et al. 2009). The centromeres are difficult to map and analyze because of the repetitive sequence (Wolfgruber et al. 2009), and these tandem sequences complicate genome assembly and alignment using short-read sequencing. Poor assembly and alignment lead to low mappability, causing normalized neutral copy numbers in read count methods. Owing to the low combination and conserved sequences in the centromere region, most of the centromeres have low genetic diversity, whereas the numbers of haplotypes were different and could be divided into three groups. Centromeres 1, 4, and 6 show very low diversity; centromere 2, 3, 8, and 9 were in the middle group; and centromeres 5, 7, and 10 had high diversity (Schneider et al. 2016). Therefore, the higher deletion frequency is high on centromere 9, or the surrounding region may be caused by the difference in genetic diversity.

Other than SNPs, SVs are also important in comparative genomics. Therefore, we transformed deletions into applicable markers to detect genetic diversity in our large population. At present, most studies on population structure are based on some classical molecular markers. Compared with bi-allelic or multiple allelic markers such as SNPs or SSRs, deletions are more complicated owing to their varied size. However, in our study, we simplified deletions to a bi-allelic marker based on the copy number status of each window and investigated its power to explain genetic diversity. Compared with SNPs located in the deletion window, the correlation was not robust in this population. There were two reasons for this: (1) the deletion is a 3000-bp window for one allele, the allele frequency of deletion reflects a relative extensive diversity compared with SNPs, which are only one base pair; the other is the deletion is PAV segment or copy number loss relative to "normal window", which has the neutral copy number across the genome in our study, but SNPs are single nucleotide polymorphism, which are not related to the copy number of the located window. In our study, the subpopulation partition was consistent with SSRs and SNPs, indicating that deletions can be used for population structure analysis. Our analysis also illustrated the SS group had the lowest number of deletions among three subgroups, which may be due to the high sequence similarity of the B73 reference line in the SS group. The frequency of deletions was determined to be positively correlated with gene density, repeats, and relative centromere distance, indicating that deletions are enriched in regions described as the far ends of chromosomes, which possess more genes and repeats. These indicate that deletions can probably affect gene structure and are sometimes redundant in the maize genome (Swansonwagner et al. 2010).

The regulation of gene expression is very complex, occurring through regulatory elements and transcription factors in different tissues and at different time points. In our study, both the deletion proportion and eQTL analysis indicated that the effect of deletion on gene expression was different in different tissues. The proportion of deletions was positively correlated with gene expression rank, indicating that larger deletion proportion can reduce gene expression, although large differences were not found. A previous study analyzed the transcriptomes of primary roots among B73, Mo17, and their F_1 hybrids. The results showed that 65 of 1124 genes were expressed in the hybrids but only in one of the parents, showing complementation of CNVs (Paschold et al. 2012). eQTL detection is a standard tool that can help us dissect regulatory pathways and elements (Fu et al. 2013; Wang et al. 2018) and investigate the association between CNV and gene expression. Local and distant eQTL are two relative concepts in eQTL mapping studies, but the cutoff of division for local and distant eQTL is not consistent across studies. Many studies used 20 kb (Fu et al. 2013; Liu et al. 2020a; Pang et al. 2019), while Miculan et al. (Miculan et al. 2021) used 1 Mb as the cutoff. Considering the deletion window size (3000 bp) is a relatively large interval compared with SNPs, the 50-kb cutoff is reasonable to divide the distant and local eQTL. In our study, distant eQTL are much more frequent than local eQTL, implying that *trans*-eQTL were more common than *cis*-eQTL. In contrast, the opposite effect was found on gene expression shows the trend. Deletions can influence gene expression by the presence and absence of *cis*-acting regulators, such as enhancers, repressors, or transcription binding sites located upstream of genes. As gene regulation is usually a complex network, changes of genes from the same network, especially transcript factors, may impact many genes, thus leading to *trans*-regulation. Using seven different tissues, we have demonstrated that the common target genes were enriched in some essential biological functions and that unique eQTL play an important role in stress or stimulus response. This suggests that deletions indirectly

participate in the plant adaption to stress, and different inbred lines may lead to the differences in stress tolerance (Maron et al. 2013) or disease resistance (Dolatabadian et al. 2017).

Flowering time is not only a complex quantitative trait but also a basic adaptative trait. As it is controlled by many small-effect QTL in maize (Buckler et al. 2009), most studies have analyzed flowering time using GWAS and QTL mapping of large populations. Presently, many classical sites such as *Vgt1* (Salvi et al. 2007), *ZmCCT* (Yang et al. 2013b), and *ZCN8* (Meng et al. 2011), and allelic diversity underlying flowering time in landraces indicated that most of the SNPs associated with flowering time are usually located within large structural variants (Romero Navarro et al. 2017). Therefore, GWAS analysis of flowering time using deletion alleles is a complementary study. In this study, an essential gene, that is, *Vgt1*, was associated with four flowering time traits. Previous studies revealed that a miniature transposon (MITE) insertion into a conserved noncoding sequence can cause differential methylation to alter *ZmRap2.7* expression (Castelletti et al. 2014; Salvi et al. 2007). However, only a few consistent genes were identified by SNPs and deletions for the same phenotypic dataset, and most of candidate genes were novel genes that cannot be found in known large-effect genes.

SNPs are the most widely used markers and are known to have essential roles in the analysis of evolution, GWAS, QTL mapping, and heterosis. In this study, we have compared the efficiency of study population structure, eQTL, and GWAS with deletions. We found that PCA analysis using deletion markers could explain population structure, probably owing to strong drift and selection. In contrast, deletions and SNPs cannot reciprocally validate the results of eQTL and GWAS. Owing to the close relative distance between genes and *cis*-eQTL, *cis*-eQTL can be more commonly detected by deletions and SNPs than *trans*-eQTL. Significant SNPs usually have larger effects than deletions, whether or not they are *cis*- or *trans*-eQTL. As we only collected all-expressed genes in this population as a dataset to avoid sequencing or low coverage errors, severe events, such as total gene or regulatory element loss are not included in this study. Therefore, more work is needed to characterize the impact of deletion in genome-wide genes or transcripts with the development of high-throughput third-generation sequencing technology.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-021-03965-1.

Acknowledgements XZ thanks Sara Miller, Xiaolei Liu, Dong Zhang, and Tao Zuo for the help of working in the Buckler lab and gratefully acknowledges China Scholarship Council (CSC) for financial support during studying in USA. The authors also thank Dan Liu, Ling Wu, and Bowen Luo for the help on analyzing data and Duojiang Gao and Shiqiang Gao for the help on data collection in Sichuan Agricultural University of China.

Author contributions XZ performed bioinformatics analysis and wrote the manuscript. YZ performed bioinformatics analysis and review the manuscript. KK participated in large-scale sample collection, RNA-seq and expression data production, and revised the manuscript. MCR managed the field work, collected tissues for next-generation sequencing of the population, and reviewed the manuscript. RB and QS performed raw sequencing data processing and data management. SG, ESB, and FL helped in manuscript discussion and writing. FL and XZ conceived the project. All authors read and approved the final manuscript.

Funding This work was supported by Sichuan Science and Technology Support Project 2021YFYZ0027, 2021YFYZ0020, and 2021YFFZ0017, China Agriculture Research System of MOF and MARA, the National Natural Science Foundation of China (31971955), and the US Department of Agriculture–Agricultural Research Service and the National Science Foundation grant IOS-1238014 to E.S.B.

Data availability The sequencing data of 270 inbred lines are available from NCBI SRA PRJNA389800 and can be also downloaded from CyVerse Data Store: /iplant/home/shared/panzea/raw_seq_282/bam/ (Bukowski et al. 2017). The VCF files of SNP genotyping data with AGPv3 coordinates can be downloaded in the directory: /iplant/home/shared/commons_repo/curated/Qi_Sun_Zea_mays_haplotype_map_2018/282_onHmp321. Sequencing data of RNA-seq have been deposited in the Sequence Read Archive under accession number SRP115041 and in BioProject under accession number PRJNA383416 (Kremling et al. 2018).

Declarations

Conflict of interest The authors declare no competing interests.

References

Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. Nat Rev Genet 16:197

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145-161.e123

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, Sofia da Silva H, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325:714

Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, Wang B, Xu D, Yang B, Xie C, Fan L, Gao S, Xu X, Zhang G, Li Y, Jiao Y, Doebley JF, Ross-Ibarra J, Lorant A, Buffalo V, Romay MC,

Buckler ES, Ware D, Lai J, Sun Q, Xu Y (2017) Construction of the third-generation Zea mays haplotype map. GigaScience 7

Castelletti S, Tuberosa R, Pindo M, Salvi S (2014) A MITE Transposon Insertion Is Associated with Differential Methylation at the Maize Flowering Time QTL Vgt1. G3 Genes Genom Genet, 4:805–812

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM, Consortium GT (2017) The impact of structural variation on human gene expression. Nat Genet 49:692–699

Consortium GP (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at Rhg1; mediates nematode resistance in Soybean. Science 338:1206

Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S (2006) Chromosome territories—A functional nuclear landscape. Curr Opin Cell Biol 18:307–316

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN (2021) How the pan-genome is changing crop genomics and improvement. Genome Biol 22:3

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). Plos One 7:e33234

Dolatabadian A, Patel DA, Edwards D, Batley J (2017) Copy number variation and disease resistance in plants. Theor Appl Genet 130:2479–2490

Fan K-H, Devos KM, Schliekelman P (2020) Strategies for eQTL mapping in allopolyploid organisms. Theor Appl Genet 133:2477–2497

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447:413

Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z, Wang B, Zhai L, Dai C, Xu J, Wang W, Li X, Zheng J, Chen L, Luo L, Liu J, Qian X, Yan J, Wang J, Wang G (2013) RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun 4:2832

Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theor Appl Genet 132:733–750

Gamazon ER, Stranger BE (2015) The impact of human copy number variation on gene expression. Brief Funct Genomics 14:352–357

Golicz AA, Batley J, Edwards D (2016) Towards plant pangenomics. Plant Biotechnol J 14:1099–1105

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009) A first-generation haplotype map of maize. Science 326:1115–1117

Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K (2012) Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res 22:1995–2007

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA (2015) Large multiallelic copy number variations in humans. Nat Genet 47:296

Hansen BG, Halkier BA, Kliebenstein DJ (2008) Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends Plant Sci 13:72–77

Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefusz-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan G-J, Moller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nurnberg P, Romano C, Malafosse A, Koeleman BPC, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41:160–162

Holloway B, Luck S, Beatty M, Rafalski JA, Li B (2011) Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genom 12:336

Huang C, Sun H, Xu D, Chen Q, Liang Y, Wang X, Xu G, Tian J, Wang C, Li D, Wu L, Yang X, Jin W, Doebley JF, Tian F (2018) ZmCCT9 enhances maize adaptation to higher latitudes. Proc Natl Acad Sci 115:E334

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, Della Coletta R, Tittes S, Hudson AI, Marand AP, Wei S, Lu Z, Wang B, Tello-Ruiz MK, Piri RD, Wang N, Dw K, Zeng Y, O'Connor CH, Li X, Gilbert AM, Baggs E, Krasileva KV, Portwood JL, Cannon EKS, Andorf CM, Manchanda N, Snodgrass SJ, Hufnagel DE, Jiang Q, Pedersen S, Syring ML, Kudrna DA, Llaca V, Fengler K, Schmitz RJ, Ross-Ibarra J, Yu J, Gent JI, Hirsch CN, Ware D, Dawe RK (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373:655

Institute B (2019) "Picard Toolkit.". GitHub Repository, http://broadinstitutegithubio/picard/

Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, Gao G (2016) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45:D1040–D1045

Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 9:289–320

Knouse KA, Wu J, Amon A (2016) Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res 26:376

Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES (2018) Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555:520

Kremling KAG, Diepenbrock CH, Gore MA, Buckler ES, Bandillo NB (2019) Transcriptome-Wide Association Supplements Genome-Wide Association in Zea mays. G3 Genes Genom Genet, 9:3023–3033

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li Y-x, Li C, Bradbury PJ, Liu X, Lu F, Romay CM, Glaubitz JC, Wu X, Peng B, Shi Y, Song Y, Zhang D, Buckler ES, Zhang Z, Li Y, Wang T (2016) Identification of genetic variants associated

with maize flowering time using an extremely large multi-genetic background population. Plant J 86:391–402

Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, Tamang TM, Lin J, Liu Y, Zhao M, Hao Y, McFraland F, Wang B, Qin Y, Tang H, McCarty DR, Wei H, Cho M-J, Park S, Kaeppler H, Kaeppler SM, Liu Y, Springer N, Schnable PS, Wang G, White FF, Liu S (2021) Chromosome-level genome assembly of a regenerable maize inbred line A188. Genome Biol 22:175

Liu H, Huang Y, Li X, Wang H, Ding Y, Kang C, Sun M, Li F, Wang J, Deng Y, Yang X, Huang X, Gao X, Yuan L, An D, Wang W, Holding DR, Wu Y (2019) High frequency DNA rearrangement at qγ27 creates a novel allele for Quality Protein Maize breeding. Commun Biol 2:460

Liu H, Luo X, Niu L, Xiao Y, Chen L, Liu J, Wang X, Jin M, Li W, Zhang Q, Yan J (2017a) Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. Mol Plant 10:414–426

Liu L, Du Y, Shen X, Li M, Sun W, Huang J, Liu Z, Tao Y, Zheng Y, Yan J, Zhang Z (2015) KRN4 controls quantitative variation in maize kernel row number. PLOS Genet 11:1005670

Liu Q, Liu H, Gong Y, Tao Y, Jiang L, Zuo W, Yang Q, Ye J, Lai J, Wu J, Lübberstedt T, Xu M (2017b) An atypical thioredoxin imparts early resistance to sugarcane mosaic virus in maize. Mol Plant 10:483–497

Liu S, Li C, Wang H, Wang S, Yang S, Liu X, Yan J, Li B, Beatty M, Zastrow-Hayes G, Song S, Qin F (2020a) Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. Genome Biol 21:163

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020b) Pan-genome of wild and cultivated soybeans. Cell 182:162-176.e113

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914

Lu Y, Shah T, Hao Z, Taba S, Zhang S, Gao S, Liu J, Cao M, Wang J, Prakash AB, Rong T, Xu Y (2011) Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. PLoS ONE 6:e24861

Lye ZN, Purugganan MD (2019) Copy number variation in domestication. Trends Plant Sci 24:352–365

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci USA 110:5241

McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken RS, Vermeesch JR, Hall IM, Gage FH (2013) Mosaic copy number variation in human neurons. Science 342:632

Meng X, Muszynski MG, Danilevskaya ON (2011) The *FT*-Like *ZCN8* gene functions as a floral activator and is involved in photoperiod sensitivity in maize. Plant Cell 23:942

Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr Opin Plant Biol 54:26–33

Miculan M, Nelissen H, Ben Hassen M, Marroni F, Inzé D, Pè ME, Dell'Acqua M (2021) A forward genetics approach integrating genome-wide association study and expression quantitative trait locus mapping to dissect leaf development in maize (Zea mays). Plant J 107:1056–1071

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65

Nuytemans K, Meeus B, Crosiers D, Brouwers N, Goossens D, Engelborghs S, Pals P, Pickut B, Van DBM, Corsmit E (2009) Relative contribution of simple mutations vs. copy number variations in five Parkinson disease genes in the Belgian population. Human Mutation 30:1054–1061

Pang J, Fu J, Zong N, Wang J, Song D, Zhang X, He C, Fang T, Zhang H, Fan Y, Wang G, Zhao J (2019) Kernel size-related genes revealed by an integrated eQTL analysis during early maize kernel development. Plant J 98:19–32

Paschold A, Jia Y, Marcon C, Lund S, Larson NB, Yeh C-T, Ossowski S, Lanz C, Nettleton D, Schnable PS (2012) Complementation contributes to transcriptome complexity in maize (Zea mays L.) hybrids relative to their inbred parents. Genome Res 22:2445–2454

Peiffer JA, Romay MC, Gore MA, Flintgarcia SA, Zhang Z, Millard MJ, Gardner CAC, Mcmullen MD, Holland JB, Bradbury PJ (2014) The genetic architecture of maize height. Genetics 196:1337

Prasad A, Merico D, Thiruvahindrapuram B, Wei J, Lionel AC, Sato D, Rickaby J, Lu C, Szatmari P, Roberts W, Fernandez BA, Marshall CR, Hatchwell E, Eis PS, Scherer SW (2012) A Discovery Resource of Rare Copy Number Variations in Individuals with Autism Spectrum Disorder. G3: Genes Genom Genet, 2:1665–1685

Qian L, Voss-Fels K, Cui Y, Jan Habib U, Samans B, Obermeier C, Qian W, Snowdon Rod J (2016) Deletion of a stay-green gene associates with adaptive selection in brassica napus. Mol Plant 9:1559–1569

Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, He Q, Ou S, Zhang H, Li X, Li X, Li Y, Liao Y, Gao Q, Tu B, Yuan H, Ma B, Wang Y, Qian Y, Fan S, Li W, Wang J, He M, Yin J, Li T, Jiang N, Chen X, Liang C, Li S (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184:3542-3558.e3516

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842

Reymond A, Henrichsen CN, Harewood L, Merla G (2007) Side effects of genome structural changes. Curr Opin Genet Dev 17:381–386

Rodgersmelnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li C, Li Y, Buckler ES (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc Natl Acad Sci USA 112:3823–3828

Romero Navarro JA, Willcox M, Burgueño J, Romay C, Swarts K, Trachsel S, Preciado E, Terron A, Delgado HV, Vidal V, Ortega A, Banda AE, Montiel NOG, Ortiz-Monasterio I, Vicente FS, Espinoza AG, Atlin G, Wenzl P, Hearne S, Buckler ES (2017) A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat Genet 49:476–480

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, Li B, Hainey CF, Radovic S, Zaina G, Rafalski JA, Tingey SV, Miao G-H, Phillips RL, Tuberosa R (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci 104:11376

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski

J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schneider KL, Xie Z, Wolfgruber TK, Presting GG (2016) Inbreeding drives maize centromere evolution. Proc Natl Acad Sci 113:E987

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC (2018) Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15:461–468

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. Plos Genet 5:e1000734

Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLOS Comput Biol 6:e1000770

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet 43:1160

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz HY (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C, Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet 50:1289–1295

Swansonwagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res, 45

Tranchant-Dubreuil C, Rouard M, Sabot F (2018) Plant pangenome: impacts on phenotypes and evolution. Annual Plant Reviews online:1–25

Wang X, Chen Q, Wu Y, Lemmon ZH, Xu G, Huang C, Liang Y, Xu D, Li D, Doebley JF, Tian F (2018) Genome-wide analysis of transcriptional variability in a large maize-teosinte population. Mol Plant 11:443–459

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, Allison J, Birchler JA, Jiang J, Dawe RK, Presting GG (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLOS Genet 5:e1000743

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C-Y, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316–W322

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE 4:e8451

Yang L, Liu B, Huang B, Deng J, Li H, Yu B, Qiu F, Cheng M, Wang H, Yang R, Yang X, Zhou Y, Lu J (2013a) A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. Hum Mol Genet 22:1886–1894

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, Wang Y, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Li W, Liu L, Jackson D, Yan J (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51:1052–1059

Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, Li K, Yang N, Li Y, Zhong T, Li J, Chen Y, Yan J, Yang X, Xu M (2013b) CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci 110:16969

Yuan Y, Bayer PE, Batley J, Edwards D (2021) Current status of structural variation studies in plants. Plant Biotechnol J. https://doi.org/10.1111/pbi.13646

Zhang X, Zhang H, Li L, Lan H, Ren Z, Liu D, Wu L, Liu H, Jaqueth J, Li B, Pan G, Gao S (2016) Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. BMC Genom 17:697

Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M (2020) AthCNV: a map of DNA copy number variations in the arabidopsis genome[OPEN]. Plant Cell 32:1797–1819

Zuo W, Chao Q, Zhang N, Ye J, Tan G, Li B, Xing Y, Zhang B, Liu H, Fengler KA, Zhao J, Zhao X, Chen Y, Lai J, Yan J, Xu M (2014) A maize wall-associated kinase confers quantitative resistance to head smut. Nat Genet 47:151

Authors and Affiliations

Xiao Zhang[1,2,3] · Yonghui Zhu[4] · Karl A. G. Kremling[3] · M. Cinta Romay[3] · Robert Bukowski[5] · Qi Sun[5] · Shibin Gao[1,2] · Edward S. Buckler[3,6] · Fei Lu[3,7,8,9]

Yonghui Zhu
yhzhu86@hotmail.com

Karl A. G. Kremling
kak268@cornell.edu

M. Cinta Romay
mcr72@cornell.edu

Robert Bukowski
bukowski@cornell.edu

Qi Sun
qisun@cornell.edu

Shibin Gao
shibingao@163.com

Edward S. Buckler
esb33@cornell.edu

[1] Maize Research Institute, Sichuan Agricultural University, Chengdu, Sichuan, China

[2] Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Ministry of Agriculture, Chengdu, Sichuan, China

[3] Institute for Genomic Diversity, Cornell University, 175 Biotechnology Building, Ithaca, NY, USA

[4] Crop Research Institute, Sichuan Academy of Agricultural Sciences, Chengdu, Sichuan, China

[5] Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY, USA

[6] USDA-ARS, R. W. Holley Center, Cornell University, Ithaca, NY, USA

[7] State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing, China

[8] CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[9] University of Chinese Academy of Sciences, Beijing, China