Theoretical and Applied Genetics
https://doi.org/10.1007/s00122-019-03326-z

ORIGINAL ARTICLE

QTL mapping for maize starch content and candidate gene prediction combined with co-expression network analysis

Feng Lin[1] · Ling Zhou[1] · Bing He[1] · Xiaolin Zhang[1] · Huixue Dai[2] · Yiliang Qian[3] · Long Ruan[3] · Han Zhao[1]

Received: 26 September 2018 / Accepted: 11 March 2019

Abstract

Key message **A major QTL *Qsta9.1* was identified on chromosome 9, combined with GWAS, and co-expression network analysis showed that GRMZM2G110929 and GRMZM5G852704 are the potential candidates for association with maize kernel starch content.**

Abstract Increasing maize kernel starch content may not only lead to higher maize kernel yields and qualities, but also help meet industry demands. By using the intermated B73 × Mo17 population, QTLs were mapped for starch content in this study. A major QTL *Qsta9.1* was detected in a 1.7 Mb interval on chromosome 9 and validated by allele frequency analysis in extreme tails of a newly constructed segregating population. According to genome-wide association study (GWAS) based on genotyping of a natural population, we identified a significant SNP for starch content within the ORF region of GRMZM5G852704_T01 colocalized with QTL *Qsta9.1*. Co-expression network analysis was also conducted, and 28 modules were constructed during six seed developmental stages. Functional enrichment was performed for each module, and one module showed the most possibility for the association with carbohydrate-related processes. In this module, one transcripts GRMZM2G110929_T01 located in the *Qsta9.1* assigned 1.7 Mb interval encoding GLABRA2 expression modulator. Its expression level in B73 was lower than that in Mo17 across all seed developmental stages, implying the possibility for the candidate gene of *Qsta9.1*. Our studies combined GWAS, mRNA profiling, and traditional QTL analyses to identify a major locus for controlling seed starch content in maize.

Introduction

Maize (*Zea mays*) is one of the most important production crops worldwide, and approximately 70% of its kernel weight is starch, which is the main energy source component supplying adequate food and feed to humans and animals, as well as a vital role in bio-ethanol production and the industrial applications. Improving maize kernel starch content may not only lead to higher maize kernel yields and qualities, but also enhance the portion used in industrial applications.

Typically, starch is a complex branched polymer wherein the D-glucose units are linked by α-(1,4) linkages and the branch points are α-(1,6) linkages, generating two main homopolymers: amylose and amylopectin. In the endosperm, amylose is synthesized by pyrophosphorylase (AGPase) synthesizing the ADP–glucose and granule-bound starch synthase (GBSS). Amylopectin biosynthesis requires a series of enzymatic reactions involving AGPase, soluble starch synthase (SS) lengthening the glucose polymer by the formation of α-(1,4) bonds, starch branching enzyme (BE) forming the α-(1,6) bonds, and starch debranching enzyme (DBE) that cleave α-(1,6) bonds (Jeon et al. 2010). Other than DBE, a series of α- and β-amylases also degrade starch during cereal germination.

In maize, a number of genes play major roles in starch metabolism and mutations of these genes can dramatically

Communicated by Albrecht E. Melchinger.

Feng Lin and Ling Zhou equal contributions to the paper.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-019-03326-z) contains supplementary material, which is available to authorized users.

✉ Han Zhao
zhaohan@jaas.ac.cn

[1] Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing, China

[2] Nanjing Institute of Vegetable Sciences, Nanjing, China

[3] Anhui Academy of Agricultural Sciences, Hefei, China

Published online: 19 March 2019

influence starch content (Hennen-Bierwagen and Myers 2013). For example, shrunken1 (*sh1*) mutant can reduce the activity of sucrose synthase to hinder the sucrose from converting to uridine diphosphate glucose and fructose (Chourey 1981; Chourey and Nelson 1976). Mutations of *brittle2* (*bt2*) and *shrunken2* (*sh2*) genes, encoding AGPase large subunit and AGPase small subunit respectively, cause significant reduction in starch content with reduced AGPase activity (Cameron and Teas 1954; Dickinson and Preiss 1969; Laughnan 1953; Tsai and Nelson 1966). Some other genes also have their roles in starch biosynthesis, including *brittle1* (*bt1*), *Waxy1* (*Wx1*), *Dull1* (*Du1*), *Amylose extender1* (*Ae1*), *Sugary1* (*Su1*), *Sugary2* (*Su2*), and *Zpu1* (Hennen-Bierwagen and Myers 2013). The mutants of these genes lead to reducing the starch content, such as *shrunken1*, *brittle2*, and *shrunken2*. However, a number of other genes may exist and subtly alter starch content variations with the artificial selection. By using a recombinant inbred line (RIL) population, seven genes detected as QTLs were considered as potential candidate genes for starch content, three of them encode enzymes in non-starch metabolism and four may act as regulators of starch biosynthesis (Wang et al. 2015). Unlike key genes involved in the pathway, these QTLs could be introduced into an elite cultivar through marker-assisted selection and facilitate the improvement for fine-tuning starch content in breeding program (Ashikari and Matsuoka 2006).

Network analysis based on similarity of gene expression pattern has been considered as a more rapidly way to reinforce the discovery of candidate genes or loci with higher confidence (Liseron-Monfils and Ware 2015). Gene co-expression networks are useful in identification of gene modules and key genes responsible for a particular condition. Weighted correlation network analysis (WGCNA) is a technique widely used for finding modules by using the gene expression levels that are highly correlated across samples (Langfelder and Horvath 2008). Through this approach, co-expression networks were constructed and gene modules were successfully detected in *Arabidopsis*, rice, maize, soybean, tomato, sugarcane, and *Populus* (Choe et al. 2016; Das et al. 2017; DiLeo et al. 2011; Downs et al. 2013; Ferreira et al. 2016; Shaik and Ramakrishna 2013; Zinkgraf et al. 2017). Based on transcript abundance data at 50 developmental stages in maize, 24 robust co-expression modules were classified and strongly associated with tissue types and related biological processes (Downs et al. 2013). In networks, those genes highly connected are called hub genes and expected to play important roles in some biological mechanism of response under particular conditions (Das et al. 2017). Through integrated network analysis, a class of hub genes had been identified to induce major transcriptome reprogramming during grapevine development (Palumbo et al. 2014). In most cases, co-expression network analysis was used to identify gene modules for a particular condition; nevertheless, it also can be used for discovering conserved modules across species under different conditions (Leal et al. 2014; Obertello et al. 2015). By using co-expression analysis between tomato and potato coupled with chemical profiling, an array of 10 genes were revealed partaking in SGA biosynthesis, a toxic substance found in some tubers and tomatoes (Itkin et al. 2013). Through analysis of a drought-tolerant maize genotype HKI1532 and a drought-sensitive genotype PC3, 174 drought-responsive genes were selected from HKI1532, and their co-expression network revealed key correlations between different adaptive pathways (Thirunavukkarasu et al. 2017).

On the other hand, co-expression networks analysis contained several limits, for example, a high level of false positive interactions based on similarity of expression patterns independent of any real physical or regulatory link. Another problem is the low correlation between mRNA and protein levels under the condition studied (Petricka et al. 2012). So network analysis should rely on the integration of these methods with conventional and molecular breeding. Combination of molecular networks, transcriptomic data, and genetic map information from QTL and genome-wide association studies analyses should efficiently reduce the false positive rate inherent to each individual method (Liseron-Monfils and Ware 2015).

In this study, QTL mapping together with genome-wide association analysis identified a genomic region controlling starch content on chromosome 9, and co-expression network analysis suggested candidate genes. By using intermated B73 × Mo17 (IBM) population, a major QTL *Qsta9.1* was mapped in a 1.7 Mb region of 5,080,913 bp to 6,826,022 bp of chromosome 9 based on B73 v3 reference genome. Allele frequency analysis in a newly constructed population validated the association of this locus with starch content variation. Genome-wide association study identified a significant SNP in the ORF region of GRMZM5G852704_T01 within *Qsta9.1* interval. Co-expressed genes were screened through maize inbred line B73 and Mo17 across six seed developmental stages. Modules of co-expressed genes were identified, and functional enrichment was performed. In this major QTL interval, one transcripts GRMZM2G110929_T01 annotated as GLABRA2 expression modulator was assigned in the module with enriched function of carbohydrate biosynthetic or metabolic process, showing the potential candidate for starch content.

Materials and methods

Plant materials

Totally 283 intermated recombinant inbred lines (RILs) derived from B73 and Mo17 (the IBM population) and the parental lines were employed in this study. The

materials were grown at Liuhe (LH), Nanjing, and Sanya (SY), Hainan, China, with three replication blocks in each location in both 2015 and 2016. This experiment was performed in a randomized complete block design with 285 rows in each block and around 20 plants per row.

A $F_{2:3}$ population was newly constructed from the cross between B73 and Mo17 in 2016, including 540 lines. Out of these individuals, two starch content tails consisting of 27 plants each were selected for further analyses.

The seeds of the 149 maize lines were obtained from the Jiangsu Academy of Agricultural Sciences, Beijing Academy of Agricultural and Forestry Sciences and Anhui Academy of Agricultural Sciences (Supplementary Table S1), and planted in the field of Liuhe (LH), Nanjing, China, in 2018, as the GWAS analysis population. The mature seeds of each line were harvested in bulk and used for starch content analysis.

Starch content measurement

The 3,5-dinitrosalicylic acid (DNS) method (Miller 1959) was used to determine starch content with ten mixed mature ears on a VECTER22/N near-infrared analyzer. Three replicates were performed for each sample.

Genotyping and QTL mapping

The genetic map of IBM population including more than 1300 markers was downloaded at the Maize Genetics and Genomics Database (http://www.maizegdb.org/data_center/qtl-data). To saturate the gaps in the IBM linkage map, more than 260 InDel markers were developed following the mInDel pipeline (Lv et al. 2016) by comparing corresponding sequences between B73 (RefGen_v3_sequence) and Mo17 sequenced by the Department of Energy Joint Genome Institute (JGI) (http://www.maizegdb.org/).

The software QTL IciMapping (v4.1) (Meng et al. 2015) was used for map construction and QTL mapping. ICIM-ADD was selected for QTL scanning, and the LOD score for declaring a QTL was set as 2.5.

Genome-wide association study

An association mapping panel composed of 149 maize inbred lines was used as plant material in this study (Supplemental Table S1). The seedlings of 149 maize lines were sampled and SNP genotyping was performed using the MaizeSNP50 (50 K) BeadChip based on Illumina platform as described by the manufacturer (Illumina, Inc. San Diego, CA). SNPs with $< 5\%$ minor allele frequency data and $> 20\%$ missing data were eliminated as redundant data. In addition, we used a linkage disequilibrium threshold (r^2) of 0.20 with a SNP window size of 50 and number of SNPs to shift window at each step of 10 (PLINK command: –indep-pairwise 50 10 0.20) (Purcell et al. 2007). Then, it was plotted with R scripts, which drew averaged r^2 against pairwise marker distances.

Phylogenetic tree was constructed by using neighbor-joining method on the basis of distance matrix calculated by using the software PHYLIP v.3.68 (http://evolution.genetics.washington.edu/phylip.html) and was presented by using interactive tree of life (iTOL) v3 (Letunic and Bork 2016). Principal component analysis (PCA) was performed with software EIGENSOFT v6.1.4 (https://www.hsph.harvard.edu/alkes-price/software/). Population substructure was calculated using ADMIXTURE (version 1.3) (Alexander et al. 2009) running with default settings for $K = 1$ to $K = 10$. And the kinship coefficient matrix that explained the most probable identity by state of each allele between individuals was estimated with the TASSEL program (Bradbury et al. 2007).

Association analyses were conducted using both general linear model (GLM) and mixed linear model (MLM) through the TASSEL program (Bradbury et al. 2007). The threshold for significant association was set based on the Bonferroni correction ($0.05/9076 = 5.51 \times 10^{-6}$). The Manhattan plot of $-\log_{10} P$ was generated using R software.

RNA-seq dataset

For co-expression network analysis, the RNA-seq data of maize whole seeds of B73 and Mo17 with six developmental stages, including 12, 16, 20, 24, 30, and 36 days after pollination, were downloaded from https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP015339. After trimmed via SolexaQA (Cox et al. 2010) with Phred score ≥ 20 longer than 20 bp, reads were mapped to the B73 v3 reference genome through Bowtie2 (v2.1.0) (Langmead et al. 2009). Only the unique hit in the reference genome was used for further analysis. Based on the coordinates of mapped reads, the read depth of each gene was computed and the fragments per kilobase of transcript per million mapped reads (FPKM) was calculated by using Cufflinks (v2.1.1) (Trapnell et al. 2010) representing the expression levels.

Co-expression network analysis

Two-way ANOVA was performed using a custom-made function in *R* to identify transcripts that were differentially expressed following different stages. The *P* values for the model were then corrected for multiple hypotheses testing using FDR correction at 5% (Benjamini and Hochberg 1995). The transcripts passing the cutoff ($P \leq 0.05$) for the model and developmental stages were deemed significant.

Following the WGCNA procedure (Langfelder and Horvath 2008), consensus co-expression network construction and module detection were conducted.

Gene functional annotation and enrichment

The Blast2GO software package v4.1.9 (Conesa et al. 2005) was used to predict gene functions in three steps: finding homologues sequences with BLASTX algorithm with a cut-off *E*-value of 1.0E−3, mapping to retrieve gene ontology (GO) terms, and annotating to predict reliable functions. The transcripts with significant GO terms were determined with an *E*-value hit filter of less than 1.0E−6 and an annotation cutoff of 55.

In addition, SEA (singular enrichment analysis) tool on the AGRIGO sever was used by searching the predicted transcripts against the *Zea mays spp.* v5a to determine significant functional enrichment among the transcripts in each module. The list of transcript names was prepared, and enrichment GO terms were collected after statistical test from background of maize genome locus with a $P \leq 0.01$ significance threshold (after FDR population correction).

Results

QTLs associated with starch content in maize kernel

Although more than 1300 markers have anchored in the IBM map, still some gaps exist. To fill in the gaps, 1859 InDel markers with unique position were developed based on B73 and Mo17 genome sequences. 265 InDel markers were integrated into the IBM linkage map and filled some of the gaps. The newly constructed linkages map had extended the coverage from 6242.7 to 6407.1 cM and reduced the average interval from 4.69 to 4.00 cM.

In the IBM population, the average kernel starch content spanned from 62.04 to 72.25, implying large variation for starch content in this population. The whole genome was scanned and eight QTLs were detected for starch content in at least two experiment conditions, distributing on chromosomes 1, 2, 3, 7, and 9, respectively (Table 1). There were two QTLs identified on chromosomes 1, 3, and 9 (Table 1).

The most stable QTL (*Qsta9.1*) located on chromosome 9 was detected in all the four experiment conditions with the LOD score as 3.17–5.30, explaining 5.45–6.84% of the phenotypic variance (Table 1). B73 contributed the higher starch content allele. This QTL was mapped between *umc1867* and *JAAS669* on the terminal region of the short arm (Fig. 1). Based on B73 RefGen_v3, *Qsta9.1* was mapped in the physical distance around 1.7 Mb which was from 5,080,913 to 6,826,022 bp. Another QTL (*Qsta9.2*) was also detected on chromosome 9 which is more than 200 cM away from *Qsta9.1* (Table 1).

The other QTLs were detected in two experiments, explaining 3.74–7.97% of the phenotypic variance. Two QTLs located on chromosomes 3 (*Qsta3.1*) and 7 (*Qsta7*) obtained the higher starch content allele from Mo17, explaining as high as 5.65% and 7.36% of the phenotypic variance, respectively.

Table 1 QTLs detected for starch content in maize kernel by using the IBM population

QTLs	Trait	Chr	Position (cM)	Left marker	Right marker	LOD	R^2 (%)	Add
Qsta1.1	2015SY	1	394.0	*umc2025*	*z22*	5.81	6.78	0.45
	2016SY	1	394.0	*umc2025*	*z22*	4.92	5.86	0.41
Qsta1.2	2015SY	1	717.0	*csh4*	*umc1446*	4.31	4.95	0.38
	2016SY	1	717.0	*csh4*	*umc1446*	4.75	5.65	0.40
Qsta2	2016LH	2	311.5	*csu1080b*	*z47*	5.12	6.69	0.40
	2015LH	2	313.5	*z47*	*mmc0401*	4.22	7.34	0.51
Qsta3.1	2015SY	3	297.0	*z58*	*psr119a*	3.09	3.74	−0.33
	2016SY	3	300.0	*psr119a*	*cdo105*	4.63	5.65	−0.40
Qsta3.2	2016LH	3	399.0	*JAAS5448*	*umc1644*	3.56	5.64	0.36
	2015LH	3	404.5	*JAAS5448*	*umc1644*	2.55	5.11	0.40
Qsta7	2016SY	7	146.0	*crt2*	*phi034*	4.68	5.78	−0.41
	2015SY	7	151.0	*bnlg1094*	*psr371b*	6.56	7.97	−0.49
Qsta9.1	2016SY	9	24.5	*umc1867*	*JAAS669*	5.30	6.84	0.44
	2015LH	9	25.0	*umc1867*	*JAAS669*	3.17	5.76	0.43
	2016LH	9	26.5	*umc1867*	*JAAS669*	3.53	5.45	0.35
	2015SY	9	28.0	*umc1867*	*JAAS669*	4.70	6.68	0.45
Qsta9.2	2015SY	9	262.0	*psr129a*	*umc1107*	4.47	5.15	0.39
	2016SY	9	262.0	*psr129a*	*umc1107*	4.66	5.54	0.39

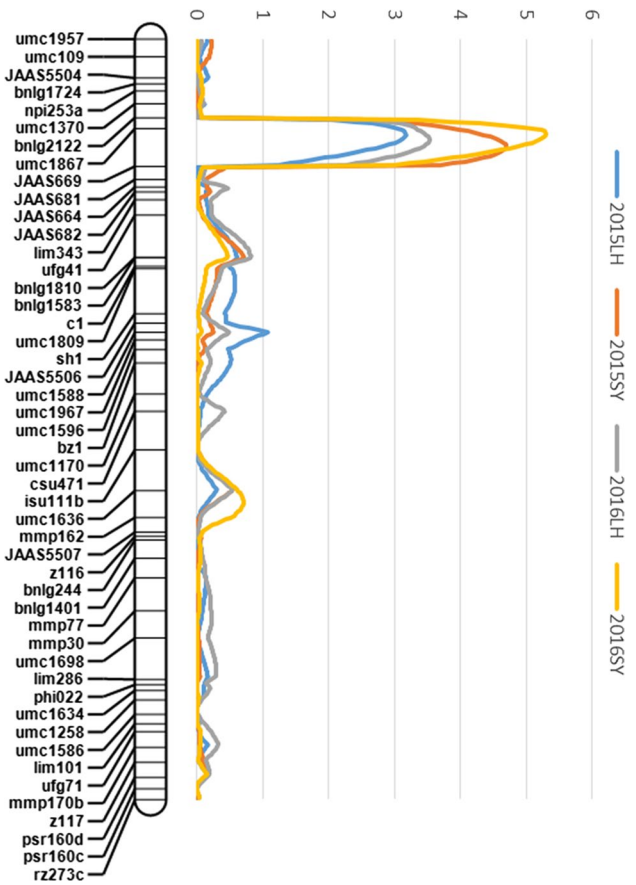

Fig. 1 The major QTL *Qsta9.1* detected on chromosome 9 in four experimental conditions

Allele frequency analysis in *Qsta9.1* region between two extreme tails of starch content

From a newly constructed $F_{2:3}$ population derived from B73 × Mo17, two extreme tail lines of starch content were selected with the average starch content of the lines of high tail of 77% and that of the low tail of 65%; each tail contains 27 lines. To further validate whether the major QTL *Qsta9.1* is associated with starch content variation, allele frequency test was conducted between the two extreme tails. Five markers anchored within *Qsta9.1* interval and eight markers adjacent to the region were used to detect the allele frequency difference between the two tails. Significant differences were detected for the five markers in the *Qsta9.1* region, showing significant deviation from 1:1 ratio (Table 2; Fig. 2). Declined allele frequency differences were observed for the flanking markers (Table 2), in agreement with the QTL analyses. Meanwhile, 50 markers evenly distributed throughout the genome were used to test the allele frequencies in the extreme tails, and 1:1 ratio was identified for each parental allele, indicating no bias exit in the individuals between two extreme tails (Fig. 2).

Genome-wide association study (GWAS) for starch content

A total of 47,479 SNPs remained after eliminating redundant data, and then, linkage disequilibrium was analyzed. Finally, 9076 SNPs were obtained for subsequent analysis. Population structure of the 149 maize lines based on the 9076 SNPs showed that the minimum peak value of cross-validation error appeared at $K = 7$ when the number of subpopulations increases from 2 to 10 (Supplementary Fig. S1; Supplementary Fig. S2A). Cross-validation results obtained by ADMIXTURE (Alexander et al. 2009) suggested $K = 7$ as the most likely number of groups (Supplementary Fig. S2A). In the PCA, seven principle components explained 68.02% of the total SNP variance, while the first and second principle components explained 16.15% and 12.20% of the total SNP variance, respectively (Supplementary Fig. S2B). A neighbor-joining tree was constructed based on *P* genetic distance showing seven clusters for this panel (Supplemental Fig. 2C), which was consistent with the results of the population structure analysis.

Considering the potential spurious associations in GWAS, the general linear model (GLM) and the mixed linear model (MLM) were compared, as shown in the quantile–quantile (QQ) plots (Supplementary Fig. S3), and both models fit to our data. GWAS for starch content was performed based on a dataset containing 149 maize lines genotyped with 9076 SNPs (Fig. 3). According to the MLM analysis, ten SNPs reached the significant level, distributing on chromosomes 1, 3, 4, and 9, respectively (Table 3; Fig. 3), which were also detected in GLM analysis (Table 3; Supplementary Fig. S4).

One SNP located at 5,877,060 bp within the *Qsta9.1* interval was identified significant association with starch content in the GWAS population (Table 3; Fig. 3;

Table 2 Allele frequency analysis in *Qsta9.1* region between two starch content extreme tails

Regions (marker number in the region)	Low tail frequency		High tail frequency		Frequency difference	Probability
	B73 allele	Mo17 allele	B73 allele	Mo17 allele		
Left *Qsta9.1*: (4)	0.39	0.61	0.69	0.31	0.30	< 0.05
Qsta9.1: (5)	0.35	0.65	0.69	0.31	0.34	< 0.05
Right *Qsta9.1*: (4)	0.40	0.60	0.62	0.38	0.22	< 0.05

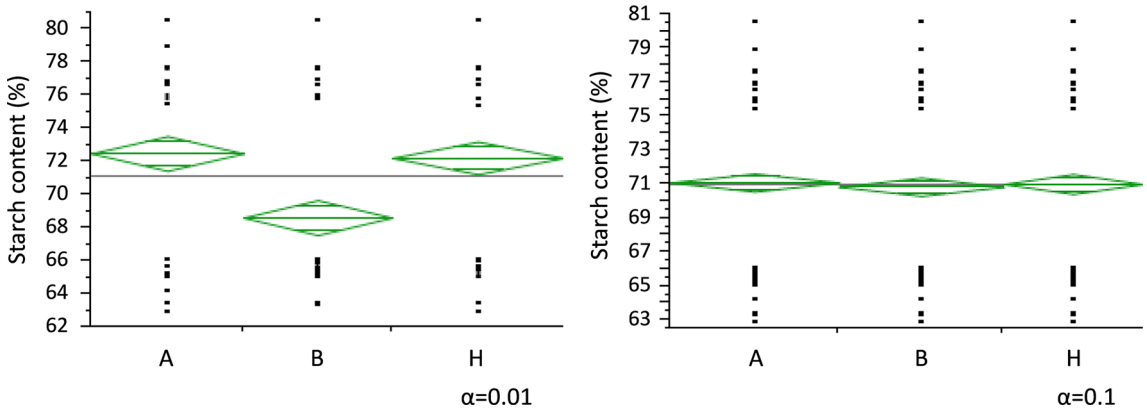

Fig. 2 Allele frequency analysis between two starch content extreme tails from a newly constructed $F_{2:3}$ population. Left: allele frequency test of nine markers surrounding *Qsta9.1* region; Right: allele frequency test of 50 markers distributed across maize genome. A: allele from B73; B: allele from Mo17; H: heterozygous

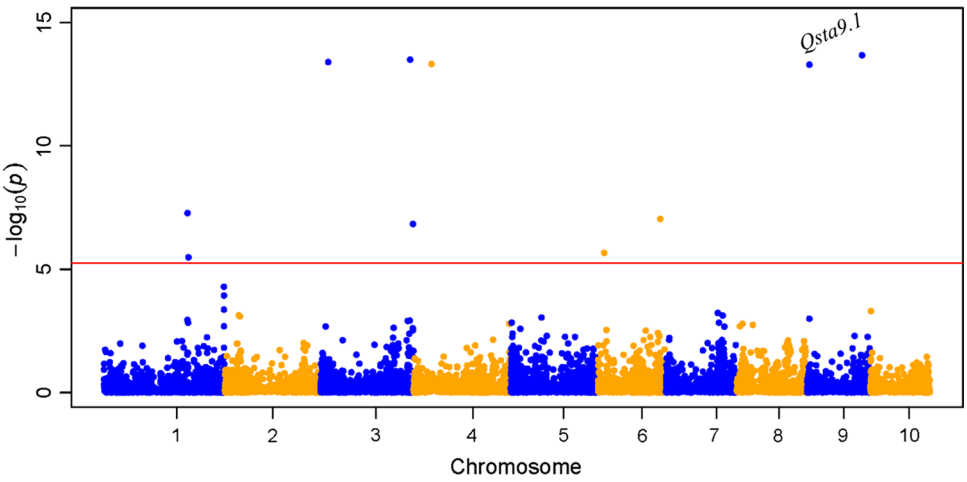

Fig. 3 Manhattan plots of the association of SNPs with starch content with MLM analysis. Genome position is shown along the *x*-axis divided by chromosome, and the $-\log_{10} P$ for association is shown on the *y*-axis

Supplementary Fig. S4). Another SNP assigned in *Qsta9.1* region did not reach the significant level in MLM analysis, but showed significant association in GLM analysis and located at 5,997,883 bp on chromosome 9 (Supplementary Fig. S4). This result confirmed the contribution of this chromosome region to starch content variance.

Annotation of the genes predicted in the major QTL *Qsta9.1* region

Within the 1.7 Mb region on chromosome 9 defined by the markers, 116 transcripts were predicted (Supplementary Table S2). Blast2GO analysis showed that they encode membrane-anchored ubiquitin-fold 3 isoform X2, peroxisome proliferator-activated receptor gamma coactivator 1-beta, TLD-domain-containing nucleolar, cycloartenol-C-24-methyltransferase, pentatricopeptide repeat-containing mitochondrial, etc. (Supplementary Table S2). The significant SNP identified in *Qsta9.1* interval with MLM in GWAS analysis

Table 3 SNPs significantly associated with starch content in GWAS analysis

SNP	Chromosome	Position	*P* value	
			GLM	MLM
S1_208321460	1	208,321,460	2.11E−12	5.40E−08
S1_210512907	1	210,512,907	2.94E−09	3.25E−06
S3_19783377	3	19,783,377	3.09E−36	3.97E−14
S3_223778724	3	223,778,724	3.86E−36	3.15E−14
S3_231079903	3	231,079,903	1.05E−14	1.49E−07
S4_44889892	4	44,889,892	2.34E−36	4.73E−14
S6_15445098	6	15,445,098	2.45E−10	2.15E−06
S6_155418930	6	155,418,930	9.37E−13	9.31E−08
S9_5877060	9	5,877,060	2.96E−36	5.09E−14
S9_137058114	9	137,058,114	3.49E−36	2.10E−14

was located in the ORF region of GRMZM5G852704_T01, which was annotated as ethylene-responsive transcription factor RAP2.24 (Supplementary Table S2).

Expression pattern of the genes in *Qsta9.1* region across seed development

By using the RNA-seq data of whole seeds at 12, 16, 20, 24, 30, and 36 days after pollination, expression patterns of those genes in the *Qsta9.1* region were detected. After removing the low-expressed transcripts with a maximum FPKM below one, 24 transcripts without GRMZM5G852704_T01, whose expression level was lower than one across all development stages, were left for further analysis. Based on two-way ANOVA, eight transcripts showed different expression patterns between B73 and Mo17 across seed developmental stages, including GRMZM2G003023_T01, GRMZM2G019291_T02, GRMZM2G019291_T07, GRMZM2G051330_T01, GRMZM2G082198_T03, GRMZM2G090226_T01, GRMZM2G103227_T01, and GRMZM2G110929_T01 (Fig. 4).

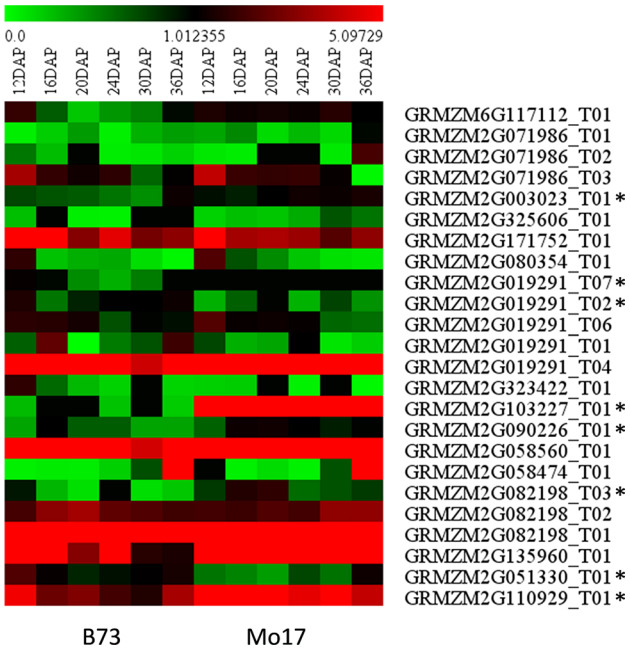

Fig. 4 Expression pattern of 24 transcripts predicted in *Qsta9.1* region compared between B73 and Mo17 across six seed developmental stages. Asterisk: transcripts showing significant different expression pattern between B73 and Mo17 according to two-way ANOVA

Identification of differentially expressed genes between B73 and Mo17 during seed development

To identify genes correlated with starch content, the transcriptome profiles of whole seed with six developmental stages of B73 and Mo17 were downloaded from NCBI, including the data of 12, 16, 20, 24, 30, and 36 days after pollination. Low-expressed transcripts with a maximum FPKM below one were removed as background noise. Two-way ANOVA was conducted with a custom-made function, and differentially expressed genes during seed development were filtered. Finally, 16,808 transcripts were selected for further analysis (Supplementary Table S3).

Construction of co-expression networks corresponding to seed developmental stages

Co-expression network was constructed based on WGCNA method, 28 modules were detected across the whole transcriptome, and 38–4133 transcripts were included in each module (Supplementary Table S3; Fig. 5).

In order to understand functional enrichment of these transcripts, predicted transcripts were searched against the *Zea mays spp* v5a on the AGRIGO server with SEA tool. Based on the $P \leq 0.01$ threshold, those transcripts in six modules got significant functional enrichment, including biological process (cellular localization, protein folding, etc.), cellular component (cytoplasm, intracellular, mitochondrion, etc.) and molecular function (nucleoside binding, transferase activity, etc.) (Supplementary Table S4).

However, enriched function carbohydrate biosynthetic process, cellular carbohydrate biosynthetic process, or cellular carbohydrate metabolic process was only detected in one module (pink) (Supplementary Table S4), suggesting that genes gathered in this module associated with carbohydrate-related processes.

Candidate gene for the major QTL *Qsta9.1*

Combined QTL mapping and co-expression network analysis, the eight differentially expressed transcripts in the *Qsta9.1* region were assigned into different modules. GRMZM2G019291_T02, GRMZM2G019291_T07, and GRMZM2G051330_T01 were put in the same module (yellowgreen), GRMZM2G110929_T01 was assigned in module pink, and the others were in module darkslateblue, palevioletred2, darkseagreen1, and navajowhite1, respectively (Supplementary Table S3). No significant functional enrichment was detected for module yellowgreen, palevioletred2, darkseagreen1 and navajowhite1. In module darkslateblue, only significant molecular functions were identified, including adenyl nucleotide binding, ATP binding, and purine nucleoside binding (Supplementary Table S4). However, significant functional enrichments were detected in module pink, including carbohydrate biosynthetic process, cellular carbohydrate biosynthetic process, and cellular carbohydrate metabolic process, suggesting important role of genes in this

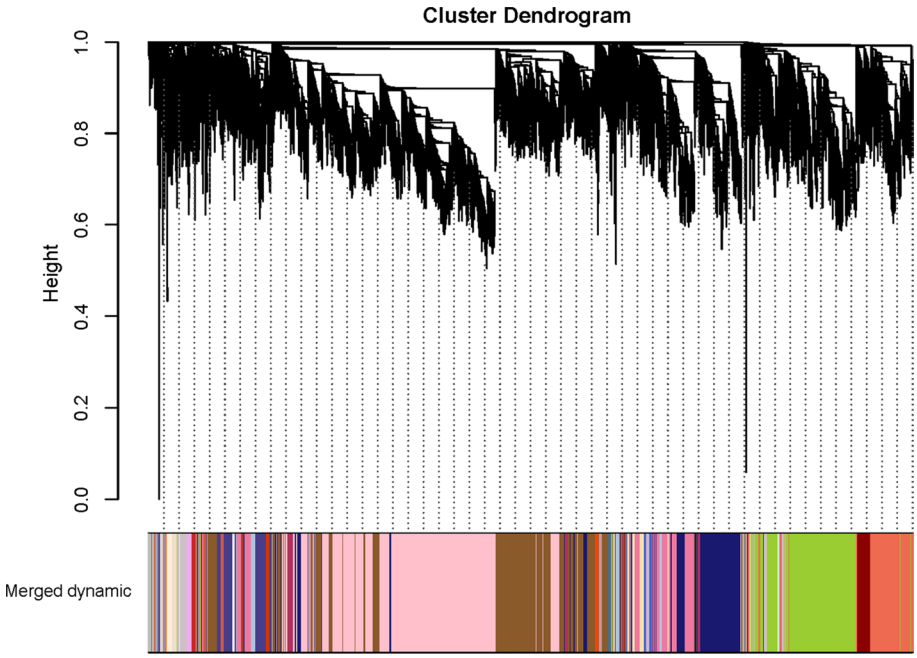

Fig. 5 Consensus modules detected through co-expression network analysis. Clustering dendrogram of genes, with dissimilarity based on topological overlap, together with assigned module colors

module for starch content (Supplementary Table S4). So GRMZM2G110929_T01 in module pink showed the strongest potential candidate for the major QTL *Qsta9.1*. It was located between 6,733,266 bp and 6,734,635 bp on chromosome 9 (B73 v3) and annotated as GLABRA2 expression modulator according to BLAST2GO.

Another transcript GRMZM5G852704_T01 could be the candidate of *Qsta9.1*, with low expression level though. Two InDels (insertion and deletion) exist between the protein sequences of GRMZM5G852704_T01 between B73 and Mo17 (Supplementary Fig. S5).

Discussion

In this paper, we combined QTL mapping, genome-wide association study and co-expression network analysis to identify candidate genes associated with starch content in maize kernel. A major QTL was identified on chromosome 9 with stable contribution to the variation of starch content. Together with co-expression network analysis, a transcript within mapped QTL interval was assigned in the module with functional enrichment associated with carbohydrate-related processes. The study reported herein we applied the classical genetics and current genomic technologies that together provide a general strategy for efficient detection of the causative genes responsible for the variation.

In previous studies, the QTLs for maize starch content have been mapped on all ten chromosomes (Clark et al. 2006; Cook et al. 2012; Dudley et al. 2007; Guo et al. 2013; Liu et al. 2016; Séne et al. 2000; Wang et al. 2010, 2015; Wassom et al. 2008; Yang et al. 2013; Zhang et al. 2008,2015). Several studies have found QTLs for starch content in maize on chromosome 9. By using an advanced recombinant inbred line population derived from Chinese line 178 and P53, two QTLs were identified in the interval of *bnlg430-dupssr19* and *umc1771-bnlg430* on chromosome 9, respectively (Zhang et al. 2015). In the CI7/K22 RIL population, a starch content QTL was found in the interval of *PZE10907827-PZE109082140* with the physical position of 126.2–130.8 Mb (Wang et al. 2015). Cook et al. (2012) detected two QTLs for starch content at 24,028,939 bp and 126,584,775 bp on chromosome 9. Through GWAS analysis, several SNPs on chromosome 9 showed significant association with amylose content (Li et al. 2018). QTLs on chromosome 9 were also reported by some other researches (Guo et al. 2013; Yang et al. 2013). Despite the declaration of location on chromosome 9 for these QTLs, their positions were not exactly the same and within large intervals. To map the loci for starch content more accurately, we used the high-density map constructed by IBM population in this study. Although this map contained more than 1300 markers, there were still some gaps left on each linkage, which require additional markers to fill in. IBM population was developed by crossing B73 and Mo17, followed by four cycles of random mating which increased the recombination rate dramatically. The main reason for gap existence is lacking adequate amount of molecular markers. Fortunately, sequencing the whole genome of B73 and re-sequencing Mo17 have produced enough sequence information for marker development. Coupled with the high recombination RIL population, we screened 1859 primer pairs InDel markers with unique position, and 265 were integrated into the IBM linkage map and filled some of the gaps. The saturated linkages

were used for QTL mapping, and the major QTL *Qsta9.1* was located into a 1.7 Mb region. Through another newly constructed $F_{2:3}$ population from B73 × Mo17, significant differences were detected for the *Qsta9.1* region between two extreme tails, validating association of the region with starch content. Also, genome-wide association study was performed and significant association was detected between SNPs in the *Qsta9.1* region and starch content variation. One hundred and sixteen transcripts were predicted in this interval, the coding proteins including fucosyltransferase, methyltransferase, kinase, transcription factor, and so forth (Supplementary Table S2). In traditional way, it is difficult to tell which one is the candidate among so many genes. But co-expression network has provided an alternative approach to do this work.

Gene co-expression network has been used as a more rapidly way in identification of gene modules and key genes responsible for particular conditions. Transcription factors associated with cell wall biosynthesis were revealed by co-expression network analysis in sugarcane (Ferreira et al. 2016). Novel regulation of potato pigmentation was revealed by network analysis of the metabolome and transcriptome (Cho et al. 2016). In maize, regulatory motifs were identified from developmental co-expression network (Downs et al. 2014). Based on transcript abundance data in maize, co-expression modules were classified association with tissue types and related biological processes (Downs et al. 2013). Except for identification of gene modules for a particular condition, comparative based co-expression approaches were also used for conserved modules across species as well as for differential modules (Du et al. 2017; Leal et al. 2014; Thirunavukkarasu et al. 2017). To identify candidate gene for starch content in maize kernel, differentially expressed genes between B73 and Mo17 during seed development were screened and co-expression network was constructed. Finally, 28 modules were detected across the whole transcriptome and functional analysis showed that module pink associate with carbohydrate-related process, suggesting that genes in this module play roles for starch content in kernel.

Together with the QTL mapping results, one transcript GRMZM2G110929_T01 located within the 1.7 Mb region of the major QTL *Qsta9.1* was assigned in module pink and annotated as the GLABRA2 expression modulator (GEM) encoding proteins contains GRAM (glucosyltransferases, Rab-like GTPase activators, and myotubularins) domain. In Arabidopsis, GEM is an ABA-responsive protein as part of the ABA signaling pathway (Mauri et al. 2016). One of the GEM family members GEM-RELATED5 (GER5) closely related to GLABRA2 expression modulator has been implicated in regulating seed development and inflorescence architecture based on its expression pattern, which was shown to be in part overlapping with that of GER1 and GEM in reproductive organs (Baron et al. 2014). Furthermore, mutant analysis showed transcripts changes in carbohydrate metabolism and catabolic processes (Baron et al. 2014). These results suggested the potential role of GRMZM2G110929 affecting starch content in maize kernel.

Based on GWAS analysis, one significant SNP was detected within the ORF region of GRMZM5G852704_T01 annotated as ethylene-responsive transcription factor RAP2.4. Although with low expression level, there were differences between the protein sequences of this gene in B73 and Mo17. In *Arabidopsis*, RAP2.4 was involved in multiple developmental processes regulated by light and ethylene, including hypocotyl elongation and gravitropism, apical hook formation and cotyledon expansion, flowering time, root elongation, root hair formation, and drought tolerance (Lin et al. 2008). Overexpression of the *Rap2.4f* transcriptional factor in *Arabidopsis* promotes leaf senescence (Xu et al. 2010). Another one RAP2.2 in the same family affected the induction of genes encoding sugar metabolism (Hinz et al. 2010). These results also made GRMZM5G852704_T01 one of the candidates of *Qsta9.1* controlling starch content in maize kernel.

Author contribution statement HZ conceived and designed the research program. FL did QTL mapping and analyzed the data and results. LZ carried out the association study and expression analysis. XZ conducted allele frequency analysis. BH performed the differential co-expression network analysis. HD, YQ, and LR collected phenotype data, and FL and HZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements This work was supported by Grant from National key research and development program of China (2017YFD0102005), Natural Science Foundation of Jiangsu Province, China (BK20160582), National Natural Science Foundation of China (31601315), Key Research and Development Program of Jiangsu Province (BE2017365), and Project of Science and Technology of Anhui Province (1604a0702021).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The experiments carried out in this study comply with the ethical standards in China.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Ashikari M, Matsuoka M (2006) Identification, isolation and pyramiding of quantitative trait loci for rice breeding. Trends Plant Sci 11:344–350

Baron KN, Schroeder DF, Stasolla C (2014) GEm-Related 5 (GER5), an ABA and stress-responsive GRAM domain protein regulating seed development and inflorescence architecture. Plant Sci 223:153–166

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300

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

Cameron JW, Teas HJ (1954) Carbohydrate relationships in developing and mature endosperms of brittle and related maize genotypes. Am J Bot 41:50–55

Cho K, Cho KS, Sohn HB, Ha IJ, Hong SY, Lee H, Kim YM, Nam MH (2016) Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. J Exp Bot 67:1519–1533

Choe E, Drnevich J, Williams MM (2016) Identification of crowding stress tolerance co-expression networks involved in sweet corn yield. PLoS ONE 11:e0147418

Chourey PS (1981) Genetic control of sucrose synthetase in maize endosperm. Mol Gen Genet 184:372–376

Chourey PS, Nelson OE (1976) The enzymatic deficiency conditioned by the shrunken-1 mutations in maize. Biochem Genet 14:1041–1055

Clark D, Dudley JW, Rocheford TR, LeDeaux JR (2006) Genetic analysis of corn kernel chemical composition in the random mated 10 generation of the cross of generations 70 of IHO × ILO. Crop Sci 46:807–819

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of illumina second-generation sequencing data. BMC Bioinform 11:485

Das S, Meher PK, Rai A, Bhar LM, Mandal BN (2017) Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: an application to aluminum stress in Soybean (*Glycine max* L.). PLoS ONE 12:e0169605

Dickinson DB, Preiss J (1969) Presence of ADP-glucose pyrophosphorylase in shrunken-2 and brittle-2 mutants of maize endosperm. Plant Physiol 44:1058–1062

DiLeo MV, Strahan GD, den Bakker M, Hoekenga OA (2011) Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS ONE 6:e26683

Downs GS, Bi YM, Colasanti J, Wu W, Chen X, Zhu T, Rothstein SJ, Lukens LN (2013) A developmental transcriptional network for maize defines coexpression modules. Plant Physiol 161:1830–1843

Downs GS, Liseron-Monfils C, Lukens LN (2014) Regulatory motifs identified from a maize developmental coexpression network. Genome 57:181–184

Du J, Wang S, He C, Zhou B, Ruan YL, Shou H (2017) Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. J Exp Bot 68:1955–1972

Dudley JW, Clark D, Rocheford TR, LeDeaux JR (2007) Genetic analysis of corn kernel chemical composition in the random mated 7 generation of the cross of generations 70 of IHP × ILP. Crop Sci 47:45–57

Ferreira SS, Hotta CT, Poelking VG, Leite DC, Buckeridge MS, Loureiro ME, Barbosa MH, Carneiro MS, Souza GM (2016) Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. Plant Mol Biol 91:15–35

Guo Y, Yang X, Chander S, Yan J, Zhang J, Song T, Li J (2013) Identification of unconditional and conditional QTL for oil, protein and starch content in maize. Crop J 1:34–42

Hennen-Bierwagen TA, Myers AM (2013) Genomic specification of starch biosynthesis in maize endosperm. In: Becraft PW (ed) Seed Genomics, pp 123–137

Hinz M, Wilson IW, Yang J, Buerstenbinder K, Llewellyn D, Dennis ES, Sauter M, Dolferus R (2010) *Arabidopsis RAP2.2*: an ethylene response transcription factor that is important for hypoxia survival. Plant Physiol 153:757–772

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, Tikunov Y, Bovy A, Chikate Y, Singh P, Rogachev I, Beekwilder J, Giri AP, Aharoni A (2013) Biosynthesis of antinutritional alkaloids in Solanaceous crops is mediated by clustered genes. Science 341:175–179

Jeon JS, Ryoo N, Hahn TR, Walia H, Nakamura Y (2010) Starch biosynthesis in cereal endosperm. Plant Physiol Biochem 48:383–392

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9:559

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:1

Laughnan JR (1953) The effect of the *sh2* factor on carbohydrate reserves in the mature endosperm of maize. Genetics 38:485–499

Leal LG, López C, López-Kleine L (2014) Construction and comparison of gene co-expression networks shows complex plant immune responses. PeerJ 2:e610

Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245

Li C, Huang Y, Huang R, Wu Y, Wang W (2018) The genetic architecture of amylose biosynthesis in maize kernel. Plant Biotechnol J 16:688–695

Lin RC, Park HJ, Wang HY (2008) Role of *Arabidopsis* RAP2.4 in regulating light- and ethylene-mediated developmental processes and drought stress tolerance. Mol Plant 1:42–57

Liseron-Monfils C, Ware D (2015) Revealing gene regulation and associations through biological networks. Curr Plant Biol 3–4:30–39

Liu N, Xue Y, Guo Z, Li W, Tang J (2016) Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. Front Plant Sci 7:1046

Lv Y, Liu Y, Zhao H (2016) mInDel: a high-throughput and efficient pipeline for genome-wide InDel marker development. BMC Genom 17:290

Mauri N, Fernández-Marcos M, Costas C, Desvoyes B, Pichel A, Caro E, Gutierrez C (2016) GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. Sci Rep 6:22660

Meng L, Li H, Zhang L, Wang J (2015) QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J 3:269–283

Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem 31:426–428

Obertello M, Shrivastava S, Katari MS, Coruzzi GM (2015) Cross-species network analysis uncovers conserved nitrogen-regulated network modules in rice. Plant Physiol 168:1830–1843

Palumbo MC, Zenoni S, Fasoli M, Massonnet M, Farina L, Castiglione F, Pezzotti M, Paci P (2014) Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative

switch genes that induce major transcriptome reprogramming during grapevine development. Plant Cell 26:4617–4635

Petricka JJ, Schauer MA, Megraw M, Breakfield NW, Thompson JW, Georgiev S, Soderblom EJ, Ohler U, Moseley MA, Grossniklaus U, Benfey PN (2012) The protein expression landscape of the *Arabidopsis* root. Proc Natl Acad Sci 109:6811–6818

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Séne M, Causse M, Damerval C, Thévenot C, Prioul JL (2000) Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines. Plant Physiol Biochem 38:459–472

Shaik R, Ramakrishna W (2013) Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. PLoS ONE 8:e77261

Thirunavukkarasu N, Sharma R, Singh N, Shiriga K, Mohan S, Mittal S, Mittal S, Mallikarjuna MG, Rao AR, Dash PK, Hossain F, Gupta HS (2017) Genomewide expression and functional interactions of genes under drought stress in maize. Int J Genomics 2017:2568706

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Tsai CY, Nelson OE (1966) Starch-deficient maize mutant lacking adenosine diphosphate glucose pyrophosphorylase activity. Science 151:341–343

Wang YZ, Li JZ, Li YL, Wei MG, Li XH, Fu JF (2010) QTL detection for grain oil and starch content and their associations in two connected $F_{2:3}$ populations in high-oil maize. Euphytica 174:239–252

Wang T, Wang M, Hu S, Xiao Y, Tong H, Pan Q, Xue J, Yan J, Li J, Yang X (2015) Genetic basis of maize kernel starch content revealed by high-density single nucleotide polymorphism markers in a recombinant inbred line population. BMC Plant Biol 15:288

Wassom JJ, Wong JC, Martinez E, King JJ, DeBaene J, Hotchkiss JR, Mikkilineni V, Bohn MO, Rocheford TR (2008) QTL associated with maize kernel oil, protein, and starch concentrations; kernel mass; and grain yield in Illinois high Oil × B73 backcross-derived lines. Crop Sci 48:243–252

Xu H, Wang X, Chen J (2010) Overexpression of the *Rap2.4f* transcriptional factor in *Arabidopsis* promotes leaf senescence. Sci China Life Sci 53:1221–1226

Yang G, Dong Y, Li Y, Wang Q, Shi Q, Zhou Q (2013) Verification of QTL for grain starch content and its genetic correlation with oil content using two connected RIL populations in high-oil maize. PLoS ONE 8:e53770

Zhang J, Lu X, Song X, Yan J, Song T, Dai J, Rocheford T, Li J (2008) Mapping quantitative trait loci for oil, starch, and protein concentrations in grain with high-oil maize by SSR markers. Euphytica 162:335–344

Zhang H, Jin T, Huang Y, Chen J, Zhu L, Zhao Y, Guo J (2015) Identification of quantitative trait loci underlying the protein, oil and starch contents of maize in multiple environments. Euphytica 205:169–183

Zinkgraf M, Liu L, Groover A, Filkov V (2017) Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in *Populus* under diverse environmental conditions. New Phytol 214:1464–1478