Theor Appl Genet
DOI 10.1007/s00122-016-2716-0

ORIGINAL ARTICLE

Genome-wide association studies of drought-related metabolic changes in maize using an enlarged SNP panel

Xuehai Zhang[1] · Marilyn L. Warburton[2] · Tim Setter[3] · Haijun Liu[1] · Yadong Xue[4] · Ning Yang[1] · Jianbing Yan[1] · Yingjie Xiao[1]

Received: 20 September 2015 / Accepted: 15 April 2016

Abstract

Key message **Genetic determinants of metabolites related to drought tolerance in maize.**

Abstract Water deficit or drought is one of the most serious abiotic stresses of plant development and greatly reduces crop production, and the plant's response to this deficit leads to many metabolic changes. To dissect the genetic basis of these metabolic traits in maize, we performed a genome-wide association analysis of drought-related traits using 156,599 SNPs in 318 maize inbred lines. In total, 123 significant SNP/trait associations ($P \leq 6.39E{-}6$) involving 63 loci were identified for related metabolic and physiological traits in multiple tissues and different environments under two irrigation conditions. Of the 63, 23 loci demonstrated a significant interaction effect between QTL and water status, indicating that these metabolite-associated loci were probably related to drought stress tolerance. To evaluate the potential utility of metabolite-associated loci applied in hybrid maize breeding, we assembled two groups of hybrid entries with high or low drought tolerance and measured the metabolic and physiological traits. In the hybrid pools, a set of 10 metabolite-associated loci identified in leaf and ear were validated as responsive to drought stress. The favorable alleles of these ten loci were significantly enriched in hybrids with high drought tolerance, which jointly explained almost 18.4 % of the variation in drought tolerance using a multivariate logistic regression model. These results provide clues to understanding the genetic basis of metabolic and physiological changes related to drought tolerance, potentially facilitating the genetic improvement of varieties with high drought tolerance in maize breeding programs.

Communicated by C. Carolin Schön.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-016-2716-0) contains supplementary material, which is available to authorized users.

✉ Yingjie Xiao
shanren0179@163.com

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

[2] USDA-ARS Corn Host Plant Research Resistance Unit, Mississippi State University, Box 9555, Starkville, MS 39762, USA

[3] Section of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

[4] Agronomy College, Henan Agricultural University, Zhengzhou 450002, China

Introduction

Maize (*Zea mays* L.) is one of the most important crops worldwide for food, animal feed and biofuel, and displays the highest global grain production (Haley 2011). Drought (i.e., water deficit) is one of the most serious abiotic stresses of plant development and greatly reduces crop production. In maize, drought stress during flowering causes an asynchrony between silk emergence and pollen shedding, which increases the anthesis-silking interval (ASI) and leads to significant yield losses (Lu et al. 2010; Ribaut et al. 1996).

Drought tolerance (DT) is a complex trait which refers to the capacity of the plant to be more productive under drought stress (Ribaut et al. 2009). Improvement of drought tolerance of released cultivars is critically important to sustain grain yield in the face of continually degrading global environments. The clear exploration of the genetic basis of

Published online: 27 April 2016

drought tolerance will speed genetic improvement of maize varieties for drought tolerance; however, it has been very difficult to identify any genes with a measurable effect on drought tolerance in maize. There are several interacting metabolic traits that vary in response to water stress (WS), and this may contribute to the sensitivity of this trait to even small environmental or climate changes (Paupière et al. 2014; Tuteja and Gill 2013).

In WS, drought sustained during flowering and early kernel development leads to many detrimental changes in plant physiology and development, including delayed days to silk emergence and increased anthesis-silking interval (Xue et al. 2013), stomatal closure, plant wilting and leaf rolling, and premature senescence of leaves and hence, a reduction in the leaf area index and a fall in photosynthetic activity, all of which cause yield reduction (Ribaut et al. 2009). The phytohormone abscisic acid (ABA) is an important plant metabolite that plays an essential role in drought response, and increased production of ABA would help the plants to survive under drought stress (Seki et al. 2007). This includes regulation of stomatal aperture and transcription levels of a large number of genes related to plant stress response (Pinheiro and Chaves 2011). Primary metabolites, such as sugars and amino acids, have also been reported to accumulate during WS (Bartels and Sunkar 2005). Sugar and amino acid levels are significantly correlated with drought tolerance (Lebreton et al. 1995; Mohammadkhani and Heidari 2008), so they could potentially be used as drought tolerance component traits to assist in the selection of elite inbred lines with drought tolerance in maize breeding.

Due to a complex genetic architecture, the study of drought tolerance in maize is difficult, and a good alternative is to study metabolic traits that vary with drought levels to unravel the genetic basis of this tolerance. In plant species, genome-wide association study (GWAS) has emerged as a powerful tool to identify QTLs and natural variation for many agriculturally important traits (Atwell et al. 2010; Buckler et al. 2009; Huang et al. 2010). With the advent of high-efficiency metabolic profiling and next-generation sequencing technologies, there have been many reports of GWAS that dissect the genetic basis of primary metabolites in multiple species (Chen et al. 2014; Luo 2015; Wen et al. 2014, 2015). This is mainly because the plant metabolome is the readout of plant physiological status and often acts as the intermediary connecting the visible agronomic phenome and the underlying genome, thus allowing identification and, ultimately, selection of causative genetic loci that are less influenced by the environment (Chan et al. 2010; Keurentjes et al. 2006; Meyer et al. 2007; Riedelsheimer et al. 2012).

Nevertheless, little is known about metabolic response to drought. In maize, Setter et al. (2011) employed the Illumina GoldenGate SNP array with 1229 informative SNPs to perform GWAS of metabolite levels under WS, and identified several significant loci and candidate genes. However, marker density in the study by Setter et al. (2011) was insufficient for total genomic coverage, and the genetic basis of varying metabolic and physiological traits under water deficit in maize remains unclear. The extent of linkage disequilibrium in specific populations generally determines the number of markers required to saturate the whole genome, and simulation and empirical studies suggest that hundreds of thousands of markers are needed in maize to have the statistical power to identify the majority of QTLs associated with complex traits via GWAS (Li et al. 2013; Yan et al. 2011; Yang et al. 2014). It is thus likely that reanalysis of the previous study by Setter et al. (2011) with an enlarged panel of markers will allow further dissection of the genetic basis of metabolic and physiological traits in response to drought.

To this aim, we made use of the metabolic and physiological traits of the 318 diverse maize inbred lines used by Setter et al. (2011) and enlarged the 1229 SNPs to 156,599 SNPs by integrating three genotype datasets generated from the GoldenGate array, the MaizeSNP50 array and RNA-sequencing, using efficient imputation methods. To evaluate the practical utility of results of this expanded GWAS, we assembled two groups of maize hybrids with high or low drought tolerance and measured the metabolic and physiological traits using the same procedures as in the GWAS study. The objectives of this study were: (i) to assess the feasibility of enhancing SNP coverage via imputation methods; (ii) to further dissect the genetic basis of metabolic and physiological response to drought in maize; (iii) to evaluate the potential use of significantly associated loci in an applied hybrid breeding program to increase drought tolerance.

Materials and methods

Plant materials and field trials

An association mapping panel of 318 diverse maize inbred lines with tropical and subtropical pedigrees was employed to perform GWAS in this study (Table S1). All materials were planted at CIMMYT's experimental station in Tlaltizapan, Mexico following an alpha lattice design in 2005 and 2006, termed TL05A and TL06A, respectively. Field trials were severe drought stress (WS) in TL05A and planted under contrasting irrigated conditions including fully irrigated (well watered, or WW) and WS in TL06A. The WW trials were irrigated every 2 weeks by furrow irrigation. The WS trials were also irrigated every 2 weeks until 20 days before anthesis, when water was withdrawn;

detailed information for the field management and trials was described previously (Setter et al. 2011).

Also, the set of 318 inbred lines of the association mapping panel were crossed with a common tester, CML312, an inbred line with a good general combining ability and poor drought tolerance. In the 2006–2007 season, the testcross population was planted in WS and WW conditions in Kenya (KARI, Kiboko), Thailand (Takfa, Nakhon Sawan), and Mexico (CIMMYT, Tlaltizapan) in an alpha lattice design with two replications and 5-m rows per plot, to measure agronomic traits (Xue et al. 2013). In this study, grain yield and ASI measured in the hybrid population in each of the three environments in the WS condition were used to obtain global rank indices for evaluating the drought tolerance of the hybrids.

For each environment, the rank index for each hybrid was calculated as:

$$\text{rank-index} = \left(Z\text{-score}_{\text{yield}} - Z\text{-score}_{\text{ASI}}\right).$$

where the *Z*-score means the normalized data of each hybrid for yield and ASI, respectively, in each environment. Given that a higher ASI is negatively associated with drought performance, this subtraction of *Z*-score gave approximately equal weight to yield and ASI. *Z*-scores for yield and ASI were calculated by the formula:

$$Z\text{-score} = (\text{hybrid} - \text{mean})/\text{SD},$$

where hybrid was the yield or ASI value of a particular hybrid, and the mean and standard deviation (SD) were calculated across the hybrid population for each environment trial. To obtain a global ranking of each hybrid, the rank indices for the three environments were averaged (Table S2). According to the distribution of the global rank indices and the drought tolerance in the field, two tails of this distribution with 71 highly tolerant and 68 intolerant hybrids were selected to form two contrasting groups, referred to hereafter as the hybrid pools (Table S2). Experimental flow of this study was shown in Figure S1.

Analysis of metabolic and physiological traits

In this study, analysis of metabolic and physiological traits was conducted on both the association panel and two hybrid pools. For the association panel, metabolic traits including abscisic acid (ABA), abscisic acid glucose ester (ABA-GE), phaseic acid (Pa), proline (Pro), sucrose (Suc), glucose (Glc), total sugars (Tsug), mole fraction of sucrose per total sugar (Fsuc), mole fraction of glucose per total sugar (%glc), starch (Str), and physiological traits including dry mass (Dw) and specific leaf weight (Slw), were previously measured in multiple tissues under WS and WW conditions at 0 and 7 days after anthesis in two consecutive years (Setter et al. 2011). These same phenotypic data were directly used in the current analysis. For the hybrid pools, two tissues (i.e., ear and leaf) were sampled for analysis of metabolic and physiological traits on plants grown under WS and WW conditions at 7 days after anthesis in 2008. Ear tips were sampled in Thailand and Mexico, while leaf disks were collected in Kenya, Thailand and Mexico. The same metabolic and physiological traits described above were measured with the same protocol documented previously (Setter et al. 2011).

For the current analysis, we treated the combination of each tissue [E (ear), L (leaf) and S (silk)], metabolic and physiological trait, sample time (0 and 7 days), irrigated condition (WW and WS), and year (2005 and 2006) as individual variables to facilitate GWAS analysis. For example, E.Str.0_WW_06 indicates the variable starch content measured in the ear sampled at 0 days after anthesis under WW condition in 2006. In total, we obtained 168 variables for all twelve metabolic and physiological traits in the association panel of 318 inbred lines and 85 variables in the hybrid pools. The abbreviated name and description of all traits measured are listed in Table S3.

Genotyping and imputation

The association panel's 318 inbred lines were previously genotyped by two SNP chips: the Illumina GoldenGate Assay with 1536 SNPs chosen from drought-related genes (Setter et al. 2011) and the Illumina MaizeSNP50 BeadChip with 56,110 random SNPs as described previously (Xue et al. 2013). A very large number of imputed SNPs were added to these directly scored SNPs as described below to generate enough SNP calls to completely cover the maize genome with a sufficiently high density for GWAS. To do this, we used the SNP calls from another panel of 368 diverse maize inbred lines, which had been genotyped by the Illumina MaizeSNP50 BeadChip (Li et al. 2012b) and with 556,809 high-quality SNPs (MAF $\geq$ 0.05) by RNA-sequencing (Fu et al. 2013; Li et al. 2013). In the present study, we imputed the high-density marker genotypes for 318 inbred lines using the data from the 368 inbred lines, following a two-step approach using the identity by descent (IBD)-based projection and k-nearest neighbor (KNN) algorithm reported previously (Yang et al. 2014).

According to the physical position of each SNP on the B73 reference sequence (RefGen_v2), we merged the three marker datasets (i.e., 1536 SNPs from the GoldenGate array, 56,110 SNPs from the MaizeSNP50 array and 556,809 SNPs from RNA-seq) into an integrated marker dataset with a total of 558,629 SNPs. There were 39 inbred lines in common between the 368 and 318 line panels, and 329 inbred lines with RNA-seq data were regarded as the reference panel to impute the genotyped panel of 318 inbred lines, which provided an access to evaluate

imputation accuracy in the current study. For each of the 318 lines (those with no RNA-seq data), including the repeated 39 lines, the genotype calls from 42,742 SNPs which were shared between the two panels were used as core markers and projected onto physical maps of the maize genome. IBD regions were delimited on these maps, and within each IBD region, the genotypes for each of the 318 lines were assigned as the alleles from the RNA-seq data of the best matched line of the 329 line panel. Imputed values were then inserted for each SNP. The imputed values of the 39 inbred lines were compared with the observed values to calculate imputation accuracy; but only the observed values for these 39 lines were used in the GWAS analysis. Imputation resulted in a total of 156,599 high-density SNPs (MAF $\geq$ 0.05) available in the 318 inbred lines for GWAS.

GWAS

GWAS of the 168 variables in the 318 maize inbred lines was performed using a compressed mixed linear model (cMLM) (Zhang et al. 2010) implemented in the software package GAPIT (Lipka et al. 2012). In GAPIT, the top six principal components and a kinship matrix were used to correct for population structure and family relatedness among the 318 inbred lines in GWAS, which were calculated automatically with 44,314 SNPs from the 50 K array by setting the parameters "pca.total" as 6 and "kinship.algorithm" as "Loiselle" (Loiselle et al. 1995). Male flowering time (which is the only available flowering time-related trait in the association panel) was also used as a fixed effect (covariate) in the mixed model, since flowering time strongly correlates with the performance of plants under drought stress.

The general decay distance of linkage disequilibrium (LD) in the association panel (~100 kb, Xue et al. 2013) was used to compare the GWAS results based on the two marker datasets at the family-wise error rate of 0.05. As the number of markers in the two datasets differed greatly, the effective number (N_e) of independent tests for the two datasets was calculated (Li et al. 2012a) and the value was used to determine the global p cutoff using the Bonferroni method for correcting multiple tests ($P \leq \alpha/N_e$) (Dunn 1959, 1961). To facilitate the interpretation of GWAS results, the adjusted Bonferroni method (i.e., $P \leq 1/N$, where N is total number of genome-wide SNPs in the analysis) was used as the genome-wide p cutoff to declare the significance of SNP-trait associations, which is currently widely used in plant GWAS studies (Li et al. 2013; Huang et al. 2010). Furthermore, we evaluated the extent of local LD for each significant SNP. The extended region where the LD between nearby SNPs and the peak SNP (that with the lowest P value) decayed to $r^2 = 0.2$ was defined as the local LD-based QTL interval. For each variable, all significant SNPs with overlapping QTL intervals were categorized as an associated locus. For each associated locus, the p value and QTL intervals of the peak SNP defined the significance and the interval of the locus; the variance explained by the locus was estimated with in a general linear model (GLM) that fits the peak SNP in the "lm" function of the R software package, and which was calculated by comparing the residual sum of squares between the full model and the reduced model excluding the peak SNP. These significant QTL intervals were used to identify overlaps between the two water regimes and search for candidate genes. To test whether the metabolite-associated loci responded to drought stress, we performed a two-way ANOVA to test the significance of interaction effect between locus and water status (WW and WS) for each significant locus. The set of loci showing significant interactions with water status were defined as QTL responsive to drought stress (drought-response loci).

Performance of metabolic associations in the contrasting hybrid pools

To evaluate the potential utility of studying metabolic and physiological traits to hybrid maize breeding for drought tolerance, a t test was used to evaluate the difference of trait means between the hybrid pools. For these traits, the drought-response loci identified in GWAS were studied in an attempt to validate their effects on drought tolerance in the hybrid pools across the three environments. For each drought-response locus, the markers within the QTL interval were tested using a logistic regression model or a linear regression model that simultaneously fits the first principle component and MFLW in hybrid pools as covariates. Each drought-response locus was validated if two or more markers within the QTL interval had a significant effect on drought tolerance in the hybrid pools ($P < 0.05$). To test whether there was an enrichment of validated loci in the hybrid pools, the observed proportion of validated SNPs for the drought-response loci was compared with the expected proportion, which was the proportion of random SNPs across the whole genome showing significant effect on drought tolerance in hybrid pools ($P < 0.05$). The significance between the observed and expected proportions was estimated using the R function 'binom.test'. The proportion of the variance of drought tolerance captured by multiple loci was defined as the goodness of fit value that was estimated in a multivariate logistic regression model. The favorable allele of each significant locus was defined as the allele that was enriched in the high drought tolerance pool compared to the low drought tolerance pool, and the number of alleles in each hybrid entry was estimated. All statistical analyses were carried out using the software R package (R Core Team 2012).

Results

Imputation for accurately increasing SNP density

In this study, we integrated two actual genotypic datasets (i.e., the GoldenGate and MaizeSNP50 array) collected on a GWAS panel of 318 lines with RNA-seq data from a different (but partially overlapping) population using the two-step imputation methods based on IBD and KNN algorithms to increase the number of SNPs available for GWAS from 1229 to 558,629. It is critically important to assess the accuracy of the imputed SNPs before using them in an analysis. To this aim, a principal component analysis (PCA) was performed to illustrate the relationship between the two panels (i.e., 318 and 368 lines) using the common SNP data of the 50 K SNP array. In Fig. 1a, it can be seen that the range of variation spanned by the 318 lines is almost completely covered by the range of the panel of 368 lines. This indicates that the majority of the haplotypes present in the panel of 318 lines should also exist in the panel of 368 lines, allowing imputation of missing haplotypes in the 318 lines from the SNP information from the 368 line panel.

Also, we compared the true RNA-seq genotype with the imputed genotype for the 558,629 SNPs in a subset of 39 inbred lines that appeared in common between the two panels. The imputation accuracy was robustly high for each inbred line, ranging from 95.84 to 99.98 % for the whole marker set and from 97.17 to 99.99 % for the subset of markers with MAF ≥ 0.05 (Fig. 1b and Figure S2A). Moreover, the 39 lines shared between the two genetic panels were largely representative of the entire range of genetic variation of the 318 lines (Fig. 1c); thus, the high imputation accuracy of the 39 lines is probably true for the panel of 318 lines as well. Imputation largely reduced the proportion of missing genotypes for each line in the 318 line panel, i.e., from 0.93 to 0.44 for the whole marker set and from 0.72 to 0.13 for the subset of markers with MAF ≥ 0.05 (Fig. 1d and Figure S2B), thus greatly increasing the available data for further statistical analysis. The results suggest that imputation can greatly increase marker coverage not only in the present study but also in similar situations, where similar genetic backgrounds and overlapping lines between two different panels allow such imputation and confirmation of accuracy (Fig. 1a, c).

GWAS with metabolic and physiological traits

A GWAS analysis of the 168 phenotypic variables for the metabolic and physiological traits was run again using the previously published GoldenGate SNP array (1229 SNPs with MAF ≥ 0.05 from Setter et al. 2011; hereafter called Dataset 1) and the new integrated SNP set (156,599 SNPs with MAF ≥ 0.05; hereafter called Dataset 2) with the compressed mixed linear model controlling the same set of population structure, male flowering time, and kinship. Because the mean LD decay across all chromosomes for this panel was ~100 kb (Xue et al. 2013), for simplicity, we defined 100 kb up- and downstream of the peak SNP

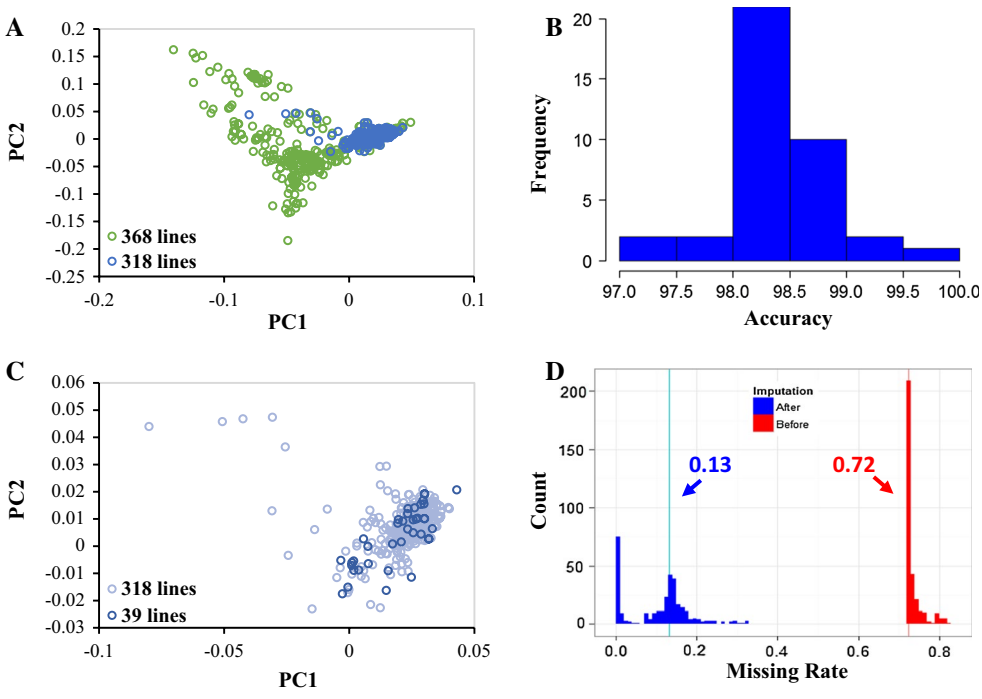
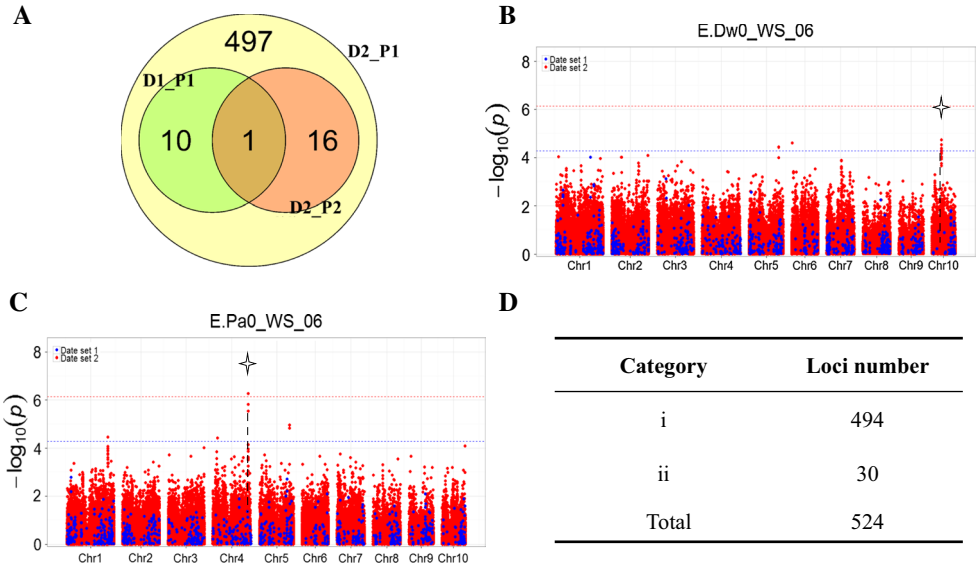


Fig. 1 Quality statistics of imputed genotypes for filtered SNPs (MAF ≥ 0.05). **a** The top two axes of variation for 368 inbred lines (*green*) and 318 inbred lines (*dark blue*); **b** Imputed accuracy, estimated using the 39 shared inbred lines for filtered SNPs (156599 SNPs; MAF ≥ 0.05); **c** PCA plot for 318 inbred lines (*Light blue*) and 39 shared inbred lines (*dark blue*); **d** Missing data rate. The *colored bars* and *numbers* represent the median missing data rate of 318 inbred lines before (*red*) and after (*blue*) imputation for filtered SNPs (156599 SNPs; MAF ≥ 0.05), respectively



Category	Loci number
i	494
ii	30
Total	524

Fig. 2 GWAS discoveries for marker sets of differing densities. **a** Venn diagram of significant loci with the two marker datasets and thresholds. D1 (Dataset 1) contained 1229 SNPs, D2 (Dataset 2) contained 156,599 SNPs; P1 and P2 represent the two significance thresholds for $P \leq 0.05/N_e$, respectively (N_e is 952 for D1 and 67607 for D2). **b** Manhattan plot of E.Dw0_WS_06. The *asterisk* and *vertical dashed line* indicate one case of a significant locus that was identified only in Dataset 2 because there were no markers available within this locus in Dataset 1. **c** Manhattan plot of E.Pa0_WS_06. The *asterisks* and *vertical dashed lines* indicate one case of a significant locus identified only in Dataset 2 because the availale markers within this locus in Dataset 1 were insufficient to reach the significance threshold. **d** A classification list of newly identified loci in Dataset 2

as a locus for testing whether the marker density influences GWAS results. At a family-wise error rate of 0.05, 11 significant loci were detected ($P \leq 0.05/N_e$, $N_e = 952$) for the 168 variables in Dataset 1, whereas 17 significant loci were identified in Dataset 2 ($P \leq 0.05/N_e$, $N_e = 67607$). Only one locus overlapped between the two Datasets. Additionally, when a less stringent p cutoff ($P \leq 0.05/952$, same as used in Dataset 1) was used in Dataset 2, a total of 524 significant loci were identified including all the 11 significant loci detected in Dataset 1 (Fig. 2a). The larger number of associated loci in Dataset 2 rather than Dataset 1 was due to the loci for which no markers (or insufficient numbers of markers) exist in Dataset 1 (i.e., 494 loci with no markers; Fig. 2b, d and 30 loci with insufficient markers; Fig. 2c, d).

To explore the genetic basis of metabolic and physiological trait variation under different water regimes, the extent of local LD was evaluated for each significant SNP identified by GWAS ($P \leq 1/156599$; Figure S3). For the 168 phenotypic variables, the 123 significant SNPs were categorized into 63 significant loci based on the local LD (Figure S4; Table 1). For each significant locus, the QTL interval ranged from 0.4 to 36 Mb with an average of 4.5 Mb (Table 1). The majority of significant loci was specifically detected in one water regime (i.e., 22 loci only in WW and 40 loci only in WS, and only one locus both in WW and WS, Table S4, Fig. 3). The percentage of phenotypic variation (R^2) that each locus could explain ranged from 3.10 to 16.14 %, with a mean of 8.34 %; 14 loci were identified that explained greater than 10 % of the variation (R^2 = 10.04–16.14 %). For ear tissue, GWAS detected a total of 23 significant loci, 8 specific to WW, in which R^2 ranged from 7 to 16.14 % with a mean of 10.17 %, and 15 specific to WS, in which R^2 ranged from 3.10 to 12.5 % with a mean of 8.14 %. For leaf tissue, GWAS detected a total of 23 loci, 6 specific to WW with an R^2 range of 4.12 to 12.57 % and a mean of 9.01 %, and 17 specific to WS and an R^2 range of 4.14–10.43 % with a mean of 7.98 %. For silk tissue, GWAS detected a total of 17 significant loci, 7 loci specific to WW with an R^2 range of 4.34–9.47 % and a mean of 7.15 %, and 9 specific to WS, with an R^2 range of 4.52–10.40 % and a mean of 7.58 %. One locus, located on chromosome 6, was significantly associated with phaseic acid concentrations in both WW and WS and explained 8.10 and 8.60 % of the phenotypic variation, respectively. Additionally, the two-way ANOVA revealed that 23 of the total 63 loci showed a significant interaction between locus and water regime (WW & WS) for the three tissues (Table 1), indicating that these QTLs of metabolites were probably to respond to drought stress.

Metabolic and physiological variation in hybrid pools

In the present study, we measured a total of 85 phenotypic variables in the hybrid pools of 139 entries (71 highly

Table 1 Summary of the metabolite-associated loci and their interactions with water regimes

Loci	Variable[a]	Peak SNP[b]	Chr.[c]	Pos (bp).[d]	QTL interval (Mb)[e]	Allele[f]	MAF[g]	GWAS[h]	Var (%)[i]	GxE[j]	Candidate gene[k]	Annotation[l]
1	L.Suc0_WS_06	chr1.S_519310	1	519310	0–1	T/C	0.316	4.38E−07	10.06	3.04E−01	GRMZM2G471814	Unknown
1	L.Tsug0_WS_06	chr1.S_519310	1	519310	0–1	T/C	0.316	3.77E−07	10.04	3.77E−01	GRMZM2G471814	Unknown
2	L.Pa0_WS_05	SYNGENTA10955	1	15170841	14–16	A/G	0.079	5.20E−06	8.76	6.49E−01	GRMZM2G035156	bHLH family protein
3	S.Tsug7_WW_06	chr1.S_23216815	1	23216815	22–24	A/G	0.112	2.32E−06	5.16	1.28E−01	GRMZM2G005435	Ethylene-responsive element-binding factor
4	E.Glc7_WS_05	chr1.S_26434284	1	26434284	25–28	G/A	0.248	4.94E−06	8.32	9.99E−01	GRMZM2G472693	Histone-lysine N-methyltrans-ferases
5	L.Glc0_WS_05	chr1.S_30259433	1	30259433	29–35	C/A	0.175	4.91E−07	9.82	2.50E−01	GRMZM2G043584	Leucine-rich receptor-like protein kinase family protein
6	E.Pa7_WS_05	PZE-101069763	1	52636504	52–54	A/C	0.089	1.38E−06	8.65	8.04E−01	GRMZM2G033570	Ethylene insensitive 3 family protein
7	L.Pa0_WS_06	chr1.S_61631338	1	61631338	54–63	A/C	0.113	3.00E−06	4.14	1.86E−01	AC204530.4_FG003	Annexin
8	L.ABA-GE0_WS_05	PZE-101128324	1	163319126	160–164	A/G	0.130	4.45E−06	6.27	3.33E−01	GRMZM2G042776	FAD-binding and arabino-lactone oxidase
9	S.ABA-GE0_WW_06	SYN11636	1	187643343	183–191	G/A	0.376	3.11E−06	7.10	*3.48E−07*	GRMZM2G132748	NADH dehydroge-nase (ubiquinone)
10	S.Pro0_WS_06	chr1.S_194573521	1	194573521	191–195	G/C	0.172	1.19E−06	8.19	*8.75E−04*	GRMZM2G025302	Ethylene induced calmodulin binding protein
11	S.Pro0_WS_06	chr1.S_267783235	1	267783235	265–273	A/T	0.104	9.80E−07	5.73	*7.64E−04*	GRMZM2G022859	Ubiquitin-conjugating enzyme
12	L.Pglc0_WS_06	chr1.S_295348610	1	295348610	293–296	C/G	0.179	4.14E−06	5.35	2.19E−01	GRMZM2G156756	Squamosa promoter-binding protein-like 12
13	L.Pglc0_WS_05	chr2.S_22172331	2	22172331	21–23	A/C	0.178	2.27E−06	7.94	4.96E−01	GRMZM2G109262	Glyoxalase I family protein
14	L.Suc7_WS_06	PZE-102052900	2	30398482	29–33	G/C	0.270	5.29E−06	6.47	4.61E−01	GRMZM2G111954	NAD(P)-linked oxidoreductase superfamily protein
14	L.Tsug7_WS_06	PZE-102052900	2	30398482	29–33	G/C	0.270	3.31E−06	5.90	3.28E−01	GRMZM2G111954	NAD(P)-linked oxidoreductase superfamily protein
15	E.Str0_WS_06	chr2.S_52318834	2	52318834	51–53	A/T	0.493	5.01E−06	7.03	2.12E−01	GRMZM2G107588	Unknown
16	E.ABA0_WS_05	chr2.S_106287422	2	106287422	106–107	C/T	0.166	8.46E−07	9.14	6.53E−01	GRMZM2G141814	Cyclin
17	L.ABA-GE0_WS_06	chr2.S_154610515	2	154610515	154–157	C/T	0.138	2.01E−06	9.14	*1.78E−06*	GRMZM2G010363	Sec14p-like phos-phatidylinositol transfer family protein

Table 1 continued

Loci	Variable[a]	Peak SNP[b]	Chr.[c]	Pos (bp).[d]	QTL interval (Mb)[e]	Allele[f]	MAF[g]	GWAS[h]	Var (%)[i]	GxE[j]	Candidate gene[k]	Annotation[l]
18	E.ABA-GE0_WS_05	chr2.S_199550110	2	199550110	198–200	A/G	0.082	5.95E−06	12.50	3.09E−01	GRMZM5G877483	F-box and Leucine-Rich Repeat domains containing protein
19	S.Tsug0_WW_06	chr2.S_205716352	2	205716352	193–209	A/T	0.132	2.76E−06	4.34	3.32E−01	GRMZM2G418415	D-Serine ammonia-lyase
20	L.Suc0_WS_06	PZE-102192647	2	235147223	235.5–236	G/A	0.270	7.79E−07	9.04	2.00E−01	GRMZM2G409771	High affinity K + transporter 5
20	L.Tsug0_WS_06	PZE-102192647	2	235147223	235.5–236	G/A	0.270	3.00E−07	9.74	2.24E−01	GRMZM2G409771	High affinity K + transporter 5
21	E.ABA0_WS_05	chr3.S_47736883	3	47736883	40–53	G/C	0.151	3.60E−06	7.05	7.13E−01	GRMZM2G133802	Tubulin beta-5 chain (Beta-5-tubulin)
22	L.Glc0_WS_06	chr3.S_59850605	3	59850605	59–61	G/A	0.293	3.45E−06	8.79	*3.56E−02*	GRMZM2G113495	Polyketide cyclase/dehydrase and lipid transport superfamily protein
23	L.Tsug0_WW_06	chr3.S_156789670	3	156789670	153–162	C/T	0.134	3.21E−06	9.67	*1.27E−03*	GRMZM2G336909	Histone-lysine N-methyltransferase
24	E.Pa0_WW_06	PZD00027.3	3	171430083	170–172	C/A	0.070	3.72E−08	11.23	*1.15E−03*	*GRMZM2G110153*	MADS16:*Zea mays* MADS16||MADS-domain transcription factor
24	E.Pa7_WW_06	PZD00027.3	3	171430083	170–172	C/A	0.070	1.10E−07	10.94	9.69E−01	*GRMZM2G110153*	MADS16:*Zea mays* MADS16||MADS-domain transcription factor
25	E.ABA7_WS_05	chr3.S_193639689	3	193639689	176–200	C/G	0.086	6.21E−06	7.59	6.83E−02	GRMZM2G042412	Copper transporter 1
26	L.ABA0_WW_06	chr3.S_201380315	3	201380315	200–202	T/C	0.145	4.43E−07	12.57	7.98E−01	GRMZM2G079727	MADS-box family gene with MIKCc type-box
27	E.Tsug7_WS_05	chr4.S_29922538	4	29922538	29–32	G/A	0.177	4.16E−06	3.10	7.24E−01	GRMZM2G141034	(ATSTP1, STP1) sugar transporter 1
28	S.Glc7_WW_06	chr4.S_67070957	4	67070957	38–74	A/G	0.141	2.10E−06	6.82	1.16E−01	GRMZM2G417402	C2H2 zinc finger protein
28	S.Glc7_WW_06	chr4.S_68066517	4	68066517	66–70	A/G	0.221	2.61E−07	9.19	*3.62E−03*	GRMZM2G126128	Unknown
29	E.ABA-GE0_WS_05	chr4.S_128263834	4	128263834	126–131	G/A	0.131	6.34E−06	9.21	6.17E−01	GRMZM2G024647	Serine/threonine-protein kinase
30	E.ABA-GE7_WW_06	chr4.S_142202245	4	142202245	140–143	A/T	0.132	3.33E−06	9.59	4.12E−01	GRMZM2G062009	NAC domain containing protein
31	L.Pa7_WW_06	chr4.S_201734181	4	201734181	201–202	A/G	0.153	5.95E−06	8.19	6.13E−01	GRMZM2G008106	Early nodulin-like protein
32	E.Pa0_WS_06	chr4.S_227275165	4	227275165	226–228	A/G	0.153	5.39E−07	5.08	9.65E−02	GRMZM2G074122	Phosphoenolpyruvate carboxylase

Table 1 continued

Loci	Variable[a]	Peak SNP[b]	Chr.[c]	Pos (bp).[d]	QTL interval (Mb)[e]	Allele[f]	MAF[g]	GWAS[h]	Var (%)[i]	GxE[j]	Candidate gene[k]	Annotation[l]
33	L.Pa0_WS_05	PZE-104157654	4	240456906	238–241	G/A	0.073	6.24E−06	8.48	8.92E−01	GRMZM2G463871	Glycosyl hydrolases
34	L.Slw0_WS_05	PZE-105011863	5	5203180	5.0–5.4	C/A	0.089	5.77E−06	6.37	4.64E−01	GRMZM2G581728	6-phosphogluconate dehydrogenase
35	E.Pa7_WS_06	chr5.S_15702954	5	15702954	15–20	T/G	0.195	1.66E−06	12.31	*4.71E−04*	GRMZM2G439195	Nicotianamine synthase
36	E.Suc7_WS_06	chr5.S_20968704	5	20968704	20–22	C/A	0.205	4.13E−06	10.04	*9.18E−03*	GRMZM2G041048	F-box domain containing protein
37	L.Pa0_WW_06	PZE-105047262	5	36593319	34–40	G/A	0.092	7.36E−07	9.93	1.67E−01	GRMZM2G110968	Tyrosine protein kinase domain containing protein
38	S.Suc7_WS_06	SYN35833	5	60996345	60–62	C/A	0.417	2.14E−06	8.40	*1.31E−02*	GRMZM2G153987	(AtGH9B5,GH9B5): glycosyl hydrolase 9B5
39	S.ABA0_WS_06	chr5.S_79707038	5	79707038	76–80	C/T	0.202	4.67E−06	4.52	*4.35E−03*	GRMZM2G472708	Expansin precursor
40	E.Pa7_WS_05	PZE-105077266	5	86343387	86–89	G/A	0.099	3.69E−07	7.26	4.38E−01	GRMZM2G052544	MYB transcription factor
41	E.FSuc7_WW_06	chr5.S_135837850	5	135837850	134–136	A/T	0.307	9.36E−07	9.51	*1.02E−02*	GRMZM2G054023	Lectin-like receptor kinase kinase
42	E.Glc0_WS_06	chr5.S_160452354	5	160452354	158–165	G/A	0.191	4.28E−07	11.95	*1.79E−02*	GRMZM2G052402	Thioredoxin family protein
43	S.Suc0_WW_06	chr5.S_188832671	5	188832671	174–190	C/T	0.163	4.85E−06	6.90	*2.44E−04*	GRMZM2G170049	MYB family transcription factor
44	L.ABA0_WS_05	SYN35256	5	212762042	212–214	A/C	0.196	6.91E−07	6.83	8.88E−01	GRMZM2G089466	RING/U-box superfamily protein:zinc finger
45	E.Suc7_WS_05	chr6.S_31709176	6	31709176	31–34	A/C	0.159	7.09E−07	11.89	5.39E−01	GRMZM2G025959	Unknown
46	S.Pa0_WW_06	PZE-106029515	6	69727443	69–70	G/A	0.275	2.86E−06	9.47	*2.91E−03*	GRMZM2G168299	Leucine carboxyl methyltransferase
47	S.Glc0_WS_05	chr6.S_85413620	6	85413620	85–88	G/C	0.270	3.40E−06	8.09	1.99E−01	GRMZM2G048561	Acyltransferase
48	S.Pa7_WW_06	chr6.S_86783674	6	86783674	85–91	A/T	0.111	3.88E−06	8.10	2.32E−01	GRMZM2G140917	ABC1 family protein
48	S.Pa0_WS_06	PZE-106041699	6	91046985	90.5–91.5	G/A	0.090	9.32E−07	8.60	*2.12E−05*	GRMZM2G027995	DEAD-box ATP-dependent RNA helicase
49	S.Dw0_WS_05	chr6.S_117227251	6	117227251	114–122	A/G	0.210	4.26E−07	10.40	*2.44E−02*	GRMZM2G037094	Unknown
50	E.ABA-GE7_WW_06	chr6.S_127786560	6	127786560	124–129	C/T	0.089	4.23E−07	7.86	9.21E−01	GRMZM2G089021	Phosphoribosyl transferase
51	S.Glc0_WS_06	chr6.S_142446368	6	142446368	142–143	C/T	0.381	4.10E−06	7.89	*3.29E−04*	GRMZM2G041961	Glycosyl hydrolase superfamily protein
51	S.Tsug0_WS_06	chr6.S_142446368	6	142446368	142–143	C/T	0.382	6.30E−06	6.67	*9.23E−04*	GRMZM2G041961	Glycosyl hydrolase superfamily protein
52	E.ABA-GE0_WW_06	chr6.S_147921280	6	147921280	143–150	T/A	0.092	5.19E−06	16.14	1.03E−01	GRMZM2G041961	Glycosyl hydrolase superfamily protein

Table 1 continued

Loci	Variable[a]	Peak SNP[b]	Chr.[c]	Pos (bp).[d]	QTL interval (Mb)[e]	Allele[f]	MAF[g]	GWAS[h]	Var (%)[i]	GxE[j]	Candidate gene[k]	Annotation[l]
53	L.Slw7_WS_05	chr7.S_20202346	7	20202346	19–21	C/T	0.405	7.13E−07	8.55	9.81E−01	GRMZM2G404688	LTP family protein
54	L.Pa7_WS_06	PZE-107030744	7	39644521	39–40	G/A	0.142	6.08E−06	7.56	*2.86E−02*	GRMZM2G075715	Auxin response factor
55	L.Pa7_WS_06	chr7.S_118311220	7	118311220	118–124	A/C	0.158	4.10E−06	10.43	*4.65E−02*	GRMZM2G041175	Senescence-associated protein DH
56	L.Pa7_WW_06	PZE-107067440	7	124218720	124–124.4	G/A	0.079	2.63E−06	4.12	5.18E−01	GRMZM2G007249	1-aminocyclopropane-1-carboxylate oxidase protein
57	E.Dw0_WW_06	chr7.S_147708162	7	147708162	147–152	C/T	0.129	2.14E−06	8.95	1.73E−01	GRMZM2G423555	Glutathione S-transferase
58	L.ABA7_WW_06	PZE-108070018	8	122934431	122–125	A/C	0.076	3.50E−06	9.56	9.46E−01	GRMZM2G035944	TCP family transcription factor
59	E.ABA0_WW_06	PZE-109091985	9	138977471	138–140	G/A	0.059	2.81E−06	7.00	5.99E−01	GRMZM5G891159	ABC transporter
60	S.Suc7_WS_05	chr10.S_24785010	10	24785010	24–26	A/G	0.157	5.40E−06	7.26	3.17E−01	GRMZM2G009196	Riboflavin synthase activity
61	S.Str.L0_WW_06	PZE-110020869	10	27545154	27–28.5	A/C	0.369	6.05E−06	7.24	*7.22E−05*	GRMZM2G073860	Ser/Thr protein phosphatase
62	S.Str.L0_WS_05	PZA02941.6	10	71361989	70.5–72	A/G	0.403	5.59E−06	7.58	5.67E−02	GRMZM2G162574	UDP-glucoronosyl/ UDP-glucosyl transferase
63	E.ABA0_WW_06	chr10.S_98698532	10	98698532	98–99	C/T	0.387	2.62E−06	10.29	8.11E−01	GRMZM2G007721	UDP-Glycosyltransferase

[a] The phenotypic variable in GWAS, represented the combination of each trait, tissue, sample time, water condition, and year

[b] Most significant SNP for each locus

[c] Chromosome

[d] Physical position of peak SNP based on the maize B73 reference sequence version 5b.60 (MaizeSequence, http://www.maizesequence.org/)

[e] The extended physical regions where the r^2 between nearby SNPs and the peak SNP decayed to 0.2 for each locus

[f] Major allele (before slash) and minor allele (after slash) for peak SNP

[g] Minor allele frequency

[h] P value of peak SNP that was estimated by the mixed linear model simultaneously controlling population structure, MFLW and kinship

[i] Phenotype variation explained by each locus

[j] The significance of the interaction between the loci and water regimes that was estimated by a two-way ANOVA analysis, the italics value indicates the locus response to water regimes is significant ($P < 0.05$)

[k] The candidate gene was defined as the gene with the most plausible annotated function or closest to the peak SNP within the local LD of peak SNP for each locus. The italics candidate genes were also detected in previous study (Setter's et al. 2011)

[l] Annotation information according to InterProScan (http://www.ebi.ac.uk/interpro/)

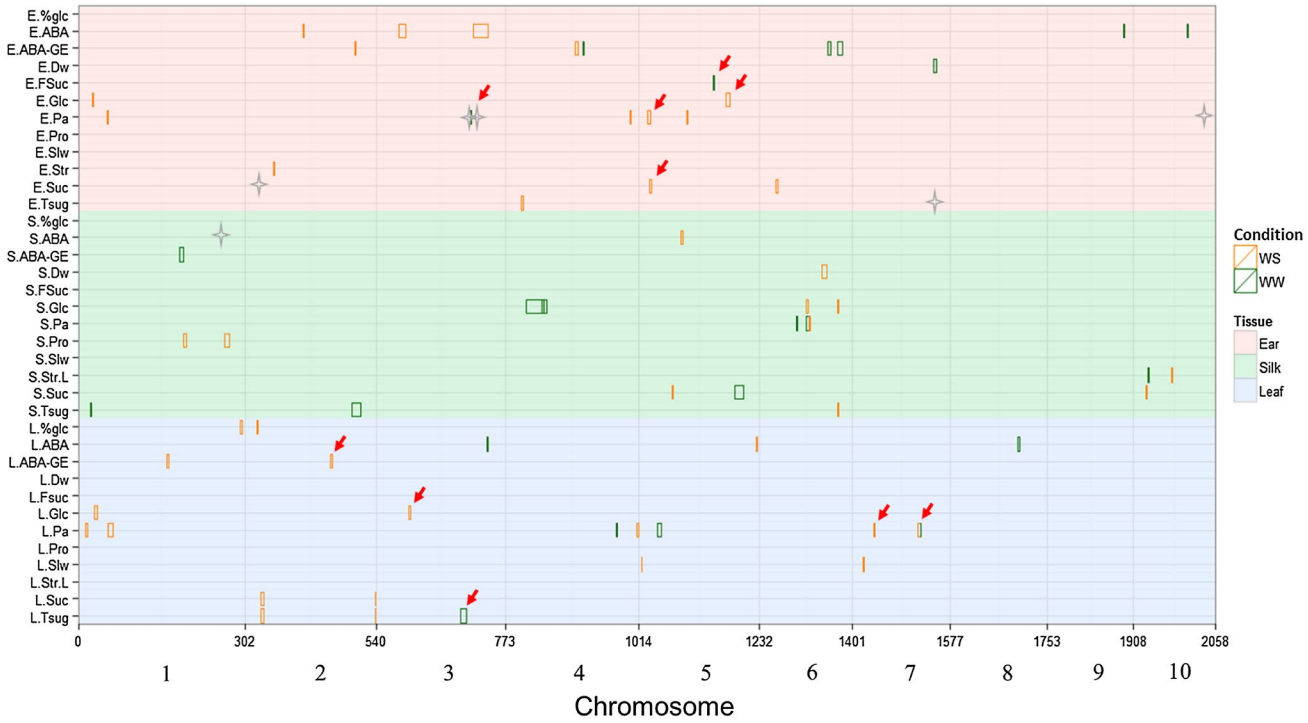


Fig. 3 Overview of metabolite-associated loci distributed across the whole maize genome. *Each box* represents a significant locus identified by GWAS ($P \leq 1/156599$) for each tissue and each drought condition. The *colored background* represents three tissues, and the *color* of the *box* indicates irrigation condition. *Gray asterisks* indicate the loci identified in Setter et al. (2011). *Red arrows* indicate the loci validated in hybrid pools

drought-tolerant and 68 intolerant hybrids) by crossing the panel with a common tester (CML312), involving the same twelve metabolic and physiological traits for leaf and ear (but not silk) in WW and WS at 7 days after anthesis (i.e., ABA, ABA-GE, Pa, Pro, Glc, Fglc, %glc, Suc, Fsuc, Tsug, Slw and Dw). A considerable range of phenotypic variation was observed for all traits in the hybrid pools (Table S5). Among the 85 variables, 22 showed significant differences between the hybrid pools with contrasting drought tolerance ($P < 0.05$), involving all twelve metabolic and physiological traits (Table S5). Thus, the hybrid pools provided an opportunity to correlate the metabolite-associated loci with drought tolerance.

From the 23 candidate drought-responsive loci chosen from GWAS for three tissues (see previous section), the set of 10 loci detected in ear and leaf tissues were selected to validate their effect on drought tolerance in the hybrid pools. The logistic regression model corroborated that all ten drought-responsive loci identified by the GWAS panel also had a significant effect on the average drought tolerance in the hybrids across three environments ($P < 0.05$). It is worth noting that nine of these ten loci (not including locus 41) were simultaneously validated in all three environments separately, indicating the reliable effects of these loci on drought tolerance in hybrids (Table 2). This was further confirmed by the higher proportion of significantly associated markers in these ten loci than the random proportion, defined as the proportion of significant random SNPs across the whole genome ($P = 4.47\text{E-}11$; Figure S5). Furthermore, the favorable alleles of these ten loci were found to be significantly enriched in the hybrids with high drought tolerance relative to the hybrids with low drought tolerance ($P = 1.87\text{E}{-}7$; Fig. 4). In total, the ten significantly associated loci jointly explained almost 18.4 % of drought tolerance variation in the hybrid pools estimated by using a multivariate logistic regression model.

Discussion

In this study, we collected data for metabolic and physiological traits across multiple tissues and environments in a subset of an association panel that had been studied previously (Setter et al. 2011). The new study presented here investigated the same set of metabolic and physiological traits over three environments in a hybrid population (Table S5) that was a subset of the most drought-tolerant and most drought-susceptible entries from a larger hybrid panel. This larger hybrid panel was previously established by test-crossing the association panel of inbred lines with inbred

Table 2 Performance of the metabolite-associated loci in the hybrid pools

Loci[a]	Trait[b]	Tissue[c]	Condition[d]	Lead SNP[e]	Chr.[f]	Pos.[g]	Allele[h]	Freq[i]	P[j]	E[k]	Candidate gene[l]	Annotation[m]
17	ABA-GE	Leaf	WS	chr2.S_156954199	2	156954199	A/C	0.27/0.08	0.0036	+/+/+	GRMZM2G010363	Sec14p-like phosphatidylinositol transfer family protein
22	Glc	Leaf	WS	PZE-103053471	3	60528865	A/C	0.3/0.19	0.0077	+/+/+	GRMZM2G113495	Polyketide cyclase/dehydrase and lipid transport superfamily protein
23	Tsug	Leaf	WW	PUT-163a-28982738-1684	3	157517780	G/A	0.39/0.14	0.0058	+/+/+	GRMZM2G336909	Histone-lysine N-methyltransferase
24	Pa	Ear	WW	chr3.S_171712818	3	171712818	T/A	0.91/0.82	0.0446	+/+/+	GRMZM2G110153	MADS16:*Zea mays* MADS16‖MADS-domain transcription factor
35	Pa	Ear	WS	PZE-105031503	5	17063973	G/A	0.84/0.62	0.0062	+/+/+	GRMZM2G439195	Nicotianamine synthase
36	Suc	Ear	WS	chr5.S_20918006	5	20918006	G/T	0.96/0.91	0.0097	+/+/+	GRMZM2G041048	F-box domain containing protein
41	Fsuc	Ear	WW	chr5.S_135837850	5	135837850	T/A	0.29/0.26	0.0061	−/−/−	GRMZM2G054023	Lectin-like receptor kinase
42	Glc	Ear	WS	PZE-105107800	5	164481392	C/A	0.37/0.14	0.0025	+/+/+	GRMZM2G052402	Thioredoxin family protein
54	Pa	Leaf	WS	PZE-107030398	7	39125424	G/A	0.96/0.83	0.0037	+/+/+	GRMZM2G075715	Auxin response factor
55	Pa	Leaf	WS	chr7.S_119958867	7	119958867	T/C	0.08/0.04	0.0056	+/+/+	GRMZM2G041175	Senescence-associated protein DH

[a] The ten loci were selected from the 23 metabolite-associated loci that showed significant interaction with water regimes, details can be seen in results; The ID of these ten loci were derived from Table 1

[b] Metabolic and physiological traits

[c] Tissues sampled in hybrids including Leaf and Ear

[d] Irrigated conditions including well watered (WW) and water stress (WS)

[e] The lead SNP means the most significant SNP within the local LD-based QTL interval for each locus estimated by the multivariate logistic regression controlling the population structure and flower time in hybrid pools

[f] Chromosome

[g] Physical position of lead SNP based on the maize B73 reference sequence version 5b.60 (MaizeSequence, http://www.maizesequence.org/)

[h] Favorable allele (before slash) and unfavorable allele (after slash)

[i] Frequency of favorable allele in the hybrid pool of high tolerance (before slash) and low tolerance (after slash)

[j] Significance value across three environments based on the multivariate logistic regression

[k] "+" and "−" indicated that the locus validated or not at Mex, Ken, and Thai based on the multivariate linear regression

[l] The candidate gene was defined as the gene with the most plausible annotated function or closest to the peak SNP within the local LD of peak SNP for each locus

[m] Annotation information according to InterProScan (http://www.ebi.ac.uk/interpro/)

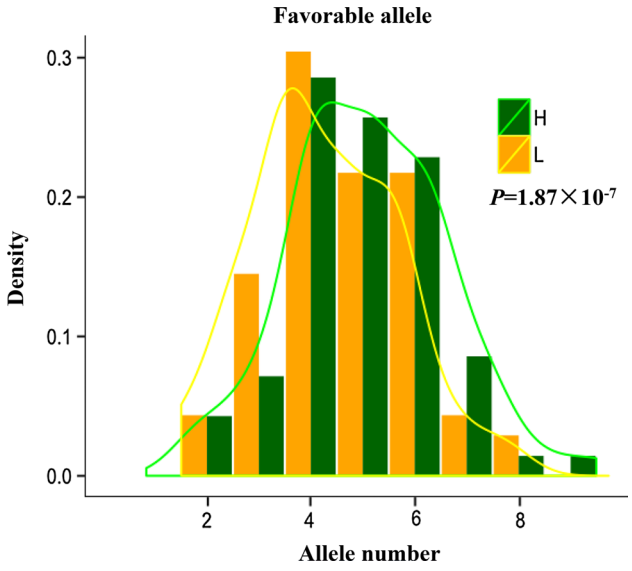


Fig. 4 Distribution of favorable alleles within the two hybrid pools. All the favorable alleles were derived from the lead SNP for the ten loci that exhibited significant differences in allele frequency between the two hybrid groups. The *bar* marked with "H" means the hybrids with high drought tolerance, while the *bar* marked with "L" means the hybrids with low drought tolerance. The *P* value declared the significance of differences between the two groups based on ANOVA

line CML312 (Xue et al. 2013). The metabolic and physiological data of the hybrid pools were then used to validate the effect of metabolite-associated loci identified in the current GWAS study on drought tolerance in hybrids.

The previous GWAS for metabolic and physiological traits was conducted using a set of 1229 SNPs derived from drought tolerance-related candidate genes (Setter et al. 2011; Yan et al. 2009), and identified only seven SNP-trait associations underlying six candidate loci (Setter et al. 2011). In the present study, 156,599 high-quality SNPs were obtained by integrating three genotypic datasets and using an efficient imputation method, leading to an enlarged SNP dataset. This enabled the identification of a considerable number of significant candidate loci potentially responsible for metabolic and physiological variation, including 26 new loci significantly associated with metabolic and physiological traits in leaf tissue, where no significantly associated SNP was detected in the previous, smaller study (Setter et al. 2011). The six loci significantly associated with drought-related metabolites in Setter et al. (2011) were in pathways that affect ABA and carbohydrate metabolism in floral tissues during drought. However, the p values of the same SNPs in the present study were different, such that only one locus was still significant (at $p \leq 6.39E{-}6$ or $p \leq 1/156599$), and others were only marginally significant at a suggestive cutoff ($p \leq 8.14E{-}4$ or $p \leq 1/1229$) in the currently expanded GWAS (Table S6). The significance of the SNPs has decreased from the previous to the current study mainly due to the decrease in sample size from 350 in the first study to 318 in the current study, which would decrease statistical power to identify loci with minor effects and unbalanced allele frequency. Moreover, the PCA of the 318 inbred lines was largely distinct when estimated with 1229 SNPs compared to 50 K SNPs (Figure S6), which probably also influenced the correction of the inflated association p values due to false positives. As the PCA of 318 inbred lines estimated using whole 50 K SNPs may have ascertainment bias due to the background in which they were discovered (Ganal et al. 2011), we re-estimated the PCA of the 318 lines using the less-biased Panzea markers extracted from the 50 K SNPs and found that the bias effect for PCA analysis is small in our association panel (Figure S7).

In the present study, GWAS results provide further insights into the genetic architecture of metabolic variation in maize under different water regimes (Table 1) and enabled the identification of 63 significant candidate loci-trait associations. Most of these were new, and, like in the previous study, found to be involved in the accumulation of carbohydrates and ABA-derived metabolites under drought stress. This accumulation has been hypothesized as a drought response in earlier studies (Seki et al. 2007; Mohammadkhani and Heidari 2008). Of particular interest, *GRMZM2G110153*, which was also identified in the previous study (Setter et al. 2011), was significantly associated with phaseic acid (Pa) at 0 days after anthesis in WW ears. This gene encodes a MADS-box transcription factor in maize (*MADS16* or *Zmm16*) that is important for the specification of floral organs in Arabidopsis, and is preferentially expressed in young floral organs in maize and millet (Wuest et al. 2012). In maize, *Zmm16* may be one of the regulators of stamen development (Whipple et al. 2004), and we can hypothesize that Pa accumulation in maize ears (as seen in Table 2) may be linked to a hormonal role in floral development regulated by *Zmm16* (Setter et al. 2011).

Another drought tolerance candidate gene identified in the current study was *GRMZM2G041048*, which encodes an F-box domain containing protein, and was significantly associated with Sucrose (Suc) at 7 days after anthesis in ears under WS condition (Table 1). Some F-box proteins, such as DROUGHT TOLERANCE REPRESSOR1 (DOR1) and Empfindlicher im Dunkelroten Licht1-Like Protein3 (EDL3), are reported to be involved in plant response to abiotic stress via ABA and drought-response pathways (Zhang et al. 2008; Koops et al. 2011). In plants, the F-box protein more axillary growth 2 (MAX2) is an important component of strigolactone signaling pathways and has been shown to regulate diverse biological processes, including plant architecture, photomorphogenesis, senescence, and

karrikin signaling (Brewer et al. 2013). In Arabidopsis, the max2 mutant is strongly hypersensitive to drought stress compared with the wild-type Arabidopsis protein. In strigolactone signaling pathways, only *MAX2* genes are involved in plant response to drought and ABA. Mutations in other genes in the family, including *MAX1*, *MAX3*, and *MAX4*, produce similar phenotypes to the wild type and do not display any defects in stress responses. These findings indicate that *MAX2* not only participates in strigolactone signaling pathways but also plays an important role in plant response to abiotic stress conditions (Bu et al. 2014).

For each maize tissue, there were large numbers of significantly associated loci specific to the irrigation condition (Table 1), and more than one-third of the loci showed significant interactions with water regime, probably reflecting the genetic determinants of metabolic response to the drought stress. The dissection of QTLs expressed under different irrigation conditions is highly beneficial to understanding the genetic basis of drought tolerance in maize and to help improve elite varieties with drought tolerance in breeding programs.

Over the last century, maize grain yield has increased nearly eightfold due to selection of superior parents for hybrid production (Duvick 2005). Thus, it is important to evaluate GWAS results in hybrid backgrounds to assess their potential utility in modern elite (and hybrid) cultivars. Heterosis of productivity and stress tolerance gained from enhanced hybrid vigor could mask the effects of individual loci identified via GWAS, as the effect of each locus is usually very small. To determine if these loci can add practical value to a hybrid breeding program, 139 contrasting hybrids (71 tolerant and 68 intolerant to drought, Figure S8) were selected from the total 318 inbred lines from the GWAS panel that had been test-crossed with CML312 (Xue et al. 2013). This hybrid population containing the high or low drought-tolerant hybrids was analyzed using ten loci that had been found to be associated with drought stress in the inbred panel. All ten loci exhibited significant effects on drought tolerance in the hybrid pools (Table 2). Together, the ten candidate drought-response genes explain almost 18.4 % of the drought tolerance variation and provide potential new targets of selection for the genetic improvement of highly drought-tolerant hybrid maize. However, much of the phenotypic variation was left unexplained due to the complex architecture of drought tolerance in maize. This "missing heritability" (Maher 2008) may be attributed to genes with minor effects and epistasis, which is hard to detect in GWAS studies of this size.

Author contribution statement MW and TS designed the study. YX and JY supervised the study. TS and MW performed the experiments. XZ, HL, YX and NY analyzed the data. XZ, YX, MW, TS, and JY prepared the manuscript and all authors read and approved the manuscript.

Acknowledgments This research was supported by the National Hi-Tech Research and Development Program of China (2012AA10A307), the National Natural Science Foundation of China (31222041, 31401389) and the Generation Challenge Program of the CGIAR.

Compliance with ethical standards

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

References

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JD, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631

Bartels D, Sunkar R (2005) Drought and salt tolerance in plants. Crit Rev Plant Sci 24:23–58

Brewer PB, Koltai H, Beveridge CA (2013) Diverse roles of strigolactones in plant development. Mol Plant 6:18–28

Bu Q, Lv T, Shen H, Luong P, Wang J, Wang Z, Huang Z, Xiao L, Engineer C, Kim TH, Schroeder JI, Huq E (2014) Regulation of drought tolerance by the F-box protein MAX2 in Arabidopsis. Plant Physiol 164:424–439

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325:714–718

Chan EK, Rowe HC, Hansen BG, Kliebenstein DJ (2010) The complex genetic architecture of the metabolome. PLoS Genet 6:e1001198

Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H, Zhang W, Zhang L, Yu S, Wang G, Lian X, Luo J (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 46:714–721

Dunn OJ (1959) Estimation of the medians for dependent variables. Ann Math, Stat, pp 192–197

Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56:52–64

Duvick DN (2005) The contribution of breeding to yield advances in maize (*Zea mays* L.). Adv Agron 86:83–145

Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z, Wang B, Zhai L, Dai C, Xu J, Wang W, Li X, Zheng J, Chen L, Luo L, Liu J, Qian X, Yan J, Wang J, Wang G (2013) RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun 4:2832

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schon CC, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One 6:e28334

Haley C (2011) A cornucopia of maize genes. Nat Genet 43:87–88

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Keurentjes JJ, Fu J, De Vos CR, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nat Genet 38:842–849

Koops P, Pelser S, Ignatz M, Klose C, Marrocco-Selden K, Kretsch T (2011) EDL3 is an F-box protein involved in the regulation of abscisic acid signalling in Arabidopsis thaliana. J Exp Bot 62:5547–5560

Lebreton C, Lazić-Jančić V, Steed A, Pekić S, Quarrie S (1995) Identification of QTL for drought responses in maize and their use in testing causal relationships between traits. J Exp Bot 46:853–865

Li MX, Yeung JM, Cherny SS, Sham PC (2012a) Evaluating the effective numbers of independent tests and significant *P*-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet 131:747–756

Li Q, Yang X, Xu S, Cai Y, Zhang D, Han Y, Li L, Zhang Z, Gao S, Li J, Yan J (2012b) Genome-wide association studies identified three independent polymorphisms associated with alpha-tocopherol content in maize kernels. PLoS One 7:e36807

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). Am J Bot 82:1420–1425

Lu Y, Zhang S, Shah T, Xie C, Hao Z, Li X, Farkhari M, Ribaut JM, Cao M, Rong T, Xu Y (2010) Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. Proc Natl Acad Sci USA 107:19585–19590

Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24:31–38

Maher B (2008) The case of the missing heritability. Nature 456:18–21

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, Altmann T (2007) The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proc Natl Acad Sci USA 104:4759–4764

Mohammadkhani N, Heidari R (2008) Drought-induced accumulation of soluble sugars and proline in two maize varieties. World Appl Sci J 3:448–453

Paupière MJ, van Heusden AW, Bovy AG (2014) The metabolic basis of pollen thermo-tolerance: perspectives for breeding. Metabolites 4:889–920

Pinheiro C, Chaves M (2011) Photosynthesis and drought: can we make metabolic connections from available data? J Exp Bot 62:869–882

Ribaut JM, Hoisington DA, Deutsch JA, Jiang C, Gonzalez-de-Leon D (1996) Identification of quantitative trait loci under drought conditions in tropical maize. 1. Flowering parameters and the anthesis-silking interval. Theor Appl Genet 92:905–914

Ribaut JM, Betran J, Monneveux P, Setter T (2009) Drought tolerance in maize. In: Handbook of maize: its biology. Springer, New York, pp 311–344

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA 109:8872–8877

Seki M, Umezawa T, Urano K, Shinozaki K (2007) Regulatory metabolic networks in drought stress responses. Curr Opin Plant Biol 10:296–302

Setter TL, Yan J, Warburton M, Ribaut JM, Xu Y, Sawkins M, Buckler ES, Zhang Z, Gore MA (2011) Genetic association mapping identifies single nucleotide polymorphisms in genes that affect abscisic acid levels in maize floral tissues during drought. J Exp Bot 62:701–716

R Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Tuteja N, Gill SS (2013) Climate change and plant abiotic stress tolerance. Wiley, New York

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, Yan J (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 5:3438

Wen W, Li K, Alseekh S, Omranian N, Zhao L, Zhou Y, Xiao Y, Jin M, Yang N, Liu H, Florian A, Li W, Pan Q, Nikoloski Z, Yan J, Fernie AR (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. Plant Cell 27:1839–1856

Whipple CJ, Ciceri P, Padilla CM, Ambrose BA, Bandong SL, Schmidt RJ (2004) Conservation of B-class floral homeotic gene function between maize and Arabidopsis. Development 131:6083–6091

Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, Lohan AJ, Loftus B, Graciet E, Wellmer F (2012) Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci USA 109:13452–13457

Xue Y, Warburton ML, Sawkins M, Zhang X, Setter T, Xu Y, Grudloyma P, Gethi J, Ribaut JM, Li W, Zhang X, Zheng Y, Yan J (2013) Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. Theor Appl Genet 126:2587–2596

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One 4:e8451

Yan J, Warburton M, Crouch J (2011) Association mapping for enhancing maize (*Zea may* L.) genetic improvement. Crop Sci 51:433–449

Yang N, Lu Y, Yang X, Huang J, Zhou Y, Ali F, Wen W, Liu J, Li J, Yan J (2014) Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. PLoS Genet 10:e1004573

Zhang Ye XuW, Li Z, Deng XW, Wu W, Xue Y (2008) F-box protein DOR functions as a novel inhibitory factor for abscisic acid-induced stomatal closure under drought stress in Arabidopsis. Plant Physiol 148:2121–2133

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360