Introduction

Maize is one of the most important crops in the world. Given the fact of ~10% protein and ~70% starch in the endosperm, it serves as a primary food for humans and feed for livestock (Nelson and Pan, 1995). There are mainly two forms of starch: the linear amylose and branched amylopectin. Amylose is composed of glucose residues linked through \( \alpha-1,4 \) bonds without any branch. Amylopectin contains both linear and branched glucose chains, 5% of which are joined via \( \alpha-1,6 \) bonds that introduce chain branches. The clusters of side chains allow amylopectin to fold into a dense and large molecule of glucose residues, which in turn increases kernel weight (James et al., 2003).

Starch in the normal maize endosperm is approximately 25% amylose and 75% amylopectin (Nelson and Pan, 1995). The ratio of amylose to amylopectin plays an important role in appearance, structure and quality of food product and processing. The amylose content determines the starchy gelling and firmness, whereas the amylopectin is primarily responsible for the formation of crystalline granules and thickening of paste (Whitt et al., 2002). In general, high amylose improves the product texture of starch and turns it into a source of slowly digestible carbohydrate (Collins, 1909; Kim et al., 1998; Stinard et al., 1993), while low amylose corresponds to higher peak of paste viscosity and strong resistance to retrogradation (Van Hung et al., 2006).

Over the past decades, enormous progress has been made in understanding of the genetics and biochemistry of starch synthesis. The key enzymes involved in starch synthesis have been elucidated (Figure 1). Basically, the starch is synthesized by a suite of enzymes, including sucrose synthase (SUS), ADP-glucose pyrophosphorylase (AGPase, the small unit encoded by brittle2 (bt2) and the large by shrunken2 (sh2)), soluble starch synthases (SSs) and granule-bound starch synthase (GBSS, encoded by waxy (wx)), starch-branching enzyme (BE, encoded by amylose extender1 (ae1)) and starch-debranching enzyme (DBE, encoded by sugary1 (sugar1) (sugary1) (Jeon et al., 2010). The starch synthesis pathway in maize endosperm is thought to begin with the cleavage of sucrose into fructose and UDP-glucose, catalysed by SUS, the products of which may be then converted into ADP-glucose (ADPG) by AGPase. Amylose and amylopectin both use ADPG as the activated glucosyl donor for synthesis, but they are synthesized by different enzymes afterwards. GBSS is responsible for synthesis of amylose. Amylopectin biosynthesis requires a well-coordinated machinery complex of enzymes including SSs, BE and DBE (Figure 1) (Hannah, 1997; James et al., 2003; Jeon et al., 2010; Keeling and Myers, 2010; Whitt et al., 2002). Lack of the BE enzyme in ae1 leads to accumulation of up to 50% amylose due to the less amylopectin production. Mutations at the waxy locus eliminate amylose synthesis, resulting in 100% amylopectin in the endosperm. Both mutants have been used in the crop breeding to create either high- or low-amylose maize to alter starch properties and utility. Reduction in amylopectin has been accomplished with ae1 and su1 mutants (James et al., 1995).

There is a dramatic variation in amylose content in the maize natural population ranging from zero (the waxy maize) to 64% (the amylo maize) (Mercier, 1973). Elucidating the genetic variation in starch biosynthesis and regulation is laborious based on traditional QTL mapping. Genome-wide association study
endosperm. In order to improve the seed quality by manipulating target genes, we selected a mapping population of 464 lines from a natural-variation germplasm pool with a wide genetic diversity. There was no obvious division of subgroups in our population, which could eliminate false positive and improve the power to recover meaningful associations. We measured the amylose content for the 464 maize lines including 454 natural inbred lines and 10 waxy lines that contain low amylose (Table S1) using a modified iodine colorimetry. There was a great variation in the amylose content ranging from 7.38% to 32.82% in the association panel. The overall distribution exhibited left-skewed due to the presence of waxy maize materials (Figure 2). The average kernel amylose content was 25.93%, close to 25% in the normal maize seed.

**Linkage disequilibrium**

Developed by Lai's group (Liu et al., 2016a), 9 007 194 single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) more than 0.05 could cover the whole maize genome and are theoretically sufficient for efficient GWAS analysis in maize with the genome size of 2 300 MB. We took advantage of these qualified SNPs as input data to calculate the genome-wide linkage disequilibrium (LD) in the association panel. We found a rapid decline in LD with the increasing physical distance on all chromosomes. The mean length of LD decay decreased rapidly to 15 kb at a cut-off of \( r^2 = 0.2 \) (Figure 3). The decay rate varied among chromosomes with a shortest LD (150 kb, \( r^2 = 0.1 \)) estimate observed on chromosome 6 and the longest LD (450 kb, \( r^2 = 0.1 \)) observed on chromosome 7. The overall LD decay distance was 250 kb (\( r^2 = 0.1 \)) across the entire genome (Figure 3).

**Loci associated with the amylose content**

GWAS was performed to associate the phenotype of amylose content with the genotype of 464 diverse inbred lines. Given the population structure and familial relatedness in the natural population, we applied the mixed linear model (MLM) to decrease the false-positive error. Quantile–quantile plots (Figure 4a) showed that the observed distribution did not deviate from the expected line, indicating that the confounders did not have significant influence. However, the curved tail represented a small number of true associations with the amylose content among thousands of unassociated SNPs. Association analysis identified 42 SNPs significantly associated with the amylose content at the minimum of \( P \leq 5 \times 10^{-8} \) (Figure 4b and Table S2). Due to the overall LD decay distance being 250 kb (\( r^2 = 0.1 \)) across the entire genome (Figure 3), a 250-kb region flanking the left and right sides of each SNP was defined as a QTL. 352 genes identified within the defined QTLs are summarized in Table S2. Of them, 39 candidate genes at 27 QTLs were implicated in carbohydrate metabolism and regulation (Table 1). The most significant SNP (chr9.S_23283117) was linked with waxy1 (~22 kb away of GRMZM2G024993) (Figure 4b), a key gene encoding GBSS for amylose biosynthesis (Shure et al., 1983). Four SNPs exist in gene regions, namely two (chr9.S_145846544) in the locus of GRMZM2G425683 (Figure 5a), one (chr4.S_160268689) located in GRMZM2G110483 (Figure 5b) and one in AC044284_FGT004 with unknown function yet. GRMZM2G425683 encodes a major facilitator superfamily protein (MFS-like) responsible for sugar transport. The two SNPs in GRMZM2G425683 create a stop codon and a non synonymous amino acid substitution (Figure 5c). The allele of C was greatly

**Results**

**Natural variation in the amylose content**

The amylose/amylpectin composition in starch is a complex quantitative trait affecting the quality and yield of maize...
correlated with a high amylose content, while the allele of G was linked with a low level of amylose (Figure 5e). Surprisingly, GRMZM2G425683 was specifically expressed in roots, but not in the seeds (Stelpflug et al., 2016). GRMZM2G110483 encodes a pentatricopeptide repeat-containing protein (PPR-like), which is highly expressed in the endosperm. The SNP in the PPR-like gene causes a C-to-T mutation leading to an Ala-to-Val amino acid substitution (Figure 5d). The allele for high amylose is C compared with the T allele for low amylose (Figure 5f). Still, further investigation is required to test whether GRMZM2G425683 and GRMZM2G110483 affect the amylose content.

**Functional annotation for candidate genes**

Besides the known waxy gene involved in amylose synthesis, we identified more genes such as GRMZM2G089836, GRMZM2G173674 and GRMZM2G069008 from the upstream of the starch synthesis pathway. For example, GRMZM2G089836 encodes an invertase. It was reported that sucrose entering the seed has to be cleaved by invertase into glucose and fructose, which are then resynthesized into sucrose (Walley et al., 2016; Wind et al., 2010). The mutant of the Arabidopsis cytosolic invertase had a high sucrose-to-glucose ratio and grew shorter primary roots (Qi et al., 2007). The maize miniature1 (mn1), a

---

**Figure 2** The frequency distribution of amylose content. x-axis shows the amylose content, and y-axis shows the population frequency.

**Figure 3** Linkage disequilibrium decay on maize chromosomes and whole genome. The red dashed horizontal line shows the LD threshold for the association panel ($r^2 = 0.1$).
mutant with lack of cell wall invertase2, reduced the sucrose transport to the seeds, which resulted in 30% loss of seed weight (Kang et al., 2009). The gene of GRMZM2G173674 encodes a phosphoglucomutase (PGM). It was studied in potato that phosphoglucomutase catalysed the conversion of glucose-6-phosphate into glucose-1-phosphate that acted as a precursor for starch biosynthesis in plastids (Van Harselaar et al., 2017). The gene was broadly expressed in all tissues, especially in immature cob and early stage of seed development (Walley et al., 2016). It was found that the plastidic phosphoglucomutase expression was up-regulated when starch began to accumulate, while its mutant impaired starch synthesis in rice pollen grains and caused male sterility (Lee et al., 2016). GRMZM2G069008 encodes a Nudix hydrolase that is widely distributed in dicots and monocots. Its main function catalyses the cleavage of ADP-glucose linked to starch biosynthesis (Kraszewska, 2008). It was reported that the pattern of ADP-glucose hydrolytic activity was inversely correlated with starch accumulation (Rodriguez-Lopez et al., 2000). One of Nudix hydrolases reduced the levels of both ADP-glucose and the amylose content when it was overexpressed (Comparot-Moss and Denyer, 2009; Keeling and Myers, 2010; Zeeman et al., 2010). Sucrose is produced from the carbon fixation of Calvin cycle in leaves through the activity of mitochondrial NAD-dependent malic enzyme (ME encoded by GRMZM2G085747), phosphoribulokinase (PRK encoded by GRMZM2G026024 and GRMZM2G143804) and other enzymes. Then, sucrose is transported to the storage organ by sucrose transporter (SUT6 encoded by GRMZM2G106741), where it is imported into the cytosolic compartment of each cell. In the cytosol, the sucrose synthase (SUS) or invertase (INV encoded by GRMZM2G089836) cleaves sucrose into fructose and UDP-glucose. After glucose 6-phosphate is transported into plastids, phosphoglucomutase (PGM encoded by GRMZM2G173674) catalyses glucose 6-phosphate into glucose 1-phosphate that is activated into ADP-glucose as the substrate for amylose synthesis. The content of ADP-glucose is also negatively regulated by Nudix hydrolases (NUDT encoded by GRMZM5G809417 and GRMZM2G069008). GBSS encoded by GRMZM2G024993 cooperates with other glycosyltransferases (AC186147.3_FG008, GRMZM2G129090, GRMZM2G164912, GRMZM2G039017) and glycosidase (GRMZM2G402368, GRMZM2G073584, GRMZM2G164912) to catalyse and elongate the sugar chains (Figure 6).

In conclusion, we propose to integrate the previous knowledge with our GWAS analysis into a model pathway of amylose biosynthesis in maize endosperm (Comparot-Moss and Denyer, 2009; Jeon et al., 2010; Stamp et al., 2016). The dynamic transcriptome analysis showed that waxy was actively expressed...
in the middle phase of maize endosperm development (Chen et al., 2014). Mutations in waxy eliminate amylose but increase the amylopectin content during the grain filling. These waxy mutants have been used in breeding to create high-amylopectin maize (Stamp et al., 2016; Tsai, 1974). Recent studies were focused on the downstream genes involved in the final product of starch biosynthesis, such as bt2, sh2, su2, ae1, su1 (Figure 1). Due to the complex of starch biosynthesis and small effect from multiple QTLs, the genetic mechanism of amylose itself rather than total starch synthesis is challengeable to investigate. Taking advantage of GWAS analysis, we were able to detect almost all QTLs at a genome-wide level with sufficient accuracy and sensitivity. Noticeably, the enzymes in the upstream of amylose pathway were deciphered, such as sucrose transporter, invertease, phosphoglucomutase, Nudix hydrolase, glycosyltransferases and glycosidases that are responsible for the precursors for amylose biosynthesis (Figure 6).

The 2,300 Mb of maize genome includes ~32,000 genes with large intergenic regions and ~84% repeats (Jiao et al., 2017). On average, every 72 kb contains a gene. Thus, it is reasonable that we found waxy being 22 kb away from one of the most significant SNPs with a P-value of 3.86e−14. The acceptable distance between SNPs and candidate genes relies on species, genotype structure and the linkage disequilibrium decay distance. For example, the GWAS for seed oil melting point in Arabidopsis identified two candidate genes FAD2 (AT3G12120) and FATB (AT1G08510) being 74 and 55 kb away from their corresponding ~value of 3.86e−14. The acceptable distance between SNPs and candidate genes relies on species, genotype structure and the linkage disequilibrium decay distance. For example, the GWAS for seed oil melting point in Arabidopsis identified two candidate genes FAD2 (AT3G12120) and FATB (AT1G08510) being 74 and 55 kb away from their corresponding

### Table 1 SNPs and candidate genes significantly associated with amylose content

| SNP ID       | Allele | P value | Associate genes | Annotation | Symbol   |
|--------------|--------|---------|-----------------|------------|----------|
| chr3.5_139436833 | C/T    | 2.06e−08 | GRMZM2G174769 | Putative galacturonosyltransferase-like 9 | GAUT-like9 |
| chr3.5_13949410 | A/G    | 2.22e−09 | GRMZM2G340756 | MYB-related-transcription factor | MYB14 |
| chr3.5_140663351 | A/G    | 4.61e−11 | GRMZM2G402368 | Alpha-galactosidase 3 | GAL3 |
| chr3.5_141284190 | C/T    | 2.77e−08 | GRMZM2G073584 | Beta galactosidase 9 | Iac29 |
| chr3.5_141284190 | C/T    | 2.77e−08 | GRMZM2G067171 | Putative GATA transcription factor | GATAA31 |
| chr3.5_149274688 | A/T    | 3.12e−08 | GRMZM2G127789 | Glutathione S-transferase GST 29 | GST29 |
| chr4.5_125184357 | A/G    | 1.43e−09 | GRMZM2G106741 | Sucrose transporter 6 | SUT6 |
| chr4.5_125294997 | G/T    | 7.84e−12 | GRMZM2G312806 | Mitochondrial transcription termination factor | mTERF |
| chr4.5_159118307 | A/G    | 1.57e−09 | GRMZM2G062024 | Phosphoribokinase | PRK |
| chr4.5_160268689 | C/T    | 2.89e−11 | GRMZM2G110483 | Pentatriopeptide repeat-containing protein | PPPR-like |
| chr4.5_160268689 | C/T    | 2.89e−11 | GRMZM2G589696 | DOF-transcription factor 43 | dof43 |
| chr4.5_160268689 | C/T    | 2.89e−11 | GRMZM2G143804 | Phosphoribokinase | PRK |
| chr4.5_163058675 | A/G    | 2.19e−08 | AC186147.3_FG008 | Alpha-6-galactosyltransferase | x34.3 |
| chr4.5_165621095 | C/T    | 4.31e−08 | GRMZM2G122846 | bZIP transcription factor | bZIP |
| chr4.5_25408662 | C/G    | 3.08e−08 | GRMZM5G1809417 | Nudix hydrolase 15 | NUDT15 |
| chr4.5_33734598 | A/G    | 4.76e−09 | GRMZM2G0108482 | Orphans transcription factor | Orphans |
| chr4.5_33734598 | A/G    | 4.76e−09 | GRMZM2G363540 | Glutathione S-transferase GST 26 | GST26 |
| chr4.5_46702131 | C/G    | 1.09e−09 | GRMZM2G100583 | NAC domain transcription factor | NAC75 |
| chr4.5_46744218 | C/T    | 2.49e−10 | GRMZM2G129090 | UDP-glycosyltransferase 91D1 | UGT91D1 |
| chr5.3_149999559 | A/C    | 3.71e−08 | GRMZM2G173674 | Phosphoglucomutase | PGM |
| chr5.3_20217324 | A/C    | 4.45e−08 | GRMZM2G164912 | Galactan beta-1,4-galactosyltransferase | GAL51 |
| chr5.3_24096652 | A/C    | 4.71e−09 | GRMZM2G085747 | Malic enzyme | ME |
| chr5.3_67474748 | A/C    | 3.28e−08 | GRMZM2G128929 | L-lactate dehydrogenase | LDH |
| chr5.3_67474748 | A/C    | 3.28e−08 | GRMZM2G089836 | Invertase 2 | INV2 |
| chr5.3_83843824 | A/G    | 1.48e−08 | GRMZM2G449843 | Alpha/beta-Hydrolases | ABHD |
| chr5.3_10374862 | C/T    | 4.87e−08 | GRMZM2G0119017 | 3-beta-glycosyltransferase | BSGALT |
| chr5.3_138336082 | C/T    | 4.52e−08 | GRMZM2G118979 | Alpha/beta-Hydrolases | ABHD |
| chr5.3_138363042 | C/G    | 2.40e−08 | GRMZM2G095727 | Two-component response regulator-like PRR73 | ARR8-like |
| chr5.3_139318861 | A/C    | 1.20e−08 | GRMZM2G165357 | UDP-glucuronic acid decarboxylase 1 isoform | UXS1 |
| chr5.3_139318861 | A/C    | 1.20e−08 | GRMZM2G161293 | Beta-1,3-galactosyl-O-glycoprotein | BGDP |
| chr5.4_145846544 | C/T    | 6.52e−12 | GRMZM2G355679 | Calmodulin-binding transcription activator 2-like | CAMTA2 |
| chr5.4_145846544 | C/T    | 6.52e−12 | GRMZM2G026742 | HSF-transcription factor 9 | hsf19 |
| chr5.4_145846544 | C/T    | 6.52e−12 | GRMZM2G126936 | NAC domain-containing protein 67-like | nac645 |
| chr5.4_145846544 | C/T    | 6.52e−12 | GRMZM2G425683 | Major facilitator superfamily protein | MF5-like |
| chr5.4_145846544 | C/T    | 6.52e−12 | GRMZM2G126834 | ARK transcription factor | ar1 |
| chr5.3_23283117 | C/T    | 3.87e−14 | GRMZM2G024993 | Granule-bound starch synthase | GBSS |
| chr5.3_23283117 | C/T    | 3.87e−14 | GRMZM2G171395 | NAC domain transcription factor | nac868 |
| chr5.3_24270258 | A/G    | 1.60e−08 | GRMZM5G811192 | Glycerolipids | GHB1 |
| chr5.3_24270258 | A/G    | 1.60e−08 | GRMZM2G69008 | Nudix hydrolase 14 | NUDT15 |

Association analysis identified 27 SNPs and 39 candidate genes implicated in carbohydrate metabolism and regulation.

© 2017 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 16, 688-695
the distance from the lead SNP was applied as 50 kb to define potential candidate genes (Li et al., 2013). The 200-kb genomic region on either side of the most significant SNP is determined as an acceptable distance for genome-wide study of drought tolerance in maize (Wang et al., 2016).

GWAS has begun to serve as a new foundation for understanding the genetic architecture of complex agronomic traits in crops, although it still has confounding limitations. GWAS is powerful to analyse traits underpinned by a small number of loci with large effect size. It should be noticed that the identified significant 42 SNPs are not always the true causative loci by the facts of linkage or error structure of data. We need to take into account of sample size, the imperfect genotyping data and other confounding factors. We maximized the genetic variance within 464 samples collected from different locations. We only used inbred lines to keep genetically identical, which are suited to GWAS analysis. We found 39 candidate genes from 27 significant loci associated with the amylose content. Surprisingly, we did not find any overlapped genes such as sh2 and bt2 in the known pathway of starch biosynthesis. Additionally, rare variants suffer from many other noncausative rare variants with strong or complete association within the genome (Korte and Farlow, 2013).

Experimental procedures

Plant material

The GWAS mapping population included 464 temperate maize inbred lines collected and genotyped by Dr. Lai’s laboratory at China Agricultural University (Liu et al., 2016a). All the inbred lines were planted in the field of Sanya (Hainan Province, 18.75N, 109.17E) in November 2013. The mature seeds of each inbred line were harvested in bulk and used for amylose content analysis.

Amylose content determination

Apparent amylose content (AAC) was measured based on the modified protocol from Gibbon et al. (Gibbon et al., 2003). Mature kernels were ground lightly in a grinding mill. Instead of total starch isolation, we filtered the kernel flour with 80-mesh nylon screen and dried the flour at 37 ºC overnight. We weighed 25 mg kernel flour into 15-mL centrifuge tube, added 4 mL 80% ethanol and mixed well, heated in a water bath at 80 ºC for 20 min with a mixing every 5 min, then centrifuged at 1500 g for 10 min, and supernatant was poured. The precipitates were gelled in 300 lL of absolute ethyl alcohol and 4.5 mL of 1M NaOH by heating to 65 ºC for 1 h; subsequently, 500 lL sample was transferred to 50-mL centrifuge tube, diluted by adding 25 mL water and neutralized in 1 mL 1M acetic acid, then added 1 mL I2/KI solution (2 mg/mL I2/20 mg/mL KI) and made constant volume to 50 mL, mixed well and incubated for 10 min. The OD was measured on a spectrophotometer (Beckman Coulter) using a 620-nm filter. The amylose content was determined by the average value from three biological replicates.

Association mapping and SNP annotation

The genotypes of 9,007,194 SNP markers filtered to retain only MAF >0.05 were generated by Dr. Lai’s laboratory at China Agricultural University (Liu et al., 2016a). The decay distance of
linkage disequilibrium in the association panel was calculated by PopLDdecay (https://github.com/BGI-shenzhen/PopLDdecay) with all SNP markers and mixed linear model (MLM) GWAS was carried out using the EMMAX suite (Kang et al., 2010). EMMAX-kin was used to calculate the Balding–Nichols kinship matrix to account for family relatedness among the samples, after which EMMAX associations with the phenotype were calculated. The significant threshold of $P$ value is $5 \times 10^{-8}$, indicated by a horizontal line in the Manhattan plot at $-\log_{10} (P \text{ value}) = 7.3$. All SNPs significantly related to the phenotype were annotated by snpEff (Cingolani et al., 2012). The SNP with the most significance within the same LD block ($r^2 < 0.1$) was selected to represent the locus. The LD between SNPs was also calculated by PopLDdecay (https://github.com/BGI-shenzhen/PopLDdecay). To seek candidate genes in the identified loci for the amylose, we used public gene annotation data sets from both maizeGDB (http://maizegdb.org) and NCBI (https://www.ncbi.nlm.nih.gov). All the annotated genes within the decay distance of linkage disequilibrium in the association panel of SNPs were retrieved. Blastp, InterProScan (Jones et al., 2014) and Pfam (Finn et al., 2014) were also used for annotation candidate genes.

Acknowledgements

We thank Dr. Jinsheng Lai and Dr. Junpeng Shi from China Agricultural University for generously providing the maize inbred lines and their genotypes for our research. We appreciate Dr. Xuehui Huang from Shanghai Normal University and Dr. Hao Gong from Institute of Plant Physiology & Ecology of CAS for their guidance of GWAS analysis. The project was supported by the Ministry of Science and Technology of China (2016YFD0100503 to WW) and Chinese Academy of Sciences Grant (XDA08020107 to YW).

Conflict of interest

The authors have no conflict of interest to declare.

References

Branham, S.E., Wright, S.J., Reba, A., Morrison, G.D. and Linder, C.R. (2016) Genome-wide association study in Arabidopsis thaliana of natural variation in seed oil melting point: a widespread adaptive trait in plants. J. Hered. 107, 257–265.
