CellPress
PARTNER JOURNAL

Molecular Plant
Research Report

The HuangZaoSi Maize Genome Provides Insights into Genomic Variation and Improvement History of Maize

Chunhui Li[1,6], Wei Song[1,6], Yingfeng Luo[2,6], Shenghan Gao[2,6], Ruyang Zhang[1], Zi Shi[1], Xiaqing Wang[1], Ronghuan Wang[1], Fengge Wang[1], Jidong Wang[1], Yanxin Zhao[1], Aiguo Su[1], Shuai Wang[1], Xin Li[3], Meijie Luo[1], Shuaishuai Wang[1], Yunxia Zhang[1], Jianrong Ge[1], Xinyu Tan[2], Ye Yuan[2], Xiaochun Bi[2], Hang He[3], Jianbing Yan[4], Yuandong Wang[1,*], Songnian Hu[2,5,*] and Jiuran Zhao[1,*]

[1]Maize Research Center, Beijing Academy of Agriculture & Forestry Sciences (BAAFS), Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding, Shuguang Garden Middle Road No. 9, Beijing 100097, China

[2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[3]School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking University, Beijing 100871, China

[4]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

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

[6]These authors contributed equally to this article.

***Correspondence: Yuandong Wang (wyuandong@126.com), Songnian Hu (husn@big.ac.cn), Jiuran Zhao (maizezhao@126.com)**

https://doi.org/10.1016/j.molp.2019.02.009

ABSTRACT

Maize is a globally important crop that was a classic model plant for genetic studies. Here, we report a 2.2 Gb draft genome sequence of an elite maize line, HuangZaoSi (HZS). Hybrids bred from HZS-improved lines (HILs) are planted in more than 60% of maize fields in China. Proteome clustering of six completed sequenced maize genomes show that 638 proteins fall into 264 HZS-specific gene families with the majority of contributions from tandem duplication events. Resequencing and comparative analysis of 40 HZS-related lines reveals the breeding history of HILs. More than 60% of identified selective sweeps were clustered in identity-by-descent conserved regions, and yield-related genes/QTLs were enriched in HZS characteristic selected regions. Furthermore, we demonstrated that HZS-specific family genes were not uniformly distributed in the genome but enriched in improvement/function-related genomic regions. This study provides an important and novel resource for maize genome research and expands our knowledge on the breadth of genomic variation and improvement history of maize.

Key words: HZS, comparative genomic analysis, tandem duplication, pedigree analysis, identity-by-descent

Li C., Song W., Luo Y., Gao S., Zhang R., Shi Z., Wang X., Wang R., Wang F., Wang J., Zhao Y., Su A., Wang S., Li X., Luo M., Wang S., Zhang Y., Ge J., Tan X., Yuan Y., Bi X., He H., Yan J., Wang Y., Hu S., and Zhao J. (2019). The HuangZaoSi Maize Genome Provides Insights into Genomic Variation and Improvement History of Maize. Mol. Plant. **12**, 402–409.

INTRODUCTION

Maize (*Zea mays* subsp. *mays*) is not only a globally important crop but also a classic genetic model plant. It has an annual production of more than 1 billion tons (Yang et al., 2018). It has been predicted that 45% of cereal demand is expected from maize by the year 2050 (Hubert et al., 2010). Maize exhibits an extremely high level of genetic diversity, and its nucleotide diversity is close to the upper limit of that estimated for crops (Tenaillon et al., 2001; Wright et al., 2005; Lai et al., 2010; Chia et al., 2012). The rich genetic diversity in maize has played important roles in its breeding improvement, which is probably the key for yield improvement to meet the demands of global population growth.

The first draft genome of maize inbred line B73 was reported in 2009 (Schnable et al., 2009), and a more complete B73

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

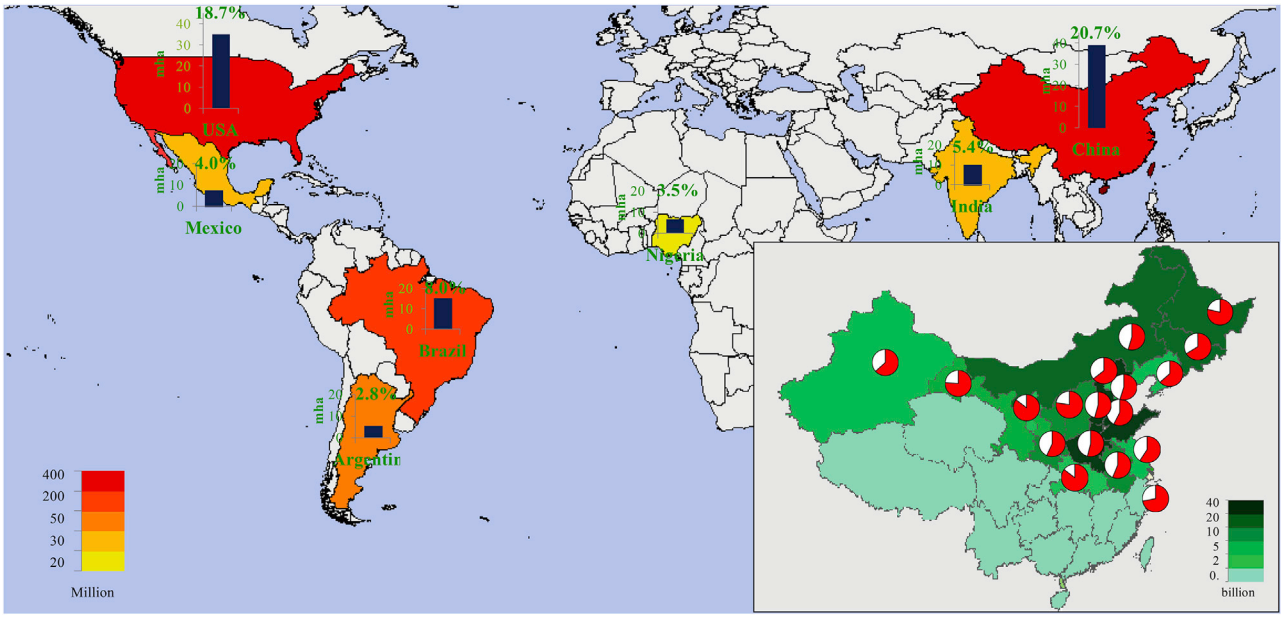

Figure 1. Geographic Distribution of Maize Planting Area and Yield in Major Maize-Growing Countries.
Pie charts (red) show the proportions of planting areas of HZS-related cultivars within each province of China.

reference genome was released two years ago (Jiao et al., 2017). The comparison between the maize inbred lines B73 and Mo17 revealed that only 50% of the sequences presented in both genotypes (Brunner et al., 2005; Morgante et al., 2007). Using the reduced representation sequencing of 14 129 maize inbred lines, another maize pan-genome study selected 4.4 million accurately mapped tags as sequence anchors, 1.1 million of which were presence/absence variations (Lu et al., 2015). Therefore, unlike other several plants (Gan et al., 2011; Li et al., 2014; Zhao et al., 2018), there are very limited genomic resources for maize, including PH207 (Hirsch et al., 2016), W22 (Springer et al., 2018), and Mo17 (Sun et al., 2018). Our study adds a new sequenced draft genome to the growing list of sequenced maize genomes.

HuangZaoSi (HZS) is considered to originate from a landrace in China and exhibits large genetic differences from other maize germplasms (Zhang et al., 2018). The HZS is one of the most important core maize germplasms in China, and hybrids bred from HZS-improved lines (HILs) are planted in more than 60% of maize fields in China, which is one of the largest maize-growing countries in the world (Figure 1). The high breeding value of HZS makes it an attractive material for genetic research (Li et al., 2015a). Therefore, deciphering the genome of HZS will facilitate our understanding of its genomic diversity and breeding improvement history.

RESULTS AND DISCUSSION

Assembly of the HZS Genome

We *de novo* assembled 10 pseudo-chromosomes of HZS using next-generation sequencing reads (Supplemental Table 1) and maize pan-genetic markers (Lu et al., 2015). The cumulative size of the final assembly was 2.2 Gb, with the contig and scaffold N50 size of 78.2 kb and 223.9 Mb, respectively (Supplemental Table 2). The smallest and largest of HZS pseudo-chromosomes were 149.8 Mb and 302.4 Mb, similar to the B73 genome (Supplemental Table 3). More than 98% of HZS RNA-sequencing (RNA-seq) reads were mapped to the assembly (Supplemental Table 4), and approximately 90% of HZS transcripts assembled from 12 tissues. RNA-seq datasets were aligned within a single scaffold with the cutoff of at least 90% length coverage and 95% identity (Supplemental Table 5). Benchmarking universal single-copy orthologs (BUSCO) analysis indicated that the completeness of the HZS assembly was comparable with that of the high-quality B73 reference genome (Supplemental Tables 6 and 7). The HZS genome encoded 40 893 protein-coding genes, with comparative genetic coverage with the proteomes of PacBio assembled maize genomes assessed by BUSCO analysis (Supplemental Table 7). These findings collectively indicated the high-quality nature of the HZS genome assembly and annotation.

Comparative Genomic Analysis between HZS and Other Maize Lines

Pairwise alignments of individual pseudo-chromosomes found that 65.73%, 65.06%, 65.16%, 55.39%, and 35.84% of the HZS assembly could be one-to-one aligned to the assemblies of B73, Mo17, W22, PH207, and teosinte (*Zea mexicana*), indicating that the alignment of HZS and B73 had longer cumulative homologous segments (Supplemental Table 8). For the comparison of HZS and B73, 1379.3 Mb (65.73%) of the HZS genome matched in one-to-one syntenic blocks with 1376.4 Mb (65.59%) of the B73 genome, and there were 15 636 251 SNPs and 1 839 995 indels within those syntenic blocks.

To confidently detect the HZS-specific genes, we retrieved HZS-specific genes in HZS-specific genomic regions. We found the number of HZS-specific genes varied across the four maize cultivar genomes (B73, 134 genes; Mo17, 266 genes; PH207,

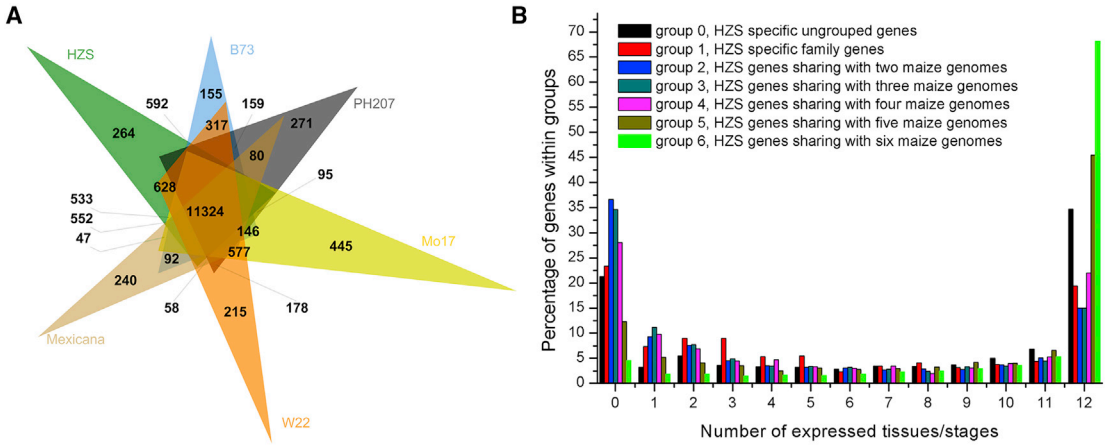

Figure 2. Features of HZS Gene Families.
(A) Comparison of the gene families among sequenced maize genomes.
(B) Distribution of the expression breadth of HZS genes among different categories of genes.

254 genes; and W22, 227 genes) (Supplemental Table 9). Only six HZS-specific genes were not detected in any other maize cultivars (B73, Mo17, PH207, or W22), and these genes exhibited tissue or development specific transcriptional patterns (Supplemental Table 10). Interestingly, at least one of the 6 HZS-specific genes presented in 89.5% of the maize wild relatives but only in 43.3% of the modern maize cultivars (Supplemental Figure 1). This finding suggests that some genes may have been lost during domestication, but others may have been inherited and maintained from ancient maize genomes, which likely facilitated adaptation of maize to diverse environments.

We then detected the conserved protein-coding genes without amino acid substitutions across the maize cultivars. The number of conserved genes between HZS and the other four maize genomes were 13 533 (B73), 12 573 (Mo17), 12 359 (PH207), and 12 736 (W22). In total, 4283 HZS genes were conserved compared with four other maize cultivars genomes. We found that these conserved genes shared by all five maize cultivar genomes were functionally enriched in the basic biological features, such as guanyl nucleotide binding function (GO:0019001, $p < 0.01$) and structural constituent of ribosome (GO:0003735, $p < 0.05$).

Analysis of the Evolution of HZS-Specific Genes and Their Expression Variations

We focused on the genetic features of the HZS genome by clustering the HZS proteome and those of other five genomes (B73, PH207, Mo17, W22, and *mexicana*). Of the 40 893 HZS proteins, 37 756 proteins were grouped into 20 767 families, and the rest of the 3137 genes were HZS-specific ungrouped genes (SUGs, group 0). A total of 24 264 HZS proteins were clustered into 11 324 core families shared by the six genomes (HZS core family genes [CFGs], group 6), whereas 638 HZS proteins were grouped into 264 HZS-specific gene families (specific family genes [SFGs], group 1) (Figure 2A and Supplemental Table 11). The number of HZS genes in groups shared with two (group 2), three (group 3), four (group 4), and five (group 5) genomes were 4666, 2248, 1958, and 3982, respectively.

Expression pattern analysis was carried out for HZS genes within different groups (Figure 2B). Overall, HZS CFGs tended to have broader expression breadth, accounting for ~68% of the genes universally expressed in all sampled tissues, whereas only ~5% of the genes were not expressed in all tissues or only expressed in one tissue. Compared with CFGs, SFGs showed a higher percentage of tissue-/stage-specific expression (Figure 2B), lower expression levels, and higher variable coefficients of expression (Supplemental Figure 2). This finding suggests that SFGs with specific expression likely confer HZS with unique characteristics from the other maize lines, and these genes may be the genetic basis of the unique adaptability and heterosis of HZS.

We further explored the origin type of HZS-specific family genes and HZS-expanded family genes, both of which may confer enhanced functions or novel functions for HZS. We found that a higher percentage of members in HZS-specific gene families come from a proximal/tandem duplication event (21.94%, 140/638) compared with the genome average (9.25%, 3785/40 893, $p < 0.01$, chi-squared test). The percentages of proximal/tandem duplicated genes within HZS groups sharing two, three, four, five, and six genomes were 8.42%, 9.25%, 11.79%, 11.03%, and 9.45%, respectively. Those 140 proximal/tandem duplicated genes were distributed in 69 HZS-specific gene families, 52 of which were solely made up of proximal/tandem duplicated genes. There was apparent difference in contributions between the HZS-specific gene families (8.46%) and the genome average (20.19%) for the whole-genome/segmental duplication event ($p < 0.01$, chi-squared test), which represents another important evolutionary force promoting gene family formation. Further exploration of the evolutionary origin of HZS-expanded gene families indicated that tandem duplication also played an important role. For instance, a gene family containing a Prp18 domain (Pfam PF02840.14) had 12 copies present in HZS, and only 7, 5, 7, 8, and 3 copies in B73, PH207, Mo17, W22, and *mexicana*, respectively. Among these HZS genes, 8 copies were located adjacently, indicating their origin was likely from a tandem duplication event, and furthermore, most copies were universally expressed (Supplemental Table 12). Collectively, these findings

suggest that tandem duplication is an important mechanism that has shaped the HZS-specific proteome among maize genomes.

We also detected the influence of the origin types of duplicated genes on the organ-enhanced expressed genes, which may play important roles in tissue/organ formation and function. We identified 8224 genes with organ-enhanced expression (defined as genes with an expression level in a single organ at least twice the level in any other) (Young et al., 2011) from 10 different organs/tissues. Gene function enrichment analysis revealed that these organ-enhanced expressed genes were involved in oxidation reduction processes (GO:0055114, $p < 0.001$) and response to chemical stimulus (GO:0042221, $p < 0.001$). Organs show distinct expression patterns for those genes, where leaf samples (eighth leaf) had the largest number of genes (1407 genes), followed by primary roots (1157 genes) and immature cob (1079 genes). Of these three organs, the primary roots had the highest proportion of proximal/tandem duplicated genes (22.38% versus the genome average of 11.73%, $p < 0.01$, chi-squared test), followed by immature cob (6.48%) and leaf (6.32%). In summary, it seems that proximal/tandem duplication events play a key role in the evolution of tissue specificity.

Pedigree Analysis of the Formation History of HZS and HILs

The high-quality HZS genome assembly allowed feasible exploration of the improvement history of HILs. We resequenced 12 Sipingtou landraces and 28 HILs with an average depth of ~15X (Supplemental Table 13), and a total of 56.4 M raw SNPs/Indels were identified. To assess the domestication relationships between the landraces and HILs, we performed both PCA (Supplemental Figure 3) and phylogenetic (Supplemental Figure 4) analyses. Results showed that the genetic relationship of landrace Tangsipingtou (TSPT) to HZS was closer than that of other Sipingtou landraces, which was also supported by the population structure (Supplemental Figure 5). These data suggested that HZS was selected from an off-type line derived from TSPT lines, which is consistent with the previous study (Zeng et al., 1997).

To determine which of the Sipingtou lines were donors, we analyzed identity-by-descent (IBD) segments transferred from the 12 Sipingtou landraces to HZS. It was shown that a number of long, continuous IBD segments were identified in both TSPT and HSPT (Figure 3A). Moreover, these segments complemented each other, and many of the recombination sites matched completely, strongly suggesting that TSPT and HSPT were the main genetic origin of HZS.

To detect the IBD segments from HZS to HILs, we found some important genomic regions in which almost all IBD segments were retained in HILs (Figure 3B). We designated these regions as IBD conserved regions. In total, we detected 862 IBD conserved regions in HZS, accounting for 29.07% of the total genome. It was inferred that these IBD conserved regions were related to the characteristics of HILs.

Selection Signals during Maize Improvement

We scanned genomic regions using a likelihood method (XP-CLR) to identify potential selective signals. Using of the top 5% of XP-CLR values as the threshold, we identified a total of 719 selective sweeps (Figure 3C), among which 437 (60.53%) regions overlapped with the IBD conserved region. These regions were probably under artificial selection during the breeding improvement of HILs and were also likely to be related to HIL characteristics, thus we defined these selected regions as HZS characteristic selected regions (CSRs).

Expectedly, we found hundreds of genes in CSRs that were likely maintained during the improvement process of HILs, and some of their functions have previously been reported in maize (Supplemental Table 14). To further annotate CSRs, we performed GWAS for agronomic and yield-related traits using 28 elite HILs and 173 other HIL germplasms. Many yield-related QTLs were enriched in CSRs (Figure 3C), including kernel weight per cob (KWPC1.1, KWPC1.2, and KWPC8.1) and cob diameter (CD2.1, CD2.2, and CD4.1), etc. Furthermore, we focused on one candidate gene in the KWPC QTL region (chromosome 8: 22.82–23.11 Mb), which was in CSRs (Supplemental Figure 6). This gene *ZmGn1A* (Chr8_23292.7.3583), was the ortholog of rice *OsCKX2* (cytokinin oxidase/dehydrogenase), which has been shown to control cytokinin accumulation and increase the number of reproductive organs (Ashikari et al., 2005). We further analyzed the haplotypes and found that lines carrying haplotype B exhibited significantly greater kernel weight per cob and kernel number than those carrying haplotype A (Supplemental Figure 7). Notably, some elite HILs, such as Jing92 (one parent of the hybrid Jingke968), carried haplotype B. Our findings suggest that yield-related genes are clustered in CSRs and may play important roles in breeding improvement. This result is consistent with a previous study that genes controlling important traits may cluster and be retained in chromosome blocks (Huang et al., 2018).

The Distribution Features of SFGs in the HZS Genome

SFGs were not evenly distributed throughout the HZS genome and clustered with improvement/function-related genes. As mentioned, (1) the characteristics of HILs have been retained by IBD conserved regions; (2) CSRs define the selective signals in IBD conserved regions; and (3) QTL regions in CSRs indicate the potential function-related loci in CSRs. Non-core family genes, which included SFGs, accounted for 32.99%, 34.84%, 36.01%, and 39.42% of all genes in whole-genome regions, in IBD conserved regions, in CSRs, and in QTLs in CSRs, respectively (Figure 4A). However, core family genes showed the opposite trend. Furthermore, SFGs was distributed at a higher proportion in yield-related QTLs, such as kernel weight per cob (KWPC) and cob weight (CW), but low in other agronomic QTLs and even lower than the average level of the whole genome (Figure 4B). These results suggest that SFGs may be related to the conditioning of the yield-related traits in breeding improvement.

SFGs presented at a low percentage in the genome, which was similar to rare alleles. As suggested in a previous study, rare alleles often existed in high proportions in QTL regions (Tennessen et al., 2012; Nelson et al., 2012; The UK10K Consortium, 2015; Keinan and Clark, 2012), and rare alleles were significantly accumulated during breeding progress (Jiao et al., 2012). In this study, more efforts were made on SFGs

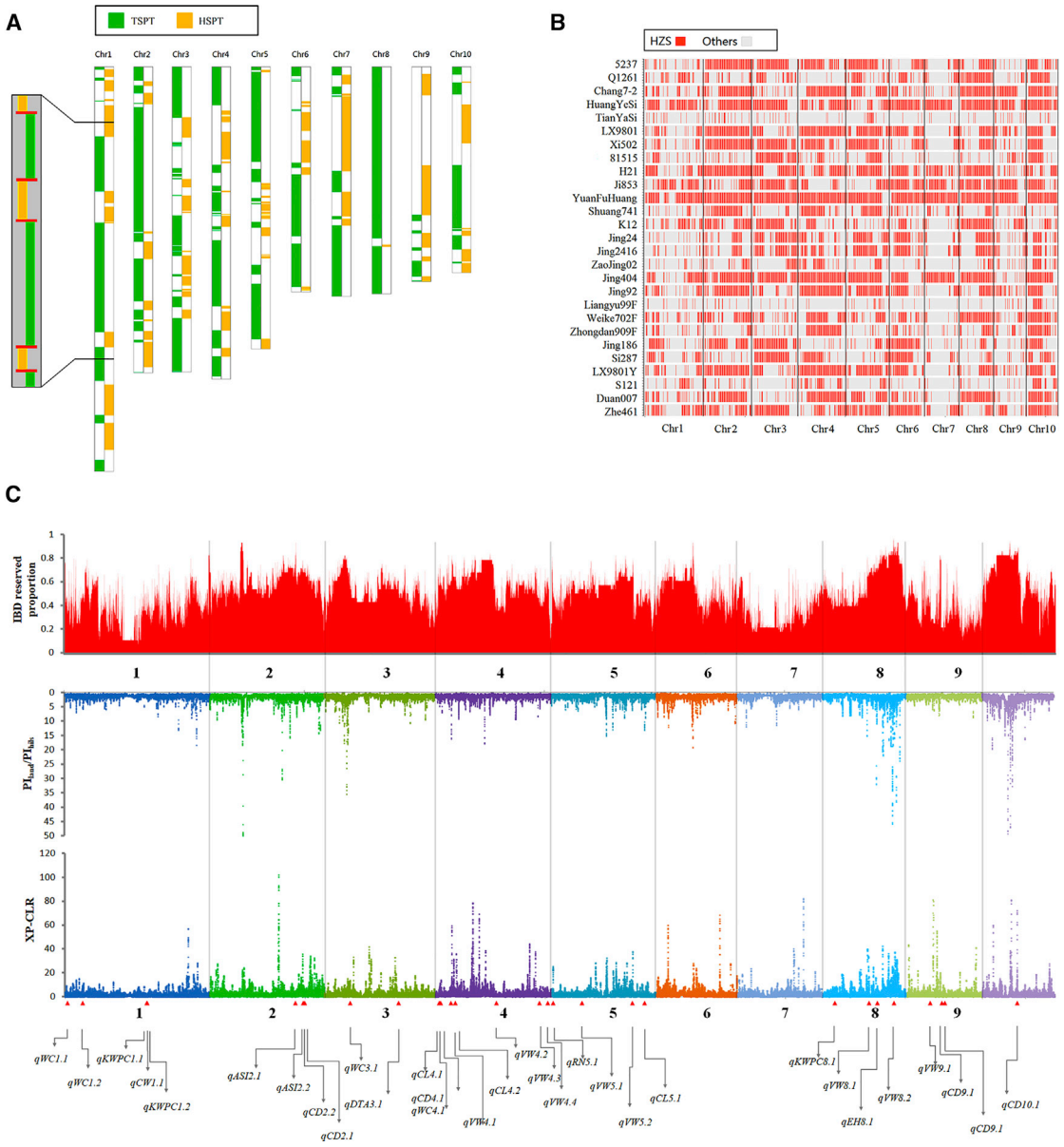

Figure 3. Distribution of Inferred IBD Segments and Genome-wide Screening of Selective Sweeps during Improvement.
(A) The IBD segments transferred from TSPT and HSPT landrace to HZS.
(B) The IBD segments transferred from HZS to HILs.
(C) Genome-wide screening and functional annotations of selected regions.

rather than allele frequencies, and we identified high proportions of SFGs in important QTL regions, similar to the high proportions observed for rare alleles. SFGs appeared in higher proportions in QTLs for yield-related traits than those for other agronomic traits, and their low expression and tissue-/stage-specific expression suggested that SFGs are probably involved in the complex functions of regulatory networks (Li et al., 2015b). For instance, one of the SFGs, Chr3_170084_8470 (Zm00001d042399) was found in a CSR and encoded a putative calcium-binding EF hand family protein to act on many metabolic pathways, and was involved in PAMP-triggered immunity (Fu et al., 2013), pollen germination, tube growth (Wang et al., 2009), and other processes. Previous research has shown that candidate genes for improvement specifically express in different tissues in a cultivar-dependent manner, for example, mature leaf tissue in B73 (Walley et al., 2016) but tassel and anther in HZS. Our finding, combined with the roles of tandem duplicated genes, which is an important mechanism that has shaped the SFGs demonstrated in case studies, such as nematode resistance in soybean (Cook et al., 2012), grain size, and japonica-indica hybrid male sterility in rice (Wang et al., 2015; Shen et al., 2017), and aluminum tolerance in maize (Maron et al., 2013), strongly suggesting that we should expand our research to focus on SFGs and tandem duplicated genes in maize and other plants.

In summary, we presented a draft genome of HZS and comprehensively compared the differences in maize genomes. We also

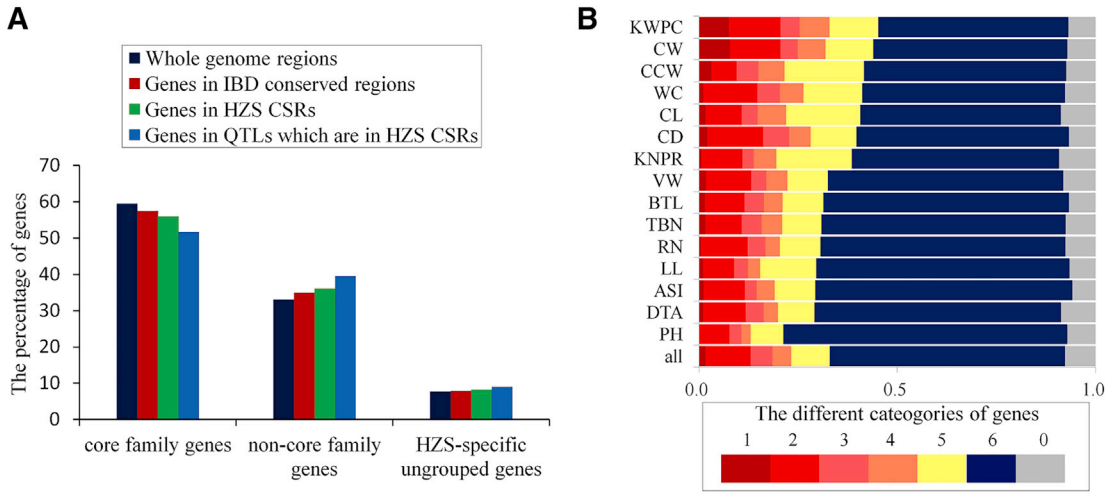

Figure 4. The Distribution Features of SFGs in the HZS Genome.

(A) The percentage of core family genes, non-core family genes, and HZS-specific ungrouped genes in different genome regions.

(B) The percentage of different categories of genes in QTLs associated with agronomic and yield-related traits. KWPC: kernel weight per cob, CW: cob weight, CCW: corncob weight, WC: water content, CL: cob length, CD: cob diameter, KNPR: kernel number per row, VW: volume weight, BTL: bare top length, TBN: tassel branch number, RN: row number, LL: leaf length, ASI: anthesis-silk interval, DTA: days to anthesis, PH: plant height.

traced the key genome blocks with HIL formation during the improvement history. Our findings expanded our knowledge on the breadth of genomic variation and improvement history of maize. This study provides an important and novel resource for maize genome research and sheds light on the underlying mechanisms of maize breeding improvement.

METHODS

Sampling and Sequencing

All genomic DNA was extracted from the young leaves of a single plant, and genomic libraries were constructed, including two pair-end (insert size, 450 bp, 750 bp) and three mate-pair (insert size, 4 kb, 7 kb, and 10 kb) libraries. Approximately 450 Gb (~200 × genomic coverage) of NGS reads were obtained using the HiSeq2500/HiSeq3000. A total of 40 HZS-related lines were selected for sequencing, including 28 HILs and 12 landrace lines.

Genome Assembly

The NGS data were initially assembled using DenovoMAGIC 2 (NRGene, Nes Ziona, Israel). In brief, the assembly process comprised the following two steps.

Reads Pre-processing and Error Correction

PCR duplicates, Illumina adaptors, and Nextera linkers (in mate-paired libraries) were trimmed from sequencing reads. For the 450 bp insert size paired-end library with 2 × 250 bp read length, the reads were merged with the minimal required overlapping of 10 bp at the end to create the merged fragment reads. After pre-processing, all reads that contained putative sequencing error (with low k-mer frequency) were filtered out.

De Novo *Assembly*

First, the assembly consisted of constructing a De Bruijn graph of k-kmer to generate contigs from the merged fragment reads. The read pairs were used to find reliable paths in the graph between contigs for repeat resolving and contigs extension. Later, contigs were linked into scaffolds with PE and MP information with SSPACE (Boetzer et al., 2011). The size of the gaps between contigs was estimated according to the distance of PE and MP links. At the final gap-filling step, PE and MP links and De Bruijn graph information were utilized to extend unique sequences at the edges to fill the gaps.

Protein Annotation

Protein annotations were carried out by BLASTP (Altschul et al., 1990) against the NCBI non-redundant protein database/SwissProt database and searching the InterPro database (Mulder and Apweiler, 2007). GO information for each protein was retrieved from InterPro annotation. KEGG annotation was carried out by BlastKOALA (http://www.kegg.jp/blastkoala/) (Ogata et al., 1999).

Comparative Genomic Analysis

The completeness of protein-coding regions of HZS and other six published maize genomes were evaluated using BUSCO analysis (v3.0.2, embryophyta_odb9) (Simao et al., 2015).

Pairwise genome alignments between HZS and the other five maize genotypes were initially carried out using the Nucmer tool (settings: –mum -c 90 -l 40) implemented in MUMmer (v3.23), followed by the delta_filter program (settings: -r and -q) in conjunction with one-to-one mapping between the references and queries. The show-snp tool was then employed to identify SNPs and Indels in the one-to-one alignment block (parameter –Clr TH). The BWA MEM module was used to detect the PAV regions as the reference (Mo17), and the HZS CDSs overlapped with PAV regions larger than 75% were considered as HZS-specific genes. We aligned HZS CDSs to those of the other five maize cultivar genomes to identify the conserved CDSs without amino acid substitutions using Gmap (gmap-2018-07-04).

For gene family construction, only one transcript-encoded protein was randomly selected to represent loci with multiple transcripts. All-versus-all BLASTP search results (m8 format with E value threshold of 1e−5) were used for gene family construction by Orthomcl (v2.0.9). A Venn diagram is displayed in the website (http://jvenn.toulouse.inra.fr/app/index.html). The classification of HZS genes was carried out by BLASTP all-against-all comparisons of all proteins (m8 format, e value ≤1e−10) using the duplicate_gene_classifier module integrated within MCScanX (m 50, s50).

RNA-Seq and Data Processing

HZS was planted in an open field in Beijing, China. Various tissues were harvested and immediately placed in liquid nitrogen and stored at −80°C. Three biological replicates were performed for each sample.

Total RNA was extracted using Trizol, and the mRNA-seq library was constructed using Kapa transcriptome kits and then sequenced with Hiseq3000. We filtered out adapter sequences and low-quality bases from RNA-seq reads and mapped them to the HZ assembly according to gene annotation models by Hisat2 (v2.0.4). The expression level for each gene was determined by Stringtie (v 1.2.3), and the average level was calculated from three biological replicates.

ACCESSION NUMBERS

The raw sequence data, genome assembly, and annotation have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under project PRJCA001247 and are publicly accessible at http://bigd.big.ac.cn/gsa.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the Science and Technology Planning Project of Beijing (D161100005716002), the Beijing Scholars Program (BSP041), the Innovative Team Construction Project of BAAFS (JNKYT201603). Y.L. acknowledges the Youth Innovation Promotion Association of Chinese Academy of Sciences (no. 2017140).

AUTHOR CONTRIBUTIONS

C.L., W.S., S.H., J.Z., R.W., F.W., J.Y., and Y.W. designed and managed the project. Shuaishuai Wang, Y. Zhao, J.G., X.T., Y.Y., X.B., and Y. Zhang prepared samples and performed biological experiments for DNA sequencing. R.Z., J.W., X.W., A.S., and M.L. collected and grew the plant material. C.L., Y.L., S.G., X.L., and H.H. performed data analyses. C.L., W.S., Y.L., Shuai Wang, S.G., and Z.S. wrote the manuscript.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: January 31, 2019
Revised: January 31, 2019
Accepted: February 17, 2019
Published: February 22, 2019

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., Angeles, E.R., Qian, Q., Kitano, H., and Matsuoka, M. (2005). Cytokinin oxidase regulates rice grain production. Science **309**:741–745.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**:578–579.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., and Rafalski, A. (2005). Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell **17**:343–360.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux[J]. Nat. Genet. **44**:803–807.

Cook, D.E., Lee, T.G., Guo, X.L., Melito, S., Wang, K., Bayless, A., Wang, J.P., Hughes, T.J., Willis, D.K., Clemente, T., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science **338**:1206–1209.

Fu, L., Yu, X., and An, C. (2013). Overexpression of constitutively active OsCPK10 increases *Arabidopsis* resistance against *Pseudomonas syringae* pv. tomato and rice resistance against *Magnaporthe grisea*. Plant Physiol. Biochem. **73**:202–210.

Gan, X.C., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature **477**:419–423.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell **28**:2700–2714.

Huang, J., Li, J., Zhou, J., Wang, L., Yang, S.H., Hurst, L.D., Li, W.H., and Tian, D.C. (2018). Identifying a large number of high-yield genes in rice by pedigree analysis, whole-genome sequencing, and CRISPR-Cas9 gene knockout. Proc. Natl. Acad. Sci. U S A **115**:7559–7567.

Hubert, B., Rosegrant, M., Boekel, M.A., and Ortiz, R. (2010). The future of food: scenarios for 2050. Crop Sci. **50**:33–50.

Jiao, Y.P., Pelusok, P., Shi, J.H., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X.H., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**:524–527.

Jiao, Y.P., Zhao, H.N., Ren, L.H., Song, W.B., Zeng, B., Guo, J.J., Wang, B.B., Liu, Z.P., Chen, J., Li, W., et al. (2012). Genome-wide genetic changes during modern breeding of maize. Nat. Genet. **44**:812–815.

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science **336**:740–743.

Lai, J.S., Li, R.Q., Xu, X., Jin, W.W., Xu, M.L., Zhao, H.N., Xiang, Z.K., Song, W.B., Ying, K., Zhang, M., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. **42**:1027–1030.

Li, C.H., Li, Y.X., Bradbury, P.J., Wu, X., Shi, Y.S., Song, Y.C., Zhang, D.F., Rodgers-Melnick, E., Buckler, E.S., Zhang, Z.W., et al. (2015a). Construction of high-quality recombination maps with low-coverage genomic sequencing for joint linkage analysis in maize. BMC Biol. **13**:78.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2015b). The impact of rare variation on gene expression across tissues. Nature **550**:239–243.

Li, Y.H., Zhou, G.Y., Ma, J.X., Jiang, W.K., Jin, L.G., Zhang, Z.H., Guo, Y., Zhang, J.B., Sui, Y., Zheng, L.T., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. **32**:1045–1052.

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T.Y., Li, Y., Li, Y.X., Semagn, K., Zhang, X.C., et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. Nat. Commun. **6**:6914.

Maron, L.G., Guimarães, C.T., Kirst, M., Albert, P.S., Birchler, J.A., Bradbury, P.J., Buckler, E.S., Coluccio, A.E., Danilova, T.V., Kudrna, D., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. U S A **110**:5241–5246.

Morgante, M., De Paoli, E., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. **10**:149–155.

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol. Biol. **396**:59–70.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science **337**:100–104.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **27**:29–34.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F.S., Pasternak, S., Liang, C.Z., Zhang, J.W., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **362**:1112–1115.

Shen, R., Wang, L., Liu, X., Wu, J., Jin, W., Zhao, X., Xie, X., Zhu, Q., Tang, H., Li, Q., et al. (2017). Genomic structural variation-mediated allelic suppression causes hybrid male sterility in rice. Nat. Commun. **8**:1310.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Springer, N.M., Anderson, S.N., Andorf, C.M., Ahern, K.R., Bai, F., Barad, O., Barbazuk, W.B., Bass, H.W., Baruch, K., Ben-Zvi, G., et al. (2018). The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. **50**:1282–1288.

Sun, S.L., Zhou, Y.S., Chen, J., Shi, J.P., Zhao, H.M., Zhao, H.M., Song, W.B., Zhang, M., Cui, Y., Dong, X.M., et al. (2018). Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. **50**:1289–1295.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F., and Gaut, B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. U S A **98**:9161–9166.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W.Q., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X.M., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science **337**:64–69.

The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. Nature **526**:82–90.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**:814–818.

Wang, G.F., Ji, J., El-Kasmi, F., Dangl, J.L., Johal, G., and Balint-Kurti, P.J. (2015). Molecular and functional analyses of a maize autoactive NB-LRR protein identify precise structural requirements for activity. PLoS Pathog. **11**:e1004674.

Wang, Y., Zhang, W.Z., Song, L.F., Zou, J.J., Su, Z., and Wu, W.H. (2009). Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in *Arabidopsis*. Plant Physiol. **148**:1201–1211.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., and Guat, B.S. (2005). The effects of artificial selection on the maize genome. Science **308**:1310–1314.

Yang, N., Xu, X.W., Wang, R.R., Peng, W.L., Cai, L.C., Song, J.M., Li, W.Q., Luo, X., Luyao Niu, L.Y., Wang, Y.B., et al. (2018). Contributions of *Zea mays* subspecies Mexicana haplotypes to modern maize. Nat. Commun. **8**:1874.

Young, N.D., Debelle, F., Oldroyd, G., Geurts, R., Cannon, S.B., Mayer, K.F., Gouzy, J., Van De Peer, Y., Schoof, H., Udvardi, M.K., et al. (2011). The medicago genome provides insight into evolution of rhizobial symbiosis. Nature **480**:520–524.

Zeng, S.S., Ren, R., and Liu, X.Z. (1997). The important position of Huang Zao Si in maize breeding and production in China. Maize Sci. **4**:1–6.

Zhang, R.Y., Xu, G., Li, J.S., Yan, J.B., Li, H.H., and Yang, X.H. (2018). Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. Theor. Appl. Genet. **131**:1–15.

Zhao, Q., Feng, Q., Lu, H.Y., Li, Y., Wang, A.H., Tian, Q.L., Zhan, Q.L., Lu, Y.Q., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. **50**:278–284.