Running title: Genetic architecture of maize kernel traits

Correspondence to: Jianbing Yan (email: yjianbing@mail.hzau.edu.cn) and Qing Li (email: qingli@mail.hzau.edu.cn)

Tel: +86-27-87280110
Fax: +86-27-87384670

National Key Laboratory of Crop Genetic Improvement
Huazhong Agricultural University
No.1 Shizishan Street, Hongshan District
Wuhan 430070, China.

Article title: The conserved and unique genetic architecture of kernel size and weight in maize and rice

Authors: Jie Liu¹, Juan Huang¹, Huan Guo¹, Liu Lan¹, Hongze Wang¹, Yuancheng Xu¹, Xiaohong Yang², Wenqiang Li¹, Hao Tong¹, Yingjie Xiao¹, Qingchun Pan¹, Feng Qiao¹, Mohammad Sharif Raihan¹, Haijun Liu¹, Xuehai Zhang¹, Ning Yang¹, Xiaqing Wang¹, Min Deng¹, Minliang Jin¹, Lijun Zhao¹, Xin Luo¹, Yang Zhou¹, Xiang Li¹, Wei Zhan¹, Nannan Liu¹, Hong Wang¹, Gengshen Chen¹, Qing Li¹* and Jianbing Yan¹*

¹ National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China.
² National Maize Improvement Center of China, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China.

performed the experiments. J.L., Q.L. and J.Y. prepared the manuscript and all authors read
and approved the manuscript.

One sentence summary: Ten segregating populations yield both conserved and
species-specific genetic architecture of kernel size and weight in maize and rice

Funding information: This work is supported by grants from National Science Foundation
of China (91635303), the National Key Research and Development Program of China
(2016YFD0101003) and Huazhong Agricultural University Scientific & Technological
Self-innovation Foundation.
ABSTRACT

Maize is a major staple crop. Maize kernel size and weight are important contributors to its yield. Here, we measured kernel length, kernel width, kernel thickness, hundred kernel weight and kernel test weight in 10 recombinant inbred line populations and dissected their genetic architecture using three statistical models. In total, 729 quantitative trait loci (QTLs) were identified, many of which were identified in all three models, including 22 major QTLs that each can explain more than 10% of phenotypic variation. To provide candidate genes for these QTLs, we identified 30 maize genes that are orthologs of 18 rice genes reported to affect rice seed size or weight. Interestingly, 24 of these 30 genes are located in the identified QTLs or within 1 Mb region of the significant single nucleotide polymorphisms (SNPs). We further confirmed the effect of five genes on maize kernel size/weight in an independent association mapping panel with 540 lines by candidate gene association analysis. Lastly, the function of ZmINCW1, a homolog of rice GRAIN INCOMPLETE FILLING 1 (GIF1) that affects seed size and weight, was characterized in detail. ZmINCW1 is close to QTL peaks for kernel size/weight (< 1 Mb) and contains significant SNPs affecting kernel size/weight in the association panel. Over-expression of this gene can rescue the reduced weight of the Arabidopsis homozygous mutant line in the AtcwINV2 gene (Arabidopsis ortholog of ZmINCW1). These results indicate that the molecular mechanisms affecting kernel/seed development are conserved in maize, rice, and possibly in Arabidopsis.
INTRODUCTION

Maize (Zea mays L.) is one of the most important crops and is cultivated world-wide as a source of staple food, animal feed and industrial materials. According to the Food and Agriculture Organization (FAO), the production of maize was 1,016.7 million tonnes in 2013 which was far more than rice and wheat (745.7 and 713.1 million tonnes, respectively) (FAO Statistical Pocketbook 2015). Yield improvement is a central goal of maize breeding. Kernel size and weight are two significant components of maize yield and many attempts have been made to elucidate the genetic basis of kernel size and weight.

Many studies have mapped QTLs for natural variations in kernel size and weight. For example, Liu et al. (2014) identified 55 QTLs for kernel size and weight in an F2 population. Raihan et al. (2016) mapped sixteen major QTLs for kernel traits in a recombinant inbred line (RIL) population. Jiang et al. (2015) mapped 28 QTLs in a testcross population. Most of these studies used two diverse inbred lines to develop the segregating population and used a limited number of genetic markers to construct the linkage map, which greatly limited the resolution and power to detect rare and/or small effect QTLs. Large-scale QTL mapping studies including more diverse genetic backgrounds and dense genetic markers would provide more insight into the number and effect of QTLs controlling the natural variations of kernel size and weight in maize.

Despite the large number of QTLs that have been identified for maize kernel size and weight, none of them has been delimited to the causative variation. On the other hand, many genes controlling maize kernel development have been cloned using kernel mutants identified from Robertson’s Mutator stocks, including emp2 (empty pericarp2), emp4 (empty pericarp4), emp5 (empty pericarp5), emp16 (empty pericarp16), dek1 (defective kernel 1), dek35 (defective kernel 35), MPPR6 (maize pentatricopeptide repeat 6), smk1 (small kernel 1), emb14 (embryo defective 14), ubl1 (U6 biogenesis-like 1) and many others (Fu et al., 2002; Lid et al., 2002; Gutiérrez-Marcos et al., 2007; Manavski et al., 2012; Liu et al., 2013; Li et al., 2014; Li et al., 2015; Chen et al., 2016; Li et al., 2016; Xiu et al., 2016). Mutations in these genes usually have severe phenotypes in kernels, such as empty pericarp where both
embryo and endosperm cannot develop properly. It’s unclear whether weak mutations (or genetic variations) of these genes exist in nature and whether such genetic variants can contribute to phenotypic diversity in maize kernel. A comparative analysis between these mutant genes and kernel size/weight QTLs will not only provide candidate genes for QTLs but also shed light on the extent to which genes identified through mutant studies can contribute to natural variations in maize kernel phenotypes.

Maize shares common ancestors with rice (Murat et al., 2017). Comparative QTL studies between species showed that similar traits were usually controlled by QTLs that are located within syntenic regions among the species (Paterson et al., 1995). This idea has been further illustrated by comparative functional studies at the single gene level. Many genes that could affect the seed shape and weight have been fine-mapped and cloned in rice, such as GS3 (Fan et al., 2006; Mao et al., 2010), GW2 (Song et al., 2007) and GS5 (Li et al., 2011). Li et al. (2010a, 2010b) isolated the maize orthologs of rice GS3 and GW2, and showed that the maize genes also control similar traits, though with different genetic variations. Similarly, Liu et al. (2015) showed that GS5 contributes to kernel size variation in maize as well as in rice. Notably, the wheat orthologs of rice GW2 and GS5 were also significantly associated with wheat kernel size and weight (Su et al., 2011; Hong et al., 2014; Qin et al., 2014; Jaiswal et al., 2015; Wang et al., 2015; Ma et al., 2016; Simmonds et al., 2016; Wang et al., 2016). In addition to GS3, GW2 and GS5, many other genes controlling rice kernel size/weight have been cloned, for example, genes involved in G-protein signaling (DEP1, D1) and genes from phytohormone pathways (DST and Gln1a for cytokinin; D11, SRS5, D61, qGL3, SMG1 for brassinosteroid; TGW6 for auxin). It remains an open question whether their maize homologous counterparts have a similar function in the phenotypic diversity of kernel size/weight. A systematic investigation of the function of these genes in maize would provide more insights into the genetic mechanisms controlling kernel development in the two closely related important crops.

In this study, we used 10 RIL populations to dissect the genetic basis of maize kernel size and weight with three models - separate linkage mapping (SLM), joint linkage mapping...
Genetic architecture of maize kernel traits

(JLM) and genome-wide association mapping (GWAS). Many QTLs with major and minor
effect were identified. A comparison between these QTLs and the genes from maize mutant
studies and maize homologs of well-known rice seed size/weight genes suggested that many
of these genes have a role in controlling natural variations of kernel size and weight. There
are also many QTLs that are not coincident with any known candidate genes or contain
candidate genes that don’t affect natural variations in kernel size and weight. These results
suggest both a conserved and species-specific genetic architecture of kernel traits between
rice and maize. Furthermore, we found that ZmINCW1, an ortholog of rice seed weight gene
GIF1, had conserved function that affects kernel/seed development in maize and Arabidopsis.
Our results help to elucidate the genetic basis of maize kernel size and weight.
RESULTS

Phenotypic variation and heritability of kernel size and weight

We used 10 RIL populations derived from 14 diverse maize inbred lines (Pan et al., 2016) to dissect the genetic architecture of kernel size and weight in maize. These lines were grown under multiple environments. Seven of them (B73 × BY804, KUI3 × B77, K22 × CI7, DAN340 × K22, ZHENG58 × SK, YU87-1 × BK and ZONG3 × YU87-1) were planted in eight environments, while the other three (DE3 × BY815, K22 × BY815 and BY815 × KUI3) were planted in four environments. Five maize kernel traits were measured for each line in these 10 RIL populations, including hundred kernel weight (HKW, weight of one hundred kernels), kernel test weight (KTW, weight of 250 ml kernels), kernel length (KL), kernel thickness (KT) and kernel width (KW) (Figure 1A). Best-linear unbiased prediction (BLUP) values of each line were used to represent the phenotypic value. Both the 14 parental lines and the 10 RIL populations showed significant variations in these five kernel traits (Figure 1B, C, and Figure S1). Broad-sense heritability ranged from 0.53 (KT in DE3 × BY815) to 0.94 (HKW in ZHENG58 × SK), with most higher than 0.80 (Table S1). This suggests that phenotypic variation is largely controlled by genetic factors and can be genetically mapped.

We calculated the correlation coefficient between HKW and the three kernel size traits, KL, KW and KT (Table S2) and found that KW and KT were significantly positively correlated with HKW in all 10 RIL populations, while KL was positively correlated with HKW in 7 RIL populations. The correlation coefficients for KL were usually smaller compared with KW and KT, suggesting that KW and KT may play more important roles for kernel weight in maize.

Dissection of the genetic architecture of maize kernel size and weight with three methods

First, we performed SLM in each RIL population with the composite interval mapping method (Zeng, 1994). In total, we identified 373 QTLs for kernel size and weight, including 90, 70, 61, 89, and 63 QTLs for HKW, KTW, KL, KW, and KT, respectively (Figure 2A, Table 1, Table S3, Figure S2, and Figure S3). The phenotypic variation explained by each
QTL ranged from 2.91% to 19.43%, with an average of 7%. Out of these 373 QTLs, 267 (72%) were identified in only one population. Some QTLs could be identified in at least two populations for the same trait, including 27 QTLs for HKW, 24 for KTW, 15 for KL, 25 for KW, and 15 for KT. The presence of population-common and specific QTLs may reflect differences in allele frequency of the underlying causative sites and suggests that populations from diverse genetic backgrounds are needed to comprehensively understand the genetic architecture of kernel size and weight. Importantly, 10, 11, 10, 17, and 9 major QTLs (R^2 > 10%, i.e. QTLs that can explain more than 10% of the phenotypic variation) were identified for HKW, KTW, KL, KW, and KT, respectively. Among these major QTLs, 17 could be detected in more than one population for the same trait. We also detected 18 major QTLs that can affect more than one trait (Tables S4 and S5). An example QTL for KW, KT, and HKW is shown in Figure 2B. Detailed information about these 373 QTLs is provided in Table S6.

We also performed JLM and GWAS by analyzing the 10 RIL populations jointly (see Methods). In JLM, we identified 56, 59, 55, 68, and 62 QTLs for HKW, KTW, KL, KW, and
In GWAS, we detected between 123 and 198 significant SNPs for each trait (Table S8). To avoid redundancy of the significant SNPs caused by linkage disequilibrium, we performed a backward regression procedure (see Methods). After this analysis, 30, 22, 26, 32, and 25 independent SNPs for HKW, KTW, KL, KW, and KT, respectively, were obtained (Table 1 and Table S9). Some SNPs can control two or more traits simultaneously. On average, each of the identified SNPs...
with GWAS could only explain a very small amount of phenotypic variation (between 1.06% and 1.35%) compared with QTLs identified with SLM, but they could jointly explain a large portion of phenotypic variation (55.85%, 59.17%, 59.91%, 75.72%, and 36.40% for HKW, KTW, KL, KW, and KT, respectively).

Notably, a considerable number of loci could be identified by more than one model (Figure 2C). For example, 61.9% of QTLs identified with SLM can also be detected using JLM or/and GWAS. Similarly, 55.0% of JLM QTLs and 69.2% of GWAS SNPs can be identified by the other two models. More importantly, 22 major QTLs can be identified in all three models. These results confirm the reliability of the identified QTLs and suggest that the three statistical models are complementary to each other. The integrated use of these models can provide more insight into the genetic architecture of phenotypic variation.

Natural variations of some maize mutant genes were significantly associated with kernel development

To identify candidate genes for the QTLs, we collected 36 maize genes that had been cloned using maize kernel mutants and were reported to be involved in maize kernel development (Table S10). Of these 36 genes, 21 were located in the QTLs identified by SLM, 7 were located in the QTLs identified by JLM, and 15 were located within a 1-Mb region of the significant SNPs identified by GWAS (Table S10). To further confirm the function of these genes, we used an independent association panel consisting of 540 lines. This panel has been genotyped with 1.25M SNPs (Liu et al., 2016). Between 1 and 209 SNPs were identified in these 36 genes and were used to identify loci that are significantly associated with variations in kernel size and weight. We found that 7 of these 36 genes affect at least one kernel trait (Table S10 and Figure S5). For example, Dek36 (GRMZM5G892151) was significantly associated with both HKW ($P = 5.13 \times 10^{-4}$) and KTW ($P = 6.30 \times 10^{-4}$, Figure S5B). Many of the most significant SNPs were either located in the untranslated region or represented synonymous substitutions. This is reasonable considering that loss-of-function alleles of most genes usually lead to defective kernels with limited/no viability, and thus, genetic variation
greatly affecting gene function would be unfavorable under natural conditions. It is also possible that the synonymous variants are in linkage disequilibrium with the causal variants which were not assayed. Nevertheless, these findings provide evidence that many of the genes identified from mutant studies contain natural genetic variations and that many of them contribute to phenotypic diversity in maize kernel size and weight.

Many rice seed size/weight genes have conserved functions in maize

To provide insight into whether rice seed size/weight genes have similar functions in maize, we investigated the function of the maize orthologs of 18 rice genes that have been shown to affect seed size or weight (Table S11). These 18 genes are involved proteasomal degradation, phytohormones (auxin, cytokinin, brassinosteroid), G-protein signaling and other processes.

Based on the MSU Rice Genome Annotation Project (Kawahara et al., 2013, http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/), we identified 30 maize orthologs of these 18 rice genes (Table S11), with 9 genes having one ortholog, 6 genes having two orthologs and 3 genes having three orthologs. Co-localization of these 30 genes and the identified QTLs highlighted several interesting findings: (1) Twenty-four of the 30 genes were located in the QTL confidence intervals or within 1-Mb flanking regions of the significant SNPs (Figure 3 and Table S11); (2) Three genes were located within 500 Kb of the peak and two genes were very close to the peak (3.4 Kb for *ZmGW7-2*, 59.9 Kb for *ZmBG2*); (3) Six genes were located in major QTLs, for example, *ZmSLG* was located in a major QTL ($R^2 = 12.40\%$) for KW in K22 × CI17 population and a major QTL ($R^2 = 12.89\%$) for HKW in BY815 × KUI3 population (Table S11); and (4) Five genes were located within QTLs identified in all three models. These results suggest that the rice orthologous genes provide good candidates for the maize QTLs.

We investigated the function of these 30 genes using the association panel consisting of 540 inbred lines. Between 9 and 174 SNPs (58.3 on average) were identified within each gene and were used for candidate gene association analysis. Out of these 30 genes, five were significantly associated with at least one kernel size or weight trait by candidate gene
association analysis (Figure 3, Figure S4, and Table S11) and all of these five genes were located in the QTLs identified in this study. ZmSLG (GRMZM2G179703) was significantly associated with KW (P = 1.06 × 10^{-3}, Figure S4A); ZmGLW7-1 (GRMZM2G113779) was significantly associated with KT (P = 9.44 × 10^{-5}) and HKW (P = 1.04 × 10^{-3}, Figure S4B); ZmGL2 (GRMZM2G034876) was significantly associated with KW (P = 2.65 × 10^{-5}) and HKW (P = 2.97 × 10^{-5}, Figure S4C); ZmGW7-2 (GRMZM2G370081) was significantly associated with KW (P = 4.57 × 10^{-4}, Figure S4D); and ZmSRS1-2 (GRMZM2G414043) was significantly associated with KL (P = 4.56 × 10^{-4}, Figure S4E).

The most significant SNPs in ZmGL2 and ZmGW7-2 are mis-sense variants that lead to amino acid changes (Gly/Ser^{211}, His/Gln^{271}), while the most significant SNPs in ZmSLG, ZmGLW7-1, and ZmSRS1-2 are located in the 5'-UTR or 3'-UTR region. GWAS for expression levels of these five genes showed that the expression levels of ZmSLG were significantly associated with SNPs located in this gene (P = 3.86 × 10^{-11}, Figure S4F). We also found a significant correlation between expression level of ZmSLG and KW (P < 0.01, r = -0.18, Figure S4G). This indicates that the cis-element near ZmSLG might regulate its expression.
expression to affect kernel development. Notably, \textit{ZmSLG} is also located within a region that has been shown to be under artificial selection during the generation of small seed and big seed populations by Hirsch et al. (2014). These results provide evidence that some maize orthologous genes might have similar and conserved functions as their rice counterparts.

\textit{ZmINCW1}, whose protein sequence has highly similarity to rice \textit{GIF1}, affects kernel development in maize

Maize gene \textit{Mn1} (GRMZM2G119689, Miniature-1, also named \textit{incw2}) is critical in maize kernel development and the \textit{mn1} seed weighs only 20% of the normal seed (Lowe and Nelson 1946). \textit{Mn1} encodes a cell-wall invertase and extremely low invertase activity is the causal basis of the mutant phenotype (Miller and Chourey, 1992). Because of the tetraploid origin of maize, \textit{Mn1} has a paralog in the maize genome, GRMZM2G095725. Both \textit{Mn1} and GRMZM2G095725 are orthologs of rice \textit{GIF1} (Figure S4), which affects rice grain filling and weight (Wang et al., 2008). Interestingly, both \textit{Mn1} and GRMZM2G095725 are located in the confidence intervals of QTLs for maize kernel traits in this study.

In addition to the two genes, we identified a third gene, GRMZM2G139300, that shares high protein sequence similarity to rice \textit{GIF1} (Identity = 71.28\%, E-value = 0) (Figure S6). GRMZM2G139300 had previously been designated as the \textit{incw1} loci on maize chromosome 5 and encodes a cell-wall invertase (Shanker et al., 1995). Hereafter, it was named \textit{ZmINCW1}. The best homolog of \textit{ZmINCW1} in maize is \textit{Mn1}, suggesting that \textit{ZmINCW1} may affect kernel size and weight in maize, which is also supported by our QTL mapping and candidate gene association analysis results. QTL mapping shows that \textit{ZmINCW1} is located in the QTLs for HKW and KW in B73 × BY804 population (Figure 4A) and the phenotypic variations explained by these QTLs are 14\% (HKW-12YN), 8\% (KW-11DHN), 3.7\% (KW-BLUP). The distances between \textit{ZmINCW1} and peaks of these 3 QTLs are less than 1 Mb (550 Kb, 417 Kb, and 786 Kb, respectively), providing a good candidate for these QTLs. We also detected QTLs covering \textit{ZmINCW1} for KL in DAN340 × K22 population (Figure S7A) and HKW in DE3 × BY815 population (Figure S7B) (phenotypic variations explained by these two QTLs...
Genetic architecture of maize kernel traits

are 8.5% and 19%, respectively). *ZmINCW1* was also significantly associated with KL (P = 7.10 × 10^{-6}), KW (P = 4.27 × 10^{-4}), KT (P = 5.71 × 10^{-4}) and HKW (P = 8.73 × 10^{-3}) (Figure 4B) by candidate gene association analysis in an independent association panel with 540 diverse inbred lines. Four SNPs, chr5.S_169457546, chr5.S_169458449, chr5.S_169458176 and chr5.S_169458817, were the most significant SNPs for KL, KW, KT, and HKW, respectively. Chr5.S_169457546 is located in the second intron of *ZmINCW1*; chr5.S_169458449 and chr5.S_169458176 are synonymous variants; and chr5.S_169458817 is located in the 3'UTR of *ZmINCW1*. It is likely that these SNPs either affect phenotypes through regulatory roles or are in linkage disequilibrium with the causal polymorphisms.
Genetic architecture of maize kernel traits

*ZmINCW1* showed abundant variations in gene expression in the association panel (Figure 4D), implying that expression changes in this gene may contribute to phenotypic diversity. To further investigate the functional mechanisms of *ZmINCW1*, we performed GWAS of the expression of *ZmINCW1* in an association mapping population consisting of 368 diverse maize inbred lines that are a subset of the 540 inbred lines. We identified four SNPs that showed significant correlations with the expression of *ZmINCW1* (P < 7.97×10⁻⁷). Notably, the most significant SNP (chr5.S_169456915) also showed significant correlation with KT (P = 2.57 × 10⁻³). Besides the significant SNPs located within *ZmINCW1*, there were some significant SNPs located ~220 Kb upstream of the gene (three SNPs in intergenic region and four SNPs in other three genes, Figure 4C). We then performed GWAS conditioning on the most significant SNPs in *ZmINCW1* and found no other significant SNPs for the expression of *ZmINCW1*. This result indicates that the expression of *ZmINCW1* is mainly regulated by nearby variations. Interestingly, expression levels of *ZmINCW1* are significantly associated with kernel traits in two of twelve environments (KL in 2011 Hainan, r = 0.17, P = 1.87 × 10⁻³; HKW in 2011 Yunnan, r = 0.16, P = 7.30 × 10⁻³, Figure 4D). The failure to detect significant associations in the other ten environments indicates an environmental effect of this gene. Together, these results support the notion that *cis*-variations around *ZmINCW1* affect its expression, which in turn can control kernel size and weight in maize.

**Over-expression of maize *ZmINCW1* in *Arabidopsis* can increase seed weight**

To further verify the function of *ZmINCW1*, we identified a T-DNA mutant of the *AtcwINV2* gene (Stock: SALK_068113C, http://www.arabidopsis.org/), which is the ortholog of *ZmINCW1* in *Arabidopsis* (Figure S5). This line has a T-DNA insertion in the fourth exon of *AtcwINV2* (Figure 5, A and B), and this insertion disturbs the glycosyl hydrolase C-terminal domain. Compared with wild-type (Col-0) controls, the homozygous mutants showed normal growth and they could produce normal seed (Figure 5, C and D). However, the thousand seed weight was significantly reduced in the mutant plant (16.85 mg vs 16.14 mg, one-way
Next, we over-expressed *ZmINCW1* in this mutant, and the expression level of *ZmINCW1* was confirmed using RT-PCR and western blot (Figure 5, F and G). We screened two positive transgenic lines (T₁) that expressed *ZmINCW1* and eight negative transgenic lines using RT-PCR and western blot (Figure 5, F and G). We screened two positive transgenic lines (T₁) that expressed *ZmINCW1* and eight negative transgenic lines using RT-PCR and western blot (Figure 5, F and G).
lines that had no detectable expression of Zm\textit{INCW1} or that didn’t have the transgene. Both positive and negative T\textsubscript{1} transgenic lines bear normal seeds (Figure 5, H and I), however, the positive T\textsubscript{1} transgenic lines had increased thousand seed weight compared with the negative T\textsubscript{1} transgenic lines by 28.35\% (21.52 mg vs 16.77 mg, one-way ANOVA, N = 6/26, P = 1.51 \times 10^{-12}, Figure 5J). We also found significant differences between wild-type and negative T\textsubscript{2} transgenic lines (N = 7/5, P = 9.91 \times 10^{-3}, Figure 5K), and between positive transgenic lines and negative transgenic lines in the T\textsubscript{2} generation (N = 6/5, P = 0.03, Figure 5K). These findings suggest that Zm\textit{INCW1} has conserved function for seed development in maize and \textit{Arabidopsis}. 
DISCUSSION

Maize has tremendous phenotypic and genotypic diversity. It has been estimated that there is a polymorphic site in every 44 bp on average, that the B73 reference genome sequence may capture only ~70% of the low-copy genome fraction represented by 27 diverse maize inbred lines (Gore et al., 2009), and that ~50% of the representative transcript assemblies identified from 503 maize inbred lines are not present in B73 (Hirsch et al., 2014). Jin et al. (2016) identified 13,382 genes with expression presence/absence (ePAV) in 368 maize inbred lines and found that 788 novel genes were associated with 487 metabolic traits and a novel gene was associated with kernel width. These pan-genome and pan-transcriptome analyses showed great diversity and the importance of PAVs or ePAVs for agronomic traits (Lai et al., 2010; Jin et al., 2016).

In this study, we mapped 373 QTLs with the SLM model for natural variation of maize kernel size and weight with 10 RIL populations. This represented 309 independent loci. The total numbers of QTLs identified by other studies, which used only one population, ranged from 12 to 55 (Liu et al., 2014; Zhang et al., 2014; Jiang et al., 2015; Chen et al., 2016; Raihan et al., 2016). Compared with the results published previously, we identified many more QTLs for kernel size and weight. There might be two main reasons for the greater number of QTLs identified in this study. First, RIL populations had more recombinant events compared with F2 populations used by other groups. Second, more diversity was introduced with more parental inbred lines (14 in this study) compared with two or four lines used by others. Neuffer and Sheridan (1980) estimated that maize kernel mutants map to 285 loci. This number is very close to the number of loci (309) identified for natural variation of kernel size and weight by this study. This estimate and our results show the complexity of the genetic basis of maize kernel size and weight.

Cell wall invertase, which hydrolyzes sucrose into glucose and fructose, plays an important role in plant growth and development. Transgenic carrot plants with reduced cell wall and vacuolar invertase activity had altered phenotypes at the very early stages of development and reduced tap root development leading to smaller organ size (Tang et al.,
Rice GIF1, a cell wall invertase, was reported to affect grain filling and weight (gif1 mutant had ~24% lighter seeds compared with wild-type) and was a domestication gene (Wang et al., 2008). In maize, Mn1 (incw2), which also encodes a cell wall invertase, was confirmed to be important for kernel development. Kernel weight of Mn1 mutant was about 20% of wild-type due to the low cell wall invertase activity (Lowe and Nelson 1946; Miller and Chourey 1992). ZmINCW1 also encoded a cell wall invertase and was located in QTLs mapped in 3 RIL populations and was significantly associated with kernel size and weight. T-DNA insertion lines of Arabidopsis ortholog (AtcwINV2) of ZmINCW1 had reduced seed weight. Transformation of ZmINCW1 into this mutant increased seed weight, which indicates that ZmINCW1 had conserved function for kernel/seed development in maize and Arabidopsis. Similar conserved function was also reported for Mn1, OsGIF1, and AtcwINV1. Constitutive expression of Mn1, AtcwINV1, and OsGIF1 via transgenic method in an elite maize inbred line (Ye478) produced larger cobs and kernels leading to up to 145.3% improvement in grain yield (Li et al., 2013). These results suggest that genes encoding cell wall invertase might be a good choice for yield improvement through marker-assisted selection or genetic engineering.

Comparative genetic analysis is a powerful method for identifying genes that have conserved function across species, such as flowering time (Ghd7 in rice and ZmCCT in maize, Xue et al., 2008; Hung et al., 2012; Yang et al., 2013) and branching regulation (tb1 in maize and OsTB1 in rice, Clark et al., 2006; Takeda et al., 2003). Seed size and weight are two of the most important agronomic traits for yield and undergo selection during domestication. In rice, many genes affecting kernel development have been cloned (Table S7), such as GS3, GW2, and GS5. Their orthologs in maize, ZmGS3 (Li et al., 2010a) ZmGW2-CHR4, and ZmGW2-CHR5 (Li et al., 2010b), ZmGS5 (Liu et al., 2015) were also found to be involved in kernel development, but with different mechanisms. Here, we found that ZmINCW1 has conserved function for kernel/seed weight in maize and Arabidopsis and that expression regulation by cis-element might be the cause of the phenotypic change. This is very different from rice GIF1 where a 1-nt deletion caused the premature termination of its open reading frame.
frame. These findings suggest that even though these genes have conserved function, the types of genetic variation important for the phenotype may be different between species.

We used comparative genetic analysis and identified 30 genes that are orthologs of 18 cloned rice genes for seed size or weight. Among these 30 genes, 26 are located in the candidate region mapped by at least one method (SLM, JLM and GWAS) in the RIL populations and 5 were found to be significantly associated with kernel traits by candidate gene association mapping in a large association panel. Given the conserved functions of many of the known genes for kernel development in maize, rice, and wheat (Su et al., 2011; Hong et al., 2014; Qin et al., 2014; Jaiswal et al., 2015; Wang et al., 2015; Ma et al., 2016; Simmonds et al., 2016; Wang et al., 2016), these genes represent additional candidates for kernel development across various species.

In summary, our findings shed light on the genetic basis of kernel size and weight in maize. We provided candidate genes for many of the loci that contribute to natural variation in maize kernel size and weight. We also provided evidence for a conserved and unique genetic architecture of kernel traits in maize compared with rice.

MATERIALS AND METHODS

Plant materials and phenotype measurements

Seven RIL populations (B73 × BY804, KUI3 × B77, K22 × CI7, DAN340 × K22, ZHENG58 × SK, YU87-1 × BK and ZONG3 × YU87-1) were planted in eight trials in Hubei, Chongqing, Henan, Yunnan and Hainan province in China during 2011 and 2012, while the other three RIL populations (DE3 × BY815, K22 × BY815 and BY815 × KUI3) were planted in four trials (Chongqing, Hubei, Henan and Yunnan province in China during 2012) because of insufficient seeds for field trials in 2011. An association mapping population consisting of 540 inbred lines (AM540) was also planted in these 8 environments during 2011 and 2012. All populations were planted with one random block replication per location. For each line, we planted 11 plants per row and selected 5 well-pollinated ears in the middle of the row to measure 5 kernel size and weight traits (i.e. kernel length, kernel width, kernel thickness,
hundred-kernel weight and kernel test weight). Before measuring traits, we first mixed kernels of these 5 ears and used a digital ruler to measure KL, KW and KT of 30 single kernels (illustrated in Figure 1A). The average of these 30 kernels was used to represent the trait measurement. We measured hundred-kernel weight 3 times for a single line and used the mean value to represent HKW for that line. We used 250 ml of kernels to measure the kernel test weight for each line and if there were not enough kernels to measure 250 ml, we used at least 50 ml of kernels to measure the weight and then converted it to the 250 ml weight.

We used best-linear unbiased prediction (BLUP) value of each line to perform data analysis, including phenotype statistics, correlation analysis, and QTL mapping. BLUP values were computed by PROC MIXED in the Statistical Analysis System (SAS Institute, 1997) and Pearson correlation coefficients were calculated with Excel. The heritability for each trait was calculated as: \[ H^2 = \frac{\delta^2_g}{\delta^2_g + \delta^2_e/n} \], where \( \delta^2_g \) is the genetic variance, \( \delta^2_e \) is the residual variance and \( n \) is the number of environments.

**Genotype**

Ten RIL populations used in this study have been genotyped with the Illumina MaizeSNP50 BeadChip, with each population having 11,360 to 15,285 polymorphic markers, and these polymorphic markers were used to construct a high-density linkage map (Pan et al., 2016). These ten populations contain 1979-3071 genetic blocks in which no recombinant events occur. The 14 inbred lines used to construct 10 RIL populations were also contained in the 368 lines that were genotyped by RNA-Seq in a previous study (Fu et al., 2013). Thus, we projected the 1.03 million SNPs genotype of the 14 parental lines onto their 1887 offspring RILs using a two-step imputation strategy. We first used a method similar to the aforementioned imputation to separately project high-density SNPs from two parents onto offspring RILs based on the linkage map for each population, and then mapped the projected genotypes of RILs to base pairs according to the parental genotypes. In total, there were 14,612 genetic blocks for JLM and 185,212 blocks for GWAS.
Separate linkage mapping (SLM), joint linkage mapping (JLM) and genome-wide association analysis (GWAS) in the RIL populations

SLM was performed by composite interval mapping (CIM) using the Windows QTL Cartographer software version 2.5 (Wang et al., 2012) in each RIL population. The program settings were as follows: CIM Model = Model 6: Standard Model, control markers numbers = 5, window size = 10.0 cM, regression method = Backward Regression Method, walk speed = 0.5 cM. We used LOD = 2.5 as the threshold and the 2-LOD interval was considered as the QTL candidate region.

We combined 10 RIL populations to perform JLM and GWAS. For JLM, a linear mixed model was used to detect significant recombination blocks. The model is: \( y = X\beta + Z\gamma + \xi + \epsilon \), where \( X\beta \) represents fixed effects, \( Z \) is an \( N \times P \) matrix for the genotype (\( N \) is the total number of SNPs; \( P \) is the number of lines used to construct RIL populations), \( \gamma \) is a vector of genetic effects for markers, \( \xi \) is a vector of polygenic effect, \( \epsilon \) is a vector of the residual errors. The restricted maximum likelihood (REML) was used to estimate the parameters and a permutation test of 500 permuted samples was used to determine the threshold of likelihood ratio test (LRT) scores. At the type I error rate of 0.05, the threshold of LRT was 2.76.

For GWAS, we used a stepwise regression method (Tian et al., 2011) with minor modification. To control the polygenic background effect, the GWAS was performed per chromosome at a time. For each chromosome, we forced population effects and the effects of QTLs detected by SLM and JLM from other chromosomes to be included in a general linear model. The residual of this model was then used as the dependent variable to test all SNPs on the current chromosome. We used both forward and backward regressions to select variables and the cutoff P-value for SNPs entering or leaving the model was determined by 500 permutations. The SNPs in the final model were regarded as significant SNPs and the P-value was calculated from the marginal \( F \)-values of the SNPs. To reduce SNP redundancy, we performed a final backward regression for the significant SNPs. For SNPs falling within the QTL regions, a backward regression was conducted one QTL at a time, where population effects and all other QTLs were fitted to the model. For the remaining SNPs falling outside...
the QTL regions, a backward regression was conducted by forcing population effects and all
QTLs in the model. The median cutoff P-value of ten chromosomes was used as the threshold
of marker resulting in the final backward model.

Overlapping analysis
To analyze overlap between QTLs identified by SLM in each of the 10 RIL populations, we
compared the confidence intervals of the mapped QTLs. When two QTLs are overlapped,
they were considered to represent a single unique QTL. For the analysis of overlapping
between genes and QTLs, we compared the positions of the genes and the confidence interval
of QTLs identified with three methods. If the candidate region of a QTL identified by SLM is
large than 5 Mb, we limited the candidate region to 2.5 Mb on each side of the peak position.
Genes that fell into the candidate regions of QTL identified by SLM and JLM or 1 Mb
flanking region of the significant SNPs identified by GWAS were considered to be located in
the mapped QTLs. A significant level was obtained by comparing number of genes falling
within QTLs in our observation with the numbers resulted from 10,000 permutations. For
each permutation, we randomly selected 30 genes and count the number of genes falling
within QTL regions. To evaluate the overlaps between QTLs identified by the three different
models, a candidate region was used for each QTL identified by SLM and JLM, while 1-Mb
flanking region of the significant SNPs was used for GWAS. When there was an overlap
between two QTLs, we considered them as one unique QTL.

Candidate gene association analysis in an association panel
We identified 30 maize orthologs of 18 cloned rice genes based on the MSU Rice Genome
Annotation Project (Kawahara et al., 2013). This rice annotation project used 232,821
representative peptide sequences from rice (release 7), Arabidopsis (release 10), poplar
(release 2.2), grapevine (release 1_12x), sorghum (release 1.4), maize (release 5b filtered set),
and Brachypodium (release 1.0) to identify orthologous groups with OrthoMCL software (Li
et al., 2003).
Since 1.25M SNPs had been previously mapped in AM540 (Liu et al., 2016), we used this genotype dataset to identify the SNPs in these 30 genes and then performed candidate gene association analysis. We used mixed linear model (MLM, Yu et al., 2006), which took population structure and kinship into consideration, to test the significance between the SNPs within candidate genes and kernel traits with TASSEL software (Bradbury et al., 2007). Threshold was determined by Bonferroni correction (P < 0.05/N, where N is the gene number) for each trait.

**Phylogenetic tree construction**

Amino acid sequences of cell wall invertase protein sequences in rice, maize, and Arabidopsis were aligned, using the CLUSTALW program (Thompson et al., 1994). A phylogenetic tree was constructed with MEGA6 (Tamura et al., 2013). The statistical method was neighbor-joining and 1,000 bootstrap replications were used to test the phylogeny. The substitution model was p-distance and the partial deletion option was selected to treat gaps/missing data. The genes used for the phylogenetic tree construction were: GRMZM2G095725, ZmINCW1 (GRMZM2G139300), ZmINCW2 (GRMZM2G119689), ZmINCW3 (GRMZM2G123633), and ZmINCW4 (GRMZM2G119941) from maize; OsCIN1 (LOC_Os02g33110), OsCIN2 (GIF1, LOC_Os04g33740), OsCIN3 (LOC_Os04g33720), OsCIN4 (LOC_Os01g73580), OsCIN5 (LOC_Os04g56930), OsCIN6 (LOC_Os04g56920), OsCIN7 (LOC_Os09g08072), and OsCIN8 (LOC_Os09g08120) from rice; AtcwINV1 (At3G13790), AtcwINV2 (At3G52600), AtcwINV3 (At1G55120), AtcwINV4 (At2G36190), AtcwINV5 (At3G13784), and AtcwINV6 (At5G11920) from Arabidopsis.

**Genome wide association analysis (GWAS) of the expression level of ZmINCW1**

Expression QTLs for ZmINCW1 were identified through genome-wide association analysis in an association mapping population consisting of 368 maize inbred lines which were subsets of the 540 diverse lines used for candidate gene association. Gene expressions were quantified in these 368 lines by RNA-Seq in a previous study (Fu et al., 2013). Since we had
genotyped 368 lines with extra methods (Affymetrix Axiom Maize Genotyping 600K Array and genotyping-by-sequencing, Liu et al., 2016), we performed expression GWAS again with TASSEL software using mixed linear model (MLM, Q+K) model (Yu et al., 2006). The threshold was determined by Bonferroni correction and it was $P < 7.97 \times 10^{-7}$ ($P < 1/n$, where $n$ is the total number of SNPs).

Transgenic analysis in *Arabidopsis*

The CTAB method was used to extract the *Arabidopsis* DNA and primers AtcwINV2_T-DNA and p745 were used to confirm the T-DNA insertion. Primer ZmINCW1_CDS was used to amplify the maize open reading frame (ORF) of ZmINCW1. The coding region of ZmINCW1 fused with the HA tag was then cloned behind the cauliflower mosaic virus 35S promoter into pCAMBIA99-1-3 vector and transformed into *Arabidopsis* by Agrobacterium-mediated transformation. Primer ZmINCW1_CDS was used to screen positive transgenic lines. Total RNA was extracted from fresh leaves using an RNA extraction kit (BioTeke, China) and cDNA was synthesized from the extracted RNA using the TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix kit (TransGen, Chian). Quantitative PCR was performed for the gene expression using ZmINCW1_RT and AtACTIN primers. All primers used in this study were listed in Table S12.

For protein expression analysis, proteins were extracted from aerial parts of 3 individual plants (6 leaves) and separated by SDS-PAGE. Proteins were transferred to nitrocellulose membrane for western blot. HA detection was performed using 1:5000 dilution of anti-HA mouse monoclonal antibody, followed by hybridization with 1:10000 dilution of goat-anti-mouse-HRP second antibody. The HRP signal was detected by the ECL substrate kit.

When we measured the *Arabidopsis* seed weight, we first took digital photos of the seeds and then measured the seed weight. With each photo, we used ImageJ (https://imagej.nih.gov/ij/) software to count the number of seeds. For each individual, we
measured weight of at least 200 seeds 3 to 5 times with replacement and then converted it to thousand seed weight.

Supplemental Data:

Supplemental Figure S1. Boxplots of kernel size and kernel weight in 10 RIL populations.

Supplemental Figure S2. QTLs detected in 10 RIL populations with separate linkage mapping.

Supplemental Figure S3. Overview of identified QTLs and significant SNPs for KTW, KL, KW, and KT.

Supplemental Figure S4. The significant associations between kernel traits and five maize genes which are orthologs of cloned rice genes for seed size or weight.

Supplemental Figure S5. The significant associations between seven maize mutant genes, which were reported to involved in maize kernel development, and kernel traits.

Supplemental Figure S6. Phylogenetic tree of cell wall invertase proteins in maize, rice and Arabidopsis.

Supplemental Figure S7. ZmINCW1 was located in candidate regions of QTLs identified in DAN340 × K22 for KL and DE3 × BY815 for HKW.

Supplemental Table S1. The mean value, standard deviations and heritability of five kernel traits in 10 RIL populations

Supplemental Table S2. Correlation coefficients between HKW and three other kernel size traits in 10 RIL populations

Supplemental Table S3. QTL numbers for kernel size and weight in 10 RIL populations

Supplemental Table S4. List of phenotypic variation explained by pleiotropic QTLs which could be detected in more than one population for same traits

Supplemental Table S5. List of phenotypic variation explained by pleiotropic QTLs which could be detected in more than one population for different traits

Supplemental Table S6. Full list of identified QTLs with SLM for kernel size and weight in 10 RIL populations
**Supplemental Table S7.** Full list of identified QTLs with JLM for kernel size and weight through combining 10 RIL populations

**Supplemental Table S8.** Full list of identified significant SNPs with GWAS for kernel size and weight through combining 10 RIL populations

**Supplemental Table S9.** Full list of candidate SNPs identified with GWAS for kernel size and weight through combining 10 RIL populations

**Supplemental Table S10.** Significant associations between maize orthologs of cloned rice genes and maize kernel size and weight

**Supplemental Table S11.** Significant associations between reported maize genes for seed development and maize kernel size and weight in this study

**Supplemental Table S12.** Primers used in this study

**ACKNOWLEDGMENTS**

We thank Mr. Xiongbing Yan for his help in fieldwork and Dr. Zhibing Lai and Dr. Hongze Wang for helping with the *Arabidopsis* transgenic experiments.
Figure legends

Figure 1. The measurements of kernel traits and variations of kernel size among 14 parental lines and representative lines in two RIL populations. (A) Measurements of kernel length, width, and thickness illustrated with a B73 kernel, bar = 1 cm. (B) 14 parental inbred lines used in this study showed considerable variations of kernel size. The arrows point from paternal lines to maternal lines. (C) The kernels of representative lines in YU87-1 × BK (left) and ZHENG58 × SK (right) RIL populations, bar = 1 cm.

Figure 2. Overview of the QTLs and significant SNPs for HKW identified with three models. (A) Top panel (Manhattan plot): the likelihood ratio test (LRT) scores from JLM. The red points under the x-axis indicate the significant SNPs identified in all three models. Middle panel: the results of SLM in each of the 10 RIL populations. The colored rectangles indicate the QTL regions in each RIL population and the color density is proportional to the LOD values. Bottom panel: the result of GWAS. The blue upward triangles indicate that the minor allele increases HKW relative to the major allele, the green downward triangles indicate the opposite effect, and red dots indicate the candidate SNPs identified by the backward regression model. (B) A pleiotropic QTL that was identified for KW, KT and HKW (R² = 12.40%, 5.13% and 6.92%, respectively) in K22 × CI7 population. (C) The number of overlapped QTLs or SNPs identified with three methods. Blue numbers are for SLM, red for GWAS, and green for JLM. For example “102/75/116” means that 102 QTLs identified with the SLM model overlapped with 75 QTLs identified with the JLM model and 116 SNPs from the GWAS model.

Figure 3. Comparative analysis of QTLs and genes identified from maize mutant studies or based on rice seed size or weight genes. A total of 21 rice genes (18 from this study and GS3, GW2 and GS5 from previous studies, shown in red) and 36 maize genes (shown in blue) reported to be involved in maize kernel development in mutant studies are shown. Points with different color and shape indicate that genes were significantly associated with maize kernel
size or weight by different methods. The heat map filled in the chromosome region indicates the density of QTLs for kernel traits. The window size is 1 Mb. SLM, separate linkage mapping; JLM, joint linkage mapping; GWAS, genome-wide association analysis.

Figure 4. **ZmINCW1 was significantly associated with maize kernel development.** (A) *ZmINCW1* is located in the QTLs identified in B73 × BY804 population for kernel size and weight. 12YN, Yunnan province in 2012; 11DHN, Hainan province in 2011. The arrow indicates the position of *ZmINCW1*. (B) SNPs in *ZmINCW1* were significantly associated with kernel size and weight in an association panel. (C) Genome-wide association analysis of the expression level of *ZmINCW1*. The red points indicate the SNPs located in *ZmINCW1*. (D) The expression level of *ZmINCW1* was significantly positively correlated with HKW in 2011 Yunnan (N = 292, r = 0.16, P = 7.30 × 10⁻³).

Figure 5. **Over-expression of ZmINCW1 rescues the reduced thousand seed weight in Arabidopsis AtcwINV2 T-DNA mutant.** (A-B) The T-DNA was inserted into the fourth exon of *AtcwINV2* and three primers were used to confirm the insertion (F + R and P + R). (C-D) Both wild-type (C) and T-DNA insertion mutant (D) had normal seeds (bar = 1 mm). (E) T-DNA insertion mutant had decreased thousand seed weight compared with wild type (N = 20/32, P = 7.34 × 10⁻⁶). (F-G) The expression levels of *ZmINCW1* in Arabidopsis were confirmed by RT-PCR (F) and western blot (G). (H-I) Both negative (H) and positive (I) transgenic lines had normal seeds (bar = 1 mm). (J) T₁ Positive transgenic lines had increased thousand seed weight compared with negative transgenic lines (N = 6/26, P = 1.51× 10⁻¹²). The combined result is shown in (E). (K) There was significant difference in thousand seed weight between positive lines and negative lines in T₂ generation (N = 6/5, P = 0.03). ***, P < 0.01; *, P < 0.05.
Table 1. QTLs or significant SNPs for kernel size and weight detected with three methods

<table>
<thead>
<tr>
<th></th>
<th>HKW</th>
<th>KTW</th>
<th>KL</th>
<th>KW</th>
<th>KT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLM&lt;sup&gt;a&lt;/sup&gt;</td>
<td>90, 27, 10, 2</td>
<td>70, 24, 11, 4</td>
<td>61, 15, 10, 2</td>
<td>89, 25, 17, 7</td>
<td>63, 15, 9, 2</td>
</tr>
<tr>
<td>JLM</td>
<td>56</td>
<td>59</td>
<td>55</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>GWAS&lt;sup&gt;b&lt;/sup&gt;</td>
<td>194, 30</td>
<td>136, 22</td>
<td>138, 26</td>
<td>163, 32</td>
<td>123, 25</td>
</tr>
</tbody>
</table>

<sup>a</sup> The number of QTLs identified in one population, more than one population, major QTLs in one population (R² > 10%) and major QTLs in at least two populations, respectively;

<sup>b</sup> The number of significant SNPs identified by GWAS and backward regression, respectively.

SLM, separate linkage mapping; JLM, joint linkage mapping; GWAS, genome-wide association study. HKW, hundred kernel weight; KTW, kernel test weight; KL, kernel length; KW, kernel width; KT, kernel thickness.


