the **genetics**society

ARTICLE

Analysis of trait heritability in functionally partitioned rice genomes

Julong Wei[1,2] · Weibo Xie [ID][3] · Ruidong Li[4] · Shibo Wang [ID][4] · Han Qu[4] · Renyuan Ma[4,5] · Xiang Zhou[2] · Zhenyu Jia [ID][4]

Abstract

Knowledge of the genetic architecture of importantly agronomical traits can speed up genetic improvement in cultivated rice (*Oryza sativa* L.). Many recent investigations have leveraged genome-wide association studies (GWAS) to identify single nucleotide polymorphisms (SNPs), associated with agronomic traits in various rice populations. The reported trait-relevant SNPs appear to be arbitrarily distributed along the genome, including genic and nongenic regions. Whether the SNPs in different genomic regions play different roles in trait heritability and which region is more responsible for phenotypic variation remains opaque. We analyzed a natural rice population of 524 accessions with 3,616,597 SNPs to compare the genetic contributions of functionally distinct genomic regions for five agronomic traits, i.e., yield, heading date, plant height, grain length, and grain width. An analysis of heritability in the functionally partitioned rice genome showed that regulatory or intergenic regions account for the most trait heritability. A close look at the trait-associated SNPs (TASs) indicated that the majority of the TASs are located in nongenic regions, and the genetic effects of the TASs in nongenic regions are generally greater than those in genic regions. We further compared the predictabilities using the genetic variants from genic regions with those using nongenic regions. The results revealed that nongenic regions play a more important role than genic regions in trait heritability in rice, which is consistent with findings in humans and maize. This conclusion not only offers clues for basic research to disclose genetics behind these agronomic traits, but also provides a new perspective to facilitate genomic selection in rice.

Introduction

It is crucial to dissect the genetic architecture of the complex traits for understanding the relationship between DNA and these traits and for benefiting genomic selection (GS) or

✉ Zhenyu Jia
zhenyuj@ucr.edu

1   College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, China

2   Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

3   National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei, China

4   Department of Botany & Plant Sciences, University of California (Riverside), Riverside, CA, USA

5   Department of Mathematics, Bowdoin College, Brunswick, ME, USA

whole-genome prediction (WGP). Genome-wide association study (GWAS) has been widely used for the identification of single nucleotide polymorphisms (SNPs), or other DNA markers, that are significantly associated with the traits of interest. It is common in GWAS that each individual trait-associated SNP (TAS) explains a very small percent of the phenotypic variation, whereas the aggregate of the effects of all associated SNPs can account for 15–45% trait variability (Manolio et al. 2009). The so-called lost trait heritability is likely ascribed to a large number of loci with small effects, many of which cannot be detected with conventional statistical approaches. To consider the loci with minor but nontrivial effects, the linear mixed models (LMMs) with a polygenic effect have been proposed to use the entire set of genomic markers, yielding increased trait heritability in GWAS or improved trait predictability in WGP (VanRaden 2008; Yang et al. 2010). In the LMMs, the genetic contributions, denoted by regression coefficients, of all the markers are treated as random effects and are assumed to follow the same distribution (Speed et al. 2012; Yang et al. 2010). Recent studies in human and crops have shown that trait-associated loci are enriched in regulatory regions, compared with the protein-coding exon

regions (Hindorff et al. 2009; Maurano et al. 2012; Schork et al. 2013). Moreover, important genetic matter, such as miRNAs and long noncoding RNAs (lncRNAs), was found to be produced in the noncoding genomic regions, suggesting a biological importance of nongenic regions in genomes (Edwards et al. 2013). Maurano et al. (2012) reported that about 93% of detected variants are located in noncoding regions, and 57.1% of these nongenic variants are positioned in DNaseI hypersensitivity sites (DHSs)—an essential gene-regulatory region. This is in agreement with other studies which also suggested that trait-associated loci tend to cluster in gene-regulatory regions, including promoter regions and transcriptional enhance elements (Hindorff et al. 2009; Schork et al. 2013). The accumulation of evidences suggested that regulatory regions in genomes generally play a more important role in dictating the phenotypes than genic regions. To compare the roles of various types of the genomic regions, analysis of heritability (analogy to ANOVA) may be performed to compare the contributions from these genomic partitions and allow for distinct distributions when LMM is used to model DNA makers in different genomic regions. For example, the functional enrichment assay showed that DHS regions account for the most heritability in human genomes (Gusev et al. 2014). Similarly, a Maize study revealed that most trait-associated variants lie in nongenic regions, which are responsible for the regulations over gene expression (Li et al. 2012). This has not been proved in rice genomes.

Rice is the most widely consumed staple food for a large part of the world population (Huang et al. 2010); therefore, understanding the roles of different parts of rice genomes will advance our knowledge of the genetic architecture of important traits and facilitate genetic improvement in breeding. Rice is an ideal model for the analysis of heritability in functional partitions of the genomes because of the following reasons: (1) the size of rice genomes is relatively small (~400 MB), (2) the high-quality reference genome with functional annotations is available (Sasaki 2005), (3) millions of genomic markers can be genotyped with cost-effective sequencing technology, (4) agronomic traits have been accurately gauged by repeated measurements in multiple projects, and (5) the major genetic determinants of agronomic traits have already been characterized (Huang et al. 2010, 2016, 2012; Yano et al. 2016). This is the first analysis of heritability on rice genomes, where we investigated five agronomic traits, i.e., yield (YD), heading date (HD), plant height (PH), grain length (GL), and grain width (GW), using a diverse population including 524 lines with 3 million SNPs. The results of the analysis of heritability based on all SNP markers showed that nongenic regions consistently account for more genetic variation in five traits than genic regions. Similar comparisons between nongenic regions and genic regions were also

performed by only using the SNPs identified by GWAS, yielding the same conclusion as the analysis of heritability. We also compared the predictions of traits using selected SNPs from genic and nongenic regions, respectively, which showed that nongenetic regions had better predictabilities than genic regions. We conclude that DNA variants in nongenic regions play a more important role than those in genic regions in determining the agronomic traits of rice. Future in-depth research in regulatory regions of rice genomes will uncover the mysteries of genetics behind these traits and eventually benefit genetic improvement in breeding.

Materials and methods

Rice population and sequencing data

In the study, we considered a diverse rice collection of 533 accessions, including 200 varieties from a core/minicore collection of *O. sativa* L. in China (Zhang et al. 2011), 132 parental lines used in the international Rice Molecular Breeding Program (Yu et al. 2003), 148 lines from a minicore subset of the US Department of Agriculture rice gene bank (Yan et al. 2009), 18 lines used for SNP discovery in the OryzaSNP project (McNally et al. 2009), and 35 lines from the Rice Germplasm Center at the International Rice Research Institute, which represent landraces and elite varieties. A total of 524 lines have been selected for the study, since the record of their phenotype and genetic variants is available. Five agnomically important traits, i.e., YD, HD, PH, GL, and GW, were considered in the study.

About 6.7 billion 90-bp paired-end reads have been generated using the Illumina HiSeq 2000 platform (Chen et al. 2014). Missing genotypes have been imputed using the genotype data of another set of 950 rice lines from Huang et al. (2012). Aligned with the rice reference genome (Nipponbare, MSU version 6.1), a total of 6,551,358 high-quality SNPs were obtained, with the minor alleles being present in at least five accessions. To simplify the computation, the SNP markers with missing data were eliminated, leaving a total of 3,616,597 SNPs in the analysis. The SnpEff software developed by Cingolani et al. (2012) was utilized for SNP annotation.

Genome partitioning and analysis of heritability

Following Speed et al's. study (Speed and Balding 2014), an LMM is used to decompose the genetic variances for a trait, i.e.,

$$y = X\beta + \sum_i^M g_i + \varepsilon, \tag{1}$$

where y is an $n \times 1$ vector representing phenotypic values with n equal to the sample size, β is the fixed effect, X is the design matrix for fixed effects (such as locations, years, etc.), g_i represents the polygenic effect of ith genomic partition type, and ε denotes the residuals which are normally distributed, i.e., $N(0, I\sigma^2)$, with I being the identity matrix and σ^2 being the variance of the residual. In this study, we sorted the rice genomes into four types of partitions, including regulatory regions, intron regions, exon regions, and intergenic regions ($M = 4$ in Eq. (1)). Different distributions are assumed for the polygenic effects in different genomic partition types, i.e., $g_i \sim N(0, K_i\phi_i^2)$, where K_i is the kinship matrix calculated using the SNPs in the ith genomic partition type and ϕ_i^2 is the genetic variance shared by the SNPs in this genomic partition type (Gusev et al. 2014). The heritability from Eq. (1) is calculated as the ratio of the sum of the total genetic variance to phenotypic variance, i.e., $h^2 = \sum_{i=1}^{M=4} \phi_i^2 / \left(\sum_i \phi_i^2 + \sigma^2\right)$. The restricted maximum likelihood method may be used to estimate the variance parameters $\phi_1^2, \phi_2^2, \phi_3^2, \phi_4^2$, and σ^2, which was similar to our previous studies (Wei et al. 2018).

The LDAK software was used for the analysis of heritability, including the estimation of genetic variance and heritability for each genomic partition type and the calculation of standard errors for the estimated parameters (Speed and Balding 2014). The enrichment score for each genomic partition type was calculated as the ratio of the percentage of the heritability explained by the SNPs in this type of genomic partitions to the percentage of SNPs falling in this partition type (the expected percentage of the heritability by these SNPs). The z-statistic was adopted to test the significance level of enrichment for each genomic partition type based on Gusev et al. (2014). The z-statistic was calculated as the difference between the percentage of h_{gi}^2 and the percentage of SNPs in the ith category divided by the analytical standard error (SE).

In addition, a simulation-based test was applied to evaluate the significance level of enrichment by contrasting the observed likelihood ratio test (LRT) statistic with the empirical NULL distribution of LRT for a calculation of a p value. The observed LRT is calculated using the below formula:

$$LRT = -2\left(L_0\left(\widehat{\beta}, \widehat{\phi}^2, \widehat{\sigma}^2\right) - L_1\left(\beta, \widehat{\phi}_1^2, \widehat{\phi}_2^2, \widehat{\phi}_3^2, \widehat{\phi}_4^2, \widehat{\sigma}^2\right)\right), \tag{2}$$

where ϕ^2 in the NULL hypothesis (L_0) is assumed to be identical across four genomic partitions. For any NULL model, we randomly partitioned the entire set of SNPs into four categories of the same sizes as the originally defined functional partition types, followed by the calculation of a LRT of the NULL model using exactly the same steps by which the observed LRT was calculated. This process was

repeated 100 times to generate a NULL distribution of the LRT, of which the 5th percentile was used for testing the NULL hypothesis that various partition types equally contribute to the phenotypic variability.

Genome-wide association study

The GWAS method with LMM, shown as follows, was used for the identification of trait-associated loci (Yu et al. 2006; Zhang et al. 2005),

$$y = X\beta + Z_k u_k + \xi + \varepsilon, \tag{3}$$

where Z_k is the genotype of the kth SNP, u_k is the fixed effect for the kth SNP, and ξ is a random effect to control the polygenic background with a multivariate normal distribution $N(0, K\phi^2)$. Note that the kinship matrix K in the GWAS analysis was calculated using the entire set of SNPs for simplicity; therefore, K did not vary when different SNPs were analyzed using the GWAS model. Other elements in Eq. (3) remain the same as Eq. (1). The LRT was used to test each SNP marker, i.e.,

$$LRT = -2\left(L_0\left(\widehat{\beta}, \widehat{\phi}^2, \sigma^2\right) - L_1\left(\widehat{\beta}, u_k, \widehat{\phi}^2, \sigma^2\right)\right), \tag{4}$$

where L_0 denotes the likelihood of the Null model (without the estimated effect of the SNP marker $\widehat{u}_k$) and L_1 denotes the likelihood of the model under evaluation (with $\widehat{u}_k$). Eigen decomposition was used for handling the algebra that involved the kinship matrix to increase the computational efficiency (Kang et al. 2008).

Genomic best linear unbiased prediction (GBLUP)

BLUP has been widely employed to predict breeding values in animals, when pedigree data are available for the calculation of kinship (VanRaden 2008). The kinship can be deduced using whole-genome variant data, yielding a more powerful method, i.e., GBLUP, for the prediction of breeding values in animals and plants. We use subscripts (0) and (1) to represent the training set and the validation set, respectively, in GBLUP. The estimated breeding value for the validation set $\widehat{g}_{(1)}$ can be described as

$$\widehat{g}_{(1)} = \sum_{i=1}^{M} \widehat{g}_{i(1)} = \sum_{i=1}^{M} K_{i(10)}\widehat{\phi}_i^2 \left(\sum_{i=1}^{M} K_{i(00)}\widehat{\phi}_i^2 + I\widehat{\sigma}^2\right)^{-1}\left(y_{(0)} - X_{(0)}\widehat{\beta}\right), \tag{5}$$

where $\widehat{g}_{i(1)}$ is the estimated breeding component for the ith genomic partition type, $K_{i(10)}$ represents the genetic covariance between the validation set and training set when the SNPs in the ith genomic partition type are considered, and $\sum K_{i(00)}\widehat{\phi}_i^2 + I\widehat{\sigma}^2$ is the $Var\ (y_{(0)})$ for the training set. The parameters, including $\widehat{\phi}_1^2, \ldots, \widehat{\phi}_M^2, \widehat{\sigma}^2$ and $\widehat{\beta}$, are
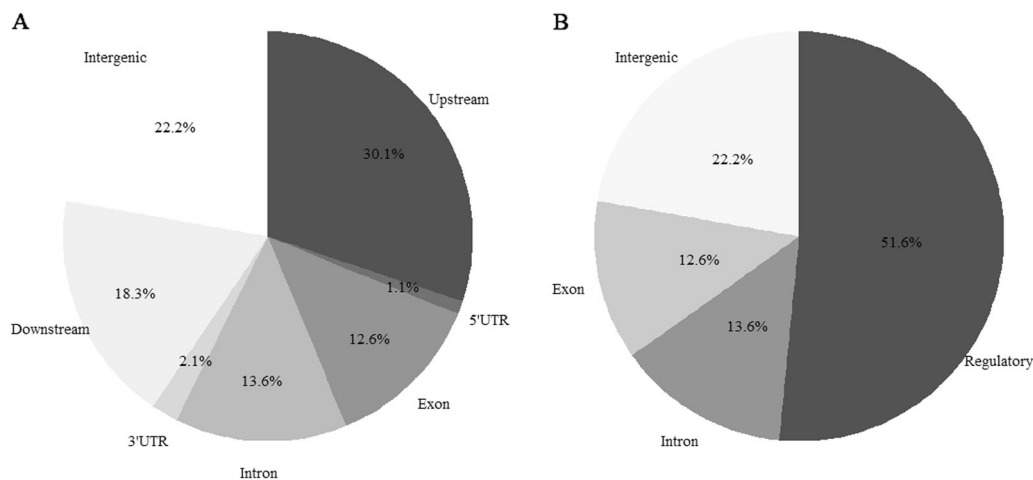
Fig. 1 Categories of 3 million of SNPs obtained based on second-generation sequence platform. **a** Distributions of the seven categories directly obtained from SnpEff software and **b** distributions of the four categories, combining upstream, downstream, 5′UTR, and 3′UTR into regulatory categories

estimated from the training set and then used for genomic prediction for the validation set. Such model is analogous to MultiBLUP by Speed et al., since multiple kinship matrices are involved in BLUP (Speed and Balding 2014), in contrast to the GBLUP models, where only one kinship matrix is used. The predicted phenotypic value for individuals in the validation sets is equal to the sum of the estimated breeding value plus fixed effects, that is, $\widehat{y}_{(1)} = X_{(1)}\beta + \widehat{g}_{(1)}$. The predictability of trait is defined as the squared correlation coefficient between the observed phenotypic values and the predicted phenotypic values (Xu et al. 2016; Xu et al. 2014), as shown in the following formula:

$$r^2 = \left(\frac{\text{cov}(y, \widehat{y}_1)}{\sqrt{\text{var}(y)\text{var}(\widehat{y}_1)}} \right)^2 \tag{6}$$

The tenfold cross-validation was implemented to calculate trait predictability for model evaluation or comparison among different genomic partition types. The data sets are equally partitioned into ten portions, in which we predicted the phenotypic value in one portion (validation) using the other nine portions (training). Each individual has a predicted value after all the portions are used as a validation set. Then we calculated the predictability as formula (6).

Results

Summary of single-nucleotide polymorphism of rice genomes

In the study, about 6.5 million high-quality SNPs distributed in the 12 chromosomes were considered (Chen et al. 2014). The SNPs with missing genotypes were eliminated, leaving a total of 3,616,597 SNPs with about 9600 SNPs per million bases on average. Since the rice genome is relatively small, such SNP density is sufficient to study the differences in heritability between functional partitions of rice genomes. Figure S1 shows that SNPs are linearly arranged in each chromosome, indicating that SNP markers are evenly distributed along the entire genome and functional partitions of the genome are equally represented in the study.

We used SnpEff tool to annotate the genetic variants based on rice reference genome (MSU version 6.1). The genome-wide variants are annotated into a total of 28 categories, the annotation for which is shown in Table S1. The 28 categories were further classified into seven genomic regions including intergenic, upstream, 5′UTR, exon, intron, 3′UTR, and downstream (Fig. 1a). Since the regulatory elements are mostly positioned in the upstream, downstream regions, and flanking untranslated regions (UTRs), we therefore further reclassified these seven genomic regions into the final four functional partitions, i.e., regulatory partition (upstream, downstream, 5′UTR, and 3′UTR), intron partition, exon partition, and intergenic partition (Table S1). Four categories are redefined based on the variant function: (1) regulatory partition, including 1,864,917 variants, about 51.6% percent of total markers, plays role in the regulation of gene expression; (2) intron partition, consisting of 491,680 variants (13.6%), represents the regions mingled with exons in the open-reading frame (ORF) and is spliced out during post transcriptions; (3) exon partition, composed of 456,341 variants (12.6%), denotes the regions that are retained in the mature mRNAs and eventually translated into protein; (4) Intergenic partition, represented by 803,659 variants (22.2%), is made up of noncoding DNA, excluding the regulatory partition (refer to Fig. 1b).

Table 1 Joint estimation of genetic variance components across the four functional regions based on linear mixed model (LMM)

Trait	Regulatory	Intron	Exon	Intergenic	Residual	Heritability
YD	0.170 ± 1.690[a]	0.000 ± 1.000	0.000 ± 1.000	0.457 ± 1.720	0.301	0.676
HD	1.142 ± 11.420	0.000 ± 1.000	0.136 ± 10.130	0.000 ± 1.000	0.053	0.960
PH	0.730 ± 1.662	0.000 ± 1.000	0.000 ± 1.000	0.000 ± 1.000	0.100	0.880
GL	0.181 ± 4.794	0.000 ± 1.000	0.000 ± 1.000	0.818 ± 5.083	0.113	0.898
GW	0.000 ± 1.000	0.000 ± 1.000	0.000 ± 1.000	0.626 ± 1.597	0.096	0.867

[a]Denotes the standard error of genetic variance components of traits

Comparison of genetic effects across functional categories

For each trait, we estimated genetic variations of the four functional categories, jointly using the LMM, where the kinship matrix for each category was calculated using the data of SNPs in that genome partition. Table 1 presents the results of the analysis of five agronomic traits, i.e., YD, HD, PH, GL, and GW. It appears that (1) genetic variation was mainly ascribed to regulatory and intergenic regions for YD and GL, (2) regulatory regions seemed to account for the most genetic variations in HD or PH, and (3) the associated genetic variation was only detected in intergenic regions in GW. Little genetic association (approximate to zero) has been identified for genetic regions including introns and exons. These results indicated that the variations of these five traits are mostly governed by the DNA variants outside gene regions rather than the variants located within genes. We also provided standard error of heritability of each functional category in Table S2, which will be used in the significant test of enrichment analysis. A three-component population structure has been disclosed by PCA (Fig. S3). We therefore reanalyzed the data by incorporating the population structure in the regression models as covariates, but there was no difference between the models with and without population structure (Table S3 vs. Table 1). We also removed individuals with strong relatedness based on cluster analysis (h parameters setting to 100 in cutree function) and the remaining 452 individuals were used to perform an analysis of partitions of heritability across functional annotations. The results are summarized in Table S4, which are very similar to the results using the entire sample.

For each trait, we calculated the value of the logarithm of likelihood based on the estimated parameters, including five variance parameters, i.e., $\widehat{\phi}_1^2$, $\widehat{\phi}_2^2$, $\widehat{\phi}_3^2$, $\widehat{\phi}_4^2$, and $\widehat{\sigma}^2$, which are marked with red triangles in Fig. S2. We also calculated logarithm of likelihood values of NULL-equivalent models by using various values, from 0 to 1.2, with an incremental step of 0.08, for these five variance parameters. Since the phenotypic values for each trait have been standardized with a standard normal distribution (N(0,1)), $\sum_{i=1}^{4} K_i \widehat{\phi}_i^2 + I\widehat{\sigma}^2$ is supposed to 1; however, the actual value of phenotypic variance may vary, so we constrained it in a bracket of 0.5–1.2. For each trait, the distribution of 14,587 values of the logarithm of likelihood for the NULL-equivalent models is shown in Fig. S2, indicating that the optimal results were always achieved by the estimated values of these five variance parameters.

The enrichment analysis (Gusev et al. 2014) was then performed on each of five traits, respectively, and the enrichment scores were computed for each genomic partition as the ratio of the observed proportion of heritability to the expected proportion of heritability, which is defined as the percentage of SNPs in this category of regions. For five traits, either regulatory or intergenic regions exhibited high enrichment in heritability (>1), whereas the exon and intron regions were significantly depleted (Fig. 2). For example, variants in intergenic regions for GW were highly enriched (an enrichment score of 4.5×) compared with the expected proportion of heritability, and explained ~100% of trait heritability (SE = 1%). Following the methods described in Gusev et al. (2014), we also performed a z-test on the enrichment score for each category, which suggested that the regulatory partition in PH and the intergenic partition in GW were extremely significant ($p < 0.01$). The remaining functional categories did not show significant enrichment, due to the large variance of estimated heritability values. The results of simulation indicated that the LRT (or likelihood) calculated from the defined functional categories is significantly larger than that of NULL models with random genomic partitions (Fig. S4).

Genome-wide association study

For each of the five traits, we examined the distribution of TASs in the four genomic categories through GWAS. The Manhattan plots in Fig. 3 showed that HD, GL, and GW are controlled by several major QTLs (peaks in the plots), whereas the variabilities of YD and PH are ascribed to many modest genetic loci (or polygenic effects) without obvious peaks. Because inclusion of test genetic variants in the kinship matrix would cause the loss of detection power, we carried out analysis of GWAS by leaving one chromosome out to boost detection power, the results of which are shown in Fig. S5. There are not significant differences in the
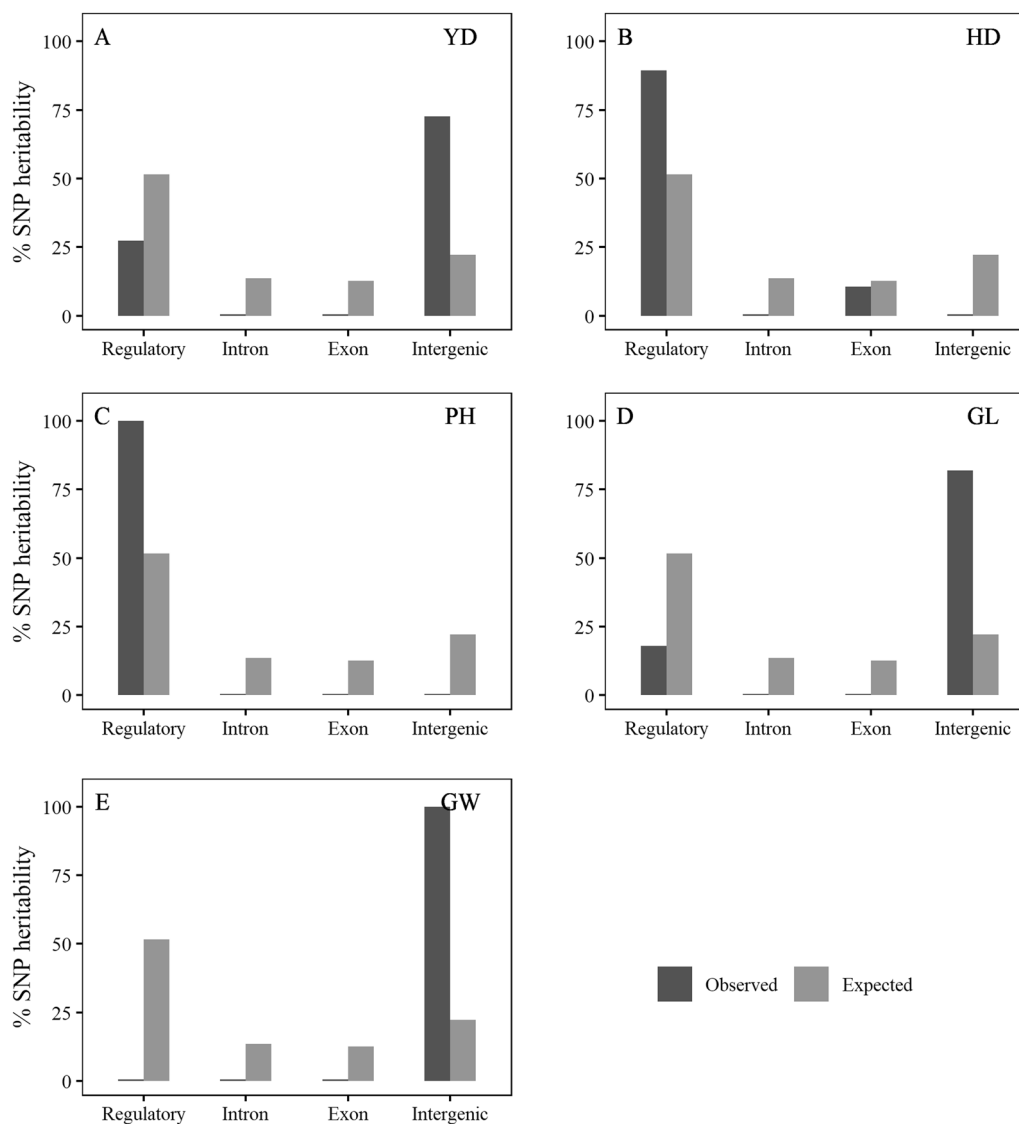
Fig. 2 Partition SNP heritability of the five agronomic traits according to functional annotations. The *y*-axis represents what proportion the functional category can account for the total genetic variation, black bars for the estimated value, and gray ones for the expectation value. **a–e** are corresponding to five traits, YD, HD, PH, GL, and GW, respectively

pattern of association between the results of two methods. Candidate genes near the major peaks that were identified for HD, GL, and GW are listed in Table 2. Important biological functions of these genes have been previously reported and some of them agree with the traits in the study. The gene *Hd3a*, which is a rice ortholog of *Arabidopsis FT* gene and plays a key role in the regulation of flowering through the multiple-signal pathway (Bian et al. 2011; Kim et al. 2008; Kojima et al. 2002), has been identified at the downstream of the HD-associated SNP (chromosome 6, 3.484 Mb). The two novel genes *OsGSTZ2* and *OSIPK* have been identified in the region near a HD-associated SNP (chromosome 12). However, neither gene has been reported to be a functionally relevant HD; rather, the gene *OsGSTZ2* has been reported to relate to seedling cold tolerance (Kim

et al. 2011), while silencing the *OSIPK* gene can cause a substantial reduction in phytate levels in seeds (Ali et al. 2013). The gene *GS3* (*GRAIN SIZE 3*), a major QTL that negatively controls rice length and weight (Fan et al. 2006; Mao et al. 2010), has been detected in proximity to the SNP (chromosome 3, 16.704 Mb) that is significantly associated with GL. The *GW5* gene is a major QTL that controls rice gain width and weight (Shomura et al. 2008; Wan et al. 2008). We have found the *GW5* gene near the SNP in chromosome 5 (5.373 Mb), which was significantly associated with GW. A new candidate gene *OsFCA* for GW has been discovered in chromosome 9 and this gene is homologous to the *Arabidopsis* flowering time gene (*FCA*) (Du et al. 2006). Previous research indicated that *OsFCA* is involved in the autonomous flowering pathway in rice (Jang
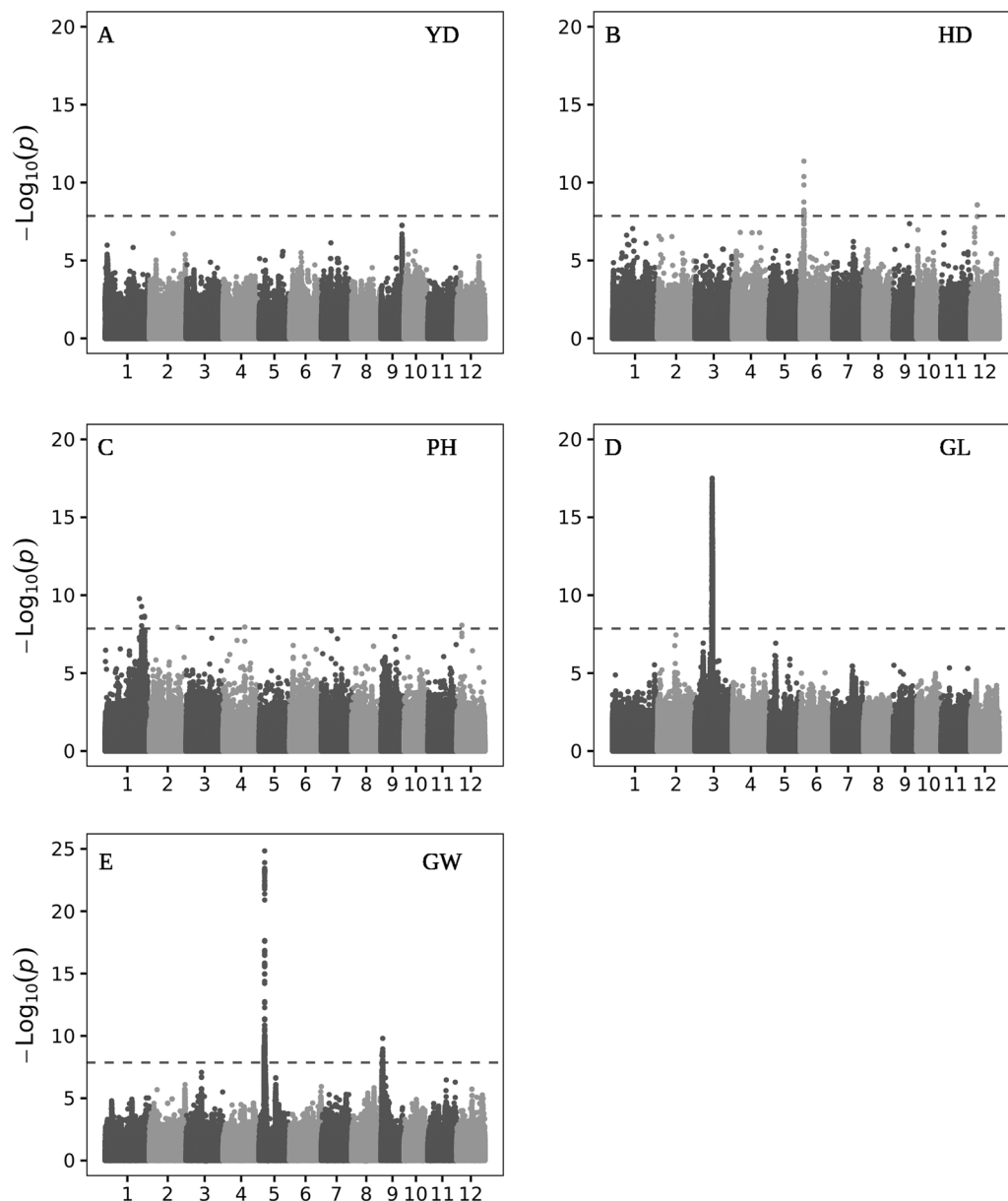
Fig. 3 Manhattan plots of the five agronomic traits from the GWAS results, where the *y*-axis is the logarithm of the *p* value and the horizontal dashed lines are the critical value at the 0.05 level after Bonferroni correction, $-\log_{10}(0.05/3{,}616{,}597) \approx 7.86$. **a–e** represent the results of YD, HD, PH, GL, and GW, respectively

et al. 2009) rather than contributing to grain shape. Further investigation is warranted to discover the broader functions of these new candidate genes for these agronomic traits in rice.

Intergenic regions and regulatory regions are treated as nongenic partition, while introns and exons are regarded as genic partition. We compared the distributions of TASs between these two partitions. We selected TASs based on *p* values using six different threshold values, i.e., 10^{-5}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-11}, and 10^{-12}, yielding 6280, 3811, 1723, 983, 695, and 465 TASs, respectively, for five traits in total. Out of these identified TASs with various selection

criterion, the majority (4716, 2857, 1333, 764, 552, and 392, respectively) were situated in nongenic regions. In order to demonstrate that TASs are enriched in nongenic regions, we performed a comprehensive simulated study. In each simulation, we randomly selected a number of SNPs from the entire genome, with the number equal to 6280, 3811, 1723, 983, 695, or 465, respectively, which corresponds to the number of identified TASs. We then found out how many of these randomly selected SNPs are located in nongenic regions. This process was repeated 100 times, for each of the six numbers of TASs selected using various thresholds, to form a NULL distribution of TASs sitting in

Table 2 Candidate genes related to agronomic traits in the identified regions

	Peak SNPs	Chromosome	$-log_{10}(p)$	Effects	Variance	Candidate gene
HD	3,484,349	6	11.371	−0.415	0.033	*Hd3a*
	6,428,603	12	8.567	−0.291	0.019	*OSIPK,OsGSTZ2*
GL	16,704,529	3	17.508	−0.491	0.116	*GS3*
GW	5,373,357	5	24.838	−0.388	0.072	*GW5*
	1,890,440	9	9.800	−0.539	0.062	*OsFCA*

nongenic regions (Fig. S6). The results showed that the TASs identified using various thresholds are significantly enriched in nongenic partition. We also compared the distributions of independent TASs selected using clump function of plink software. Due to less independent TASs identified by the stringent threshold values, we only focused two threshold values (10^{-5} and 10^{-6}) to identify 661 and 292 TASs for the five traits. Out of these identified TASs, the majority (499 and 216, respectively) were situated in the nongenic regions but not significantly higher than expectation values. This indicates that no enrichment was observed in genic or nongenic regions. This might be due to the loss of power if only an SNP was picked from a group of effective nongenic variants.

We then compared the genetic effects of the SNPs, represented by the genetic variances (proportional to heritability values) of the SNPs, between genic and nongenic partitions. The genetic variance is defined as $2p(1-p)a^2$ (Falconer et al. 1996), where p represents the minor allele frequency (MAF) and a is the additive effect of each marker estimated from LMM. We focused on the most significant SNPs, i.e., the top 50, 100, or 150 SNPs for each genomic partition among independent SNP sets selected using plink software. Figure 4 shows the distributions of the genetic effects of these top SNPs for each of the five traits. It appeared that the significant SNPs in nongenic partition generally had larger genetic effects than those in the genic partition. These results are supportive of our hypothesis that nongenic regions in the genome are the main source of genetic variations that account for the variability in complex traits.

Whole-genome prediction

We compared the predictive performance of different sets of SNPs for the five agronomic traits based on 100 replicates of tenfold cross-validation. First, we carried out the prediction using the BLUP method, in which genetic effects of SNPs were treated as random effects. The kinship matrix (or relatedness matrix) was inferred from all the SNPs included in the regression model, either the set of SNPs in each of the four functional partitions or any combination of these sets of SNPs. The results in Table 3 showed that SNPs in nongenic partition (regulatory or intergenic) produced better prediction than those in genic partition (intron or exon), when these four sets of SNPs were used in the BLUP model individually, but the differences between their prediction performances were not phenomenal. For example, in predicting HD, regulatory partition had the highest predictability (0.562), which was slightly but significantly higher than that of exon partition (0.544) based on t test ($p < 0.01$). The similar results were observed when genic partition (introns and exons) and nongenic (regulatory or intergenic) were compared using the BLUP, where a single kinship matrix was calculated from SNPs in two combined categories.

We also compared the predictive performance of various models using a multiBLUP framework. These models use the entire set of SNPs in the genome but different numbers of kinship matrices based on how the four functional SNP sets were grouped, yielding models with two kinship matrices (genic and nongenic), four kinship matrices, variable number of kinship matrices determined by the adaptive approach, or a single kinship matrix for all genomic SNPs. The results in Table 4 showed that although the differences between various models were not large, models with two genomic partitions yielded the best predictabilities. Conversely, adaptive strategies did not achieve the highest performance in most traits, except for GL.

We carried out an additional simulated study to demonstrate the prediction based on the SNPs in nongenic partition outperformed that based on the SNPs in genic partition. Each time, we randomly sampled 10^3, 5×10^3, or 10^4 independent SNP markers, respectively, from genic partition (model I), from nongenic partition (model II), or from the entire genome (model III—reference model), and used these selected SNPs in the prediction analysis through tenfold cross-validation. Each scenario was replicated 100 times. Figure 5 shows the relative changes in predictability for model I or model II by referencing model III, i.e., the values along the y-axis were calculated as $(R_i^2 - R_0^2)/R_0^2$, where R_i^2 was the predictability of model I/II when $i = 1/2$ and R_0^2 was the predictability of model III. Positive/Negative values along the y-axis indicate better/worse prediction performance compared with the reference model III. The results indicated that (1) model II generally had higher predictabilities than model I in most of the cases, especially when fewer markers were used for prediction. (2) Prediction
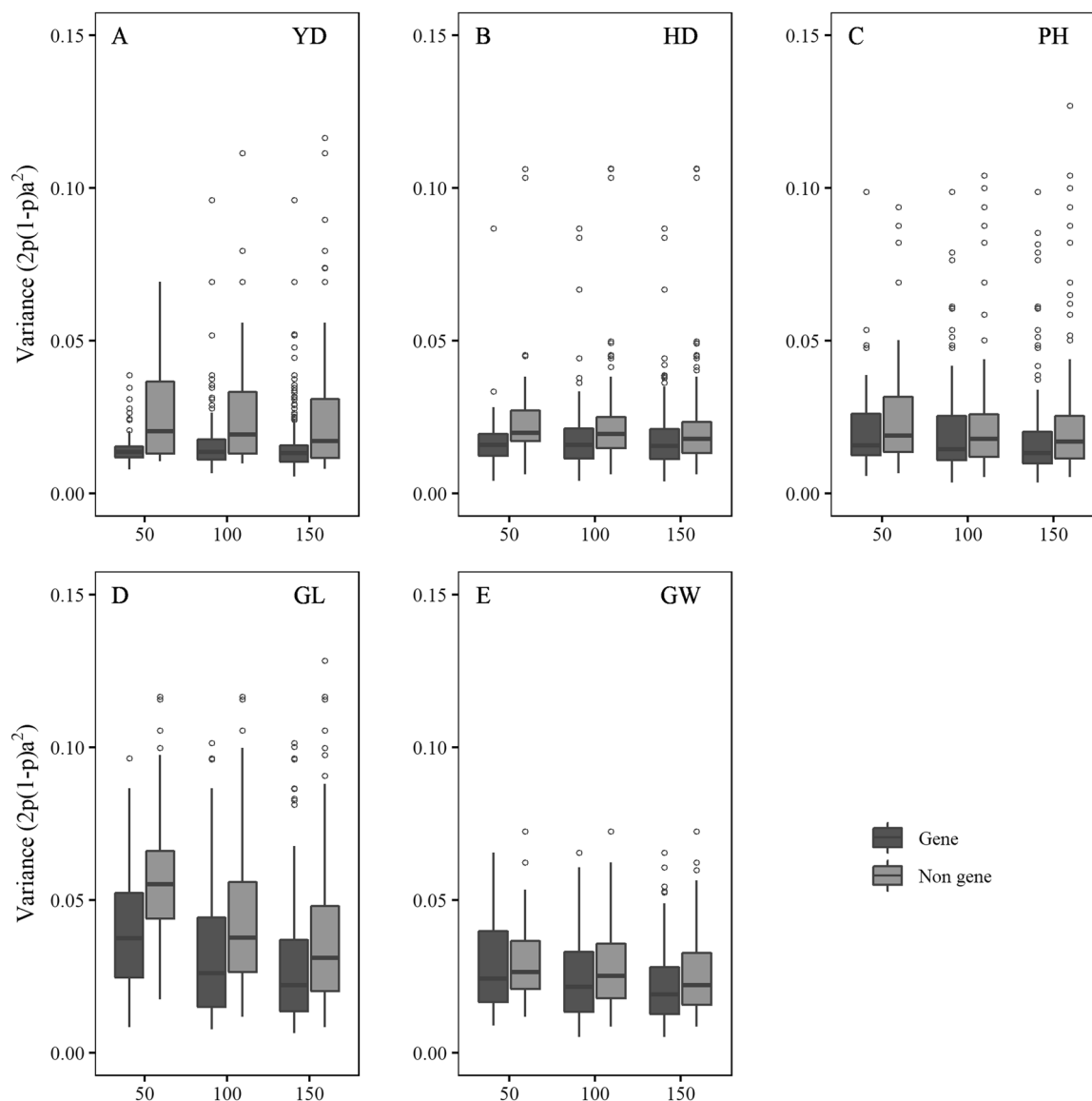
Fig. 4 Comparison of the effects of the top-ranked SNP (50, 100, and 150) in genic (black) and nongenic regions (gray). The y-axis represents how much one SNP can explain genetic variance, which is calculated based on the formula $2p(1-p)a^2$. **a–e** are corresponding to the five traits, YD, HD, PH, GL, and GW, respectively

using genic partition (model I) was not as good as that using the entire set of SNPs in the genome (model III), whereas prediction using nongenic partition (model II) outperformed the prediction using all SNPs (model III) in most of the cases. (3) The prediction performance for three prediction models (I, II, and III) became less variable and more reliable when more SNPs were included in these models.

Discussion

Mounting studies have been conducted to investigate the genetic contributions to agronomic traits across functionally partitioned genomic regions (Li et al. 2012; Xue

et al. 2016). In this study, we analyzed the trait-associated genetic variations in different partitions in rice genomes and found that nongenic regions played a key role in five agronomic traits of a natural rice population. The study population is ideal for GWAS, because this rice collection includes genetically diverse varieties from different areas, including 200 lines from a core collection in China, 132 varieties from the International Rice Molecular Breeding program, and 148 varieties from the US Department of Agricultural rice gene bank. Data for over 3 millions of high-quality SNPs are available and these SNPs are evenly distributed in the 400-Mb rice genomes (about one SNP per 130 bp), which is suitable for an analysis of heritability across genomic partitions. The analysis of

Table 3 Predictability of the five traits based on the BLUP model, just inclusion of one type of kinship matrix

SNP sets	YD	HD	PH	GL	GW
Regulatory	0.451 ± 0.008[a]	0.562 ± 0.011	0.697 ± 0.006	0.551 ± 0.007	0.693 ± 0.006
Intron	0.447 ± 0.008	0.545 ± 0.011	0.687 ± 0.008	0.545 ± 0.008	0.691 ± 0.006
Exon	0.443 ± 0.008	0.544 ± 0.011	0.692 ± 0.007	0.536 ± 0.008	0.685 ± 0.006
Intergenic	0.455 ± 0.008	0.556 ± 0.011	0.697 ± 0.006	0.552 ± 0.008	0.697 ± 0.006
Gene	0.445 ± 0.008	0.542 ± 0.011	0.687 ± 0.007	0.539 ± 0.008	0.688 ± 0.006
Non-gene	0.453 ± 0.008	0.559 ± 0.011	0.695 ± 0.007	0.552 ± 0.008	0.695 ± 0.006
GBLUP (whole)	0.451 ± 0.008	0.555 ± 0.011	0.693 ± 0.007	0.549 ± 0.008	0.694 ± 0.006

Regulatory, intron, exon, intergenic, gene and non-gene represents the kinship matrix inferred from genetic variants within this kind of category

GBLUP (whole) represents that all the variants are used to infer the kinship matrix

[a]Denotes the standard error of predictability of traits, calculated from 100 replicates of tenfold cross-validation

Table 4 Predictabilities of the five traits based on MultiBLUP framework using various combinations of the genetic relatedness matrix

Trait	mBLUP (2)[a]	mBLUP (4)	mBLUP (adapt)	GBLUP (whole)
YD	0.451 ± 0.010[b]	0.451 ± 0.008	0.445 ± 0.015	0.451 ± 0.008
HD	0.560 ± 0.008	0.556 ± 0.011	0.550 ± 0.004	0.555 ± 0.011
PH	0.701 ± 0.005	0.696 ± 0.007	0.695 ± 0.009	0.693 ± 0.007
GL	0.553 ± 0.004	0.550 ± 0.008	0.576 ± 0.005	0.549 ± 0.008
GW	0.696 ± 0.004	0.697 ± 0.006	0.695 ± 0.008	0.694 ± 0.006

[a]Denotes the number of individual genetic relatedness, such as 2 representing the two relatedness matrices was used, "adapt" representing the number of matrices automatically calculated by LDAK software and "whole" denotes that whole genetic variants were incorporated into one matrix

[b]Denotes the standard error of predictability of traits, calculated from 100 replicates of tenfold cross-validation

heritability, or variance-component method, which was proposed by Gusev et al. (2014), outperforms the regular testing approaches based on summary statistics because linkage disequilibrium (LD) between functional categories is taken into account. In this study, we used both methods, i.e., analysis of heritability across genomic partitions and GWAS, to analyze the rice data, leading to a consistent conclusion that variation of SNPs in nongenic regions contribute most to the phenotypic variation. The analysis of heritability using individual SNP sets (intron, exon, regulatory, or intergenic) showed that the genetic contributions from these four SNP sets are very similar, with the contributions from regulatory regions or intergenic regions more than those from the introns or exons (Table S5). Similarly, predictions using SNPs in intron regions or exon regions only were almost as good as the predictions made by using SNPs in regulatory regions or intergenic regions only (Table 3). However, when the four SNP sets were used simultaneously in the analysis of heritability, the genetic contributions were mainly detected from regulatory regions or intergenic regions, whereas the contributions from introns or exons were close to zero (Table 1). In summary, a larger proportion of trait heritability was explained by SNPs in nongenic partition, based on the results of the analysis of heritability.

Consistently, GWAS showed that top TASs identified in nongenic partition had larger genetic effects than those in genic partition.

Several studies of human data revealed the similar results of genetic architectures underlying complex traits. For example, Maurano et al. (2012) systematically investigate the spatial distribution of TASs and found that trait-associated variants tend to cluster in regulatory regions marked by DHSs. Gusev et al. (2014) analyzed the original GWAS data using the variance-component analysis and also reported that DHS-marked regions explained most of the genetic variations of the phenotypes. Consistent results were reported in plant studies, for example, in maize (Li et al. 2012; Rodgers-Melnick et al. 2016; Xue et al. 2016). It is not surprising to detect the importance of regulatory regions in terms of genetic contribution, because lots of *cis*-regulatory elements, such as enhancers and silencers, in these regions directly influence gene expressions by interacting with the promoters. Moreover, we found that, like the SNPs in regulatory regions, the SNPs in intergenic regions are also responsible for a large proportion of trait heritability. This was supported by recent research where the important role of these regions was recognized (Edwards et al. 2013), and by the fact that abundant functionally important molecules, such as small RNAs, micro RNAs, or
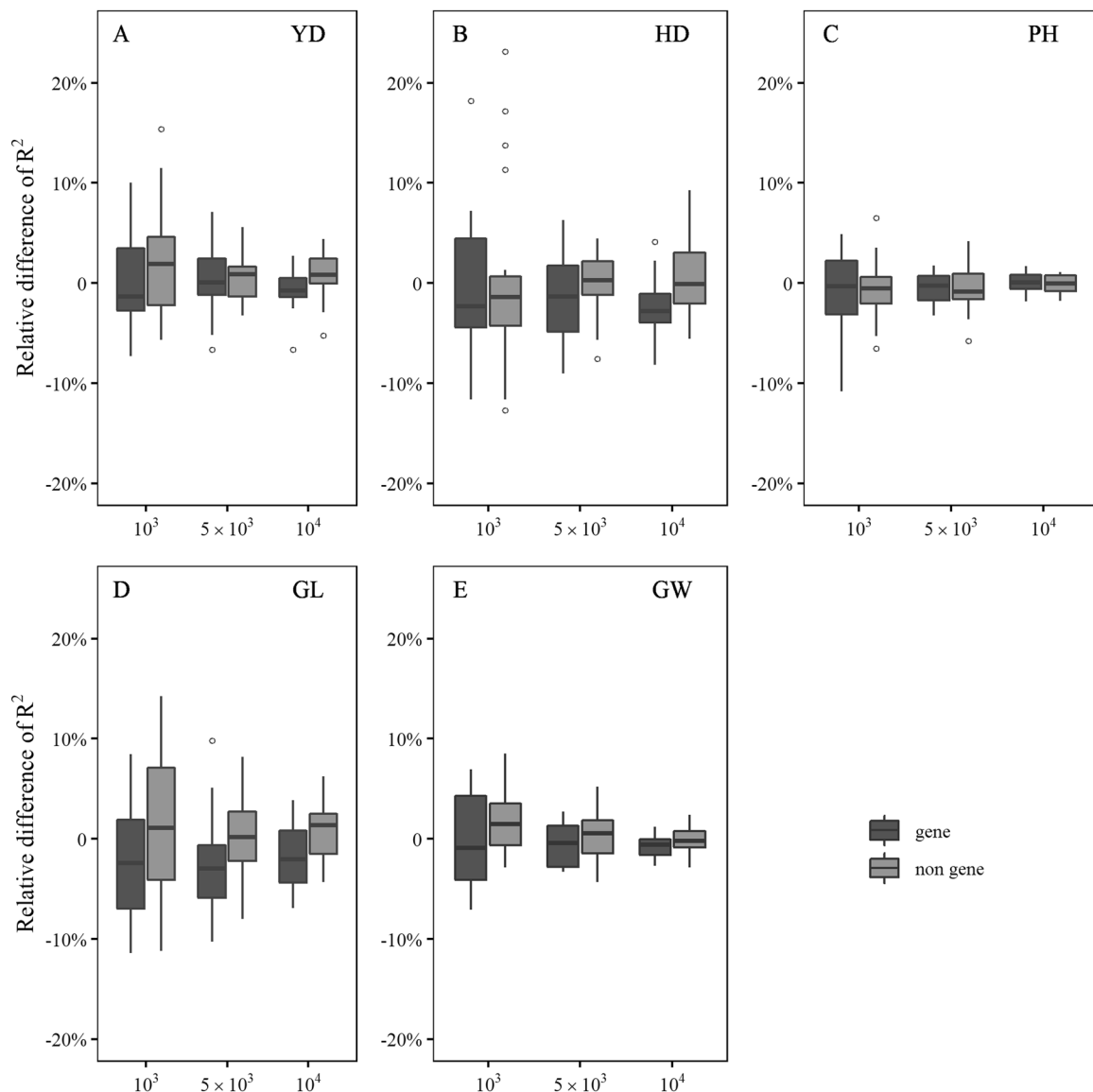
Fig. 5 Comparison of the predictabilities derived from the SNPs of genic (black) and nongenic region (gray). The y-axis represents the relative difference of predictability between model I (inclusion of SNPs within the genic region) or model II (inclusion of SNPs within the nongenic region) and the reference model in which SNPs are randomly sampled from the whole genome. The x-axis represents the number of SNPs included in the model. We performed 100 replicates for each scenario. **a–e** are corresponding to the five traits, YD, HD, PH, GL, and GW, respectively

lncRNAs, are transcribed from these regions (Bartel 2009; Cunnington et al. 2010; Ghoussaini et al. 2008; Jendrze-jewski et al. 2012). For example, miRNAs typically inter-fere gene expressions by binding to 3′UTRs of target mRNAs and repress the transcription process (Bartel 2009). The genome does not exist as a linear entity within cells, where the DNA blueprint is actually utilized. Current research shows that both *cis*- and *trans*-interactions between genetic variants, which depend on chromosome 3D con-figuration, influence gene expression and provide mechan-istic insights into functions associated with phenotypes. The study indicated that variants in intergenic regions played a more important role than those regulatory regions for YD, GL, and GW; however, this was not the case for HD and PH. These results suggested distinct regulatory patterns for various complex traits, i.e., *trans*-eQTL is key player con-trolling YD, GL, and GW, whereas, *cis*-eQTL is dominant factor for HD and PH. In this study, we conclude that the genetic variants in nongenic partition, including regulatory regions and intergenic regions, are more important than the genetic variants in genic partition in determining the traits in rice.

The positive correlation between the per-SNP heritability and the MAF associated with the SNP was reported in the

analysis of heritability in human schizophrenia (Loh et al. 2015). In the study, we followed this strategy to check the possible relationship between the genetic contribution and MAF for each of the five traits. The entire set of genomic SNPs were grouped into six MAF bins based on the frequency of the minor allele (Table S6), the number of SNPs in these bins ranging from 11.1 to 23.6% of the total SNPs. Table S6 does not show any significant association between the explained heritability and the binned SNPs, with various MAFs in the rice data. We examined the distribution of MAF in the four functional categories (Fig. S7), which shows that the four categories have similar allele frequency spectrum across different MAFs. Our data and results did not support the hypothesis that higher heritability in non-genic regions region is due to enrichment of either high or low MAF SNPs.

Several literatures had used the genomic annotation data to assist a fine mapping, including genomic position information (Kichaev et al. 2014; Pickrell 2014; Yang et al. 2017). Our study leveraged these annotation data to classify the genetic variants into functional partitions for a delicate analysis, using GS with MultiBLUP strategy (Speed and Balding 2014) or GWAS. The data, results, and conclusions in our study will advance our understanding of the genetic architectures behind agronomic traits in rice. We assume that the effects of genetic variants follow the normal distribution with different variances, when they fall into different functional partitions. These variance parameters need to be first estimated from the training set and then can be applied to GS. Our Multi-BLUP strategy is analogous to the popular software BayesR, in both of which each genetic variant follows a mixture of normal distributions and the posterior membership may be inferred from data (Erbe et al. 2012). Our results indicated that considering different variance parameters for various functional partitions did not significantly improve prediction performance. This may be owing to the fact that each functional partition already includes more than enough genetic variants (in millions) that can accurately capture the genetic relatedness of individuals in the sample. A simulation study, in which small numbers of SNPs were randomly selected from these functional partitions for the analysis, showed that SNPs in nongenic partition outcompeted the SNPs in the genic partition in predictability. Therefore, to reduce the genotyping cost in breeding if microarrays are used, we only need to place the SNP markers in nongenic regions onto the genotyping chips.

A limitation of the study is that only 524 individuals were used. Estimating genetic parameters from such small sample may be a problem, because of the potential sampling bias. Combining other available rice samples (Huang et al. 2010, 2016, 2012; Yano et al. 2016) in the analysis of heritability may increase power. However, various studies may be quite different in genotyping method, data quality, number of variants, etc, posing challenges to a combined study or meta-analysis. A desirable property of this study sample is that it is a highly diverse collection because each individual represents a subpopulation. Compared with other species like human, such a study sample enjoys a larger effect size for testing genetic parameters. When other omics data become available in the future, we may leverage these additional data to tune the analysis of heritability based on genomic annotations.

Author contributions ZYJ, JLW, WBX, and XZ conceived and designed the experiments. JLW, RDL, SBW, HQ, and RYM conducted the experiments and analyzed the data. JLW wrote the program. JLW and ZYJ wrote the manuscript. All authors have read and approved the final manuscript.

Compliance with ethical standards

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ali N, Paul S, Gayen D, Sarkar SN, Datta K, Datta SK (2013) Development of low phytate rice by RNAi mediated seed-specific silencing of inositol 1, 3, 4, 5, 6-pentakisphosphate 2-kinase gene (IPK1). PLoS ONE 8:e68161

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233

Bian XF, Liu X, Zhao ZG, Jiang L, Gao H, Zhang YH et al. (2011) Heading date gene, dth3 controlled late flowering in O. Glaberrima Steud. by down-regulating Ehd1. Plant Cell Rep 30:2243–2254

Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W et al. (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 46:714–721

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strainw1118; iso-2; iso-3. Fly 6:80–92

Cunnington MS, Koref MS, Mayosi BM, Burn J, Keavney B (2010) Chromosome 9p21 SNPs associated with multiple disease phenotypes correlate with ANRIL expression. PLoS Genet 6: e1000899

Du X, Qian X, Wang D, Yang J (2006) Alternative splicing and expression analysis of OsFCA (FCA in Oryza sativa L.), a gene homologous to FCA in Arabidopsis. DNA Seq 17:31–40

Edwards SL, Beesley J, French JD, Dunning AM (2013) Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet 93:779–797

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density

single nucleotide polymorphism panels. J Dairy Sci 95:4114–4129

Falconer DS, Mackay TF, Frankham R (1996) Introduction to quantitative genetics (4th edn). Trends Genet 12:280

Fan CH, Xing YZ, Mao HL, Lu TT, Han B, Xu CG et al. (2006) 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 112:1164–1171

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE et al. (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. J Natl Cancer Inst 100:962–966

Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet 95:535–552

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106:9362–9367

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q et al. (2016) Genomic architecture of heterosis for yield traits in rice Nature 537:629–633

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32–39

Jang YH, Park H-Y, Kim S-K, Lee JH, Suh MC, Chung YS et al. (2009) Survey of rice proteins interacting with OsFCA and OsFY proteins which are homologous to the arabidopsis flowering time proteins, FCA and FY. Plant Cell Physiol 50:1479–1492

Jendrzejewski J, He H, Radomska HS, Li W, Tomsic J, Liyanarachchi S et al. (2012) The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc Natl Acad Sci 109:8646–8651

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. Plos Genet 10: e1004722

Kim S-I, Andaya VC, Tai TH (2011) Cold sensitivity in rice (Oryza sativa L.) is strongly correlated with a naturally occurring 199V mutation in the multifunctional glutathione transferase isoenzyme GSTZ2. Biochem J 435:373–380

Kim S-K, Yun C-H, Lee JH, Jang YH, Park H-Y, Kim J-K (2008) OsCO3, a CONSTANS-LIKE gene, controls flowering by negatively regulating the expression of FT-like genes under SD conditions in rice. Planta 228:355–365

Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T et al. (2002) Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. Plant Cell Physiol 43:1096–1105

Li X, Zhu C, Yeh C-T, Wu W, Takacs EM, Petsch KA et al. (2012) Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res 22:2436–2444

Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ et al. (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet 47:1385

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al. (2009) Finding the missing heritability of complex diseases. Nature 461:747

Mao H, Sun S, Yao J, Wang C, Yu S, Xu C et al. (2010) Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. Proc Natl Acad Sci USA 107:19579–19584

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1222794

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ et al. (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci 106:12273–12278

Pickrell JK (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet 94:559–573

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open chromatin reveals the functional maize genome. Proc Natl Acad Sci 113:E3177–E3184

Sasaki T (2005) The map-based sequence of the rice genome. Nature 436:793

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF et al. (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 9: e1003449

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S et al. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nat Genet 40:1023–1028

Speed D, Balding DJ (2014) MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 24:1550–1557

Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91:1011–1021

VanRaden PM (2008) Efficient methods to compute genomic predictions. J dairy Sci 91:4414–4423

Wan X, Weng J, Zhai H, Wang J, Lei C, Liu X et al. (2008) Quantitative trait loci (QTL) analysis for rice grain width and fine mapping of an identified QTL allele gw-5 in a recombination hotspot region on chromosome 5. Genetics 179:2239–2252

Wei J, Wang A, Li R, Qu H, Jia Z (2018) Metabolome-wide association studies for agronomic traits of rice. Heredity 120:342–355

Xu S, Xu Y, Gong L, Zhang Q (2016) Metabolomic prediction of yield in hybrid rice. Plant J 88:219–227

Xu S, Zhu D, Zhang Q (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci USA 111:12456–12461

Xue S, Bradbury PJ, Casstevens T, Holland JB (2016) Genetic architecture of domestication-related traits in maize. Genetics 204:99–113

Yan WG, Li Y, Agrama HA, Luo D, Gao F, Lu X et al. (2009) Association mapping of stigma and spikelet characteristics in rice (Oryza sativa L.). Mol Breed 24:277–292

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565

Yang J, Fritsche LG, Zhou X, Abecasis G, Int Age-Related Macular D (2017) A scalable bayesian method for integrating functional information in genome-wide association studies. Am J Hum Genet 101:404–416

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P-c, Hu L et al. (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48:927–934

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu S, Xu W, Vijayakumar C, Ali J, Fu B, Xu J et al. (2003) Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. Theor Appl Genet 108:131–140

Zhang H, Zhang D, Wang M, Sun J, Qi Y, Li J et al. (2011) A core collection and mini core collection of Oryza sativa L. in China. Theor Appl Genet 122:49–61

Zhang YM, Mao YC, Xie CQ, Smith H, Luo L, Xu SZ (2005) Manpping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (Zea mays L.). Genetics 169:2267–2275