A Multilayered Screening Method for the Identification of Regulatory Genes in Rice by Agronomic Traits

Young-Joo Seol[1], So Youn Won[1], Younhee Shin[2], Jong-Yeol Lee[3], Jong-Sik Chun[4], Yong-Kab Kim[5] and Chang-Kug Kim[1]

[1]Genomics Division, National Institute of Agricultural Sciences, Jeonju, Republic of Korea. [2]Codes Division, Insilicogen Inc., Suwon, Gyeonggi-do, Korea, [3]Functional Biomaterial Division, National Academy of Agricultural Science, Jeonju, Korea. [4]School of Biological Sciences and Bioinformatics Institute, Seoul National University, Seoul, Korea. [5]School of Electrical Information Communication Engineering, Wonkwang University, Iksan, Korea.

ABSTRACT: We developed a multilayered screening method that integrates both genome and transcriptome data to effectively identify regulatory genes in rice (*Oryza sativa*). We tested our method using eight rice accessions that differed in three important nutritional and agricultural traits, anthocyanin biosynthesis, amylose content, and heading date. In the genome resequencing of eight rice accessions with 24 RNA sequencing experiments, 98% of the preprocessed reads could be uniquely mapped to the reference genome, resulting in the identification of 42,699 unique transcripts. Comparison between black and white rice cultivars showed evidence of intensive selective sweeps in chromosomes 3, 10, and 12. A total of 131 genes were differentially expressed among the black rice cultivars and found to be associated with three Gene Ontology terms (secondary metabolic process, biosynthetic process, and response to stimulus). We identified nonsynonymous Single Nucleotide Polymorphism (SNP) that likely play an important role in determining the agronomic traits differences, two upregulated and three downregulated genes in the black cultivars, and two downregulated genes in the white cultivars. The three agronomic traits were clearly grouped together by the developmental stages, regardless of any other traits, suggesting that the developmental stage is the most important factor that triggers global changes in gene expression. Interestingly, glutinous and nonglutinous black rice cultivars were distinguished from one another by different heading dates.

KEYWORDS: gene expression, selective sweep, transcriptome, whole-genome resequencing

CITATION: Seol et al. A Multilayered Screening Method for the Identification of Regulatory Genes in Rice by Agronomic Traits. *Evolutionary Bioinformatics* 2016:12 253–262 doi: 10.4137/EBO.S40622.

TYPE: Original Research

RECEIVED: July 26, 2016. **RESUBMITTED:** September 22, 2016. **ACCEPTED FOR PUBLICATION:** September 30, 2016.

ACADEMIC EDITOR: Liuyang Wang, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1016 words, excluding any confidential comments to the academic editor.

FUNDING: This study was conducted with support from the Research Program for Agricultural Science & Technology Development (Project No. PJ010112), Cooperative Research Program (Project No. PJ01043805) of the National Institute of Agricultural Sciences, and the Next-Generation BioGreen 21 Program (SSAC, Grant No. PJ011650), Rural Development Administration. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: chang@korea.kr

Introduction

The recent development of next-generation sequencing has led to an explosion in the number and diversity of different types of sequencing data, which have provided insight into numerous biological questions.[1] Genome resequencing can reveal genomic variation, evolutionary history, and population structure and can identify genomic loci responsible for phenotypic and physiological differences.[2] Similarly, RNA sequencing (RNA-seq) has been used to identify expressed genes, the networks among expressed genes, genetic variation, and regulatory genes.[3] Although these data types can be quite informative on their own, their integration is required for in-depth understanding of complex systems.[4]

Rice (*Oryza sativa* L.) is a major staple food. Its cultivars are distinguished by various characteristics such as nutritional quality (eg, amylose content, antioxidant activity, etc.) and agronomic traits (eg, heading date). At the highest level, rice cultivars are typically divided into nonglutinous rice or glutinous rice, based on the proportion of glucose polymers present, such as linear polymer amylose and highly branched amylopectin. Nonglutinous rice grains contain a large amount of amylose, whereas glutinous rice grains (sticky or waxy rice) are distinguished by the presence of amylopectin, rather than amylose. Amylose synthesis is governed mainly by an allelic series of waxy (Wx) genes (including granule-bound starch synthase)[5] and by some unidentified non-Wx genes such as SSII–3.[6] Quantitative trait loci associated with amylose content have also been reported, but their molecular function has not been identified.[7]

Black rice is a functional cultivar that is valued as a health-promoting food. Its black color is due to the accumulation of anthocyanin, the antioxidant properties of which play an important role in anticarcinogenic and anti-inflammatory activity, obesity control, and the alleviation of diabetes.[8] The anthocyanin biosynthetic pathway, which utilizes the middle steps of the flavonoid biosynthetic pathway,

is one of the most extensively studied pathways involved in plant secondary metabolism.[9,10] Various genes (eg, *MYB*, *HLH*, *DFR*, etc.) have been identified as important regulators of grain color in black rice.[11–13] The Gramene portal (http://www.gramene.org/) reports that the rice genome contains 13 genes involved in anthocyanin biosynthesis, and the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) database reports that 14 orthologous gene groups are involved in the anthocyanin pathway. Heading date is a critical trait that affects the adaptation of a cultivar to different cropping locations and growing seasons and is thus the most important breeding objective. Genetic, molecular, and genomic approaches have uncovered more than 20 genes and quantitative trait loci involved in the regulation of heading time in multiple rice cultivars.[14,15]

Although the integration of genome resequencing data with RNA-seq data is expected to pinpoint regulatory genes more effectively, a methodology is required. To develop such a methodology, we chose rice cultivars with diversity in three important traits, anthocyanin biosynthesis, amylose content, and heading date. Here, we report the generation of genome resequencing and RNA-seq data and describe an analytic approach for integrating these two types of next-generation sequencing data for identifying the genes involved in the three important rice characteristics.

Materials and Methods

Rice materials and experimental design. Eight rice accessions were obtained from the RDA-Genebank (http://www.genebank.go.kr/). Their characteristics are shown in Table 1. The six black rice accessions are categorized by the two most important genetic characterizations, namely, heading date (ie, early, medium, or late maturing) and physiochemical seed properties (ie, glutinous or nonglutinous). The white rice *Dongjin* and *DJ_chal* differ in amylose content (Table 1). Genome resequencing was performed with eight accessions. For RNA-seq, a total of 24 nonreplicated experiments were conducted with eight rice accessions at three developmental time points (ie, 5, 10, and 15 days after heading). Total RNA was extracted from the seed tissue at three developmental time points. We compared the sequence variations between eight accessions and the rice reference sequence (*Nipponbare*). *Nipponbare*, a white and nonglutinous rice, was used to identify the sequence variations between each accession and the reference.

Genome resequencing and reads mapping. For genome resequencing, a paired-end sequencing library was constructed using the TruSeq™ DNA Sample Preparation Kit v2 (Illumina Inc.) and sequenced by an Illumina HiSeq 2000 (Illumina Inc.). The quality score distribution of each library was checked using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). If a sequenced base had a quality score lower than Q20, which indicates an accuracy of 99% for the base call, it was changed to "N". Reads with less than 90 bp or with "N"s at more than 10% of their total base positions were removed. The filtered reads were mapped to the reference genome MSU Rice Genome Annotation Project Release 7[16] using CLC Assembly Cell software version 4.1 (CLC Bio) with the advanced options (95% identity and 90% coverage by high-scoring base pairs).

Detection of SNPs and InDels. SNPs and small insertions and deletions (InDels) in eight rice accessions were identified against the rice reference genome using the "find variations" function of the CLC Assembly Cell software with the following parameters: minimum depth = 5, minimum mismatch count = 3, and limit fraction ≥20%. To reduce erroneous SNP calls caused by mapping of reads to multiple regions in the reference genome, we filtered out the low-quality mapping regions that lacked 95% sequence identity and 100% coverage by high-scoring base pairs.

Population structure analysis with SNPs. We performed a population structure analysis using variation in SNPs with the FRAPPE program, which is based on the maximum likelihood method.[17] We generated PED files for the eight rice accessions using 10,000 iterations, considering cluster numbers (K) from 2 to 6.[2] To construct the phylogenetic tree, we calculated the genetic distances between the different accessions using SNPs. The neighbor-joining method was applied to construct the tree based on the distance matrix calculated by the PHYLIP version 3.695 program (http://evolution.genetics.washington.edu/phylip/). The graphical view of the phylogenetic tree was generated with FigTree v1.4 software (http://tree.bio.ed.ac.uk/software/figtree/).

Linkage disequilibrium analysis. Linkage disequilibrium (LD), the nonrandom association of alleles at two or more different loci, is a sensitive indicator of the population genetic forces that structure a genome.[18] We measured LD levels in each accession by calculating the correlation coefficient (r^2) of alleles with the Haploview program.[19] Parameters used were as follows: maxdistance, 1,000; dprime, 0.6; minGeno, 0.6;

Table 1. Rice accessions used in this study.

TYPE	WHITE RICE		EARLY		MEDIUM		LATE	
	CULTIVAR	ID*	CULTIVAR	ID	CULTIVAR	ID	CULTIVAR	ID
Non-glutinous	Dongjin	IT212532	Heugjinju	IT191964	Heugseol	IT218587	Heugnam	IT212512
Glutinous	DJ_chal	IT196275	JH_chal	IT214862	BH_chal	IT235306	SH_chal	IT284610

Note: ID*, Accession number of RDA-Genebank, Korea (http://www.genebank.go.kr/).

minMAF, 0.1; and hwcutoff, 0.001. We then used scripts written in R to plot averaged r^2 against pairwise marker distances.

Detection of selective sweeps. To detect genomic areas with selective sweeps driven by artificial selection, we calculated the reduction of diversity (ROD) based on the ratio of diversity between black and white rice accessions. ROD values were calculated with variables π_{black} (π value of black rice accessions) and π_{white} (π value of white rice accessions) using the following equations:

$$\mathrm{ROD} = 1 - \left(\pi_{black} / \pi_{white}\right) \quad (1)$$

where population parameter π is the average number of nucleotide differences between any two DNA sequences.[2] We divided the entire genome into 10, 50, 100, 200, and 500 kb windows and calculated the ROD value in each window. The candidate genes in selective sweep regions were screened based on the significance level $P \leq 0.01$ (1.0%) of the right tail in the ROD distribution. In addition, black accessions (π_{black}), white accessions (π_{white}, control), and the ROD of two groups were compared using Circos diagrams.[20]

Detection of putative genes using selective sweeps. At each detected SNP position, we counted the number of reads corresponding to the most and least frequently observed allele in each group.[21,22] Manhattan plots for visualizing selective sweeps were generated using allele counts of identified SNP positions in 40 kb sliding windows along the genome, with a step of 20 kb. In order to display the putatively selected regions from the extreme tails, we applied a threshold of $ZH_p \leq -3$, where ZH_p is the Z-transformed heterozygosity.

The pooled heterozygosity (H_p) was calculated with variables N_{max} (reads number of most observed allele) and N_{min} (reads number of least observed allele) using the following equations:

$$H_p = 2\sum N_{max} \sum N_{min} \Big/ \left(\sum N_{max} + \sum N_{min}\right)^2 \quad (2)$$

where $\sum N_{max}$ = sums of N_{max}, $\sum N_{min}$ = sums of N_{min}.

ZH_p for selective sweep values was calculated using the following equation:

$$ZH_p = \left(H_p - \mu H_p\right) \Big/ \sigma H_p \quad (3)$$

where μ = average, σ = standard deviation.

Gene expression analysis. For the RNA-seq experiments, the quality of raw reads was checked with the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Preprocessing of Illumina RNA-seq raw reads was conducted with the same method as that used for genomic resequencing.

The preprocessed reads were mapped to a rice reference genome, MSU Rice Genome Annotation Project Release 7, using the TopHat program, and gene expression levels were identified using Cufflinks with the default parameters.[23] The fragments per kilobase of transcript per million mapped reads (FPKM) score were calculated with the transcribed fragments. To compare the gene expression levels among samples, FPKM scores were subjected to global normalization and were used for further analysis. Principal component analysis (PCA) was performed with GeneSpring GX software 12.5 (Agilent Technologies). Functional associations were computed using the singular enrichment analysis (SEA) method with the normalized FPKM values and Gene Ontology (GO) categories. In the GO enrichment analysis, genes located in the selected regions were extracted from the rice gene annotation file (MSU_osa1r7, http://rice.plantbiology.msu.edu/index.shtml). GO enrichment analysis was conducted through agriGO (http://bioinfo.cau.edu.cn/agriGO/) with a significance level of $P \leq 0.05$ and the *O. sativa* species option.[24]

Accession codes. The resequencing data have been deposited in EMBL-EBI (http://www.ebi.ac.uk) under the accession numbers: ERP009995 *(Heugjinju),* ERP009996 *(Heugseol),* ERP009997 *(Heugnam),* ERP009998 *(JH_chal),* ERP009999 *(BH_chal),* ERP010000 *(SH_chal),* ERP010001 *(Dongjin),* and ERP010002 *(DJ_chal).* In addition, the 24 RNA-seq data sets have been deposited in EMBL-EBI with accession numbers from ERP009858 to ERP009904.

Results and Discussion

Genome resequencing. Aiming to more effectively and precisely identify genes that regulate a particular trait, we sought to establish an analysis process that integrates both genome resequencing and RNA-seq. Considering the growing interest in its nutritional value and the lack of knowledge about it relative to white rice, several black rice cultivars, with variation in agronomic and nutritional features, were selected to test our screening procedure. Six black rice accessions were categorized by their two most important characteristics, namely, heading date (ie, early, medium, or late maturing) and grain amylose content (ie, glutinous or nonglutinous), as shown in Table 1. Two white rice accessions with either high or low grain amylose content were also included. The proposed multilayer screening procedure was then used to pinpoint regulatory genes for black color (ie, anthocyanin biosynthesis), amylose content, and heading date.

The genomes for all eight rice accessions were sequenced with an Illumina HiSeq 2000. On average, 14,973,528,684 bp of raw reads were generated per accession, which corresponded to 40.0× coverage of the rice reference genome (Supplementary Table 1). Raw sequences were preprocessed by trimming out raw quality reads and adaptors and by removing bacterial contamination (Supplementary Tables 1 and 2). The preprocessed reads were mapped to the rice reference genome.[16] For all rice accessions, 98.0% of the preprocessed reads were uniquely mapped onto the reference genome, and the average mapping depth was 40.0× of the rice genome, suggesting that our resequencing data were thorough enough for subsequent analysis (Supplementary Table 3).

Variation among the rice genomes. Genomic variation was determined across the eight rice accessions. On average, SNPs accounted for 88.6% of all polymorphisms, followed by deletions and insertions (Table 2). In terms of SNP and InDel genotypes, homozygous genetic variation (80.4%) was more abundant than heterozygous variation (Table 2). White rice accessions (*Dongjin* and *DJ_chal*) showed less genomic variation against the rice reference genome, compared with the six black rice accessions (Table 2). We focus here on the predominant type of genomic variation, SNPs, and not on InDels. We identified a total of 1,241,646 unique SNPs (Table 2). Most SNPs were located in intergenic regions, and 15.5% of the SNPs were located in coding sequence (CDS) regions (Supplementary Table 4). Among the latter, synonymous SNPs (ie, SNPs that do not change the protein sequence) were more common than nonsynonymous SNPs (ie, SNPs that change the protein sequence), and synonymous SNPs represented 52.1% of the total number of SNPs (Supplementary Table 4). Although the total number of SNPs was higher in black rice than in white rice, the number of nonsynonymous SNPs was not proportionately increased in black rice (Supplementary Table 4).

Inference of population structure. To accurately investigate rice population structure using variation in SNPs, we removed SNPs that were located in the genomic regions with low-depth sequencing coverage in each rice cultivar. From the 2,130,187 initial high-quality SNPs, we selected 981,808 SNPs for population analysis. We estimated the population structure of the individual rice cultivars using the FRAPPE program.[17] With the assumption that K clusters exist, individuals within populations were divided into K groups based on a maximum likelihood method. Here, ancestry was analyzed by increasing K from 2 to 6. From $K = 4$, white rice accessions (*Dongjin* and *DJ_chal*) were distantly separated from black rice accessions (Fig. 1A). Black rice cultivars with late heading dates (*Heugnam* and *SH_chal*) were clustered together (Fig. 1A). Although *JH_chal* and *Heugseol* differ in heading date and grain amylose content, the two black accessions formed a separated group from $K = 3$ (Fig. 1A). Phylogenetic tree analysis resulted in similar conclusions about population structure. The white accessions had the greatest distance from the black accessions (Fig. 1B). In addition, JH_chal, Heugseol, and SH_chal/Heugnam showed a high degree of relatedness (Fig. 1B).

Table 2. Statistics of identified SNPs and InDels.

GENOTYPE				VARIATION TYPE		
SAMPLE	TOTAL	HOMO	HETERO	SUBSTITUTION	INSERTION	DELETION
Heugjinju	317,987	257,419	60,568	282,439	17,584	17,964
Heugseol	316,859	255,925	60,934	278,668	18,877	19,314
Heugnam	318,621	249,956	68,665	283,892	17,415	17,314
Dongjin	264,847	212,508	52,339	237,237	13,599	14,011
JH_chal	313,393	257,386	56,007	274,706	19,267	19,420
BH_chal	304,929	246,522	58,407	269,866	17,405	17,658
SH_chal	310,146	247,379	62,767	275,342	17,457	17,347
DJ_chal	257,873	207,620	50,253	228,037	14,763	15,073
Total	1,241,646	985,473	302,672	1,115,907	66,103	65,707

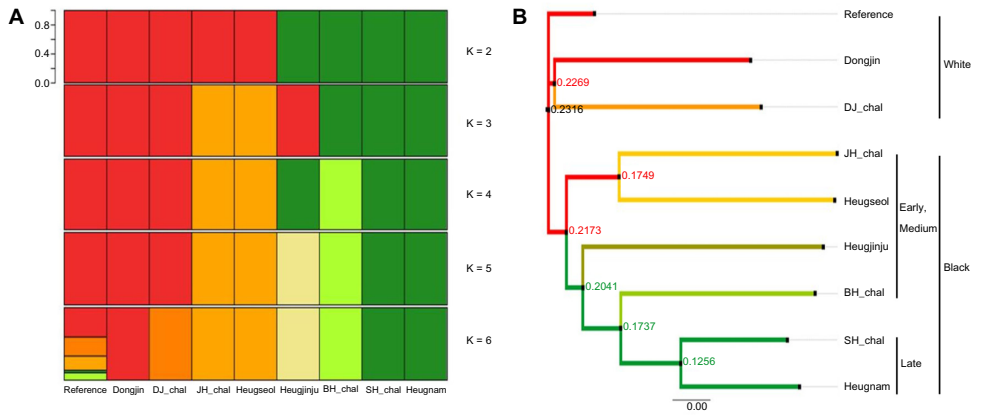

Figure 1. Population structure of eight rice accessions and the rice reference genome. (**A**) Results of a population structure analysis using the FRAPPE program with 981,808 high-quality SNPs. Each accession is represented by a vertical bar. (**B**) A neighbor-joining phylogenetic tree of the rice genomes based on SNPs. The phylogenetic tree was generated with 1,000 bootstrap repetitions, and all nodes were clustered with bootstrap values.

To identify statistical associations between alleles at different loci, we estimated the LD for the distinct groups. For LD plots, we calculated the r^2 value, defined as the correlation coefficient of SNP frequencies, between pairs of SNPs. We observed that LD decayed more slowly in the white group than in the black group, and in the glutinous group than in the nonglutinous group. However, we did not obtain significant results for the population genetics analysis due to our small sample size. The genetic distance among rice accessions decreases with increasing sample sizes. The coefficient of variation for estimates of genetic distance decreases with increasing sample sizes.[25] Hence, accurate analyses of population genetics are not possible with a small sample size.[26]

Detection of putative genes using selective sweeps. It has been suggested that genomic regions that confer agriculturally favorable traits have low levels of variation and biased allele frequency spectra, due to prolonged domestication and breeding.[2] Similarly, by comparing SNP levels, for example, between black and white rice, genomic regions that show low levels of single nucleotide variation only in black rice cultivars would be expected to contain genes that are favorably selected in black rice, such as anthocyanin biosynthesis genes. That is, they would show evidence of selective sweeps. To detect selective sweeps driven by artificial selection, we used SNPs to calculate the ROD score in every nonoverlapping window of 40 kb along the entire genome, with a step of 20 kb. We then plotted the scores only for the genomic regions that showed positive values for π_{black}, π_{white}, and ROD (Fig. 2A). π_{black} and π_{white} represent summary statistics for measuring genetic diversity in black and white rice populations, respectively. The significant outlier regions (ie, with a right tail $\pi \leq 0.01$) were identified as our putative loci. The genomic regions in chromosome 10 showed the biggest differences in outlier genes between white and black rice accessions (Fig. 2A). Our catalogued regions possibly include either key genes responsible for black rice-specific characteristics or genomic regions selectively kept in black rice. The efficiency of this method was verified by conducting the same analysis with additional comparisons. Glutinous vs. nonglutinous black rice accessions were compared (Supplementary Fig. 1A), as were black rice cultivars with different heading dates (Supplementary Fig. 1B).

In order to effectively present the regions showing evidence of selective sweeps across all chromosomes, the ROD distribution was replotted in a Circos diagram using 5-kb windows, along with gene density (Fig. 3). Comparison between black and white rice accessions showed that intensive selective sweeps occurred in chromosomes 3, 10, and 12 (Fig. 3). Amylose content was analyzed by comparing glutinous and nonglutinous black rice accessions with the respective white accessions (Supplementary Fig. 2). Glutinous rice showed a very strong signal for a selective sweep in chromosome 11, whereas nonglutinous rice showed evidence of selective sweeps of varying intensities across all chromosomes (Supplementary Fig. 2). Finally, black rice cultivars with a late heading date showed evidence of selective sweeps

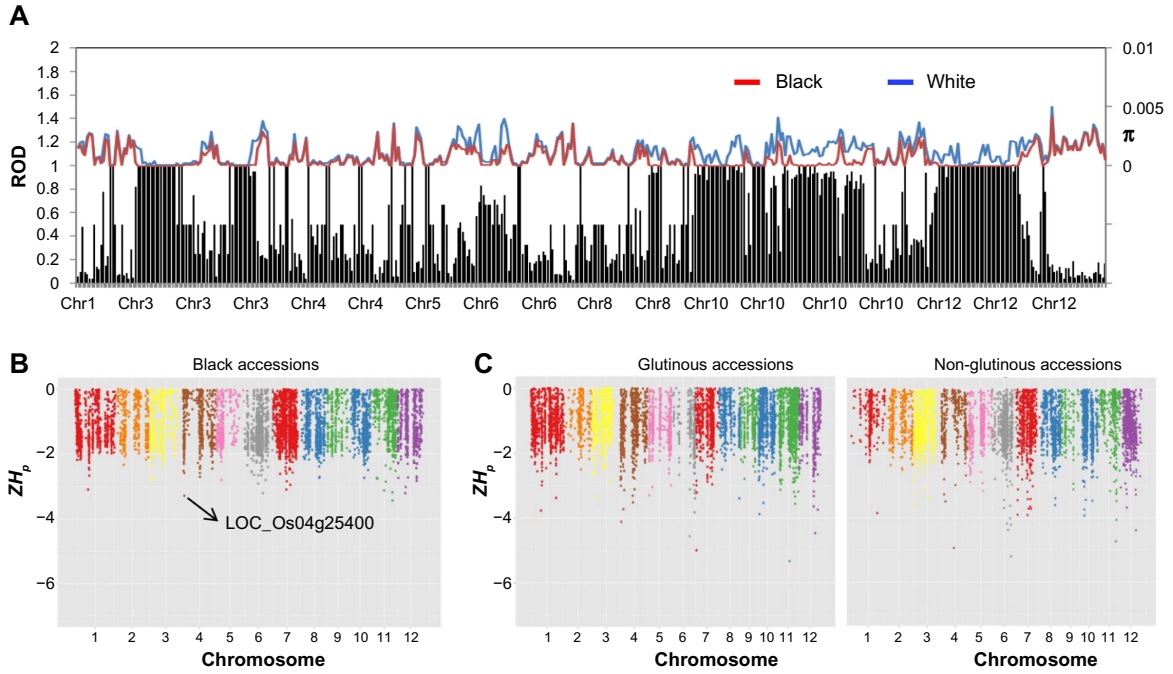

Figure 2. Variation in heterozygosity in selective sweep regions in chromosomes 1–12. (**A**) The ROD score was calculated for the rice group in chromosomes 1–12. The significant outlier regions (ie, with right tail $\pi \leq 0.01$) are shown in blue for the white group and red for the black group. (**B**) The ZH_p score for the black group against the white group is shown in this Manhattan plot. (**C**) The ZH_p of the glutinous and nonglutinous groups against the white group is shown in this Manhattan plot.

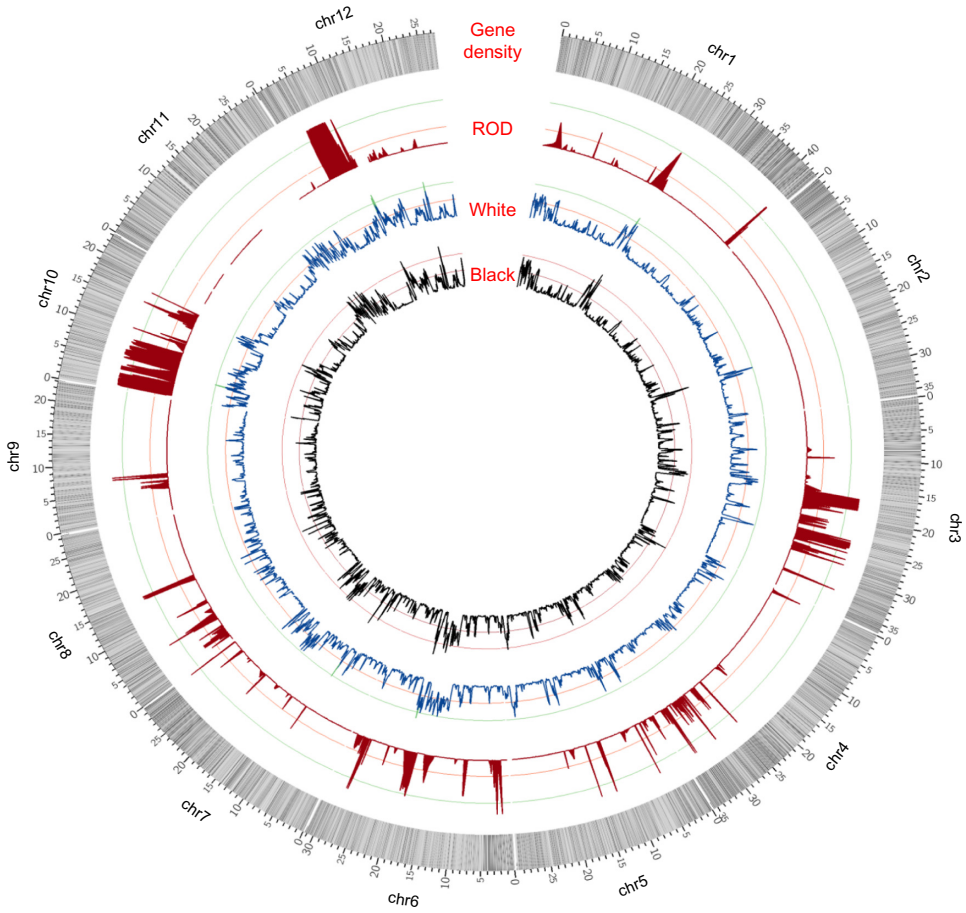

Figure 3. Circos diagram showing the ROD pattern among rice groups. The Circos diagram was generated with a 5-kb window. The outer ring shows gene density calculated across all chromosomes. Regions with significant ROD scores are shown in the next ring with their ROD values. ROD = 1 – (π_{black}/π_{white}), where π_{black} is the π value of the black group and π_{white} is the π value of the white group.

throughout all chromosomes, compared with their white control (Supplementary Fig. 3).

To detect the putative genes located in the genomic areas that showed evidence of selective sweeps, the distribution of ZH_p was plotted along chromosomes 1–12 (Fig. 2B and C). Because the genes in the populations are complex and an extremely low ZH_p value indicates a selective sweep, we focused on genomic windows with a ZH_p score of −3 or lower. In the black rice population, we observed that LOC_Os04g25400, an uncharacterized protein, was the only locus that is located in the significant outlier regions (Fig. 2B). For glutinous and nonglutinous rice accessions, we found that 21 and 18 genes passed the threshold ZH_p score, respectively (Fig. 2C and Supplementary Table 5). When the threshold score for ZH_p was set to −4 (ie, $ZH_p \leq -4$), we did not find candidate genes that showed a strong selective sweep signal, except for some genes in the glutinous population. Presumably, the small population size resulted in the underestimation of the actual heterozygosity level. Because the detection of selective sweeps based on genomic SNP data is complicated by the intricacies of the schemes used to discover the SNPs, the possible mechanisms of the selective sweep, and the effects of the sample size, the small sample size in our analysis likely makes the results of the analysis flawed.[27] Those findings demonstrate that the selection pressures associated with crop–domestication regimes can exceed by one or two orders of magnitude than those observed for genes under even strong selection in natural systems.[28] Nevertheless, these selected gene candidates, found through selective sweep analysis, provide useful guidance for rapidly identifying genes with agronomic significance.

Functional association using gene expression data. We performed a total of 24 nonreplicated RNA-seq experiments with eight rice accessions at three developmental stages (ie, 5, 10, and 15 days after heading). Using TopHat/Cufflinks, the expression levels for 42,699 unique transcripts in rice were quantified with FPKM values for 24 samples. In order to globally compare the effects of four factors (rice grain colors, amylose content, natural heading date, and developmental stages) on gene expression, PCA was conducted using FPKM values. In the PCA, most of the samples were divided into three groups. Developmental stage (ie, 5, 10, or 15 days after heading) was the most important factor distinguishing the three groups (Fig. 4). Samples from the earliest developmental stage (five days after heading) were clearly grouped together, regardless of any other traits of the rice accessions,

whereas those from other stages were somewhat intermixed together (Fig. 4), suggesting that developmental stage is the most significant factor that triggers global changes in gene expression.

By a comparison between black and white groups, we identified differentially expressed genes (DEGs) with more than a two-fold change between groups. After DEGs in all developmental stages were searched separately in glutinous and nonglutinous subgroups, common DEGs in both analyses were identified (Supplementary Table 6). Of the 131 DEGs, we ultimately found five upregulated and four downregulated genes in black rice accessions (Table 3 and Supplementary Table 6). As expected, due to their high accumulation of anthocyanin, the LOC_Os04g47040 gene, which encodes anthocyanin regulatory Lc protein, was upregulated in black cultivars (Table 3).

To provide valuable insight into the functional associations of the 131 DEGs for black rice cultivars, enrichment analysis was conducted with FPKM values and GO categories using the SEA method (P value ≤ 0.05). Three GO terms – secondary metabolic process, biosynthetic process, and response to stimulus – were found to be significantly enriched in the upregulated genes in black rice accessions, while no enrichment was detected in the downregulated genes (Supplementary Table 7).

Comparison of SNP structure and gene expression. In order to integrate genome resequencing and RNA-seq

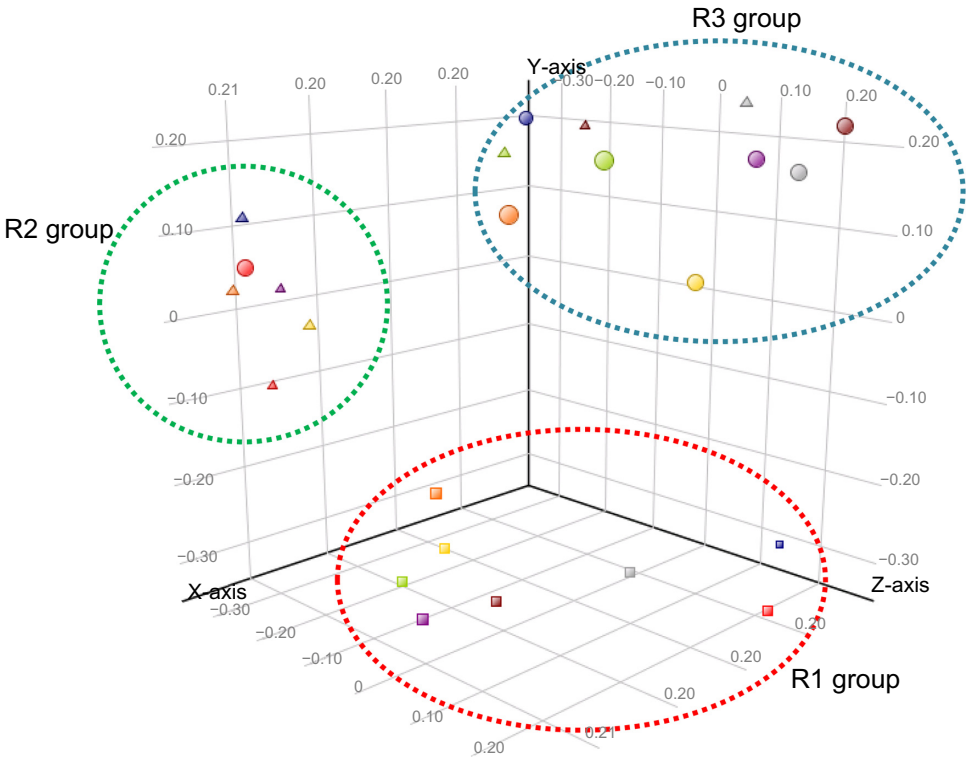

Figure 4. Three-dimensional PCA with RNA-seq data. Global expression analysis resulted in 24 samples that were clustered into three groups. Each color represents a different rice accession. The shape represents different developmental stages (ie, square: 5 days, triangle: 10 days, circle: 15 days after heading date).

Table 3. Differentially expressed genes in black rice.

GROUP	GENE ID	LOCUS	DESCRIPTION
Up	LOC_Os11g32810	Chr11:19385004–19388767	Leucine rich repeat family protein, expressed
	LOC_Os03g07270	Chr3:3707179–3708517	Glycine-rich cell wall protein, putative, expressed
	LOC_Os01g57250	Chr1:33082076–33083412	Expressed protein
	LOC_Os04g47059	Chr4:27915597–27939824	TRANSPARENT TESTA 8, putative, expressed
	LOC_Os04g47040	Chr4:27881154–27890773	Anthocyanin regulatory Lc protein, putative, expressed
Down	LOC_Os03g45220	Chr3:25537495–25538099	Expressed protein
	LOC_Os11g11920	Chr11:6606426–6616027	Resistance protein, putative, expressed
	LOC_Os03g32330	Chr3:18495805–18501143	Expressed protein
	LOC_Os11g34824	Chr11:20402883–20405478	Expressed protein

data, we analyzed the expression levels of genes containing genetic variation in the form of SNPs. First, the number of CDS-located SNPs that were shared in rice subgroups with a particular feature was counted. For example, all of the black and all of the white rice accessions contained 496 and 1,127 SNPs in CDS regions, respectively. Second, SNPs were divided into synonymous and nonsynonymous SNPs. In black rice cultivars, the total number of synonymous and nonsynonymous SNPs was 248 (Supplementary Tables 8 and 9). In black and white rice cultivars, synonymous SNPs were more abundant in white rice varieties, whereas nonsynonymous SNPs were more frequent in black rice varieties (Supplementary Tables 8 and 9). Although the above analysis for population structure revealed that the six black accessions were more distant from the reference rice than the two white rice accessions, counting SNPs that co-occurred in six cultivars but not in two cultivars would result in fewer synonymous SNPs in black accessions. Third, we analyzed the functional associations of genes with nonsynonymous SNPs, motivated by the significant influence that nonsynonymous SNPs can have on protein structure and function, and in turn, on phenotype.[29] With the SEA method, six GO terms were found to be enriched (P value ≤ 0.05) for the black rice group (Supplementary Table 10), while 21 relevant GO terms were enriched for the white rice group (Supplementary Table 11). Finally, we examined the transcript expression of genes with nonsynonymous SNPs in all eight rice cultivars during three developmental time points.

SNPs can be used in association analyses to identify candidate genes for use as functional markers of stress tolerance in rice. SNPs can also be used to define the genetic structure of populations as well as the diversity and differentiation between rice populations, particularly with regard to adaptation to different ecological habitats.[30] The public availability of genomic and expressed sequence tag sequences of multiple rice genotypes has enabled the identification of many SNPs.[31] Using whole-genome resequencing data and the RNA-seq data in rice, we identified candidate genes that might be related to macronutrient transport and flavonoid pathways including anthocyanidin biosynthesis.[32]

When plotted with the normalized expression level, the LOC_Os01g43980 and LOC_Os01g43990 genes, which contained black rice-specific SNPs, were observed to be upregulated at the studied developmental periods in the black group (Supplementary Fig. 4A). Conversely, the LOC_Os03g44870, LOC_Os11g11770, and LOC_Os11g11810 genes were downregulated in the black group (Supplementary Fig. 4B). Among genes with white rice-specific nonsynonymous SNPs, the LOC_Os01g57310 and LOC_Os07g02490 genes were downregulated at all time points only in the white rice group (Supplementary Fig. 5). We did not detect any upregulated genes in the white rice group.

We identified seven DEGs with selected nonsynonymous SNPs in the black and white groups (Supplementary Table 12). Although hundreds of genes with nonsynonymous SNPs were detected in the black rice group, a relatively small number of genes exhibited the unique expression patterns shown in Supplementary Figures 4 and 5, whereby changes in amino acids did not change expression levels. Rather, genetic variation in protein-coding genes appears to have influenced posttranslational events, such as protein modification, protein–protein interaction, and turnover, which would cause the phenotypic differences between black and white rice accessions, such as in anthocyanin biosynthesis. Although we did not address SNPs located in intergenic regions, such as promoters, these likely play an important role in black and white rice phenotypic differences, both via the recognition of cis-acting elements in promoters by transcription factors and via changes in the transcription of black rice-specific genes.

Conclusion

We developed a multilayered screening method that combines both whole-genome resequencing and RNA-seq to identify genes that regulate specific phenotypes in plants. For eight rice accessions with differences in anthocyanin biosynthesis, amylose content, and heading date, we generated genome sequence and RNA-seq data at three developmental stages. A total of 98% of our preprocessed genome resequencing data could be uniquely mapped to the reference genome, producing an average coverage depth of 40.0×. The 24 nonreplicated RNA-seq experiments identified the 42,699 unique transcripts.

We were able to detect genomic variation, understand the population structure and LD, and detect genes that likely regulate the phenotypic differences between black and white rice varieties. Comparison between black and white rice cultivars showed that intensive selective sweeps occurred in chromosomes 3, 10, and 12. We identified three GO terms (secondary metabolic process, biosynthetic process, and response to stimulus) associated with 131 DEGs in the black rice cultivars. Our comparisons of the SNP structure and gene expression among the eight rice cultivars suggested that SNPs likely play an important role in determining the phenotypic differences between black and white rice cultivars. The expression patterns of genes containing selected nonsynonymous SNPs revealed two upregulated genes at three developmental time points and three downregulated genes at three developmental time points in the black cultivars and two downregulated genes at three developmental time points in the white cultivars. Integration of transcriptomic data further narrowed down the catalogue of genes involved in anthocyanin biosynthesis. Our method was additionally tested by searching for genes responsible for amylose content and heading date.

The three agronomic traits (rice grain colors, amylose content, and natural heading date) were clearly grouped together by the developmental stages, regardless of any other traits, suggesting that the developmental stage is the most significant factor that triggers global changes in gene expression. Especially, glutinous and nonglutinous black rice cultivars were

identified by different heading dates. Although the identified genes require experimental validation and an increased sample size would be helpful for analyzing population structure, our study demonstrates the potential of this screening method to identify genes underlying biological phenomena.

Author Contributions

Developed and wrote the code for the manuscript: C-KK, Y-JS. Verified and advised on the design and features of manuscript: SYW, YS, J-YL. Provided overall scientific and technical guidance and assisted with the creation of the manuscript: J-SC, Y-KK. All the authors contributed to the writing and improvement of the manuscript and read and approved the final version.

Abbreviations

LD: linkage disequilibrium
ROD: reduction of diversity
ZH_p: Z-transformed heterozygosity
PCA: principal component analysis
FPKM: fragments per kilobase of transcript per million mapped reads
SEA: singular enrichment analysis.

Supplementary Material

Supplementary Table 1. Generation and preprocessing of raw sequencing data. Note: [a]The percentage of reads left after preprocessing.

Supplementary Table 2. Preprocessing of raw sequencing reads.

Supplementary Table 3. Summary of genome resequencing and read mapping onto reference genome. Note: [a]The rate of zero coverage on one site.

Supplementary Table 4. Statistics of genomic locations and types for SNPs found from genome resequencing of eight rice accessions. Notes: [a]The synonymous SNPs in the CDS region. [b]The nonsynonymous SNPs in the CDS region.

Supplementary Table 5. The list of genes with the significance level of $ZH_p \leq -3.0$ for glutinous and nonglutinous rice accessions.

Supplementary Table 6. The number of upregulated and downregulated genes in black rice groups.

Supplementary Table 7. Singular enrichment analysis with 73 upregulated genes in black rice accessions. Note: P, biological processes.

Supplementary Table 8. The number of genes with synonymous SNPs detected in each phenotypic category. Notes: 1* means the SNP exists, and 0 means the SNP does not exist.

Supplementary Table 9. The number of genes with nonsynonymous SNPs detected in each phenotypic category. Notes: 1* means the SNP exists, and 0 means the SNP does not exist.

Supplementary Table 10. Singular enrichment analysis with 248 black rice-specific genes containing nonsynonymous SNPs using GO terms. Notes: P, biological processes; F, molecular function.

Supplementary Table 11. Singular enrichment analysis with 80 white rice-specific genes containing nonsynonymous SNPs using GO terms. Notes: P, biological processes; F, molecular function; C, cellular component.

Supplementary Table 12. The list of differentially expressed genes with selected nonsynonymous SNPs in the black and white groups.

Supplementary Figure 1. The graph shows the ROD values that were calculated on chromosomes 1–12. (a) The significant outlier regions (ie, those with right tail $\pi \leq 0.01$) for the glutinous and nonglutinous groups. (b) The significant outlier regions (ie, those with right tail $\pi \leq 0.01$) for the early, medium, and late developmental-stage groups.

Supplementary Figure 2. Circos diagram of the ROD patterns between the glutinous and nonglutinous groups. Abbreviations: G, glutinous group; W, white group (control); N, nonglutinous group; ROD (G/W), ROD between glutinous and white groups; ROD (N/W), ROD between nonglutinous and white groups.

Supplementary Figure 3. Circos diagram of the ROD patterns between the late stages among three developmental time points in the rice groups.

Supplementary Figure 4. Patterns of gene expression determined with selected nonsynonymous SNPs in the black group. The *X*-axis represents the 24 rice samples. B1–B18: six black rice accessions with three developmental stages, W1–W6: two white rice accessions with three developmental stages. (a) The two upregulated genes are shown at three developmental time points in the black group. (b) The three downregulated genes are shown at three developmental time points in the black group. B, six black rice with three time stages; W, two white rice with three time stages.

Supplementary Figure 5. The graph of the specific pattern of gene expression using selected nonsynonymous SNPs of the white group. The two downregulated genes show three developmental time points in the white group. Notes: B, six black rice with three time stages; W, two white rice with three time stages.

REFERENCES

1. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15:121–32.
2. Xu X, Liu X, Ge S, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol*. 2012;30:105–11.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63.
4. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet*. 2010;11:476–86.
5. Mikami I, Uwatoko N, Ikeda Y, et al. Allelic diversification at the wx locus in landraces of Asian rice. *Theor Appl Genet*. 2008;116:979–89.
6. Tian Z, Qian Q, Liu Q, et al. Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. *Proc Natl Acad Sci U S A*. 2009;106:21760–5.
7. Fasahat P, Rahman S, Ratnam W. Genetic controls on starch amylose content in wheat and rice grains. *J Genet*. 2014;93:279–92.

8. He J, Giusti MM. Anthocyanins: natural colorants with health-promoting properties. *Annu Rev Food Sci Technol*. 2010;1:163–87.
9. Winkel-Shirley B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol*. 2001;126:485–93.
10. Xie DY, Dixon RA. Proanthocyanidin biosynthesis – still more questions than answers? *Phytochemistry*. 2005;66:2127–44.
11. Shih CH, Chu H, Tang LK, et al. Functional characterization of key structural genes in rice flavonoid biosynthesis. *Planta*. 2008;228:1043–54.
12. Du H, Zhang L, Liu L, et al. Biochemical and molecular characterization of plant MYB transcription factor family. *Biochemistry (Mosc)*. 2009;74:1–11.
13. Shao Y, Jin L, Zhang G, Lu Y, Shen Y, Bao J. Association mapping of grain color, phenolic content, flavonoid content and antioxidant capacity in dehulled rice. *Theor Appl Genet*. 2011;122:1005–16.
14. Dogara AM, Jumare AI. Origin, distribution and heading date in cultivated rice. *Int J Plant Biol Res*. 2014;2:1–6.
15. Liang G, Zhang Z, Zhuang J. Quantitative trait loci for heading date and their relationship with genetic control of yield traits in rice (*Oryza sativa*). *Rice Sci*. 2013;20:1–12.
16. Kawahara Y, de la Bastide M, Hamilton JP, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013;6:1–10.
17. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*. 2005;28:289–301.
18. Slatkin M. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008;9:477–85.
19. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21:263–5.
20. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19:1639–45.
21. Rubin CJ, Zody MC, Eriksson J, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*. 2010;464:587–91.
22. Axelsson E, Ratnakumar A, Arendt ML, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*. 2013;495:360–4.
23. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7:562–78.
24. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res*. 2010;38:W64–70.
25. Kalinowski ST. Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity (Edinb)*. 2005;94:33–6.
26. McNally KL, Childs KL, Bohnert R, et al. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A*. 2009;106:12273–8.
27. Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data. *Genome Res*. 2005;15:1566–75.
28. Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD. Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics*. 2006;173:975–83.
29. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 2002;30:3894–900.
30. Parida SK, Mukerji M, Singh AK, et al. SNPs in stress-responsive rice genes: validation, genotyping, functional relevance and population structure. *BMC Genomics*. 2012;13:426.
31. Shirasawa K, Maeda H, Monna L, Kishitani S, Nishio T. The number of genes having different alleles between rice cultivars estimated by SNP analysis. *Theor Appl Genet*. 2007;115:1067–74.
32. Kim CK, Seol YJ, Shin Y, et al. Whole-genome resequencing and transcriptomic analysis to identify genes involved in leaf-color diversity in ornamental rice plants. *PLoS One*. 2015;10:e0124071.