Genome-wide association studies of ionomic and agronomic traits in USDA mini core collection of rice and comparative analyses of different mapping methods

Shuai Liu
Mississippi State University https://orcid.org/0000-0003-2244-2216
Hua Zhong
Wuhan University
Xiaoxi Meng
Mississippi State University
Tong Sun
Wuhan University
Yangsheng Li
Wuhan University
Shannon R. M. Pinson
USDA-ARS Dale Bumpers National Rice Research Center
Sam Chang
Mississippi State University
Zhaohua Peng (zp7@bch.msstate.edu)
https://orcid.org/0000-0001-6047-8434

Research article

Keywords: rice, ionomic traits, agronomic traits, multivariate GWAS

Posted Date: May 5th, 2020

DOI: https://doi.org/10.21203/rs.2.22202/v2

Version of Record: A version of this preprint was published on September 24th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02603-0.

Abstract

Background: Rice is an important human staple food vulnerable to heavy metal contamination due to its unique physiology and growth environment. High yield with low heavy metal contamination is a common but highly challenging goal for rice breeders worldwide due to lack of genetic knowledge. To identify candidate QTLs for rice yield and heavy metal content, sixteen ionomic traits and thirteen agronomic traits of the USDA Rice mini-core collection were analyzed using both univariate and multivariate GWAS methods in this study. The USDA Rice Mini-Core Collection contains about 1% of the whole Rice Collection of the National Small Grains Collection (NSGC), USA.

Results: Using the p-value $<1.53\times10^{-8}$, this criterium p-value was determined by the Bonferroni correction for p-value of 0.05, 106, 47, and 97 QTLs were identified for ionomics in flooded environment, unflooded environment, and agronomic traits, respectively. A large number of QTLs coincide well with previous report results while many of the QTLs are new QTLs, suggesting the efficiency of GWAS methods and the reliability of this study. Our results further showed that each of the four GWAS methods can identify unique as well as common QTLs. When univariate methods failed to identify QTLs for a trait, the multivariate methods frequently detected QTLs. However, when many QTLs were detected by univariate methods, the number of QTLs detected by multivariate methods were reduced in many cases. These analyses suggest that using multiple GWAS methods can complement each other in QTL identification. In addition, several candidate genes involved in ionomic and agronomic traits control were identified by analyzing the sequences of the candidate QTL regions.

Conclusions: Significant QTLs for heavy metal, mineral, and agronomic traits are presented in the rice genome and some of them have been fine mapped in the rice genome in this study. This research provides novel insights into the genetic basis of both ionomic and agronomic variations in rice, establishing an important foundation for further studies on reducing heavy-metal contamination and improving crop yields. In addition, the comparison analysis of the GAWS methods showed that each method has unique feature and different method can complement each other.

Background

Rice is an important cereal which feeds more than half the world's population [1]. With the rapid expansion of global population, food security has become a highly challenging task. Meanwhile, anthropogenic activities such as mining, smelting, chemical engineering, energy related industry, and broad application of pesticides & fertilizers in agriculture has led to frequent heavy metal contamination in soil, including Cadmium (Cd), Manganese (Mn), Nickel (Ni), and metalloid Arsenic (As) [2]. Soil with excessive heavy metal elements represses plant germination and growth, resulting in a decrease of crop yield [3, 4]. Meanwhile, plants uptake the toxic heavy metal elements from contaminated soil and accumulate them in edible plant tissues, leading to food contamination.

The anaerobic growing conditions of flooded rice paddies and the unique physiology of the rice plant allow rice to take up some heavy metals from water and soils in a highly efficient manner and sequester it in different organs within the plants, including the grain consumed by humans. The As concentration in a rice grain is roughly about 10 times higher than other crops grown in the same region even if the soil has no or limited anthropogenic contamination [5]. Rice has been reported to contribute substantially to inorganic and organic arsenic [6–8] intake by the human population in many regions of the world. As was ranked on the top of the US Agency for Toxic Substances and Disease Registry (ATSDR) Priority List of Hazardous Substances since 1997 (https://www.atsdr.cdc.gov/spl/index.html#2017spl). It has also been listed as a toxic component by many other countries and treated as a critical contaminant during food safety inspection. Cd is one of the most toxic heavy metals, and can easily reach the food chain due to strong assimilation by crops [9, 10]. Once absorbed, Cd is efficiently retained in the human body, may causing it to stay throughout the life span with an estimated half biology life between 6 to 38 years in kidney and between 4 to 19 years in the liver (ATSDR, 1999).

In contrast to heavy metals, many mineral elements are essential to humans but deficient in rice grains, for example, zinc, calcium, and iron [11]. To increase the concentrations of these minerals can improve the nutritional value of rice thus promoting human health for those using rice as the staple food. However, it is highly challenging to either increase the essential minerals or reducing the heavy metal due to lack of understanding of the genetic bases and molecular mechanisms of the related traits. Further, it is still poorly explored whether the concentration of mineral or heavy metal is associated with agronomic traits. Although there are multiple rice association mapping studies with specific minerals, heavy metals, and agronomic traits, respectively, these studies used either different mapping populations or different statistical analyses [12–14]. Therefore, each of the studies reveals some but not all facets of the genetic bases of rice variations. Recent accessibility to comprehensive sequence data, and development of software facilitating use of more powerful statistical analytics, opens the opportunity for more comprehensive study and understanding of the genetic bases of these traits.

The USDA Rice Core Collection, containing about 10% of the whole NSGC (National Small Grains Collection) Rice Collection, was assembled by stratified random sampling method in 2002, which has been evaluated comprehensively for 25 characteristics and proven to be highly representative of the whole collection [15]. The Rice Mini-Core Collection contains approximately 10% of the Core Collection [16]. The grain mineral concentrations have been analyzed under flooded and unflooded growth conditions [17], and the agronomic traits have also been evaluated for the Core Collection [18]. But most of these researches were done before the genome sequencing data was available.

Biparental genetic mapping and Genome-wide-association-study (GWAS) are the two different tools for mapping Quantitative Trait Locus (QTL). GWAS involves studying a natural population thus reflecting historical recombination events, which is critical for crop improvement but cannot be revealed by studying the offspring of biparental crosses in linkage map [19]. GWAS has been applied successfully to a variety of plants, including Arabidopsis [20], maize [21, 22], barley [23], wheat [24], rice [25, 26], soybean [27], and cotton [28].

Univariate GWAS is a mapping method that has been successfully used for gene mapping in plants and animals. However, a large number of genes may not be detected (false negative QTLs) due to the confounding problems between population structure, kinship, and markers. The population structure causes genome-wide linkage disequilibrium between unlinked loci, which leads to statistical confounding in genome-wide association studies. Mixed models have been shown to deal well with the confounding effects of a large number of small effect loci in the diffusion background, but they do not always account for large effect loci [29]. Multivariate GWAS method considers the confounding problem between covariates and test marker to detect more QTLs and previous reports showed that multivariate GWAS had lower FDR when using the same threshold compared with univariate GWAS method [29]. In recent years, a large number of multivariate GWAS methods have been developed, including MLMM (multi-locus mixed-model) [29], FarmCPU (Fixed and random model Circulating Probability Unification) [30], mrMLM (multi-locus random-SNP-effect MLM) [31], FASTmrMLM (fast mrMLM) [32], FASTmrEMMA (fast multi-locus random-SNP-effect efficient mixed model analysis) [33], pLARmEB (polygenic-background-control-based least angle regression plus empirical Bayes) [34], pKWmEB (integration of Kruskal-Wallis test with empirical Bayes) [35], ISIS EM-BLASSO (iterative modified-sure independence screening expectation-maximization-Bayesian least absolute shrinkage and selection operator) [36], and GPWAS (Genome-Phenome Wide Association Study) [37]. The MLMM [29] uses forward-backward stepwise linear mixed-model regression, forward stepwise uses the most significant associated SNP as a new fixed-effect covariate (cofactor) and creates a new model until reaching a pre-specified maximum number of forward steps, backward stepwise means to remove least significant SNP and create a new smaller model until only one selected marker is left. Whereas, FarmCPU [30] performs marker tests with associated markers as covariates in a Fixed Effect Model (FEM), and then optimization on the associated covariate markers in a Random Effect Model separately. These multivariate GWAS methods were successfully applied to several different crop species, including cotton [38], rice [39, 40], foxtail millet [41], soybean [42, 43], maize [44, 45], and wheat [46, 47].

Results

Characteristics of SNPs

High-quality re-sequencing raw data of 191 accessions comprising the USDA Rice Mini Core, itself collected from 71 countries (**Supplementary Table S1**), was retrieved from NCBI SRA database (Accession: PRJNA301661) [51]. Genotyping of the 191 accessions were performed by GATK software. A total of 3,259,478 SNPs was obtained after filtration by minor allele frequencies (≥0.05) and integrity (≥0.4). Imputed SNPs, which were generated by Beagle 5.0 software [52], were used for further analyses. Distribution of these SNPs in the genome is summarized in **Table 1 and Figure 1a**, and the overall SNP density in the genome was 114.51 (bp/SNP). The number of SNPs ranged from 212,238 to 375,296 across the twelve rice chromosomes. Chromosome 4 held the minimum marker density with 127.23 (bp/SNP), while chromosome 11 exhibited a maximum marker density with one SNP per 100.40 bp.

Population Structure and Linkage Disequilibrium

Admixture analysis divided the 191 accessions into four ancestries, including Indica (63 accessions), Aus (37 accessions), Temperate Japonica (28 accessions), and Tropical Japonica (31 accessions) under the best K model (K = 4) **(Fig. 1b)**, which was determined by the lowest CV (cross-validation error) score **(Fig. 1c)**. Thirty-two accessions are classified as admixture (ADM) since the ratio from each single subpopulation is below 70%.

In order to reduce the amount of calculations, high-quality SNPs (SNPs integrity above 0.8) was selected to construct a maximum likelihood (ML) tree to illustrate the phylogenetic relationship of the 191 rice accessions **(Fig. 1d)**. The population was divided into four subpopulations and the color for each clade was determined according to the Admixture analysis results. The relationship obtained from phylogenetic tree is in line with the Admixture analysis.

Principal component analysis (PCA) was performed based on the 3,259,478 SNPs. Four conceivable subpopulations were separated by PC1, PC2, and PC3. The first three principal components (PCs) explained over 50% of the genetic variation. The first PC separates Indica and Japonica subpopulations (35.70%), the second PC distinguishes the Aus and Indica varieties, while the third PC separates Temperate Japonica and Tropical Japonica varietal groups **(Fig. 1e and 1f)**. Based on the results from the Admixture analysis, phylogenetic tree and PCA, the population was divided into four subgroups. In addition, the decay of LD with the physical distance between SNPs occurred at 191 kb (r^2= 0.2) **(Fig. 1g)**. Indica subpopulation exhibited the most rapid LD decay and Temperate Japonica showed the most extended LD.

Correlation of different traits

The correlation between grain ionomics in flooded environment and agronomic traits **(Supplementary Fig. S1a)**, grain ionomics in unflooded environment and agronomic traits **(Supplementary Fig. S1b)**, and between grain ionomics in flooded versus unflooded growth conditions **(Supplementary Fig. S1c)** were conducted. The results showed that days to flowering has strong correlation with Rb in flooded (0.53) and unflooded (0.57) **(Supplementary Fig. S1a and S1b)** environments. The accumulation of Cd, Mo, and Rb in rice grain in flooded environment and unflooded environment are correlated at r^2 = 0.52, 0.81, and 0.60, respectively **(Supplementary Fig. S1c)**.

Genome-wide association study by univariate GWAS and multivariate GWAS

Sixteen grain ionomic traits (As, Ca, Co, Cd, Cu, Fe, K, Mg, Mn, Mo, Ni, P, Rb, S, Sr, and Zn) under flooded and unflooded conditions were the same as reported [17]. Thirteen agronomic traits, including AMYLOSE, AWNTYPE, DAYSFLOWER, HULLCOLOR, HULLCOVER, KERNELLEN, KERNELWID, KERNELRAT, KERNELWT, LODGING, PANICLETYPE, PLANTHT, and PLANTTYPE [18], were shared by Yan as reported [53, 54] and recorded using the methods described by Li *et al* [55–57]. All these traits were analyzed using two univariate GWAS (GLM and MLM) and two multivariate GWAS (MLMM and FarmCPU) methods to identify QTLs. A total of 106 significant QTLs (p-value <1.53×10^{-8}) were detected to be associated with 9 ionomic concentration (Cd, Co, Cu, K, Mo, Ni, Rb, Sr, and Zn) in rice grain under flooded condition, in which

63, 68, 17, and 44 significant QTLs were identified by GLM, MLM, MLMM, and FarmCPU, respectively **(Fig. 2 and Supplementary Fig. S4b)**. For Cd, twenty-eight significant QTLs were identified. Three of them located near published genes (*CAL1* [14], *OsHMA2* [58], *rgMT* [59]) which have shown to be related to Cd accumulation or resistance. Seven of them were identified in previous mapping studies using univariate methods **(Supplementary Table S2)**. All of the seven QTLs were identified by univariate GWAS methods (GLM and MLM) but only two of the seven were also detected by multivariate methods (MLMM and FarmCPU) in our study. For Co, a total of eleven significant QTLs were detected. Two (one was identified by univariate methods and the other was detected by FarmCPU) of them co-located with previous reported QTLs. Nine of them were new QTLs discovered in the current study, MLMM method discovered 2 significant QTLs and FarmCPU method identified 7 QTLs, respectively. Three QTLs were detected to be significantly associated with K, one of which (only detected by FarmCPU) was also detected in previous studies [60]. For Zn, ten significant QTLs were identified, three of which co-located with previously reported loci [12, 13, 60, 61]. Among them, one significant QTL posited around 1,8001,929 bp of Chromosome 7 was detected by both univariate and multivariate methods, which located near reported QTL *qZN-7* [13]. The other two QTLs were detected by FarmCPU method only.

In the unflooded environment, only 47 QTLs were detected to be significantly associated with Cd, Fe, Mo, Ni, and Zn concentration. 29, 25, 10, and 20 significant QTLs were identified by GLM, MLM, MLMM, FarmCPU, respectively **(Fig. 2 and Supplementary Fig. S4c)**. Twenty-three identified QTLs were related to Cd, one of which located near *CAL1* gene [14], eight QTLs co-located with previous studies **(Supplementary Table S2)**. Among the eight co-localization QTLs, five were detected by univariate methods and three were identified by multivariate methods. For Fe, seven significant QTLs were identified, two of which were also reported by previous studies [60, 61], and both were detected by FarmCPU only in the current study. We noticed that for the traits that many QTLs were identified using GLM and MLM methods, the numbers of QTLs identified by MLMM and FarmCPU were less as shown in the case of Cd and Mo. When GLM and MLM method failed to identify or only identified a few significant QTLs, QTLs were successfully identified by MLMM and FarmCPU methods as shown in the case of identifying QTLs for Co, Fe, K, and Zn concentration regulation **(Supplementary Fig. S2)**. The QQ-plots in Fig. S2 (c) showed the power of MLMM and FarmCPU, which show no evidence for inflation but strong evidence for real effects. In contrast, the QQ-plots of GLM and MLM in Fig S2 (c) shows the tendency of false positive peaks. This observation is further confirmed when mining the key candidate genes controlling ionomic and agronomic traits as shown the section below. Interestingly, only 3 of the 106 (ionomic) QTLs identified in flooded growth condition were shared with the QTLs identified in unflooded condition. The three QTLs (QTLs marked with an asterisk on Chromosome 5, 6, and 12; **Figure 2, Supplementary Fig. S2, and Supplementary Table S2**) share by both growth condition were associated with Cd and Mo concentration regulation, indicating that the traits of these three QTLs were not impacted by water conditions. This result was consistent with the correlation analysis results among the ionomic traits that accumulation of Cd and Mo in rice grain in flooded and unflooded environment are correlated **(Supplementary Fig. S1c),** and the common QTLs for Cd and Mo under different irrigation conditions, respectively, are probably the genetic base for the correlation. In addition, two Cd concentration related

QTLs were detected in a proximate region on chromosome 2, one around the 25 M region and the other around the 24.9 M region. The first QTL is for flooded environment and the second is for unflooded environment. It is unknown if the encoded genes are similar. Furthermore, several loci were shown to be associated with more than one trait, indicating these QTLs may be pleiotropy. For example, the region around 15.5 Mb on chromosome 2 is associated with Cd and Mo **(Fig. 2)**.

For agronomic traits, a total of 97 significant QTLs (p-value <1.53×10^{-8}) were detected for the thirteen agronomic traits described above except for KERNELWT and PLANTTYPE (**Fig. 3, Supplementary Table S2)**. 39, 16, 29, and 50 significant QTLs were identified by GLM, MLM, MLMM, and FarmCPU, respectively **(Supplementary Fig. S4a)**. *Wax* [62] and *ALK* [63] genes were shown to be significantly associated with amylose content, which is consistent with previous reports. Gain size is a key agronomic trait that strongly linked to yield and quality. Many QTLs have been reported associating with rice grain size, which is decomposed into grain length, width, and thickness (*GS3* [64], *GS5* [65], *GW5* [66], *GW8* [67] , *GL7* [68], *TGW6* [69], *etc*.). In this study, four types of rice grain size-related traits included kernel length (KERNELLEN), kernel width (KERNELWID), kernel rate (KERNELRAT), and kernel weight (KERNELWT) were analyzed. A total of 13 QTLs were detected by univariate and multivariate GWAS methods. Among them, three were detected by univariate (GLM or MLM) GWAS methods and twelve of them were detected by multivariate (MLMM and FarmCPU) GWAS methods. 6, 2, and 5 of the 13 QTLs were found to be associated with KERNELLEN, KERNELWID, and KERNELRAT, respectively. No significant QTL was shown to be associated with KERNELWT. Remarkably, one QTL (Chromosome 3 position 16,733,441) was detected by all the four methods **(Supplementary Fig. S3f)**. The QTL locates on gene *GS3* [64], which is a major gene regulating grain size and organ size. It is worth to note that, five more significant SNPs were identified by FarmCPU and one of them located on chromosome 4 situated nearby the *NAL1* gene, which has been shown to be related to rice yield [70]. For the trait of KERNELWID, two significant QTLs were only detected by MLMM method and other methods failed to identify candidate QTLs. One of the identified QTLs positioned around 5,364,561 bp of chromosome 5, which is about 0.56 kb apart from gene *GW5* [71], which is known to control rice grain width **(Supplementary Fig. S3g)**. These results demonstrated the power of GWAS, especially the power of the multivariate (MLMM and FarmCPU) GWAS methods.

Mining candidate genes of agronomic-related traits

Lodging and Plant height are both related to cell wall properties, which could impact rice yield. Appropriate plant height and the strong stem are required for stable production [72]. A cluster of SNPs around 33.4 Mb on chromosome 1 (Lodging: 33,010,693 to 33,975,764 with leading SNP at 33,469,251; Plant height: 33,181,529 to 33,730,067 with leading SNP at 33,363,796) is shown to be significantly associated with lodging and plant height **(Supplementary Fig. S3j and S3k)**. Through LD block analysis, we defined a 72.37 kb blocks (33,458,683- 33,531,049) containing 12 genes to be the candidate locus. Among these genes, *OsPME6* (*Os01g0788400*) is annotated as pectin methylesterase 6, which is related to cell wall modification process. We further conducted blastP analysis with *Arabidopsis thaliana* and found that it shares high homology (E value = 3E-178) with Arabidopsis gene *PME18* (*AT1G11580*) **(Supplementary Table S3)**. The expression of *PME18* increased dramatically under hyper gravity

stimulus. It is speculated that pectin esterases induced pectin demethylation of carboxyl groups which increased the rigidity of pectin gel in the cell wall through calcium bridges [73]. Therefore, it is worth to test if *OsPME6* regulates rice lodging and height.

Flowering time is another important trait critical to rice production. Rice is a typical short-day (SD) flowering plant whose flowering is greatly affected by day length. A number of genes [74–77] have been reported to regulate rice flowering-time. In the current study, a total of three QTLs were significantly associated with the flowering time. Two of them were detected by FarmCPU exclusively on Chromosome 7 and 10. The other QTL on Chromosome 6 was detected with all the four different GWAS methods **(Supplementary Fig. S3c)**. The haplotype analysis showed that this region only harbors 2 genes (*OsPLL9* and *OsPLL10*). Among them, *OsPLL9* (*Os06g0583900*) located 7.15 kb away from the leading SNP. This gene is a homolog of pectate lyase gene, which may play crucial roles during rice panicle development [78]. *OsPLL9* is highly expressed in Stamen (one day before flowering), Palea (one day before flowering), and Panicle5 (heading stage) **(Supplementary Fig. S5)**. Thus, *OsPLL9* has the potential to be a candidate gene with a critical role in rice flowering.

Mining candidate genes of ionomic traits

28 and 23 significant QTLs were detected to be associated with Cd concentration in the flooded and unflooded environment, respectively. QTLs near *CAL1* (Chr2:25,190,487-25,191,188) were associated with rice grain Cd accumulation in both flooded (Leading SNP; Chr2: 24,968,588) and unflooded condition (Leading SNP; Chr2: 25,143,071). *CAL1* was annotated as a defensin-like protein, which could regulate Cd accumulation of rice leaves through translocating Cd from cytosol into extracellular spaces, but not rice grains [14]. We then further analyzed the genes around the QTLs and found there is an ABC transporter (*Os01g0121700*), its phosphorylation level was up-regulated under high Cd treatment (100 μM $CdCl_2 \cdot 2.5H_2O$) and it has been shown that the transporter reduces the concentration of Cd^{2+} through transporting PCs-Cd into vacuole [79]. Another QTL (Chr6: 29,733,715) was also showed strongly related to Cd concentration in rice grain. This QTL located near a known gene *OsHMA2*, which may decrease rice grain Cd concentration through suppressing the expression level of *OsHMA2* [58]. In addition, significantly associated SNP (Chr11: 29,014,045) posited near *rgMT* gene, which was a metallothionein protein responded to the Cd stress in *E. coli* [59]. Comparing the QTLs detected in this study with previously reported studies, we found that over fifteen QTLs were co-localized with reported loci. The details were shown in **Supplementary Table S2**. Meanwhile, thirty-two new QTLs were identified.

A significant QTL was identified on Chromosome 1 around nucleotide 4,348,829 with p-value 3.37E-10 (MLM method). This QTL posited within a 9.97 kb block (Chr1: 4,345,517 - 4,355,489) containing only one candidate gene *OsWRKY102* (*Os01g0182700*) **(Fig. 4a and 4b)**. BlastP analysis showed that the *OsWRKY102* (*Os01g0182700*) has high homology (1.00E-58) with *Arabidopsis WRKY13* (*AT3G39410*) gene **(Supplementary Table S4)**. *WRKY13* activates the expression of gene *PDR8* that encodes a Cd^{2+} extrusion pump, resulting in reducing Cd accumulation [80]. The expression profile from public data showed that *OsWRKY102* is intensively higher expressed in stem comparing to other tissues

(Supplementary Fig. S6a). When treated with a high concentration of cadmium, the expression level of *OsWRKY102* increased rapidly in both shoot and root **(Supplementary Fig. S6b)**. Overall, the results suggested that *OsWRKY102* responds at high-level cadmium treatment and regulates cadmium uptake and accumulation in rice. Another QTL (Chromosome 5 posited around 14,941,717) was identified in a flooded environment. Through LD analysis, we defined an 18.65 kb block (Chr5: 14,930,444 - 14,949,090) containing two genes, *Os05g0321600* and *Os05g0321900*. Among them, *Os05g0321900* (*OsWRKY75*) was annotated as DNA-binding WRKY domain-containing protein **(Fig. 5a)**. BlastP analysis found that this gene shares high homology (4E-52) with *WRKY55* (*AT2G40740*) of *Arabidopsis thaliana*, which regulates gold uptake and tolerance. Remarkably, one QTL (Chromosome 6 around position 11,906,590) was identified in both growth environments **(Fig. 5a and 5b)**. *OsMan07*, is an Endo-Beta-Mannanase, only 24.82 kb away from the leading SNP and were also detected by the previous study [81]. BlastP analysis found this gene has a high similarity (6E-108) to *Man3* (*AT3G10890*) gene of *Arabidopsis thaliana* **(Supplementary Table S4)**. Overexpression of *MAN3* enhanced Cd accumulation and tolerance, whereas loss-of-function of *MAN3* led to decreased Cd accumulation and tolerance [82]. All the genes' expression patterns located in the haplotype region associated with Cd were showed in **Supplementary Figure S7**. Overall, thirty-two new QTLs were identified in addition to precise identification of the loci reported in previous Cd studies.

PIP2;6 has been reported to be involved in arsenic concentration control in rice. Although no SNP with Bonferroni-corrected significant thresholds -log 10(*p*) above 7.81 was discovered, there was an SNP peak with -log(*p*) around 6 on the chromosome 4 near the published gene *PIP2;6***(Supplementary Fig. S2a)**, Suggesting that *PIP2;6* is located near a significant QTL revealed by GWAS analysis

Comparison of univariate and multivariate GWAS Methods

Our results demonstrated that there was not a single method that was able to detect all the QTLs while many loci were detected by all methods. The *GS3 gene* was shown to be associated with grain length by all the four tested methods. However, the *GW5 gene* was detected to be related to grain width only by multivariate GWAS. Similarly, the Cu related QTLs in flooded conditions and Fe related QTLs in unflooded condition were also detected by multivariate method only. Interestingly, when a large number of QTLs were identified by univariate method, the QTLs identified by multivariate method were substantially reduced. For example, over 29 QTLs for Cd were identified by univariate methods in flooded and unflooded environment but only six QTLs were identified by multivariate. Further, it appeared that the multivariate methods were able to pinpoint the location of the QTLs more precisely on the chromosome compared with the univariate methods in many cases. As shown in **Supplementary Figure 3f**, the peaks identified by univariate method were much broad than the peaks identified by multivariate methods. On the other hands, the univariate methods also identified many loci exclusively. In the case of LODGING, the QTL (candidate gene: *OsPME6*) located on chromosome 1 was only detected by GLM **(Supplementary Figure 3j)** and this gene was related with cell wall formation biological process, indicating that some of candidate QTLs might be ignored when we pursued lower FDR in multivariate methods.

Discussion

The reliability of the GWAS analyses

In this study, many previous reported loci important for agronomic and ionomic traits were rediscovered. The amylose controlling locus *Wax* gene was rediscovered by all four analysis methods and the *ALK* gene was mapped by the two univariate analysis methods. The hull cover genes *OcTCL2* and *OsWOX3* were mapped by the GLM and MLM methods. The kernel length genes *GS3* were mapped by all four methods and the *NAL1* was mapped by FarmCPU method. The kernel rate gene *GS3* was mapped by GLM, MLMM, and FarmCPU methods and the kernel wide gene *GW5* was mapped by MLMM method. The three known Cd related genes (*CAL1, OsHMA2, and rgMT*) were mapped by GLM and MLM methods. Further, the two lodging genes (*OsSPL14 and OsCESA9*) were also successfully mapped by univariate methods. Since a large number of genes/loci with known functions were successfully remapped, the results confirmed the accuracy of the imputed SNP dataset and power of mapping QTLs with GWAS. In addition, these observations also suggested that the new QTLs discovered in this study are worth to be further validated. Given that many new candidate genes were identified in the gene mining analyses, the function of these new genes can be directly tested by generating mutant or overexpression lines.

Different GWAS analysis methods posse unique features and complement each other

While many loci were mapped by all of the four tested GWAS methods, each of the four methods identified some of the function known loci exclusively. The kernel length gene *NAL1* on chromosome 4 was only identified by FarmCPU method. The kernel width gene *GW5* on chromosome 5 was only mapped by MLMM method. The lodging gene *OsSPL14* on chromosome 8 was identified by GLM method alone, and the lodging gene *OsCESA9* was only mapped by FarmCPU. There were many genes were mapped by only two or three of the four tested methods. For example, the amylose gene *ALK* on chromosome 6 was mapped by univariate methods GLM and MLM; The kernel rate gene *GS3* on chromosome 3 was mapped by GLM, MLMM, and FarmCPU. These results described above clearly demonstrated that all four methods can be effectively used to perform GWAS analysis and able to identify some of the known loci. However, none of the methods identified all the previously reported loci. Meanwhile, every method successfully mapped some loci which the other three methods failed to identify. Therefore, it is worth to test every method if our goal is to identify all related loci. In addition, it is worth to note that the differences within univariate or within multivariate methods are smaller compared to the differences between univariate and multivariate methods. Therefore, our results suggest that it is better to include at least one univariate and one multivariate method in GWAS analyses for best coverage of the QTLs.

Conclusion

In this study, a comprehensive GWAS analyses for ionomic and agronomic traits based on 3,259,478 SNPs were performed using two univariate methods and two multivariate methods. Under the criterium p-

value <1.53×10^{-8}, 106, 47, and 97 QTLs were identified for ionomics in flooded environment, unflooded environment, and agronomic traits, respectively. Under flooded environment, 28, 11, 4, 3, 40, 3, 4, 3, and 10 significant QTLs were shown to be associated with Cd, Co, Cu, K, Mo, Ni, Rb, Sr, and Zn, respectively. In unflooded condition, 23, 7, 7, 7, and 3 significant QTLs were detected to be associated with Cd, Fe, Mo, Ni, and Zn, respectively. In addition, 18, 3, 5, 19, 6, 5, 2, 28, 4, and 7 significant QTLs were tightly associated with amylose concentration, flowering time, hull color, hull cover, kernel length, kernel rate, kernel width, lodging, panicle type, and plant height, respectively. Detailed analysis of the QTLs revealed that many of the identified QTLs are co-localized with the QTLs reported in prior ionomic and agronomic studies or posited near the genes with known functions in the related traits, suggesting that the results are reliable. Our results showed that each of the four GWAS methods can identify its unique as well as common QTLs. Our results suggest that using multiple GWAS methods can complement each other in QTL identification and some methods may be more powerful with less false discovery rate. Since we used multiple GWAS analysis methods and large-scale DNA sequencing results instead of SSR markers for this study, more QTLs were identified, and the QTLs were defined to more precise regions compared with prior reports. Using the functional annotation results of the *Arabidopsis thaliana* orthologous, 3 plausible candidate genes (*OsWRKY102*, *OsWRKY75*, and *OsMan07*) are shown to be tightly associated with Cd concentration in rice **(Supplementary Table S3)**. In addition, our results showed that *OsPME6* or nearby gene may regulate plant height and *OsPLL9* or its nearby gene may play a role in flowering time control.

Our comprehensive GWAS analysis of the ionomic and agronomic traits with large scale DNA sequencing data of the USDA mini core collection sets a foundation for further genetic and molecular biology studies on mineral, heavy metal, and agronomic trait regulation.

Methods

Plant Materials

Grain ionomic traits [17] and agronomic traits [18] of the mini core collection were the same as reported. Diverse rice accessions were grown over 2 years in Beaumont, Texas under both flooded (anaerobic) and unflooded (aerobic, flush irrigated) irrigation schemes. The planting, field management, and harvest methods were as reported [53–57]. The correlations of the traits were calculated by Pearson's correlation and visualized with R corrgram package [83]. The details of the samples are listed in (**Supplementary Table S1**).

Genotyping

In order to obtain high-quality sequencing data, the reads were first filtered by NGS QC Toolkit (v2.3.3) with default settings [84]. Then, the high-quality sequence was mapped to Nipponbare MSU7.0 genomic reference (http://rice.plantbiology.msu.edu/index.shtml, Release 7) with bwa program (version 0.7.17) using default parameters [85]. PCR duplicates were marked by Picard (version 2.18). Then, HaploypeCaller of GATK was used to identify SNPs. The raw SNPs were filtered by PLINK software with

parameter '–maf 0.05 –geno 0.6 –snps-only'. Genotype imputation was performed for the remaining 3,259,478 SNPs with Beagle 5.0 [85] for further analysis.

Population structure, genetic analysis, and Linkage disequilibrium analysis

The raw SNP with integrity higher than 0.8 (181,448 SNPs) were extracted for estimating individual ancestries and constructing a phylogenetic tree. A PLINK software tool [86] was used to calculate the potential unlinked SNPs with parameter –indep-pairwise 50 10 0.2. The potentially unlinked SNPs was submitted to ADMIXTURE [87] to assess the population structure with varying K from 2 to 10. Cross-validation error was calculated for each K, and the clustering model with the lowest cross-validation error (K = 4) was selected. Population structure was displayed using online software Pophelper (http://pophelper.com/). Each individual was assigned to one of the four subpopulations based on having ≥ 70% genetic ancestry derivation, the accessions that had < 70% ancestry from one specific subpopulation were assigned to a fifth group called 'Admix'. The matrix of pairwise genetic distance was used to construct phylogenetic trees using the software SNPhylo [88] with parameters set to 'default'. Principal component analysis (PCA) and kinship matrix (K matrix) were performed with 3,259,478 SNPs using default parameter by GAPIT [89]. The decay distance of LD (linkage disequilibrium) in each subpopulation and in the whole mini-core population were determined by software PopLDdecay [90].

Genome-Wide Association Study (GWAS)

GWAS was performed among 191 rice accessions derived from USDA mini-core collection with 3,259,478 high-quality SNPs. Univariate GWAS methods (GLM and MLM) and multivariate GWAS methods (MLMM and FarmCPU) were employed to evaluate the trait-SNP associations for grain ionomic and agronomic traits using the Genomic Association and Prediction Integrated Tool (GAPIT) [89]. Principal component analysis (PCA) result was used as covariates to correct population structure due to subpopulations in the Mini Core. The genome-wide significant thresholds of the GWAS (p-value = 1.53×10^{-8}) was determined by 0.05/n (n is the number of markers)[91] and a higher significant threshold was set as 3.06×10^{-9} (0.01/n) [92]. The Manhattan and QQ plots for GWAS were visualized using the R package CMplot (https://github.com/YinLiLin/R-CMplot). Leading SNPs of each significant SNPs cluster (in 200 kb) were selected to display the location of the QTLs.

Haplotype block estimation

Haplotype blocks containing at least two SNPs were calculated with all imputed SNP using the PLINK software [86] with the following parameters: ' –blocks no-pheno-req –blocks-max-kb 2000 –blocks-inform-frac 0.95 –blocks-strong-highci 0.98 –blocks-recomb-highci 0.9'. The haplotypic blocks of each significant SNP were determined by Confidence Intervals described by Gabriel [93]. The LD heatmap was visualized by software Haploview [94]. The annotated genes located in each haplotype block were extracted from RAP-DB (https://rapdb.dna.affrc.go.jp/) **(Supplementary Table S5)**.

Gene expression data

The gene expression profile across 15 tissues (Endosperm, Callus, Seed, Radicle, Root, Plumule, Stem, Seedling, Shoot, Sheath, Leaf, Panicle, Spikelet, Lemma, and Stamen) was obtained from CREP (Collection of Rice Expression Profiles): http://crep.ncpgr.cn/crep-cgi/home.pl [95]. Gene expression data of rice plants treated with different cadmium concentration [96, 97] was adopted from TENOR (Transcriptome ENcyclopedia Of Rice): https://tenor.dna.affrc.go.jp/.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The genotype datasets analyzed during the current study are available in the NCBI SRA database (Accession: PRJNA301661), the phenotype traits analyzed are available in published article “Worldwide Genetic Diversity for Mineral Element Concentrations in Rice Grain” and website https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx

Competing interests

The authors declare that they have no competing interests

Funding

This project is partly supported by the USDA-ARS SCA agreement No. 5860667081 and Mississippi Rice Promotion Board. The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

SL, and ZP designed the work; SL, XM, TS, and YL collected the dataset; SL and HZ performed the analysis and finished the draft work; Each author substantively revised it.

all authors have read and approved the manuscript.

Acknowledgements

Not applicable

Abbreviations

ADM	admixture
AMYLOSE	amylose
As	Arsenic
ATSDR	Agency for Toxic Substances and Disease Registry
AWNTYPE	awn type
Cd	Cadmium
CREP	Collection of Rice Expression Profiles
DAYSFLOWER	flowering time
FarmCPU	Fixed and random model Circulating Probability Unification
FASTmrEMMA	fast multi-locus random-SNP-effect efficient mixed model analysis
FASTmrMLM	fast mrMLM
FEM	Fixed Effect Model
GAPIT	Genomic Association and Prediction Integrated Tool
GPWAS	Genome-Phenome Wide Association Study
GWAS	Genome-wide-association-study
HULLCOLOR	hull color
HULLCOVER	hull cover
ISIS EM-BLASSO	iterative modified-sure independence screening expectation-maximization-Bayesian least absolute shrinkage and selection operator
KERNELLEN	kernel length
KERNELRAT	kernel rate
KERNELWID	kernel width
KERNELWT	kernel weight
LD	linkage disequilibrium
LODGING	lodging
ML	maximum likelihood
MLMM	multi-locus mixed-model
Mn	Manganese
mrMLM	multi-locus random-SNP-effect MLM
Ni	Nickel
NSGC	National Small Grains Collection
PANICLETYPE	panicle type
PCA	Principal component analysis
PCs	principal components
pKWmEB	integration of Kruskal-Wallis test with empirical Bayes
PLANTHT	plant height
PLANTTYPE	plant type
pLARmEB	polygenic-background-control-based least angle regression plus empirical Bayes
QTL	Quantitative Trait Locus
SD	short-day
TENOR	Transcriptome ENcyclopedia Of Rice

References

1. Izawa T, Shimamoto K. Becoming a model plant: The importance of rice to plant science. Trends Plant Sci. 1996;1:95–9.

2. Fasani E, Manara A, Martini F, Furini A, DalCorso G. The potential of genetic engineering of plants for the remediation of soils contaminated with heavy metals. Plant Cell Environ. 2018;41:1201–32.
3. Panaullah GM, Alam T, Hossain MB, Loeppert RH, Lauren JG, Meisner CA, et al. Arsenic toxicity to rice (Oryza sativa L.) in Bangladesh. Plant Soil. 2009;317:31–9.
4. Rizwan M, Ali S, Adrees M, Rizvi H, Zia-ur-Rehman M, Hannan F, et al. Cadmium stress in rice: toxic effects, tolerance mechanisms, and management: a critical review. Environ Sci Pollut Res. 2016;23:17859–79. doi:10.1007/s11356-016-6436-4.
5. Williams PN, Villada A, Deacon C, Raab A, Figuerola J, Green AJ, et al. Greatly enhanced arsenic shoot assimilation in rice leads to elevated grain levels compared to wheat and barley. Environ Sci Technol. 2007;41:6854–9.
6. Gilbert-Diamond D, Cottingham KL, Gruber JF, Punshon T, Sayarath V, Gandolfi AJ, et al. Rice consumption contributes to arsenic exposure in US women. Proc Natl Acad Sci U S A. 2011;108:20656–60.
7. Azizur Rahman M, Hasegawa H, Mahfuzur Rahman M, Mazid Miah MA, Tasmin A. Arsenic accumulation in rice (Oryza sativa L.): Human exposure through food chain. Ecotoxicol Environ Saf. 2008;69:317–24.
8. Tuli R, Chakrabarty D, Trivedi PK, Tripathi RD. Recent advances in arsenic accumulation and metabolism in rice. Mol Breed. 2010;26:307–23.
9. Meharg AA, Norton G, Deacon C, Williams P, Adomako EE, Price A, et al. Variation in rice cadmium related to human exposure. Environ Sci Technol. 2013;47:5613–8.
10. Liu J, Li K, Xu J, Liang J, Lu X, Yang J, et al. Interaction of Cd and five mineral nutrients for uptake and accumulation in different rice cultivars and genotypes. F Crop Res. 2003;83:271–81.
11. Bashir K, Takahashi R, Akhtar S, Ishimaru Y, Nakanishi H, Nishizawa NK. The knockdown of OsVIT2 and MIT affects iron localization in rice seed. Rice. 2013;6:1–6.
12. Kumar J, Jain S, Jain RK. Linkage mapping for grain iron and zinc content in F2population derived from the cross between PAU201 and Palman 579 in rice (Oryza sativa L.). Cereal Res Commun. 2014;42:389–400.
13. Lu K, Li L, Zheng X, Zhang Z, Mou T, Hu Z. Quantitative trait loci controlling Cu, Ca, Zn, Mn and Fe content in rice grains. J Genet. 2008;87:305–10.
14. Luo JS, Huang J, Zeng DL, Peng JS, Zhang G Bin, Ma HL, et al. A defensin-like protein drives cadmium efflux and allocation in rice. Nat Commun. 2018;9. doi:10.1038/s41467-018-03088-0.
15. Yan W, Rutger JN, Bryant RJ, Bockelman HE, Fjellstrom RG, Chen MH, et al. Development and evaluation of a core subset of the USDA rice germplasm collection. Crop Sci. 2007;47:869–78.
16. Agrama HA, Yan WG, Lee F, Robert F, Chen MH, Jia M, et al. Genetic assessment of a mini-core subset developed from the USDA rice genebank. Crop Sci. 2009;49:1336–46.
17. Pinson SRM, Tarpley L, Yan W, Yeater K, Lahner B, Yakubova E, et al. Worldwide Genetic Diversity for Mineral Element Concentrations in Rice Grain. Crop Sci. 2015;55:294.

18. Yan W. Genetic Characterization of Global Rice Germplasm for Sustainable Agriculture. Food Prod - Approaches, Challenges Tasks. 2012;:243.
19. Nordborg M, Weigel D. Next-generation genetics in plants. Nature. 2008;456:720–3.
20. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43:956–65.
21. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet. 2011;43:159–62.
22. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2013;45:43–50.
23. Cai S, Yu G, Chen X, Huang Y, Jiang X, Zhang G, et al. Grain protein content variation and its association analysis in barley. BMC Plant Biol. 2013;13.
24. Jiang Y, Schmidt RH, Zhao Y, Reif JC. Quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. Nat Genet. 2017;49:1741–6.
25. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016;48:927–34.
26. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide asociation studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42:961–7.
27. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33:408–14.
28. Du X, Huang G, He S, Yang Z, Sun G, Ma X, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat Genet. 2018;50:796–802. doi:10.1038/s41588-018-0116-x.
29. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30.
30. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLoS Genet. 2016;12:e1005767.
31. Wang S-B, Feng J-Y, Ren W-L, Huang B, Zhou L, Wen Y-J, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep. 2016;6:19444. doi:10.1038/srep19444.
32. Tamba CL, Zhang Y. A fast mrMLM algorithm for multi-locus genome-wide association studies. 2018; October.
33. Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform. 2018;19:700–12.

34. Zhang J, Feng JY, Ni YL, Wen YJ, Niu Y, Tamba CL, et al. PLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. Heredity (Edinb). 2017;118:517–24.
35. Ren WL, Wen YJ, Dunwell JM, Zhang YM. PKWmEB: Integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. Heredity (Edinb). 2018;120:208–18.
36. Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PLoS Comput Biol. 2017;13:1–20.
37. Liang Z, Qiu Y, Schnable JC. Genome-phenome wide association in maize and Arabidopsis identifies a common molecular and evolutionary signature. Mol Plant. 2020.
38. Liu R, Gong J, Xiao X, Zhang Z, Li J, Liu A, et al. Gwas analysis and qtl identification of fiber quality traits and yield components in upland cotton using enriched high-density snp markers. Front Plant Sci. 2018;9 September:1–15.
39. Naveed SA, Zhang F, Zhang J, Zheng TQ, Meng LJ, Pang YL, et al. Identification of QTN and candidate genes for Salinity Tolerance at the Germination and Seedling Stages in Rice by Genome-Wide Association Analyses. Sci Rep. 2018;8:1–11. doi:10.1038/s41598-018-24946-3.
40. Wu Y, Sun Y, Sun S, Li G, Wang J, Wang B, et al. Aneuploidization under segmental allotetraploidy in rice and its phenotypic manifestation. Theor Appl Genet. 2018;131:1273–85. doi:10.1007/s00122-018-3077-7.
41. Jaiswal V, Bandyopadhyay T, Gahlaut V, Gupta S, Dhaka A, Ramchiary N, et al. Genome-wide association study (GWAS) delineates genomic loci for ten nutritional elements in foxtail millet (Setaria italica L.). J Cereal Sci. 2019;85:48–55. doi:https://doi.org/10.1016/j.jcs.2018.11.006.
42. Zhao X, Teng W, Li Y, Liu D, Cao G, Li D, et al. Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5.7. BMC Genomics. 2017;18:1–10.
43. Ziegler G, Nelson R, Granada S, Krishnan HB, Gillman JD, Baxter I. Genomewide association study of ionomic traits on diverse soybean populations from germplasm collections. Plant Direct. 2018;2:e00033.
44. Sanchez DL, Liu S, Ibrahim R, Blanco M, Lübberstedt T. Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (Zea mays L.). Plant Sci. 2018;268 November 2017:30–8. doi:10.1016/j.plantsci.2017.12.004.
45. Hu S, Sanchez DL, Wang C, Lipka AE, Yin Y, Gardner CAC, et al. Brassinosteroid and gibberellin control of seedling traits in maize (Zea mays L.). Plant Sci. 2017;263 July:132–41. doi:10.1016/j.plantsci.2017.07.011.
46. Kumar J, Saripalli G, Gahlaut V, Goel N, Meher PK, Mishra KK, et al. Genetics of Fe, Zn, β-carotene, GPC and yield traits in bread wheat (Triticum aestivum L.) using multi-locus and multi-traits GWAS. Euphytica. 2018;214:1–17. doi:10.1007/s10681-018-2284-2.
47. Jaiswal V, Gahlaut V, Meher PK, Mir RR, Jaiswal JP, Rao AR, et al. Genome wide single locus single trait, multi-locus and multi-trait association mapping for some important agronomic traits in

common wheat (T. aestivum L.). PLoS One. 2016;11:1–25.

48. Yan W, Jackson A, Jia M, Zhou W, Xiong H, Bryant R. Association Mapping of Four Important Traits Using the USDA Rice Mini-Core Collection. Rice - Germplasm, Genet Improv. 2014;:105–42.
49. Cui Y, Zhang F, Zhou Y. The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. Front Plant Sci. 2018;9.
50. Sun D, Cen H, Weng H, Wan L, Abdalla A, El-Manawy AI, et al. Using hyperspectral analysis as a potential high throughput phenotyping tool in GWAS for protein content of rice quality. Plant Methods. 2019;15:54. doi:10.1186/s13007-019-0432-x.
51. Wang H, Vieira FG, Crawford JE, Chu C, Nielsen R. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. Genome Res. 2017;27:1029–38.
52. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48. doi:10.1016/j.ajhg.2018.07.015.
53. Yan WG, Rutger JN, Bockelman HE, Tai TH. Agronomic evaluation and seed stock establishment of the USDA rice core collection. BR Wells Rice Res Stud Stuttgart Univ Arkansas, Agri Exp Sta Res Ser. 2005;:63–8.
54. Yan WG, Rutger JN, Bockelman HE, Tai TH. Evaluation of kernel characteristics of the USDA rice core collection. BR Wells Rice Res Stud Stuttgart Univ Arkansas, Agric Exp Stn Res Ser Agri Exp Sta Res Ser. 2005;:69–74.
55. Li X, Yan W, Agrama H, Hu B, Jia L, Jia M, et al. Genotypic and phenotypic characterization of genetic differentiation and diversity in the USDA rice mini-core collection. Genetica. 2010;138:1221–30.
56. Li X, Yan W, Agrama H, Jia L, Shen X, Jackson A, et al. Mapping QTLs for improving grain yield using the USDA rice mini-core collection. Planta. 2011;234:347–61.
57. Li X, Yan W, Agrama H, Jia L, Jackson A, Moldenhauer K, et al. Unraveling the complex trait of harvest index with association mapping in rice (Oryza sativa L.). PLoS One. 2012;7:1–10.
58. Satoh-Nagasawa N, Mori M, Nakazawa N, Kawamoto T, Nagato Y, Sakurai K, et al. Mutations in rice (oryza sativa) heavy metal ATPase 2 (OsHMA2) restrict the translocation of zinc and cadmium. Plant Cell Physiol. 2012;53:213–24.
59. Jin S, Cheng Y, Guan Q, Liu D, Takano T, Liu S. A metallothionein-like protein of rice (rgMT) functions in E. coli and its gene expression is induced by abiotic stresses. Biotechnol Lett. 2006;28:1749–53.
60. Zhang M, Pinson SRM, Tarpley L, Huang XY, Lahner B, Yakubova E, et al. Mapping and validation of quantitative trait loci associated with concentrations of 16 elements in unmilled rice grain. Theor Appl Genet. 2014;127:137–65.
61. Du J, Zeng D, Wang B, Qian Q, Zheng S, Ling HQ. Environmental effects on mineral accumulation in rice grains and identification of ecological specific QTLs. Environ Geochem Health. 2013;35:161–70.
62. Wang Y, Wan L, Zhang L, Zhang Z, Zhang H, Quan R, et al. An ethylene response factor OsWR1 responsive to drought stress transcriptionally activates wax synthesis related genes and increases wax production in rice. Plant Mol Biol. 2012;78:275–88.

63. Gao Z, Zeng D, Cheng F, Tian Z, Guo L, Su Y, et al. ALK, the Key Gene for Gelatinization Temperature, is a Modifier Gene for Gel Consistency in Rice. J Integr Plant Biol. 2011;53:756–65.
64. Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, et al. GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor Appl Genet. 2006;112:1164–71.
65. Li Y, Fan C, Xing Y, Jiang Y, Luo L, Sun L, et al. Natural variation in GS5 plays an important role in regulating grain size and yield in rice. Nat Genet. 2011;43:1266–9.
66. Liu J, Chen J, Zheng X, Wu F, Lin Q, Heng Y, et al. GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. Nat Plants. 2017;3:17043.
67. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, et al. Control of grain size, shape and quality by OsSPL16 in rice. Nat Genet. 2012;44:950–4. doi:10.1038/ng.2327.
68. Wang Y, Xiong G, Hu J, Jiang L, Yu H, Xu J, et al. Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat Genet. 2015;47:944–8.
69. Ishimaru K, Hirotsu N, Madoka Y, Murakami N, Hara N, Onodera H, et al. Loss of function of the IAA-glucose hydrolase gene TGW6 enhances rice grain weight and increases yield. Nat Genet. 2013;45:707–11.
70. Fujita D, Trijatmiko KR, Tagle AG, Sapasap MV, Koide Y, Sasaki K, et al. NAL1 allele from a rice landrace greatly increases yield in modern indica cultivars. Proc Natl Acad Sci U S A. 2013;110:20431–6.
71. Weng J, Gu S, Wan X, Gao H, Guo T, Su N, et al. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. Cell Res. 2008;18:1199–209.
72. Rebetzke GJ, Richards RA. Gibberellic acid-sensitive dwarfing genes reduce plant height to increase kernel number and grain yield of wheat. Aust J Agric Res. 2000;51:235–45.
73. Tamaoki D, Karahara I, Nishiuchi T, De Oliveira S, Schreiber L, Wakasugi T, et al. Transcriptome profiling in Arabidopsis inflorescence stems grown under hypergravity in terms of cell walls and plant hormones. Adv Sp Res. 2009;44:245–53. doi:10.1016/j.asr.2009.03.016.
74. Gao H, Zheng XM, Fei G, Chen J, Jin M, Ren Y, et al. Ehd4 Encodes a Novel and Oryza-Genus-Specific Regulator of Photoperiodic Flowering in Rice. PLoS Genet. 2013;9.
75. Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, et al. Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat Genet. 2008;40:761–7.
76. Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, et al. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell. 2000;12:2473–83.
77. Takahashi Y, Shomura A, Sasaki T, Yano M. Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the α subunit of protein kinase CK2. Proc Natl Acad Sci U S A. 2001;98:7922–7.

78. Zheng Y, Yan J, Wang S, Xu M, Huang K, Chen G, et al. Genome-wide identification of the pectate lyase-like (PLL) gene family and functional analysis of two PLL genes in rice. Mol Genet Genomics. 2018;293:1317–31. doi:10.1007/s00438-018-1466-x.
79. Zhong M, Li S, Huang F, Qiu J, Zhang J, Sheng Z, et al. The phosphoproteomic response of rice seedlings to cadmium stress. Int J Mol Sci. 2017;18.
80. Sheng Y, Yan X, Huang Y, Han Y, Zhang C, Ren Y, et al. The WRKY transcription factor, WRKY13, activates PDR8 expression to positively regulate cadmium tolerance in Arabidopsis. Plant Cell Environ. 2019;42:891–903.
81. Ishikawa S, Ae N, Yano M. Chromosomal regions with quantitative trait loci controlling cadmium concentration in brown rice (Oryza sativa). New Phytol. 2005;168:345–50.
82. Chen J, Yang L, Gu J, Bai X, Ren Y, Fan T, et al. MAN3 gene regulates cadmium tolerance through the glutathione-dependent pathway in Arabidopsis thaliana. New Phytol. 2015;205:570–82.
83. Friendly M. Corrgrams: Exploratory displays for correlatigon matrices. Am Stat. 2002;56:316–24.
84. Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS One. 2012;7.
85. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
86. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
87. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
88. Lee TH, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15:1–6.
89. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: Genome association and prediction integrated tool. Bioinformatics. 2012;28:2397–9.
90. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35:1786–8.
91. Yang M, Lu K, Zhao FJ, Xie W, Ramakrishna P, Wang G, et al. Genome-wide association studies reveal the genetic basis of ionomic variation in rice. Plant Cell. 2018;30:2720–40.
92. Zhao Y, Zhao W, Jiang C, Wang X, Xiong H, Todorovska EG, et al. Genetic architecture and candidate genes for deep-sowing tolerance in rice revealed by non-syn GWAS. Front Plant Sci. 2018;9 March:1–14.
93. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science (80-). 2002;296:2225–9.
94. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.

95. Wang L, Xie W, Chen Y, Tang W, Yang J, Ye R, et al. A dynamic gene expression atlas covering the entire life cycle of rice. Plant J. 2010;61:752–66.
96. Oono Y, Yazawa T, Kawahara Y, Kanamori H, Kobayashi F, Sasaki H, et al. Genome-wide transcriptome analysis reveals that cadmium stress signaling controls the expression of genes in drought stress signal pathways in rice. PLoS One. 2014;9.
97. Kawahara Y, Oono Y, Wakimoto H, Ogata J, Kanamori H, Sasaki H, et al. TENOR: Database for comprehensive mRNA-Seq experiments in rice. Plant Cell Physiol. 2016;57:e7.

Tables

Table 1 Summary of the SNPs across 12 chromosomes of *Oryza sativa*

Chromosome	Number of SNPs	Length of Chromosome (bp)	Density of SNP (bp/SNP)
1	375,296	43,270,923	115.30
2	301,111	35,937,250	119.35
3	294,312	36,413,819	123.73
4	279,049	35,502,694	127.23
5	253,001	29,958,434	118.41
6	287,238	31,248,787	108.79
7	253,651	29,697,621	117.08
8	261,070	28,443,022	108.95
9	212,238	23,012,720	108.43
10	222,521	23,207,287	104.29
11	289,053	29,021,106	100.40
12	230,938	27,531,856	119.22
Total	3,259,478	373,245,519	114.51

Figures

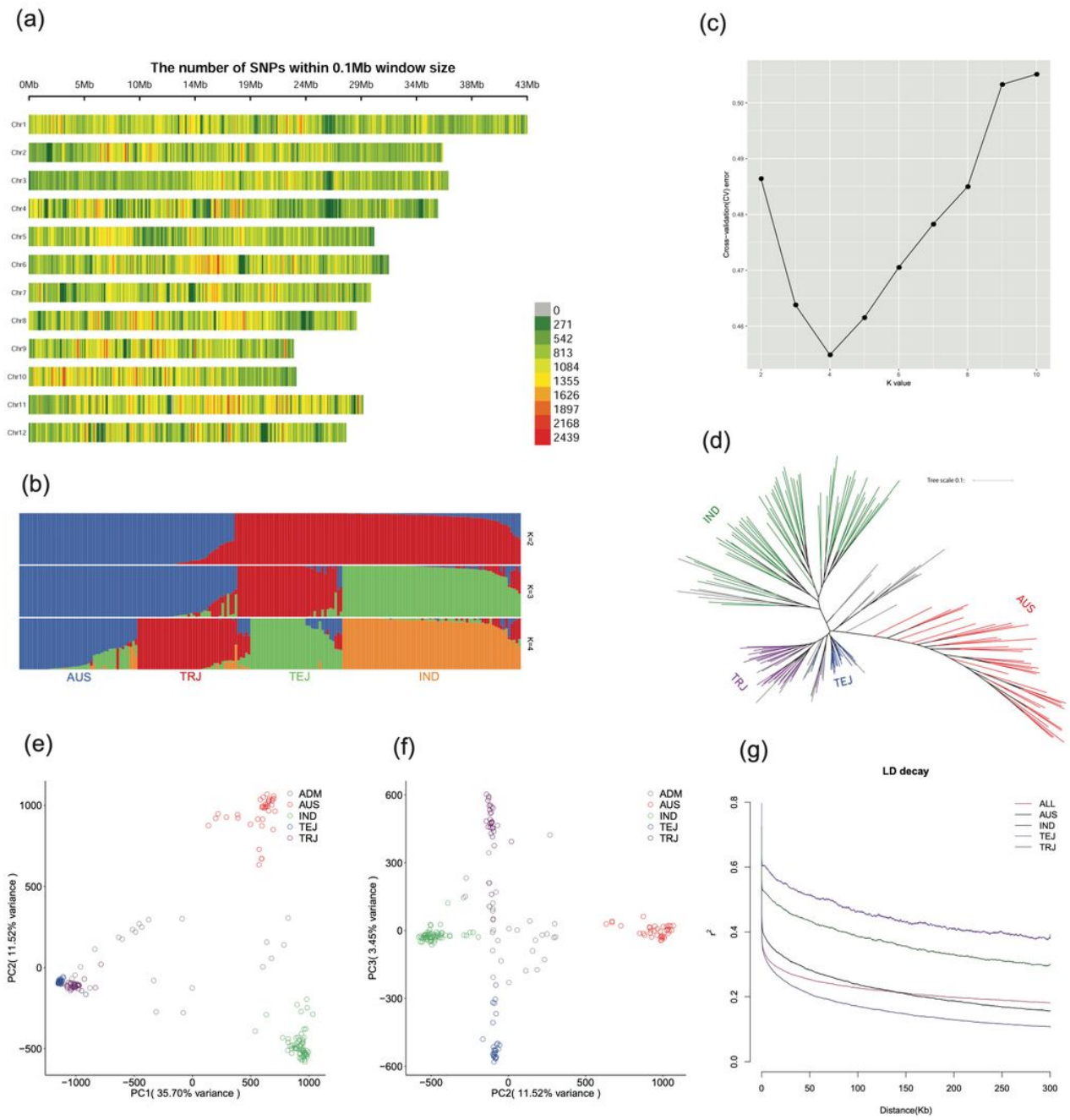

Figure 1

Sequence and structure analysis of USDA mini core collection. (a) Distribution of SNPs on the rice chromosomes. Number of SNPs per 0.1Mb window was shown as a color index (bottom right), (b) Ancestries analysis for each individual was inferred using admixture, (c) Cross-validation error (CV) score across different K value. The best K value (K=4) was chosen according to the lowest CV score for the admixture analysis, (d) Phylogenetic tree of 191 rice accessions. Green indicated Indica (IND) rice, Red indicated Aus (AUS) rice, Purple represented Tropical Japonica (TRJ) rice; Blue represented Temperate

Japonica (TEJ) rice, (e) PCA showing genetic variation in the rice accessions with first and second PCs, the color was defined by current Admixture analysis.(f) PCA showing genetic variation in the rice accessions with second and third PCs, the color was defined by current Admixture analysis. (g) Genome-wide average LD decay estimated from the whole population and each subpopulation.

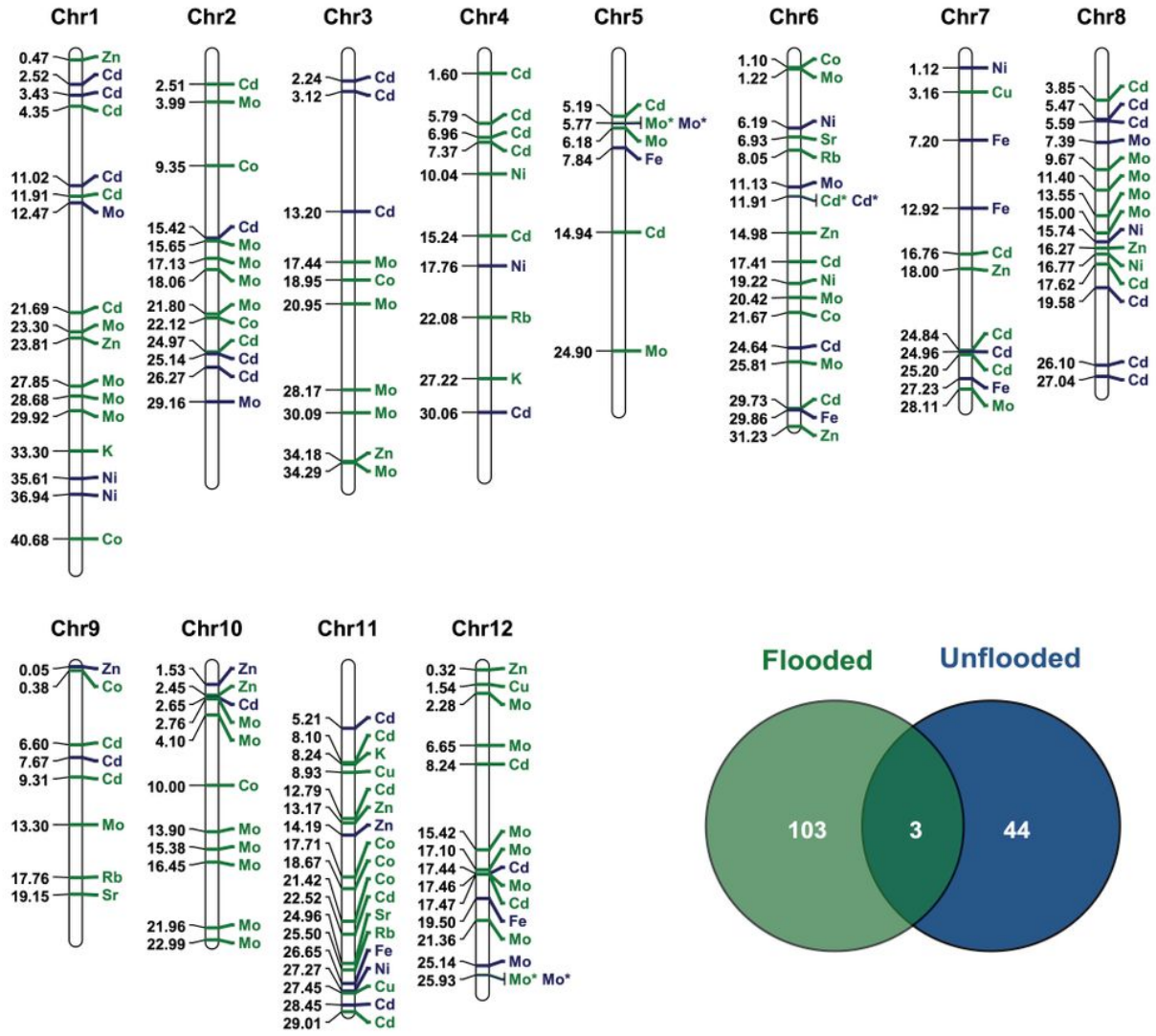

Figure 2

QTLs related to ionomic traits. (a) Distribution of significant QTLs for ionomic traits across the 12 chromosomes of rice under flooded and unflooded environment. Leading SNP was mapped to the chromosome to represent the QTLs' physical location (Mb). The physical position of each lead SNP was shown on the left side and the corresponding ionomic traits displayed on the right side. QTLs from different growth conditions were distinguished by different colors: green, flooded condition; blue, unflooded condition. An asterisk indicates the locus which was detected from both conditions, (b) The Venn diagram shows the numbers of overlapped loci within or between different conditions.

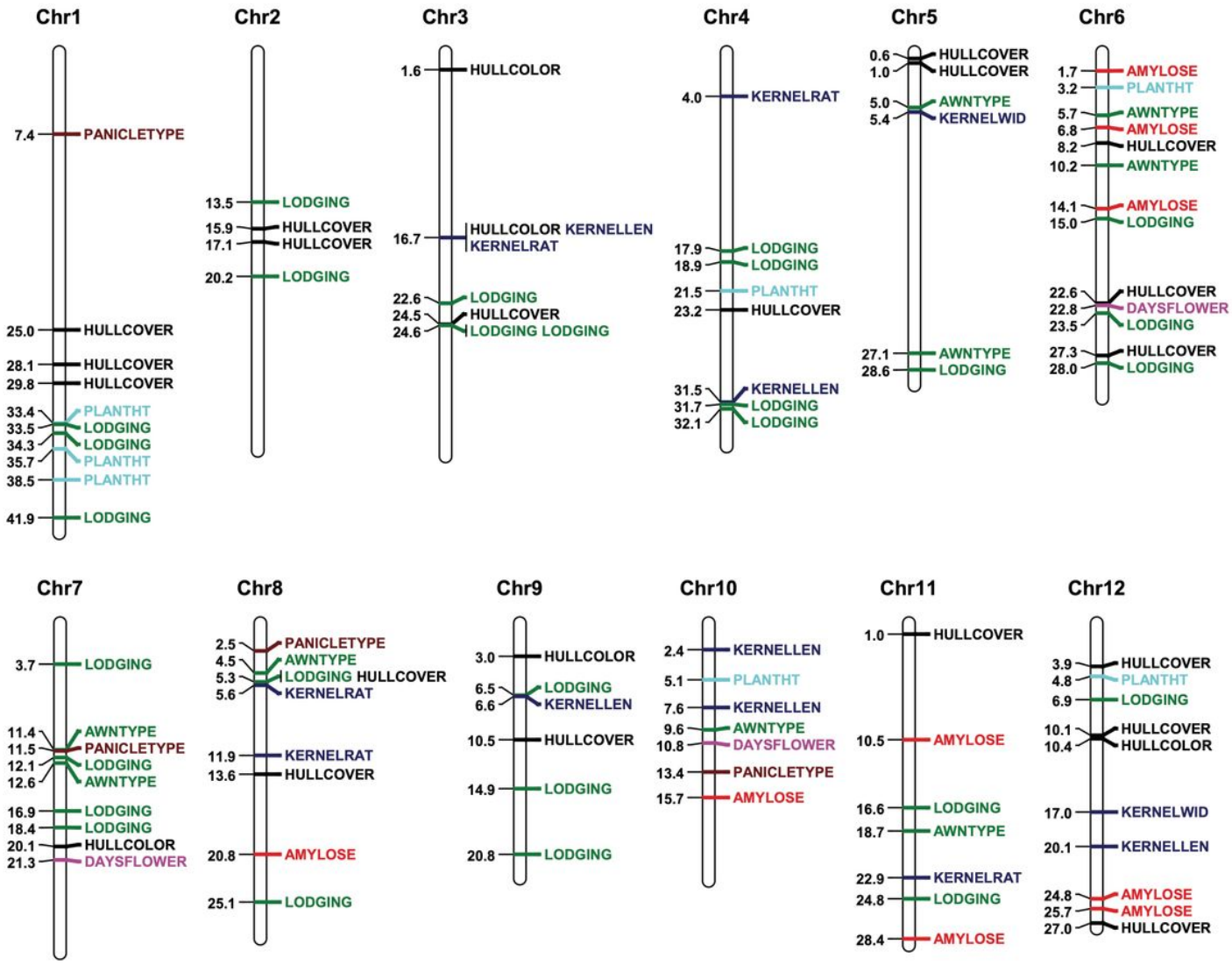

Figure 3

Distribution of significant QTLs for agronomic traits on 12 Chromosomes. Leading SNP was mapped to the chromosome to represent the QTLs' physical location (Mb). The physical position of each lead SNP was shown on the left side and the corresponding agronomic traits displayed on the right side. QTLs of different type of agronomic traits were distinguished by different colors: red, amylose; blue, grain size (kernel length, kernel width, and kernel rate); black, hull cover and hull color; purple, days to flower; brown, panicle type; green, lodging and awn type.

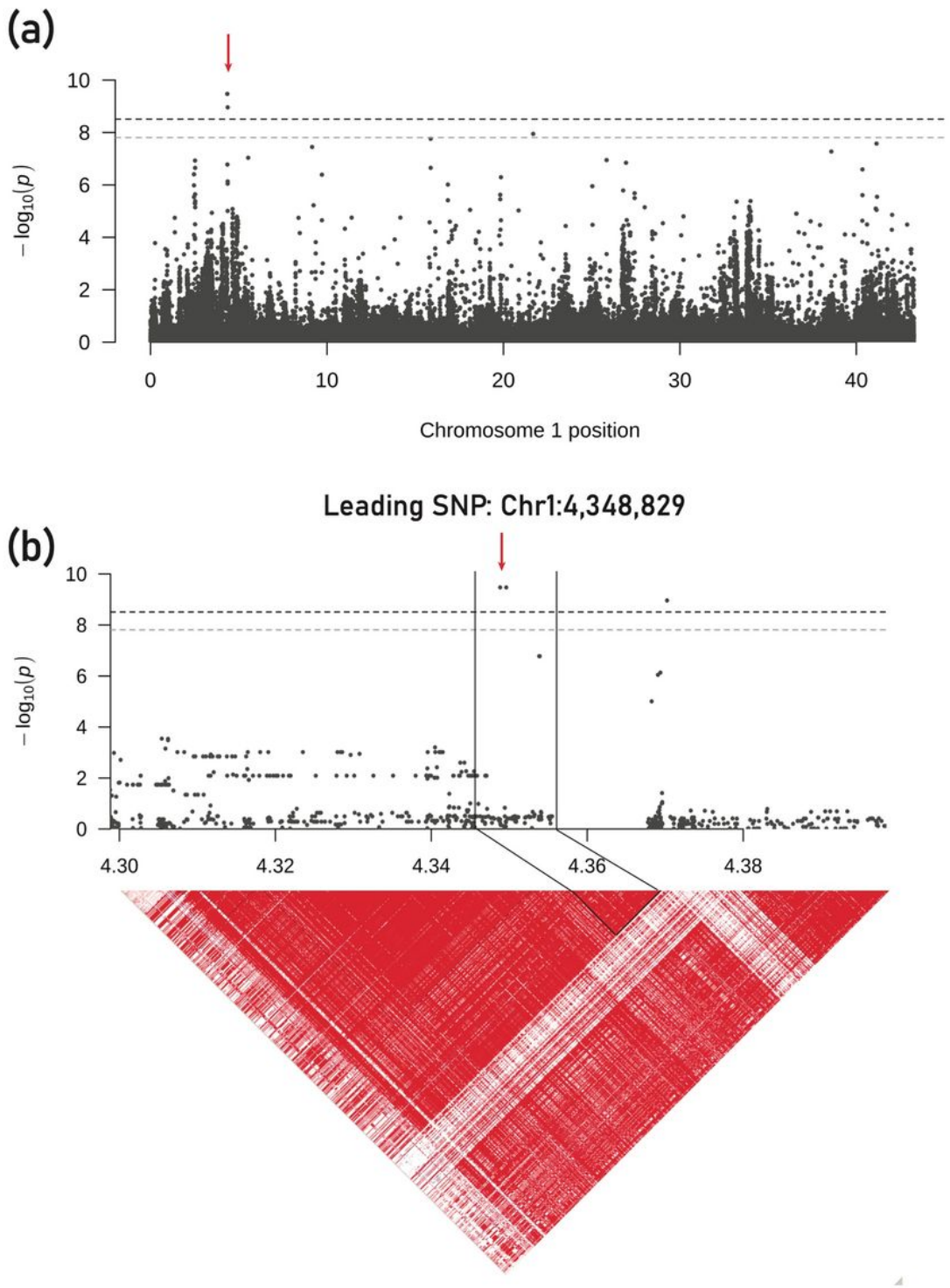

Figure 4

Identification of OsWRKY102 as a Cadmium concentration QTL in rice grain in flooded condition using MLM method. (a) Genome-wide association signals on chromosome 1, (b) Genome-wide association signals in the region at 3.29 - 4.39 Mb on chromosome 1 and LD heatmap (bottom).

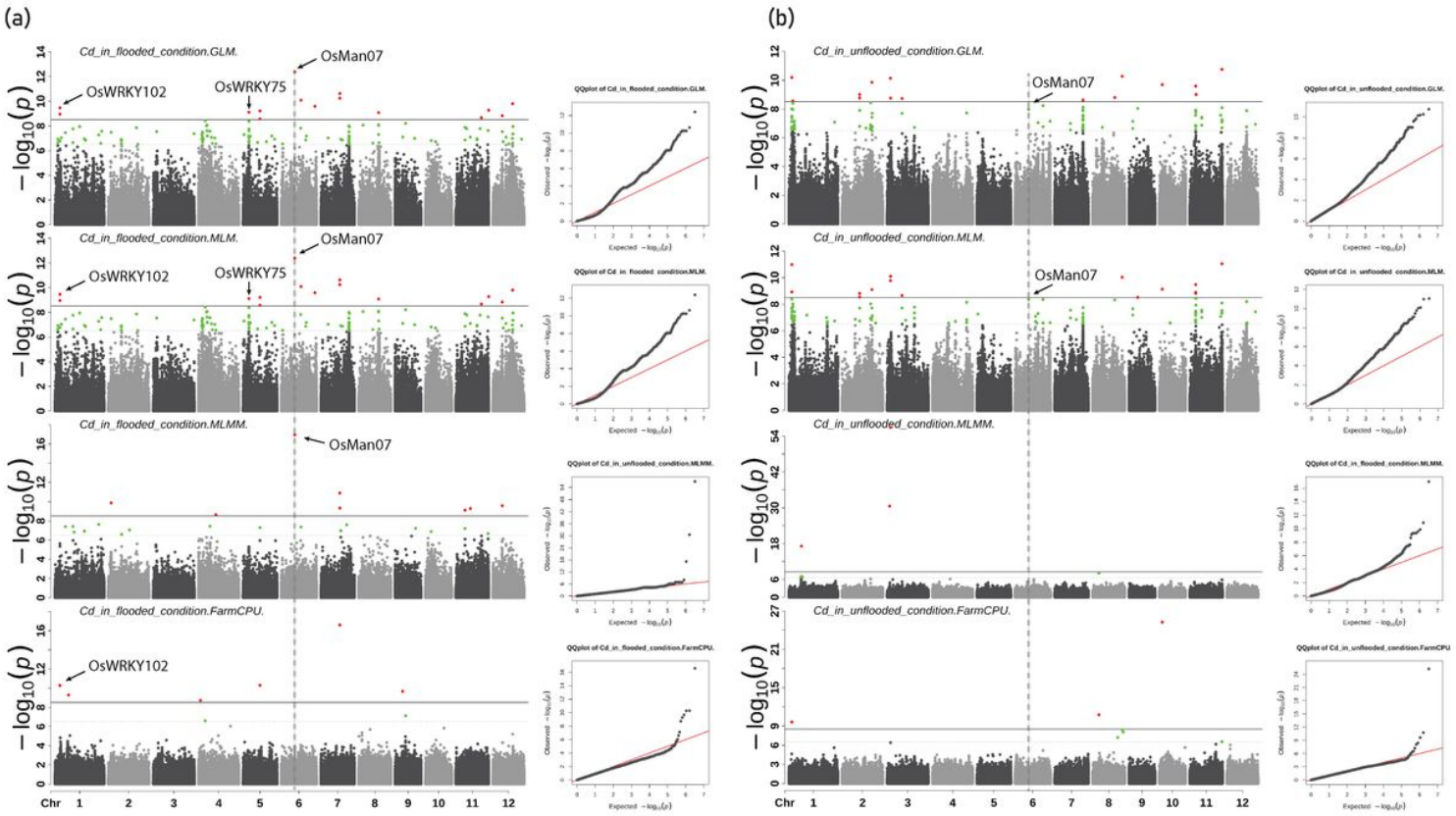

Figure 5

Genome-wide association analysis of Cd concentration with GLM, MLM, MLMM, and FarmCPU methods. (a) in flooded condition and, (b) unflooded condition. Quantile-quantile plot of each model. Black arrows indicated candidate genes. The horizontal dot grey line and green dots indicated the Bonferroni-corrected significance thresholds and SNPs at -log10(p) = 7.81. The horizontal solid grey line and red dots indicated the Bonferroni-corrected significance thresholds and SNPs at -log10(p) = 8.51. The vertical dash grey lines indicate the common QTL detected in flooded and unflooded condition.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.xlsx
- SupplementaryFigureS2.pdf
- SupplementaryTableS5.xlsx
- SupplementaryFigureS1.pdf
- SupplementaryFigureS3.pdf
- SupplementaryFigureS5.tif
- SupplementaryTableS3.xlsx
- SupplementaryFigureS6.tif

- SupplementaryFigureS7.tif
- SupplementaryFigureS4.tif
- SupplementaryTableS4.xlsx
- SupplementaryTableS2.xlsx