

RESEARCH

Open Access



# A pan-Zea genome map for enhancing maize improvement

Songtao Gui<sup>1</sup>, Wenjie Wei<sup>1</sup>, Chenglin Jiang<sup>1</sup>, Jingyun Luo<sup>1</sup>, Lu Chen<sup>1</sup>, Shenshen Wu<sup>1</sup>, Wenqiang Li<sup>1</sup>, Yuebin Wang<sup>1</sup>, Shuyan Li<sup>1</sup>, Ning Yang<sup>1,2</sup>, Qing Li<sup>1,2</sup>, Alisdair R. Fernie<sup>3</sup> and Jianbing Yan<sup>1,2\*</sup>

\*Correspondence:  
jianbing@mail.hzau.edu.cn

<sup>1</sup> National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

<sup>2</sup> Hubei Hongshan Laboratory, Wuhan 430070, China

<sup>3</sup> Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Golm, Germany

## Abstract

**Background:** Maize (*Zea mays* L.) is at the vanguard facing the upcoming breeding challenges. However, both a super pan-genome for the *Zea* genus and a comprehensive genetic variation map for maize breeding are still lacking.

**Results:** Here, we construct an approximately 6.71-Gb pan-Zea genome that contains around 4.57-Gb non-B73 reference sequences from fragmented de novo assemblies of 721 pan-Zea individuals. We annotate a total of 58,944 pan-Zea genes and find around 44.34% of them are dispensable in the pan-Zea population. Moreover, 255,821 common structural variations are identified and genotyped in a maize association mapping panel. Further analyses reveal gene presence/absence variants and their potential roles during domestication of maize. Combining genetic analyses with multi-omics data, we demonstrate how structural variants are associated with complex agronomic traits.

**Conclusions:** Our results highlight the underexplored role of the pan-Zea genome and structural variations to further understand domestication of maize and explore their potential utilization in crop improvement.

**Keywords:** Pan-Zea genome, Gene presence/absence variation (gPAV), Structural variation (SV), Narrow-sense heritability, GWAS

## Background

The increasing population and unpredictability evoked by global climate change have brought new demands to increase the productivity and quality of our crops [1]. Food production must increase 70% by 2050 to feed the increase in the world's population [2]. The past few decades have witnessed a rapid evolution of sequencing and marker technologies alongside the widespread adoption of genome-based breeding approaches [3]. These technological revolutions have promoted innovations in crop breeding from conventional phenotype-based selection to genomics-assisted breeding and genetic engineering [4, 5].

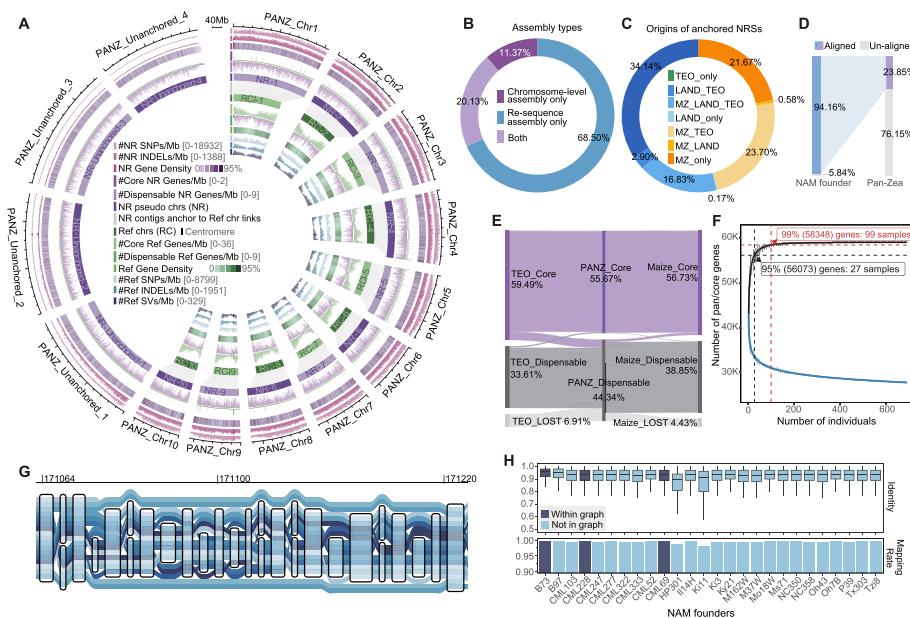
While they harbor great potential, the development of breeding technologies and the explosive growth of biological information have also highlighted the insufficiencies



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

in conventional genomics-assisted breeding strategies. The first of these insufficiencies is the use of a single reference genome. More and more evidences have shown that mapping reads onto a single reference genome can result in reference bias and missing information in highly polymorphic regions and regions that are not present in the genome [6–8]. Thus, a more comprehensive way is to replace the single reference genome with a pan-genome, which represents the complete genetic repertoire of a species. With reduced sequencing costs in recent years, the desirability to construct pan-genomes has spread from *Streptococcus agalactiae* [9] to eukaryotic species [10–12], including many major crops, such as rice, bread wheat, soybean, and tomato [13–16]. Secondly, the conventional genomics-assisted breeding strategies majorly rely on single nucleotide polymorphisms (SNPs) and short insertions/deletions (InDels, hereafter representing insertions/deletions < 50 bp) because they could be easily acquired from low-depth resequencing of cultivated lines. However, SNPs/InDels do not represent the complete genetic repertoire of a species [17, 18]. Other genetic variations, such as structural variations (SVs), also play important roles in plant genetics [19, 20], and their potential should be harnessed for crop breeding and improvement. Besides, applying multi-omic (e.g., transcriptomic, proteomic, metabolomic, and epigenetic) bio-data to reveal genetic mechanisms is becoming more practical [21]. It is highly conceivable that systematic integration of multi-omics data could accelerate crop breeding and improvement [22, 23]. Given these considerations, it follows that to aid in increasing the productivity and quality of crops from the perspectives of genomics and genetics, we should (i) construct a genus-level crop pan-genome, or “super-pan-genome” [24], that includes both cultivated and wild accessions within a genus; (ii) include more genetic variations (e.g., SVs) in addition to SNPs/InDels into genomics-assisted crop breeding, and (iii) systematically integrate multi-omics evidence to accelerate crop breeding.

Maize is a staple crop and a model organism for genetic research [25]. Since the first release of the maize B73 reference genome in 2009 [26], more than 40 maize genomes have been released to date. Moreover, multi-omics maize data, including DNA resequencing [27–31], transcriptomic [32, 33], metabolomic [34, 35], proteomic [36, 37], and epigenomic [38] data, have accumulated at the population scale. Recently, pan-maize gene sets have been constructed from the genome assemblies of the 26 founder lines of the Nested Association Mapping (NAM) population [39] and the population-level transcripts of hundreds of diverse lines [40, 41]. The potential effects of SVs on maize phenotypes have also been investigated [17, 20]. However, a pan-genome of the genus *Zea* (pan-*Zea* genome), including maize and wild taxa, and its graphical representation is still lacking. Here, we (i) constructed a pan-*Zea* genome from 11 public genome assemblies and de novo draft assemblies of 721 accessions, including 507 modern maize, 31 landraces, and 183 teosintes; (ii) revealed the patterns of genes and presence/absence variations in the genus *Zea*; and (iii) identified SVs among the maize population and systematically analyzed the potential role of the pan-*Zea* genome and SVs in maize phenotype variations. These resources and analyses will allow us to more comprehensively understand the genetic bases of complex agronomic traits in maize and provide valuable information for future improvements in maize.



**Fig. 1** Pan-Zea genome, gene, orthologous group PAV, and variant graph genome representation. **A** The genomic landscape of the pan-Zea genome. The pan-Zea genome (PANZ) comprised the AGPV4 reference genomes (RefChr\_1–10) and non-reference sequences (NRINS\_1–10 and NRINS\_Unanchored\_1–4). The reference sequence did not belong to the 10 chromosome-level genome assemblies and the related anchored non-reference sequences were not plotted. See the legend in the circle diagram for details. **B** Proportions of different assembly types in the anchored non-reference sequences. “Chromosome-level assembly only”, non-reference sequences that were only found in the 11 chromosome-level genome assemblies; “Re-sequence assembly only”, non-reference sequences that were only found in the 721 WGS de novo assemblies. **C** Proportion of anchored non-reference sequences with different sub-group origins. TEO, LAND, and MZ represent teosinte, landrace, and modern maize origins, respectively. **D** Comparison of non-reference sequences from this study to those generated from the founders of maize nested association mapping population (NAM founder). **E** Sankey plot of the proportions of the core and dispensable genes in pan-Zea (PANZ), the teosinte sub-group (TEO), and the maize sub-group. **F** Distribution of the number of pan (black) and core (blue) genes along with different numbers of sequenced individuals. See also Additional file 1: Fig. S8D–E. **G** Schematic of the variant graph genome representation for AGPV4 Chr2:171064–171220, with the SNP paths, short InDels, and a large deletion. **H** The identity and mapping rate distribution of the simulated short reads from the genomes of the 26 NAM founders against the variant graph. Dark blue individuals are presented on the variant graph, whereas light blue individuals are not.

## Results

### Pan-Zea genome construction and characterization

The pan-Zea genome was constructed from the alignments against the maize B73 reference genome V4 (AGPV4) of the de novo draft assemblies from 721 individuals, including 507 diverse maize inbred lines [42], 31 landrace individuals [43], and 183 teosinte individuals [44]. An additional 11 chromosome-level assemblies of the *Zea* genus from previous studies were also included (Additional file 1: Fig. S1–S3, Additional file 2: Table S1, Additional file 3: Table S2, Additional file 11: Supplementary Text S1). The resulting pan-Zea genome, with a total length of about 6.71 Gb, comprised ~2.14 Gb of the B73 AGPV4 reference genome (31.83%) and ~4.57 Gb of the non-B73 reference sequences (NRSs, 68.17%). More than half of the NRSs (58.86%) were anchored to the B73 AGPV4 reference genome (Additional file 11: Supplementary Text S1 and Additional file 12: Supplementary Materials and Methods) (Fig. 1A). Of the anchored NRSs, 68.50% were only found in the 721 re-sequence assemblies, and not in the B73 reference genome

(Fig. 1B). Interestingly, more than one-third of the anchored NRSs were not present in the modern maize sequence pool, including 34.14% of the teosinte-specific sequences, 0.17% of the landrace-specific sequences, and 2.90% of the sequences shared by teosinte and the landraces (Fig. 1C). Alignments between our pan-*Zea* NRSs and the NRSs generated from the 26 NAM founders (NAM-NRSs, Additional file 11: Supplementary Text S1 and [39]) showed that the pan-*Zea* NRSs included almost all (98.76%) of the NAM-NRSs, as well as many (69.52%) additional NRSs that were not identified in the NAM-NRSs (Fig. 1D). This result indicates that our pan-*Zea* NRSs considerably enlarged our genetic catalog of the maize gene pool.

The gene models and functional annotations for the pan-*Zea* genome were next generated by merging the AGPv4 reference gene annotations with the non-reference genes that were annotated based on a combination of transcript evidence, homologous protein evidence, and ab initio gene predictions (Additional file 1: Fig. S4 and Additional file 12: Supplementary Materials and Methods), resulting in 58,944 genes (39,591 AGPv4 genes and 19,353 non-reference genes, see Additional file 4: Table S3) and 21,167 orthologous groups (Additional file 12: Supplementary Materials and Methods). About 85.82% of the genes were assigned to at least one functional annotation (Additional file 1: Fig. S5).

The gene presence/absence (gPAV) patterns for each maize inbred and teosinte individual (landrace individuals were excluded from the downstream analysis to avoid bias, leaving 691 genotypes for subsequent analysis, see Additional file 12: Supplementary Materials and Methods) were estimated using a read-mapping-based method that maintained robustness among different read depths (Additional file 1: Fig. S6A). The resulting gPAV patterns followed previous reports (Additional file 1: Fig. S6B–D), with an estimated genotyping accuracy of ~99.71% and ~95.84% for true presence and true absence, respectively (Additional file 1: Fig. S6E). Principal component analysis and linkage disequilibrium (LD) rank analyses of the gPAVs revealed that the gPAVs were related to the population structure and were well represented by SNPs with ~97.37% gPAVs displaying high LD with nearby SNPs (Additional file 1: Fig. S7 and Additional file 12: Supplementary Materials and Methods). Next, to investigate the PAVs of genes and the orthologous groups from a population perspective, we identified the “core” (with population-level loss rate not significantly greater than 1%) and “dispensable” (with population-level loss rate significantly greater than 1%) genes and orthologous groups based on the gPAV and the derived orthologous group PAV (oPAV) matrices (Additional file 1: Fig. S8A, Additional file 12: Supplementary Materials and Methods). The results of these analyses revealed that ~44.34% of the pan-*Zea* genes were dispensable, while only ~7.42% of the pan-*Zea* orthologous groups were dispensable (Fig. 1E and Additional file 1: Fig. S8B). An average of 6020 genes displayed PAV patterns between two individuals, with larger differences for two inter-subspecies individuals (~6779 between one teosinte and one maize) than inner-subspecies (~5520 between two maize and ~5635 between two teosintes) (Additional file 1: Fig. S8C). Given our knowledge of the core and dispensable gene/orthologous groups, we estimated the gene/orthologue-group set size for the pan-*Zea* genome and the core genome. The in silico simulation showed that the pan-*Zea* genome (Fig. 1F), as well as the subspecies pan-genomes (pan-maize genome and pan-teosinte genome, Additional file 1: Fig. S8D–E), displayed characteristics of a “closed pan-genome” [45] with plateaus in the size curves, suggesting that we identified almost

all of the genes in maize and teosinte. The results show that 27 individuals represented an average of 95% of the pan-*Zea* gene-set (range ~91.9 to ~97.3%), while 99 individuals covered 99% of the pan-*Zea* genes (Fig. 1F).

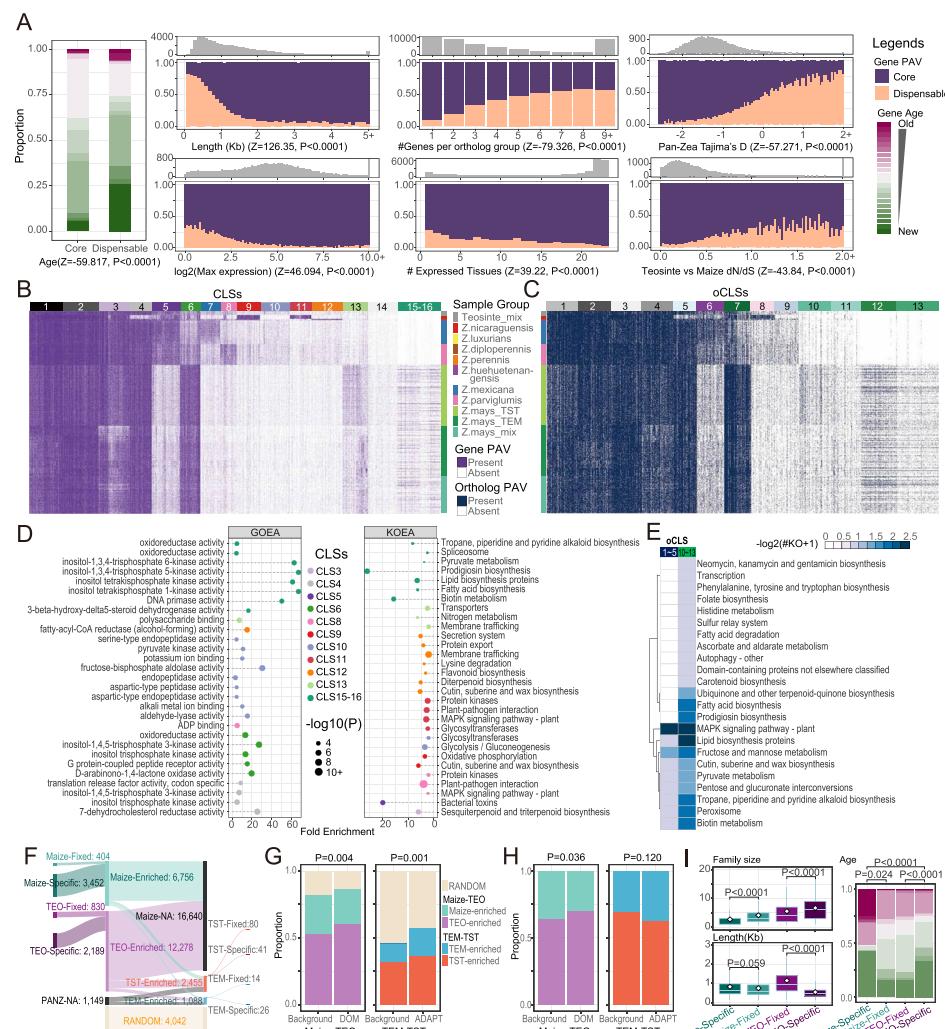
In addition to the linear representation of the pan-*Zea* genome, we also constructed a graph-based genome (Fig. 1G), including the SNPs, short InDels (<50 bp in length), and SVs (>50 bp in length) (see sections below and Additional file 12: Supplementary Materials and Methods for details of the SVs). Furthermore, we estimated a representation of the variant graph genome using reads simulated from the 26 NAM founder genome assemblies. The results showed that reads simulated from the “not-in-graph” NAM founders (23/26) had compatible mapping rates (99.40% on average) and alignment identities (91.00% on average) with those of the “within-graph” NAM founders (3/26, including B73, with 99.44% and 91.14% for the average mapping rate and alignment identity, respectively) (Fig. 1H), indicating that the variant graph represented the vast majority of the maize genetic repertoire.

#### The presence/absence patterns of pan-*Zea* genes and the orthologous groups

An interesting question is which genes are more likely to show gPAVs in the genus *Zea*. To address this question, we investigated the associations between the gPAVs and genic features, including the sub-genome origin, gene age, gene length, orthologue group size, expression levels, and selective constraints (Additional file 12: Supplementary Materials and Methods and Additional file 5: Table S4). The results of these analyses revealed that the gPAV was significantly associated with the genic features under investigation (Fig. 2A). Specifically, they showed that the absence of pan-*Zea* genes was more prevalent among genes that were newly derived, within large orthologous groups, and with genes that were either minimally or tissue specifically expressed. The dispensable genes were more likely to be evolving under relaxed selective constraint than the core genes. These findings are consistent with previous findings that older genes are more essential [46] and associated with higher expression levels and stronger purifying selection [47].

The 691 pan-*Zea* individuals were divided into three maize sub-populations and eight teosinte sub-populations (Fig. 2B and C, [42, 44]). To investigate whether these sub-populations lost genes and orthologous groups evenly, we clustered the dispensable genes into 16 clusters (CLS1–16 in Fig. 2B and Additional file 5: Table S4) and the orthologous groups into 13 clusters (oCLS1–13 in Fig. 2C). The distributions of the gPAVs and oPAVs were associated with the topology of the species tree. The distal-to-maize teosinte subspecies (*Zea nicaraguensis*, *Zea luxurians*, *Zea diploperennis*, *Zea perennis*, and *Zea mays* ssp. *huehuetenangensis*) had more subspecies-enriched genes and orthologous groups than close relatives of maize (*Zea mays* ssp. *mexicana* and *Zea mays* ssp. *parviglumis*).

Enrichment analysis suggested that the gPAV and oPAV clusters may reflect distinct molecular functions among the sub-populations (Fig. 2D–E, Additional file 6: Table S5, Additional file 7: Table S6 and Additional file 1: Fig. S9). Specifically, the teosinte concentrated genes (CLS5, 8, 9, 10, 11, and 12 in Fig. 2A) had enrichment signals, including plant-pathogen interactions, bacterial toxins, biosynthesis of flavonoids, di-/tri- or sesqui-terpenoids, cutin, suberin and wax, and the mitogen-activated protein kinase signaling pathway (Fig. 2D). These enriched pathways are all

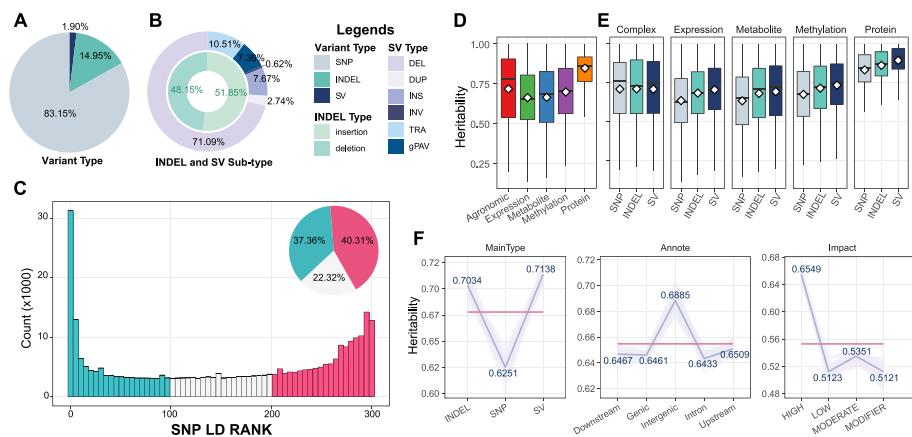


**Fig. 2** The features of the gene and orthologous group PAVs. **A** Distributions of the core and dispensable genes along with features, including gene age, gene length, number of genes per orthologous group, maximum expression, number of expressed tissues, pan-Zea Tajima's  $D$  value, and dN/dS between the teosinte and maize populations. The gray histograms in **C–H** are the distributions of counts for each feature item. **B, C** Distributions of the PAV matrices for the dispensable genes (**B**) and the orthologous groups (**C**) according to the clusters and populations. X- and Y-axes represent the individuals and genes (or orthologous groups), with the top-most and right-most bars indicating the cluster information and sub-population groups, respectively. **D** The lollipop plots of the GO enrichment analyses (GOEA) and KO enrichment analyses (KOEA) of each gene cluster, with the X- and Y-axes indicating the fold enrichment and each enriched item, respectively. The point size indicates the  $P$ -value (only records with corrected  $Q$  values  $< 0.05$  were plotted). **E** The heatmap plot indicates the KEGG pathways enriched in maize and concentrated in the orthologous group clusters (oCLS10–13) when compared to the pan-Zea scattered clusters (oCLS1–5). The color gradient indicates the number of KOs related to the current pathway. **F** Sankey plot of the proportions of unbalanced sub-group genes. **G, H** The distributions of random (**G**) and unbalanced (**H**) genes along the domestication (DOM) and adaptation (ADAPT) selective sweep regions. Background, regions that were not in the top 5% of the selective sweep signals. **I** Features of maize-specific and teosinte-specific genes. Shared legends with **A, F**, and **G**. The Z-scores and  $P$ -values were calculated from 10,000 permutations of the Wilcoxon-Mann-Whitney test. TEO, teosinte; TST, tropical/subtropical maize; TEM, temperate maize; RANDOM, balanced across groups; NA, not available due to core or lost in the entire group; the enriched genes were further divided into two categories: "Fixed", present in all individuals of the current group, and "Specific", present specifically in the current group (absent in all individuals of the other groups)

related to abiotic and biotic stress responses [48]. These findings corroborate the previous finding that reduced genetic diversity during crop domestication leads to the loss of several loci related to the stress response [49], suggesting the potential role of gene loss in the stress-susceptible changes in crops. Comparatively, maize-enriched genes (CLS6, 13, 15, and 16 in Fig. 2A) and orthologous groups (oCLS10-13 in Fig. 2B) were more likely to be related to germination, nutrition, and flavor-related pathways [50, 51]. For example, the maize concentrated orthologous group oCLS7 was enriched in amino acid and protein-related pathways when compared to the teosinte concentrated orthologous group oCLS6 (Additional file 1: Fig. S10), and the orthologous groups that are rarely present in teosintes (oCLS10–13) were enriched in metabolite pathways related to folate, fatty acids, ascorbate, carotenoids, biotin, and various carbohydrates (Fig. 2E).

To further address the contents of teosinte-specific genes and maize-specific genes, we investigated the sub-population gPAV distribution differences in teosinte versus maize (TEO-Maize) and tropical maize versus temperate maize (TST-TEM) in more detail. As results, 51.09% (3452/6756) of the maize-enriched genes were absent in all teosinte individuals (hereafter referred to as maize-specific genes), while 17.83% (2189/12,278) of the teosinte-enriched genes were lost in maize (teosinte-specific genes) (Fig. 2F). We detected 3543 TST-TEM unbalanced (enriched or diminished) genes, with only 1.93% of the sub-group-specific items. A total of 4042 genes were found with no specific distribution preference in any sub-group (random genes, Fig. 2F). Further analysis revealed that the random genes were significantly under-represented in the domestication and adaptation selective sweep regions (Fig. 2G), suggesting that sub-group unbalanced genes were selected during maize domestication and adaptation. Thus, we compared the proportion of different sub-group-enriched genes between selected and background regions (Fig. 2H). This analysis revealed that although the distribution of sub-group-enriched genes was not significantly different between TEM and TST maize, the teosinte-enriched genes were more likely to appear in domestication regions (Fig. 2H), suggesting that some teosinte genes were selected to be lost during domestication.

Another notable question is the pattern of gain-or-loss of teosinte/maize-specific genes, considering that a gene can become group-specific either through (i) loss of all members of the other group or (ii) gained from exogenous sources that were not available to the other group. While the gPAVs showed a predominance of teosinte-enriched genes, the oPAV clusters showed the opposite trend in that more orthologous groups were enriched in maize than in teosinte (Additional file 1: Fig. S11A and B), and the maize-enriched orthologous groups, particularly the maize-specific groups, tended to be smaller (Fig. 2I and Additional file 1: Fig. S11C). The comparisons of group-specific gene features indicated that the teosinte-specific genes showed a typical “easy-to-lose” pattern within larger families of shorter and newer. However, the maize-specific genes showed an opposite pattern of smaller family size and longer and considerably older genes (Fig. 2I). These analyses suggest that the teosinte-specific genes most likely resulted from gene loss, while at least a subset of the maize-specific genes were derived from resources outside of the *Zea* genus, perhaps through horizontal gene transfer from bacteria [52], fungi [53], or pests [54] or from lateral gene transfer with other grasses [55].



**Fig. 3** Features of the maize genetic variation map. **A** Pie plots of the proportions of SNP, InDel, and SV. **B** Pie plots of the proportions of InDel and SV sub-types. **C** Distribution of the number of SV  $r^2$  ranks (0–300) that are above the SNP-based median  $r^2$  value (referred to as SNP LD rank value) for common SVs. **D** Distribution of heritability within different omics trait classes. The white diamond within each box indicates the mean value, and hereinafter. **E** Distribution of heritability of the different variant types among the different omics trait classes. **F** Trend lines of heritability among the different genetic variant features. Light purple lines indicate the feature item's mean heritability of all traits for each randomization. Blue lines indicate each feature item's mean heritability for all 100 randomizations. Red lines indicate the mean heritability of all of the feature items. See also Additional file 1: Fig. S13

### Complementing the maize genetic variation map with structural variations

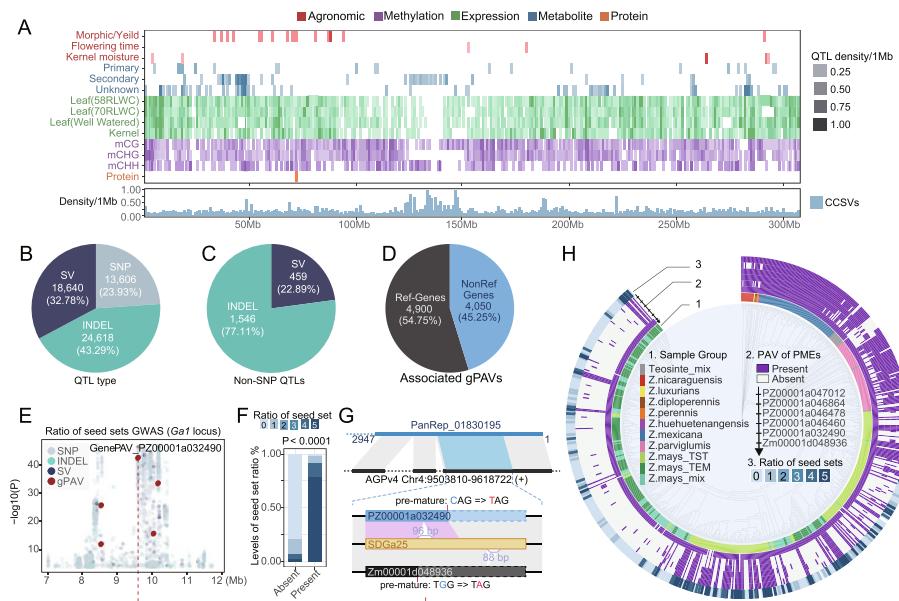
SVs have received significant attention and are responsible for various complex traits in many species [13, 18, 56]. Previous research has constructed a high-density haplotype map of the genus *Zea* using the same populations by mapping with the B73 reference genome [44]. To complement the genetic variation map and estimate the impact of SVs on the maize phenotypic variations, we constructed a comprehensive SV map by integrating evidences from variant graphs, comparative genomics, and short-read alignments. We filtered the SVs with a set of strict conditions, and only the common (MAF  $> 0.05$ ) SVs within the maize population were retained (Additional file 1: Fig. S12 and Additional file 12: Supplementary Materials and Methods). We detected 274,649 common SVs, including 181,874 deletions (DELs), 19,628 insertions (INSs), 26,894 translocations (TRAs), 7020 duplications (DUPs), and 1577 inversions (INVs), as well as 18,828 common gPAVs from the aforementioned gPAV matrix. A total of 11,208,912 SNPs and 2,015,663 InDels (1,045,218 short-insertions and 970,445 short-deletions) that were common in the maize population were extracted from the *Zea* haplotype map and were combined with the common SVs identified in the current study to form a maize common genetic variation map for downstream analyses (Fig. 3A and B and Additional file 1: Fig. S13A-B). Considering the repeat-rich nature of maize, we also assigned the SVs to their closest transposable elements (TEs) according to physical overlap and sequence similarity (Additional file 12: Supplementary Materials and Methods) and found that ~60.03% of the SVs were TE-related (Additional file 1: Fig. S13B), indicating the TE origin of a sizeable proportion of SVs in maize. The size of most of the genotyped common SVs was smaller than 5 Kb (Additional file 1: Fig. S13C), which may have been caused by the limitation in WGS short reads [56]. The SNPs, InDels, and SVs showed similar MAF distribution patterns that skewed toward rare variants (Additional file 1: Fig. S13D). The estimate of the representativeness of each SV by nearby SNPs revealed that 37.36% of

the SVs showed low LD levels with nearby SNPs (Fig. 3C and Additional file 12: Supplementary Materials and Methods), suggesting that these SVs could harbor information that cannot be represented by nearby SNPs.

With the representative genetic variation map that covered the typical genetic variation types, we conducted comprehensive investigations on the differences in the partitioning of narrow-sense heritability ( $h^2$ ) among the different genetic variation types. To address this, we estimated the phenotypic variance explained by the genetic variations from a collection of complex agronomic phenotypes and multi-omics molecular trait data, including gene expression, metabolites, protein contents, and DNA methylation (Additional file 8: Table S7 and Additional file 12: Supplementary Materials and Methods). This analysis revealed that each class of the multi-omics traits displayed high heritability (average values of 0.72, 0.66, 0.67, 0.70, and 0.85 for agronomic traits, expression, metabolites, DNA methylation, and proteins respectively; see also Fig. 3D), indicating their capability to characterize the heritability patterns among the genetic features. The  $h^2$  values of the SVs were higher than those of the SNPs and InDels in the multi-omics molecular traits (Fig. 3E) even though there were significantly more SNPs (43.8 times) and InDels (7.9 times) than SVs. To estimate the  $h^2$  differences in an unbiased manner, we binned, or partitioned and randomized, the genetic variations to keep each of the compared features, including MAF, the SNP LD rank, variant types/sub-types, genomic locations, and impacts on the genes, in the same volume (see Additional file 12: Supplementary Materials and Methods). As a result, the SVs showed a more clearly decreasing  $h^2$  pattern with the increase in the MAF than SNPs/InDels. The  $h^2$  values of the SVs were negatively correlated with the SNP LD rank values, in contrast to those of InDels (Additional file 1: Fig. S13E). When came into the same volume, SVs (with an average  $h^2$  of 0.71, range 0.70–0.72 for each randomization) explained an average of 14.19% and 1.48% more phenotypic variance than the SNPs (with average  $h^2$  of 0.63, range 0.60–0.65 for each randomization) and InDels (with an average  $h^2$  of 0.70, range 0.68–0.72 for each randomization), respectively (Fig. 3F). Higher  $h^2$  values were found in the intergenic variants, genic variants with a high impact on genes, gPAV, and TRA type of SVs, as well as the LTR and helitron-related SVs (Fig. 3F and Additional file 1: Fig. S13F). These findings suggest that SVs are more likely to lead to functional changes than other variants.

### The impact of the pan-Zea genome and structural variations on maize phenotypic variations

The comprehensive genetic variation map and the multi-level phenotypes provided opportunities to further investigate the potential effects of different genetic and genomic features on the phenotypes. To this end, we performed genome-wide association analyses for the complex traits (agronomic traits, metabolites, and protein contents) and local association analyses for the molecular traits (gene expression and DNA methylation) (see Additional file 12: Supplementary Materials and Methods for details). A total of 21,255 non-redundant QTLs with a median QTL interval of ~152.77 Kb were identified for 21,206 different traits (Fig. 4A, Additional file 1: Fig. S14 and Fig. S15). About 32.78% of the identified QTLs were SV-QTL (Fig. 4B). Among them, 459 QTLs were SV-specific (could only be identified by SVs, Fig. 4C and Additional file 9: Table S8). The proportion



**Fig. 4** Characteristics of the phenotype associated QTLs, genes, and causal variants. **A** Distribution of the associated QTLs, genes, and causal variations (example with chromosome 1, see Additional file 1: Fig. S14 for the whole genome). The heatmap represents the QTL density within each 1-Mb window, while the histograms indicate the density of the candidate causal variants that were normalized with the number of all variants within the 1-Mb window. **B** Proportions of QTL types (QTLs lead with different variant types). **C** Proportions of INDEL/SV-specific QTLs (QTLs that cannot be detected by SNPs). **D** Proportions of associated gPAVs for the reference genome genes (Ref-Genes) and non-reference genome genes (NonRef Genes). **E** Manhattan plot of the association result of the *Ga1* locus related to the ratio of seed sets, with gPAVs highlighted in red. **F** The proportion of different levels of seed set ratio related to the absence/presence of PZ00001a032490; the larger the number, the higher the seed set ratio. **G** The genome alignment indicates the anchoring of the NRS (PanRep\_01830195) on the AGPV4 genome, and the schematic plot illustrates the differences between the three *PME* genes (PZ00001a032490, SDGa25, and Zm00001d048936). Solid rectangles indicate the gene coding sequence, while the dashed rectangles indicate the missing coding part related to SDGa25. Gray ribbons indicate the matched blocks. Pink ribbons indicate the matched CDS blocks. **H** Distribution of the PAV patterns (track 2) of the six *PME* genes and the ratio of seed sets (track 3) according to the structure tree of pan-Zea individuals (track 1). **I** Distribution of the number of presented *PME* genes (# PMEs) related to the levels of the seed set ratio. The gray histogram is the distribution of total sample numbers (count) according to the X-axis, while the colored histogram indicates the proportions. The *P*-value was calculated from 10,000 permutations of the Wilcoxon-Mann-Whitney test

of SV-QTLs was much higher than the proportion of SVs in all of the variants, indicating that SVs are more likely to lead to functional changes.

About 47.54% (8950) of the tested gPAVs had associated signals (Fig. 4D and Additional file 12: Supplementary Materials and Methods), suggesting the functional importance of these gPAVs. An excellent example is shown in Fig. 4E–H in which a premature pectin methylesterase (*PME*) gene (Zm00001d048936) at the maize *gametophyte factor1* (*Ga1*) locus was reported to be causative of unilateral cross-incompatibility [163]. In our results, the major QTL contained at the *Ga1* locus was also detected to underlay the unilateral-crossing seed set ratio (a representation of the unilateral cross-incompatibility-related trait, see also Additional file 12: Supplementary Materials and Methods). Additionally, the presence of a non-reference gene (PZ00001a032490) was significantly related to a high ratio of seed setting ( $P = 1.91\text{E}-43$ ,  $R^2 = 0.51$ , Fig. 4E and F). The non-reference gene PZ00001a032490 was anchored to the position where the premature

*PME* gene Zm00001d048936 was located in AGPv4. Further alignment showed that PZ00001a032490 was highly similar to the intact *PME* gene in SDGa25 reported by Zhang et al. [163] and the Zm00001d048936 flanking sequence, with a 96-bp DEL and a premature nonsense point mutation (Fig. 4G). Although both were premature, the gene body of PZ00001a032490 had an additional 126 bp coding sequence that was absent in Zm00001d048936 (Additional file 1: Fig. S16), which contributed to the PAV polymorphism that failed to be detected in Zm00001d048936 (Fig. 4H). These results indicate that although most of the gPAVs were represented by nearby SNPs, leveraging the pan-Zea genome and gPAVs was highly useful in identifying candidate genes, which could not be directly detected using a single reference genome. Notably, in addition to the two *PME* genes (Zm00001d048936 and PZ00001a032490) in the *Ga1* locus, we also found four more non-reference *PME* genes (PZ00001a047012, PZ00001a046864, PZ00001a046478, and PZ00001a046460) (Additional file 1: Fig. S17). These *PME* genes all showed similar presence/absence patterns with PZ00001a032490, and the presence of these *PME* genes was enriched in teosintes (with presence ratio of ~85.97% in teosintes and ~40.35% in maize, Fisher's exact test *P*-value < 2.2e-16, Fig. 4H). The function of these newly identified *PME*-like genes merits further study.

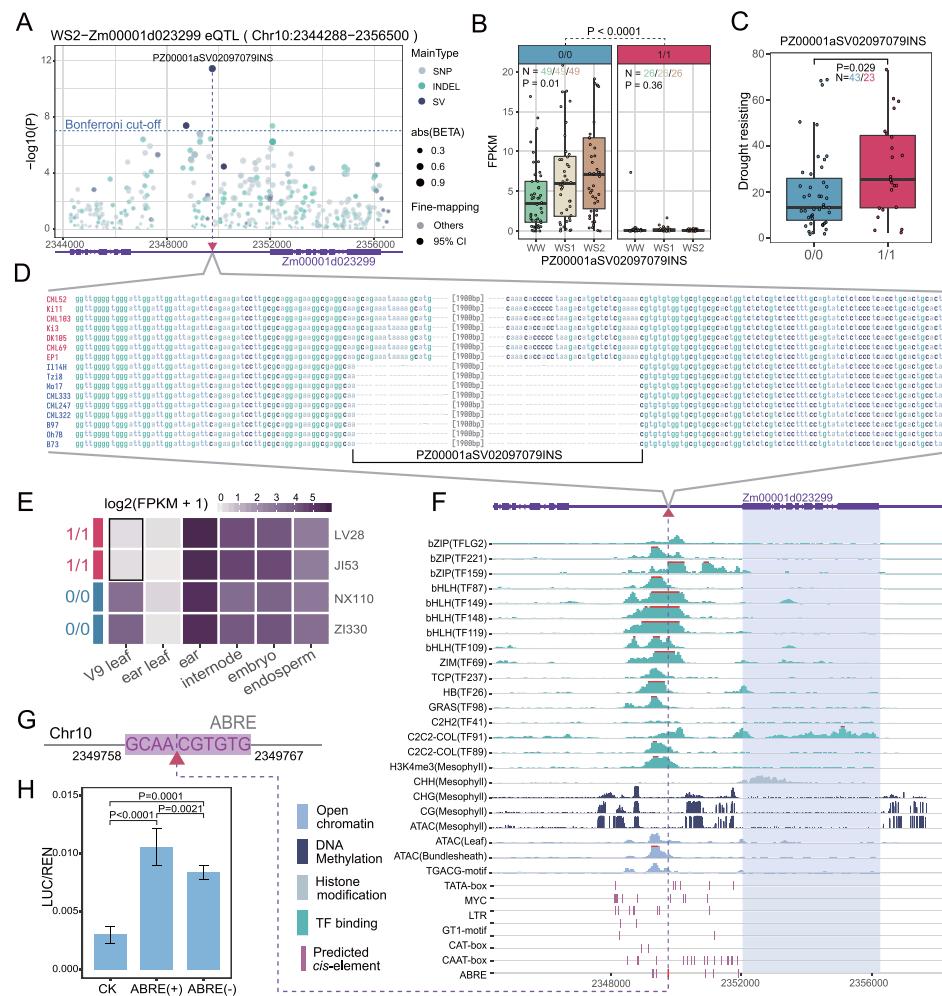
Another question of concern is to estimate the feature priorities to identify the causative genetic variations underlying the phenotypic variations. Thus, we identified the candidate causal variant (CCV) set using a Bayesian-based statistical fine-mapping algorithm [146]. A total of 807,787 genetic variations (3.25% SVs, 15.73% InDels, and 81.02% SNPs) were kept as CCVs, as they were within the 95% confidence interval of the causal variant set for at least one trait. On average, the statistical fine-mapping kept ~18 variants as CCVs from ~229 nominally associated variants (with *P* < 0.001, see Additional file 12: Supplementary Materials and Methods) for each QTL. The number of CCVs was poorly correlated with the QTL quality score, the significance of the leading variant, and the number of genetic variations within the QTL (Additional file 1: Fig. S18A), indicating that CCVs could reflect additional information that cannot be represented by using the leading variants alone. The estimate of the effect sizes of the CCVs showed that SVs and gPAVs had a larger effect size than that of the SNPs/InDels (Additional file 1: Fig. S18B). The general feature enrichment analyses between the CCVs and the nominally associated variations (see Additional file 12: Supplementary Materials and Methods) showed that the SVs, particularly INSSs and gPAVs, were more likely to be enriched in the causal variant sets than SNPs or InDels (Additional file 1: Fig. S18C). Specifically, INSSs were enriched in expression, metabolites, and methylation, while gPAVs were only found enriched in expression (Additional file 1: Fig. S19). For SVs related to different TE classes, the helitron and TIR-related SVs were more likely to be causal than the LTR-related SVs. Genic variants, particularly those with a high impact on genes, were more likely to be causal (Additional file 1: Fig. S18C).

To further investigate the effect of different genetic variations on gene expression, we estimated the enrichment of CCVs in the *cis*-eQTLs along their distance to the transcription start site (TSS). The results showed that the CCVs were enriched in TSS-nearby regions ("Causal variants" track in Additional file 1: Fig. S20), following previous results [164]. Further investigation revealed that the SVs displayed the waviest trend for the fold enrichment changes along distances to the TSS than INDELs and SNPs ("SNP"

“INDEL,” and “SV” tracks in Additional file 1: Fig. S20). This pattern indicated that the TE-related SVs were more likely to be enriched in the upstream regions of TSSs (“TE-related SVs” track in Additional file 1: Fig. S20). These findings suggest that rather than directly affect TSSs or gene body regions, SVs (particularly TE-related SVs) would be more likely to affect gene expression by affecting nearby upstream regions of the gene, where most *cis*-regulatory elements (CREs) are located [165].

A detailed example is illustrated in Fig. 5. The impact of a SV on the expression of Zm00001d023299, a zinc finger CCCH domain-containing protein ([ZEAMAP](#) [57]) that has been previously proposed to be a candidate QTL (marked as IDP103 in [Maize GDB](#)), related to drought and ultraviolet stressors [58] and was highly expressed in response to various stressors in maize (Additional file 1: Fig. S21). In the current study, an SV-specific eQTL related to the expression of Zm00001d023299 in drought-stressed leaves harbored a CCV named PZ00001aSV02097079INS (Fig. 5A). The expression of Zm00001d023299 in leaves [33] responded to different drought treatment levels, and the presence of PZ00001aSV02097079INS suppressed gene expression in leaves (Fig. 5B). The presence of PZ00001aSV02097079INS could also increase the survival rate of maize under drought stress [59] (Fig. 5C). Comparisons among maize genomes indicated that PZ00001aSV02097079INS is a 1947-bp Harbinger-transposon-like sequence (Additional file 1: Fig. S22) inserted 2269 bp upstream of Zm00001d023299 (Fig. 5D). This evidence suggests that PZ00001aSV02097079INS could be one of the causes of maize drought resistance by suppressing the expression of Zm00001d023299 in leaves. An investigation into the expression patterns in different tissues of four maize founder individuals from the Complete-diallele design plus Unbalanced Breeding-like Inter-Cross (CUBIC) population [60] with/without PZ00001aSV02097079INS showed that the impact of the PZ00001aSV02097079INS was restricted to the elongation stage (V9) of leaves (Fig. 5E). These associations were also validated in CUBIC offspring (Additional file 1: Fig. S23). These findings suggest that the suppressed expression was most likely caused by affecting tissue-specific CREs. Further epigenetic evidence and TF binding sites in maize leaves [61] revealed typical patterns of active regulatory elements near the inserted region of PZ00001aSV02097079INS, which lacked DNA methylation, and signals of several TFs, particularly the basic leucine zipper (bZIP) and basic/helix-loop-helix (bHLH) TF families (Fig. 5F). There were numerous predicted CREs within the upstream region of the Zm00001d023299 target gene (Fig. 5F), and remarkably the insertion of PZ00001aSV02097079INS was located exactly within a predicted abscisic acid responsive element (ABRE) motif (Fig. 5G). The activity of the predicted ABRE was validated by a luciferase experiment demonstrating that ablating the function of the ABRE significantly reduced the expression of its downstream target gene (Fig. 5H, Additional file 1: Fig. S24, and Additional file 10: Table S9). This finding suggests that PZ00001aSV02097079INS may have affected the expression of Zm00001d023299 in maize leaves by transposing into an ABRE motif region and blocking the binding of some tissue-specific TFs (bZIPs and/or bHLHs), which suppressed the tissue-specific expression of Zm00001d023299 in leaves, and contributed to drought tolerance in maize.

In conclusion, we revealed the genomic feature priorities that were more causatively associated with multi-omics-level phenotype variations. We showed that leveraging SVs and pan-*Zea* genome-based gPAVs can be used to detect causative associations related



**Fig. 5** An example showing that inserting a transposon suppressed gene expression by destroying the tissue-specific TF binding element. **A** Manhattan plot of an eQTL (Chr10:2344288–2356500) of Zm00001d023299 expression in drought-stressed leaves. Note that the eQTL has no associated SNPs that passed the Bonferroni-corrected  $P$ -value cut-off, the leading variant had a large effect size (BETA), and the highest probability for the cause (fine-mapping) was an insertion (PZ00001aSV02097079INS) located 2269 bp upstream of the target gene. **B** The influence of PZ00001aSV02097079INS on the Zm00001d023299 expression related to different levels of drought stress. WW, well-watered; WS1, 70% relative leaf water content (RLWC); WS2, 50% RLWC; 0/0, the reference allele; 1/1, the alternative allele. The significant differences were evaluated using the Kruskal-Wallis test. **C** The influence of PZ00001aSV02097079INS on the drought resistance-related trait (survival rate under drought stress). The significant differences were evaluated using the Wilcoxon-Mann-Whitney test. **D** The validation of PZ00001aSV02097079INS alleles by multiple genome alignment of the maize reference-level genomes. The genomes with reference and alternative alleles are blue and red, respectively. **E** Comparison of the Zm00001d023299 expression patterns in different tissues of four CUBIC parents with (red) or without (blue) the PZ00001aSV02097079INS insertion. The expression difference in the tenth leaf at the V9 stage is highlighted by the black rectangle. **F** The epigenetic patterns and TF binding sites of maize leaves, and the predicted cis-regulatory elements in the targeted eQTL region. Only items showing distinct peaks near the PZ00001aSV02097079INS were plotted for the epigenetic patterns and TF binding sites. Only the elements located in the positive strand and 4 Kb upstream of the target gene were plotted for the predicted CREs. **G** An illustration showing that the PZ00001aSV02097079INS inserted in a predicted cis-acting element involved in the abscisic acid responsive element (ABRE). **H** The relative promoter activities represented by the ratio of luciferase (LUC) to *Renilla* luciferase (REN) activity (LUC/REN in Y-axis) for the treatments (X-axis), including pGreenII 0800-LUC without a minimal promoter (CK), with ABRE (ABRE+) and without ABRE (ABRE−), see also Additional file 1: Fig. S23. The  $P$ -values were calculated from 10,000 permutations of the t-test.

to important agronomic phenotypic variations that could not be directly identified using single genome-based SNPs.

## Discussion

### Community resources to accelerate maize molecular breeding

To meet the upcoming food production challenge, considerable efforts are required in maize genomics and genetics to further understand the genetics of agronomic traits and increase the efficiency of molecular breeding. In this study, we constructed a genus-level pan-genome for *Zea*, identified the pan-*Zea* gene and orthologous group presence/absence patterns, complemented the maize genetic variation map by including common SV types, and systematically investigated the potential genetic and genomic feature priorities for multi-omics traits by identifying the trait-associated QTLs and candidate causal variants from GWAS summary statistics data. These resources will provide useful information for maize breeding in the following three ways.

First, the wild relatives of maize have shown their potential for increasing stress tolerance and overall yield [62–64]. The de novo fragmented assemblies of the 183 teosinte individuals that covered all seven teosinte subspecies, the pan-*Zea* genome, and the functional annotation of the pan-*Zea* genes and orthologous groups could all be useful alternatives. The pan-*Zea* gene-set size analyses indicated the ability to cover ~95% of the pan-*Zea* genes with a random sampling of 27 individuals, which provides general guidance for individual selection. Furthermore, the gene and orthologous group PAVs were categorized into sub-population-enriched clusters, and these clusters were often enriched with different biological functions. These results will enable breeders to select specific individuals for target traits based on genomic evidence.

Next, we assessed the potentiality of using SVs and the high-density genetic variation map for maize genome-assisted breeding. We constructed a comprehensive SV genotype matrix including all common SV types by combining the evidence from whole-genome comparisons, NGS mapping, and graph-based genotyping. The SV genotypes, along with the gPAV matrix and the previously reported SNPs and InDels from the same association mapping panel were composed into a comprehensive genetic variation map for maize genome-assisted breeding. Based on this genetic variation map and the multi-omics trait variation data of agronomic phenotypes, metabolites, expression, proteins, and methylations, (i) many of the SVs (~37.36%) were not well represented by nearby SNPs; (ii) SVs explained more heritability than SNPs and InDels in the same volume; (iii) SVs were more likely to be the cause of phenotype variation than SNPs and InDels; and (iv) SVs can represent QTLs that cannot be detected by SNPs only. These findings will enhance maize genome-based breeding in the future.

We also analyzed the practical value of the summary association statistics, QTL, associated genes, and causal variants for the multi-omics traits. The summary association statistics have shown their forces by being broadly used for humans in analyses involving gene-based association tests, fine-mapping, polygenic prediction, and cross-trait analyses [65]. While GWAS have been successful in decoding genotype-phenotype associations in maize [66], there is still a lack of comprehensive public release of summary association statistics for this crop. Here, a comprehensive genetic variation map was

developed based on the summary association statistics for the multi-omics traits from a widely used maize association mapping panel [42, 66]. This map could be a resource for more specific analyses, such as imputing genotypes or discovering variant associations with small effects using meta-analyses. A challenge for genome-assisted breeding in the big data era is to explain or predict the biochemical and macroscopic level phenotypes from the underlying genomic and genetic information under different environmental conditions, which have spawned various machine learning applications to improve crops [67]. Thus, the associated QTL and candidate causal variants could be useful resources for optimal weighting of the marker information into genomic selection models [68].

#### **Untangling the genotype-phenotype relationship by leveraging the pan-Zea genome and structural variation map**

A substantial proportion of the challenges and rewards in crop genetics are dependent upon understanding the genetic architecture of complex agronomic traits. Considerable progress has been made to untangle the puzzle of maize phenotypic variations [34, 38, 69–71], yet it still lacks global insight from the perspective of multi-omics integration. Our surveys on heritability of genetic variant feature partitions and the genetic feature priorities for the causes of the phenotypic variations provide an overview of the genetic architecture of agronomic traits in maize.

Estimates and randomization of narrow-sense heritability have shown high levels of maize omics phenotypic variance explained by genome-wide additive genetic factors. Intergenic variations, which were the majority of GWAS association hits, explained most of the phenotypic variations. However, enrichment of candidate causative variations showed that intergenic variants were less likely to be the cause of the GWAS QTL (which need to have large enough effects to be detected) than genic variants. These findings suggest the ubiquitous polygenic nature of maize agronomic traits, a largely additive genotype-to-phenotype relationship in maize, and that non-coding sequences may more likely contribute by adding multiple variants with small effects.

In addition to the general genotype-phenotype association patterns, we have also shown the potential of leveraging the pan-*Zea* genome, the comprehensive genetic variation map, and population-level multi-omics data to reveal genotype-phenotype relationships. Extra efforts would be needed to determine the underlying cause within the QTL when using a single reference genome, even if the association QTL could be detected through SNP-based analyses, particularly when the causative genes are absent in the reference genome. In our results, at least 32.83% of the pan-*Zea* genes were absent in the commonly used maize reference genome. We detected the associations between the causative gPAV and the unilateral cross-incompatibility in maize by harnessing the pan-*Zea* genome. In addition, by leveraging multi-omics data and statistical fine-mapping, we propose the potential mechanism of how a TE-derived SV affected drought resistance in maize by tissue-specific changes in gene expression. These cases have not only shown the practicability of harnessing the pan-*Zea* genome and SVs to substantially reduce the workload of genome-assistant breeding but have highlighted the potential of the pan-*Zea* genome and the SV map to better understand the internal mechanism

of the associations between genetic and phenotypic variations, thus facilitating maize breeding and improvement.

## Conclusions

In summary, we constructed a pan-*Zea* genome, analyzed the gene presence/absence patterns, and investigated the impact of the pan-*Zea* genome and different genetic variant features on maize biochemical and phenotypic variations. These findings will provide useful information for unraveling the genetic architecture of maize complex agronomic traits, accelerate maize molecular breeding, and improve our understanding of maize domestication and adaptation procedure. Still, some limitations have hampered more informative pan-*Zea* genomic and genetic results in the current study.

Perhaps the most obvious holding-back of the pan-*Zea* genome in the current study is the lack of teosinte reference genomes and population-level deep long-read sequencing data for teosintes, landraces, and elite maize. With the fragmental assemblies generated from deep-WGS reads, we had to perform additional data polishing steps to ensure credibility, which inevitably affected the volume of the results, and led to an underestimate of the teosinte sequences and genes within the pan-*Zea* genome, the large and rare SVs, and the untangled variations in the highly repetitive regions.

Foreseeable actions that could address these limitations and boost our understanding of maize breeding and improvement would be to (i) enrich the pan-*Zea* genomic information pool with reference genomes and population-level long-read sequences of teosintes and maize landraces, (ii) enlarge the pan-*Zea* genetic variation matrices and their associations with more biochemicals, (iii) refine the genetic interactions by investigating causation and pleiotropy, and (iv) decrypt the regulatory network of maize phenotypic variation by combining the genetic interactions with other bio-networks, such as interactomes [72] and cell-cell communications [73]. The concept of the pan-genome has expanded from the whole gene and sequence set to the whole set of genomic and genetic variations within a genus. Moreover, with the development of reference-free whole-genome alignments [74] and genome graphic representations [8], it may be time to change the maize reference genomes [7]. We have many reasons to be optimistic when facing the imminent challenges of increasing the productivity and quality of crops with our current resources and the anticipated upcoming progress in crop genomics and genetics.

## Methods

The detailed materials and methods, including (i) collection of genomic and transcriptomic data, (ii) de novo assembly, whole-genome comparison, and pan-genome construction, (iii) pan-genome gene annotation, (iv) pan-*Zea* gene and ortholog group analyses, (v) characterizing gene features, (vi) genotyping and characterizing the maize genetic variation map, (vii) variant graph constructing, reads simulation, and mapping, (viii) phenotype data collection and normalization, (ix) estimating of narrow-sense heritability, (x) identifying trait-associated QTLs, genes, and variants, (xi) analyses of gPAV and SV cases, and (xii) miscellaneous statistical analyses and visualizations, are available in Additional file 12: Supplementary Materials and Methods.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-022-02742-7>.

Additional file 1: Figure S1. Statistics of pan-Zea NGS *de novo* assemblies. Figure S2. Schematic diagram illustrated the filtering of assembly-to-assembly alignment. Figure S3. Sketch of the non-reference sequence anchoring and clustering pipeline. Figure S4. Pan-Zea gene annotation and functional annotation pipeline. Figure S5. Statistics of the proportions of functionally annotated genes. Figure S6. Validation of the gene present and absent variations (gPAVs). Figure S7. PCA and SNP LD rank analysis of gPAV. Figure S8. Core and dispensable genes and ortholog groups. Figure S9. KEGG pathways of each dispensable ortholog group cluster. Figure S10. The KEGG pathways enriched in maize concentrated ortholog groups when compared with teosinte concentrated ortholog groups. Figure S11. Distribution of sub-population enriched genes and ortholog groups. Figure S12. The sketch of SV calling and genotyping pipeline. Figure S13. Additional features of the maize genetic variation map. Figure S14. Distribution of the associated QTLs and causal variations along the chromosome. Figure S15. Distribution of the features of associated QTLs, genes and causal variations. Figure S16. Multiple sequence alignment of the PME genes in the *Ga1* locus. Figure S17. Multiple sequence alignments of the protein sequences of pan-Zea PME genes. Figure S18. Statistics of candidate causal variants. Figure S19. Enrichment of causal variants features within different omics trait classes. Figure S20. Enrichment of cis-eQTL causal variants along different distance to TSS. Figure S21. Expression patterns of the candidate genes of the example SV-QTLs. Figure S22. Blast graphic summary of PZ00001aSV02097079INS. Figure S23. Distribution of the Zm00001d023299 expression patterns in the tenth leaf of V9 stage in the CUBIC offspring population related to the haplotypes flanking 500Kb of Zm00001d023299. Figure S24. The diagram of vectors used in the luciferase reporter assays of Zm00001d023299 promoter with (A) or without (B) the predicted ABRE.

Additional file 2: Table S1. Estimating WGS *de novo* assemblies by whole genome comparison between WGS *de novo* assemblies and reference genome sequences of B73, SK, Mo17 and HZS.

Additional file 3: Table S2. Public genome assemblies used in the pan-Zea genome construction.

Additional file 4: Table S3. Statistics of pan-Zea gene annotations

Additional file 5: Tables S4. Pan Zea gene features.

Additional file 6: Table S5. Enrichment of molecular function GO items for each gene PAV cluster.

Additional file 7: Table S6. Enrichment of KEGG ontology for each gene PAV cluster.

Additional file 8: Table S7. Statistics of the complex and molecular trait variation data in this study.

Additional file 9: Table S8. Information of QTLs that cannot be detected by SNPs.

Additional file 10: Table S9. Primers used in the luciferase reporter assay of the activity of the predicted ABRE positioned upstream of Zm00001d023299.

Additional file 11: Supplementary Text S1.

Additional file 12: Supplementary Materials and Methods [88–159].

Additional file 13. Review history.

### Acknowledgements

The authors would thank Dr. Xin Yang and Mr. Yabing Zhu (Beijing Genomics Institute-Shenzhen) for their help in the functional annotations of the pan-Zea genome. The authors would also thank Dr. Yingjie Xiao, Dr. Yuanyan Chen, Dr. Gengshen Chen, and Mr. Gang Lu (National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University) for providing valuable advices. The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

### Review history

The review history is available as Additional file 13.

### Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Authors' contributions

Conceptualization, S.G., N.Y., and J.Y.; methodology, S.G., W.W., and C.J.; software, S.G. and W.W.; resources, J.L., L.C., W.L., and Y.W.; formal analysis, S.G., W.W., C.J., S.W., and S.L.; writing — original draft, S.G.; writing — review and editing, S.G., N.Y., Q.L., A.R.F., and J.Y.; funding acquisition, S.G. and J.Y.; supervision, J.Y. The authors read and approved the final manuscript.

### Funding

This research was supported by the National Key Research and Development Program of China (2020YFE0202300) and National Natural Science Foundation of China (U1901201; 32000396).

### Availability of data and materials

The accessions of the whole-genome sequencing reads of the pan-Zea individuals could be retrieved from GenBank: the maize association mapping panel lines [75], the teosinte lines [76], and the maize landrace lines [77]. The accessions of the 11 publicly available genome assemblies are listed in Additional file 2: Table S1. The pan-Zea RNA sequencing data (RNA-Seq) were retrieved from GenBank: the RNA-Seq of the kernels of the AMP lines [78] and the RNA-Seq of the

leaves of teosinte lines [79]. The genome assemblies of the maize NAM founders were downloaded from MaizeGDB [80]. The datasets generated in this paper, including the fragmental de novo assemblies of the 701 pan-*Zea* individuals, the pan-*Zea* genome and annotations, the maize structural variation map, and the gene PAV matrices, have been deposited in the China National GeneBank Sequence Archive (CNSA) [81] and figshare [82]. The above data, along with the graphic representation of pan-*Zea* genome and the GWAS summary data for QTL regions, could also be accessed in the download page of the ZEAMAP database [83] and the China National Center for Bioinformation (CNCB) [84]. The related analysis scripts, pipelines, and source codes are available at Github under MIT license and Zenodo: pan-*Zea* genome construction pipeline [85, 160], SV calling and genotyping scripts [86, 161], and miscellaneous analysis scripts [87, 162].

## Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors declare no competing interests.

Received: 11 March 2022 Accepted: 27 July 2022

Published online: 23 August 2022

## References

- Hickey LT, Hafeez AN, Robinson H, Jackson SA, Leal-Bertioli SC, Tester M, et al. Breeding crops to feed 10 billion. *Nat Biotechnol.* 2019;37:744–54.
- Hunter MC, Smith RG, Schipanski ME, Atwood LW, Mortensen DA. Agriculture in 2050: recalibrating targets for sustainable intensification. *Bioscience.* 2017;67:386–91.
- Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 2009;27:522–30.
- Wallace JG, Rodgers-Melnick E, Buckler ES. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu Rev Genet.* 2018;52:421–44.
- Wei X, Qiu J, Yong K, Fan J, Zhang Q, Hua H, et al. A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat Genet.* 2021;53:243–53.
- Yang X, Lee W-P, Ye K, Lee C. One reference genome is not enough. *Genome Biol.* 2019;20:104.
- Balloz S, Dobin A, Gillis JA. Is it time to change the reference genome? *Genome Biol.* 2019;20:1–9.
- Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffari A, Hickey G, et al. Pangenome graphs. *Annu Rev Genomics Hum Genet.* 2020;21:139–62.
- Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci.* 2005;102:13950–5.
- Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature.* 2018;556:339–44.
- Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet.* 2019;51:30–5.
- 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell.* 2016;166:481–91.
- Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell.* 2020;182:145–161 e123.
- Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. *Cell.* 2020;182:162–176 e113.
- Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, et al. The pangenome of hexaploid bread wheat. *Plant J.* 2017;90:1007–13.
- Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature.* 2018;557:43–9.
- Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 2009;5:e1000734.
- Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. *Nat Genet.* 2017;49:692–9.
- Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, et al. Structural variants in 3000 rice genomes. *Genome Res.* 2019;29:870–80.
- Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet.* 2019;51:1052–9.
- Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights.* 2020;14:1177932219899051.
- Scossa F, Alseekh S, Fernie AR. Integrating multi-omics data for crop improvement. *J Plant Physiol.* 2021;257:153352.

23. Li Q, Yan J. Sustainable agriculture in the era of omics: knowledge-driven crop breeding. *Genome Biol.* 2020;21:154.
24. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.* 2020;25:148–58.
25. Wallace J, Larsson S, Buckler E. Entering the second century of maize quantitative genetics. *Heredity.* 2014;112:30–8.
26. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science.* 2009;326:1112–5.
27. Gore MA, Chia JM-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. *Science.* 2009;326:1115–7.
28. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation Zea mays haplotype map. *Gigascience.* 2017;7:gix134.
29. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet.* 2012;44:803–7.
30. Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, et al. Genome-wide genetic changes during modern breeding of maize. *Nat Genet.* 2014;46:1039–40.
31. Wang B, Lin Z, Li X, Zhao Y, Zhao B, Wu G, et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat Genet.* 2020;52:565–71.
32. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet.* 2013;45:43–50.
33. Liu S, Li C, Wang H, Wang S, Yang S, Liu X, et al. Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol.* 2020;21:163.
34. Wen W, Li D, Li X, Gao Y, Li W, Li H, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat Commun.* 2014;5:3438.
35. Wen W, Jin M, Li K, Liu H, Xiao Y, Zhao M, et al. An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. *Plant J.* 2018;93:1116–28.
36. Jiang L-G, Li B, Liu S-X, Wang H-W, Li C-P, Song S-H, et al. Characterization of proteome variation during modern maize breeding. *Mol Cell Proteomics.* 2019;18:263–76.
37. Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, et al. Integration of omic networks in a developmental atlas of maize. *Science.* 2016;353:814–8.
38. Xu J, Chen G, Hermanson PJ, Xu Q, Sun C, Chen W, et al. Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol.* 2019;20:243.
39. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science.* 2021;373:655–62.
40. Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, et al. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci Rep.* 2016;6:18936.
41. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell.* 2014;26:121–35.
42. Yang X, Gao S, Xu S, Zhang Z, Prasanna BM, Li L, et al. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol Breed.* 2011;28:511–26.
43. Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* 2017;18:1–13.
44. Chen L, Luo J, Jin M, Yang N, Liu X, Peng Y, et al. Portrait of a genus: the genetic diversity of Zea. *bioRxiv.* 2021:2021.2004.2007.438828. <https://doi.org/10.1101/2021.04.07.438828>.
45. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol.* 2015;23:148–54.
46. Chen W-H, Trachana K, Lercher MJ, Bork P. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol.* 2012;29:1703–6.
47. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci.* 2009;106:7273–80.
48. Tohge T, de Souza LP, Fernie AR. Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. *J Exp Bot.* 2017;68:4013–28.
49. Warschefsky E, Penmetsa RV, Cook DR, Von Wettberg EJ. Back to the wilds: tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives. *Am J Bot.* 2014;101:1791–800.
50. Shi J, Wang H, Hazebroek J, Ertl DS, Harp T. The maize low-phytic acid 3 encodes a myo-inositol kinase that plays a role in phytic acid biosynthesis in developing seeds. *Plant J.* 2005;42:708–19.
51. Phillipi BQ. Identification of inositol 1,3,4-trisphosphate 5-kinase and inositol 1,3,4,5-tetrakisphosphate 6-kinase in immature soybean seeds. *Plant Physiol.* 1998;116:291.
52. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell.* 2019;179:1057–1067 e1014.
53. Wang H, Sun S, Ge W, Zhao L, Hou B, Wang K, et al. Horizontal gene transfer of *Fhb7* from fungus underlies Fusarium head blight resistance in wheat. *Science.* 2020;368:eaba5435.
54. Xia J, Guo Z, Yang Z, Han H, Wang S, Xu H, et al. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell.* 2021;184:1693–1705 e1617.
55. Hibdige SGS, Raimondeau P, Christin PA, Dunning LT. Widespread lateral gene transfer among grasses. *New Phytol.* 2021;230:2474–86.
56. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet.* 2020;21:171–89.
57. Gui S, Yang L, Li J, Luo J, Xu X, Yuan J, et al. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience.* 2020;23:101241.
58. Richbourg HL. QTL analysis for genes conferring tolerance to drought stress and damage from UV-B radiation. University of North Carolina Wilmington; 2008.

59. Wang X, Wang H, Liu S, Ferjani A, Li J, Yan J, et al. Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. *Nat Genet.* 2016;48:1233–41.
60. Liu H-J, Wang X, Xiao Y, Luo J, Qiao F, Yang W, et al. CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biol.* 2020;21:20.
61. Tu X, Mejia-Guerra MK, Valdes Franco JA, Tzeng D, Chu P-Y, Shen W, et al. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat Commun.* 2020;11:5089.
62. Maxted N, Kell S. Establishment of a global network for the in situ conservation of crop wild relatives: status and needs. Rome: FAO commission on genetic resources for food and agriculture; 2009. p. 266.
63. Prischmann D, Dashiell K, Schneider D, Eubanks M. Evaluating *Tripsacum*-introgressed maize germplasm after infestation with western corn rootworms (Coleoptera: Chrysomelidae). *J Appl Entomol.* 2009;133:10–20.
64. Tian J, Wang C, Xia J, Wu L, Xu G, Wu W, et al. Teosinte ligule allele narrows plant architecture and enhances high-density maize yields. *Science.* 2019;365:658–64.
65. Pasanuic B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet.* 2017;18:117–27.
66. Xiao Y, Liu H, Wu L, Warburton M, Yan J. Genome-wide association studies in maize: praise and stargaze. *Mol Plant.* 2017;10:359–74.
67. van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D. Machine learning in plant science and plant breeding. *iScience.* 2021;24:101890.
68. Ren D, An L, Li B, Qiao L, Liu W. Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits. *Heredity.* 2021;126:320–34.
69. Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, et al. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science.* 2017;357:512–5.
70. Riedelsheimer C, Liseic J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, et al. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc Natl Acad Sci.* 2012;109:8872–7.
71. Deng M, Li D, Luo J, Xiao Y, Liu H, Pan Q, et al. The genetic architecture of amino acids dissection by association and linkage analysis in maize. *Plant Biotechnol J.* 2017;15:1250–63.
72. Wu L, Han L, Li Q, Wang G, Zhang H, Li L. Using interactome big data to crack genetic mysteries and enhance future crop breeding. *Mol Plant.* 2021;14:77–94.
73. Shao X, Lu X, Liao J, Chen H, Fan X. New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell.* 2020;11:866–80.
74. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature.* 2020;587:246–51.
75. BIG: Deep DNA resequencing of the association mapping panel. NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA531553> (2019).
76. HZAU: Zea mays raw sequence reads (teosinte). NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA641489> (2020).
77. ISU: Zea mays subsp. mays Raw sequence reads. NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA300309> (2015).
78. CAAS: Zea mays transcriptome or gene expression. NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA208608> (2013).
79. HZAU: Zea mays raw sequence reads (RNA-seq). NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA645739> (2020).
80. MaizeGDB: The NAM parents sequencing project. MaizeGDB. [https://maizegdb.org/NAM\\_project](https://maizegdb.org/NAM_project) (2021).
81. Gui S: The pan-Zea genome assembly and annotation CNGBdb. (2022). <https://db.cngb.org/search/project/CNP0002007>.
82. Gui S: The pan-Zea\_genome supporting data. FigShare. <https://doi.org/10.6084/m9.figshare.19097447.v1> (2022).
83. Gui S: The ZEAMAP database: Pan-Zea genomic data and GWAS summary data. ZEAMAP. [https://ftp.cngb.org/pub/CNSA/data3/CNP0001565/zeamap/06\\_Pangenome/](https://ftp.cngb.org/pub/CNSA/data3/CNP0001565/zeamap/06_Pangenome/) (2022).
84. Gui S: The pan-Zea genome project. CNCB. <https://ngdc.cncbac.cn/bioproject/browse/PRJCA010654> (2022).
85. Gui S: pan-Zea genome construction pipeline. Github. [https://github.com/songtaogui/pan-Zea\\_construct/tree/v1.0.0](https://github.com/songtaogui/pan-Zea_construct/tree/v1.0.0) (2021).
86. Gui S: pan-Zea SV calling and genotyping scripts. Github. [https://github.com/songtaogui/pan-Zea\\_construct/tree/v1.0.0](https://github.com/songtaogui/pan-Zea_construct/tree/v1.0.0) (2021).
87. Gui S: pan-Zea miscellaneous analysis scripts. Github. [https://github.com/songtaogui/pan-Zea\\_utilities/tree/v1.0.0](https://github.com/songtaogui/pan-Zea_utilities/tree/v1.0.0) (2021).
88. Bushnell B. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. Joint Genome Institute. 2018.
89. Peng Y, Leung HC, Yiu S-M, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28:1420–8.
90. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics.* 2018;34:i142–50.
91. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–100.
92. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013;1303.3997. <https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L>.
93. Kehr B, Melsted P, Halldórsson BV. PopIns: population-scale detection of novel sequence insertions. *Bioinformatics.* 2016;32:961–7.
94. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28:3150–2.
95. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29:644–52.
96. Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2018;47:D506–15.
97. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12:59–60.

98. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:1–11.
99. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 2003;31:5654–66.
100. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34:525–7.
101. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24:637–44.
102. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoScan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*. 2005;33:W686–9.
103. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol*. 2008;9:1–22.
104. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236–40.
105. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21:3674–6.
106. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42:D222–30.
107. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Res*. 2018;46:W200–4.
108. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*. 2015;44:D286–93.
109. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. 2017;34:2115–22.
110. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2017;46:D624–32.
111. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–2.
112. Golicz AA, Martinez PA, Zander M, Patel DA, Van De Wouw AP, Visendi P, et al. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Funct Integr Genom*. 2015;15:189–96.
113. Hu Z, Sun C, Lu K-c, Chu X, Zhao Y, Lu J, Shi J, Wei C. EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics*. 2017;33:2408–9.
114. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat*. 1996;5:299–314.
115. Tang J, Liu J, Zhang M, Mei Q. Visualizing large-scale and high-dimensional data. In proceedings of the 25th international conference on world wide web; 2016. p. 287–97.
116. Zhang YE, Vibranovski MD, Landbeck P, Marais GA, Long M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*. 2010;8:e1000494.
117. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
118. Nelson CW, Moncla LH, Hughes AL. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics*. 2015;31:3709–11.
119. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
120. Freed D, Aldana R, Weber JA, Edwards JS. The Sentieon genomics tools-a fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*. 2017:115717. <http://biorexiv.org/content/early/2017/05/12/115717.abstract>.
121. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallies C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017;8:14061.
122. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31:2032–4.
123. Rausch T, Zichner T, Schlatl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28:i333–9.
124. Cameron DL, Schröder J, Pennington JS, Do H, Molanía R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res*. 2017;27:2050–60.
125. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220–2.
126. Sibbesen JA, Maretty L, Krogh A. The Danish Pan-genome C: accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet*. 2018;50:1054–9.
127. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 2015;12:966–8.
128. Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. 2017;33:2759–61.
129. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17:122.
130. Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, O'Connor CH, et al. Transposable elements contribute to dynamic genome content in maize. *Plant J*. 2019;100:1052–65.
131. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife*. 2016;5:e20777.

132. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol.* 2018;36:875–9.
133. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* 2020;21:1–28.
134. Yang N, Lu Y, Yang X, Huang J, Zhou Y, Ali F, et al. Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* 2014;10:e1004573.
135. Liu J, Huang J, Guo H, Lan L, Wang H, Xu Y, et al. The conserved and unique genetic architecture of kernel size and weight in maize and rice. *Plant Physiol.* 2017;175:774–85.
136. Li W, Yu Y, Wang L, Luo Y, Peng Y, Xu Y, et al. The genetic architecture of the dynamic changes in grain moisture in maize. *Plant Biotechnol J.* 2021;19:1195–205.
137. Jin M, Liu X, Jia W, Liu H, Li W, Peng Y, et al. ZmCOL3, a CCT gene represses flowering in maize by interfering with the circadian clock and activating expression of ZmCCT. *J Integr Plant Biol.* 2018;60:465–80.
138. Li Z, Chen J, Han L, Wen J, Chen G, Li H, et al. Association mapping resolving the major loci Scmv2 conferring resistance to sugarcane mosaic virus in maize. *Eur J Plant Pathol.* 2016;145:385–91.
139. Chen G, Wang X, Hao J, Yan J, Ding J. Genome-wide association implicates candidate genes conferring resistance to maize rough dwarf disease in maize. *PLoS One.* 2015;10:e0142001.
140. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 2020;49:D1388–95.
141. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81:1084–97.
142. Meyer PE, Meyer MPE. Package ‘infotheo’. R Package. 2009;1.
143. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet.* 2012;91:1011–21.
144. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics.* 2007;23:2633–5.
145. Haller T, Tasa T, Metspalu A. Manhattan harvester and cropper: a system for GWAS peak detection. *BMC Bioinformatics.* 2019;20:1–8.
146. Lee Y, Luca F, Pique-Regi R, Wen X. Bayesian multi-SNP genetic association analysis: control of FDR and use of summary statistics. *bioRxiv.* 2018;316471. <http://biorexiv.org/content/early/2018/05/08/316471.abstract>.
147. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
148. Wang Y, Li W, Wang L, Yan J, Lu G, Yang N, et al. Three types of genes underlying the *Gametophyte factor 1* locus cause unilateral cross incompatibility in maize. *Nat Commun.* 2022;13(1):4498. <https://doi.org/10.1038/s41467-022-32180-9>.
149. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32:1792–7.
150. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 2021;49:W293–6.
151. Luo J, Wei C, Liu H, Cheng S, Xiao Y, Wang X, et al. MaizeCUBIC: a comprehensive variation database for a maize synthetic population. Database. 2020;2020.
152. Lescot M, Déhais P, Thijss G, Marchal K, Moreau Y, Van de Peer Y, et al. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 2002;30:325–7.
153. Luo Y, Zhang M, Liu Y, Liu J, Li W, Chen G, et al. Genetic variation in YIGE1 contributes to ear length and grain yield in maize. *New Phytol.* 2022;234(2):513–26. <https://doi.org/10.1111/nph.17882>.
154. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: a python library for gene ontology analyses. *Sci Rep.* 2018;8:10872.
155. Hothorn T, Hornik K, Van De Wiel MA, Zeileis A. A lego system for conditional inference. *Am Stat.* 2006;60:257–63.
156. Gu Z, Gu L, Eils R, Schlesner M, Brors B. CircLize implements and enhances circular visualization in R. *Bioinformatics.* 2014;30:2811–2.
157. Wickham H. ggplot2. Wiley interdisciplinary reviews. *Comput Stat.* 2011;3:180–5.
158. Kolde R, Kolde MR. Package ‘pheatmap’. R package. 2015;1:790.
159. Wilke CO, Wickham H, Wilke MCO. Package ‘cowplot’. R Package. 2019.
160. Gui S: pan-Zea genome construction pipeline. Github. <https://doi.org/10.5281/zenodo.6864803> (2021).
161. Gui S: pan-Zea SV calling and genotyping scripts. Github. <https://doi.org/10.5281/zenodo.6864803> (2021).
162. Gui S. Pan-Zea miscellaneous analysis scripts. Github. <https://doi.org/10.5281/zenodo.6864805> (2021).
163. Zhang Z, Zhang B, Chen Z, Zhang D, Zhang H, Wang H, Zhang Ye, Cai D, Liu J, Xiao S, et al. A PECTIN METHYL-ESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility. *Nat Commun.* 2018;9:3678.
164. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* 2014;10:e1004722.
165. Bilas, Szafrań K, Hnatuszko-Konka K, Kononowicz AK. Cis-regulatory elements used to control gene expression in plants. *Plant Cell Tissue Organ Culture (PCTOC).* 2016;127:269–87.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.