Master's Thesis of Science

Genome-Wide Association Study of Rice Grain Morphology Combined with Micro-CT Phenotyping

Micro-CT를 이용한 현미 구조 전유전체
연관분석

February 2023

Graduate School of Agriculture and Life Science
Seoul National University
Crop Science and Biotechnology Major

Ji-Young Nam

Genome-Wide Association Study of Rice Grain Morphology Combined with Micro-CT Phenotyping

Under the direction of Dr. Hee-Jong Koh
Submitting a master's thesis of Science

February 2022

Graduate School of Agriculture and Life Science
Seoul National University
Crop Science and Biotechnology Major

Ji-Young Nam

Confirming the master's thesis written by
Ji-Young Nam
FEBRUARY 2023

Chair	_____
	Kwangsoo Kim, Ph.D
Vice Chair	_____
	Hee-Jong Koh, Ph.D
Examiner	_____
	Hyungsuk Kimm, Ph.D

Abstract

Rice (*Oryza sativa* L.) is one of the most essential stable crops in the world, providing food for over half of the global population. To enhance crop productivity and profitability in response to rising consumer demand, yield, disease resistance, and grain quality are regarded as the primary breeding objectives. Grain morphology is a crucial aspect to understand in rice breeding, as it is closely related to yield and grain quality.

Brown rice comprises three parts: endosperm, embryo, and bran layer. Despite the importance of internal grain structure for plant breeding and crop improvement, research in this area has been limited compared to the extensive studies conducted on grain shape and size. This is primarily due to the delicacy, small size, and opaque nature of the grain. In our study, to overcome the limitation of studying grain internal structure, a non-destructive 3D imaging technique, micro-CT, was employed to study the internal morphology of the grain. The phenotype data obtained from the imaging was then combined with genotype data to conduct a Genome-Wide Association Study (GWAS) to identify quantitative trait loci (QTLs) that control the trait.

As a result of the GWAS, 128 lead SNPs were detected and s genes such as *OsYUC9*, and *OsPUP7* were found to be within the QTL

region. The findings of the study indicated that the gene *OsPUP7*, which has previously been documented as having an association with grain size, may have a significant effect on grain size as observed in the panel used in this study. Moreover, the study proposed a novel relationship between *OsYUC9*, a gene regulating Indole-Acetic Acid (IAA) content in the developing grain, and the size of the rice embryo.

The study analyzed the frequency of alleles for each morphological part of the grain by selecting the most frequently occurring SNPs. The analysis of two lead SNP sites for the grain revealed significant differences between the alleles, with the A allele from the 3:26977492 site having a favorable effect on the grain morphology, resulting in longer and larger grain volume while having a narrower grain width. For the bran layer, three SNP sites were analyzed, and the results showed that the allele 4, which has three positive alleles (T, T, C), results in the largest bran size when compared to the grain for traits such as BGCR, BGVR, and BGLR. For the embryo, the results showed that the allele with G at the 1:9223183 site has a significant impact on the height and volume. These findings provide further insights into the genetic basis of rice morphological traits and suggest that these information could be utilized in breeding programs to produce high quality rice cultivars with improved grain size and nutrients.

Keyword : micro-CT, GWAS, grain morphology, QTL, Grain size,
　　　　Haplotype

Student Number : 2021-25309

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Micro-CT	micro-computed tomography
QTL	quantitative trait loci
LD	linkage disequilibrium
SNP	single nucleotide polymorphism
GL	grain length,
GW	grain width,
GT	grain thickness,
GCSA	grain cross-sectional area,
GSA	grain sagittal area,
GLA	grain longitudinal area,
GV	grain volume,
GLWR	grain length-width ratio,
EnV	endosperm volume,
EL	embryo length,
EW	embryo width,
EH	embryo height,
BCSA	bran cross-sectional area,
BSA	bran saiggital area,
BLA	bran longitudinal area,
BV	bran volume,
BGLAR	bran to grain longitudinal area ratio,
BGVR	bran to grain volume ratio

INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most important stable crops in the world, providing food for over half of the global population. According to the United States Department of Agriculture (USDA), global rice production in 2022/2023 is forecasted to reach 503.27 million tons, with more than 90% produced and consumed in Asia. To enhance crop productivity and profitability in response to rising consumer demand, yield, disease resistance, and grain quality are regarded as the primary breeding objectives. Out of these primary breeding targets, increasing yield and production has traditionally been the top priority, but there has been a shift in focus toward improving the rice grain quality(Fischer and Edmeades 2010). This trend is driven by the changing needs of consumers and the increasing awareness of the health and environmental impact of rice cultivation.

The grain morphology is a crucial aspect to understand in rice breeding, as it is closely related to both yield and grain quality. Typically, the grain yield is determined by the number of panicles per plant, the number of grains each panicle, and the grain weight (Huang, Jiang et al. 2013). Grain shape such as length(GL), width(GW), thickness(GT), and length-width ratio(LWR) is closely related to grain

weight and, consequently, yield, . Therefore, the genetic basis of grain size and shape in rice has been extensively studied and several associated quantitative trait loci (QTLs) have been identified.

The developmental stages of the spikelet hull primarily shape the size and shape of rice grains. This structural component acts as a protective layer around the grain, shielding it from external damage while providing support. The coordination of cell proliferation and expansion within the hull determines the grain's filling capacity and subsequently regulates the final size and shape of the grain. During the cell proliferation stage, cells undergo division, leading to an increase in cell number, whereas during the cell expansion stage, cells increase in size. According to a recent review, several genes, including *GS3*, *DEP1*, *GSE5*, *SG1*, *GL3.1*, *FUWA*, *GW2*, *GSE5*, and *GL3.1*, regulate cell proliferation in the spikelet hull, and *GLW7*, *APG1*, *GL4*, *SLG*, *WTG1*, and *BUL1* are involved in cell expansion, both playing crucial roles in determining grain size (Li, Xu et al. 2018).

The gene *GS3* has been first identified as the major size associated QTL. It is a well-known example of a gene that negatively regulates grain size and weight through its impact on cell proliferation during the grain development stage(Fan, Xing et al. 2006). The *GS3* encodes a transmembrane protein with four putative domains, OSR, TM, TNFR/NGFR, and VWFCS, that regulate grain size differently

(Mao, Sun et al. 2010). A loss of function or deletion in the *GS3* domain results in longer grain length.

Rice grain quality includes characteristics such as appearance, milling, cooking, nutrition and eating quality. The internal structure of grain is closely linked to the quality characteristics. The grain is composed of four main parts: the hull, the bran, the embryo, and the endosperm (Fig. 1). The composition of paddy rice can be roughly estimated to consist of 20% hull, 8% bran, 2% embryo, and 70% endosperm (Chang and Bardenas 1965). The composition and proportion of different parts can significantly impact the quality, which are essential factors for consumer acceptability and farmer profitability (Li, Chen et al. 2022).
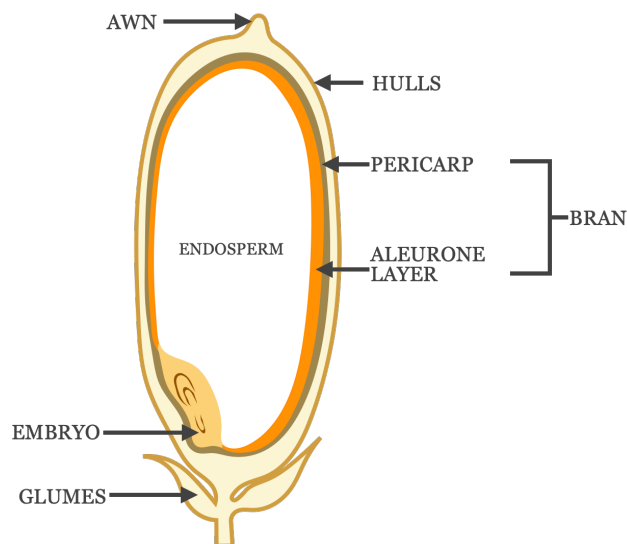
Figure 1. Morphological structure of paddy rice

The outermost portion of brown rice, the bran layer, is comprised of two distinct layers, the pericarp and aleurone layers. Despite its small proportion, the bran layer is considered to be the most nutritious part of the grain, along with the embryo, as it contains a rich blend of essential nutrients, including lipids (5-22%), carbohydrates (34-52%), fiber (7-11%), ash (6-10%), moisture (8-12%), and protein (10-16%)(Bodie, Micciche et al. 2019). In contrast, the starchy endosperm, the primary source of energy in white rice, consists primarily of starch and a small amount of proteins (Wu, Liu et al. 2016). Regardless of the nutritional importance of the bran layer, it is frequently removed during milling process to improve the eating quality. This causes a reduction of nutritional benefits. Therefore, there is a tradeoff between increasing the nutritional value of rice by preserving the aleurone layer and improving its eating quality by eliminating it. Hence, it is essential to investigate the QTLs affecting the size of the bran layer to determine the optimal proportion.

Aleurone layer, the largest part of the bran, is composed of a variable number of cells. The rice aleurone layer is typically composed of a single cell layer, but in some instances, it can consist of up to three or four cell layers at the dorsal side of the grain. The number of cell layers present in the rice aleurone layer can impact the volume of the layer and subsequently the nutritional value of the grain.

genes such as *OsROS1*, *OsCR4*, and *ADL1* are found to be associated with the number of aleurone cell layers in rice. According to previews research done by Liu, the gene *OsROS1* plays a crucial role in negatively controlling the number of aleurone cell layers in rice. The point mutation in the 14th intron of the gene *OsROS1* in the *thick aleurone 2-1* (*ta2-1*) mutant resulted in an increase in the average number of cell layers to 4.8, with some regions exhibiting up to 10 layers (Liu, Wu et al. 2018). The *Rice Crinkly4* (*OsCR4*) plays a critical role in the differentiation of epidermal cells in the palea and lemma, contributing to the proper interlocking structure of the spikelet and ultimately leading to fertile seed formation and seed setting. RNA interference-mediated suppression of mRNA expression in *OsCR4i* seeds resulted in discontinuous cuticles, uneven growth of the embryo and endosperm, as well as disruptions in the aleurone layer in the form of missing patches (Pu, Ma et al. 2012). The research found that intermediate and strong alleles of the *ADL1* gene in rice result in shootless and globular-arrested embryos, respectively. The strong allele also leads to partial loss of the aleurone layer on the ventral side. The *adaxialized leaf1* (*adl1*) mutants in rice, characterized by a mutation in the phytocalpain gene, exhibit phenotypic effects such as the presence of shootless and globular-arrested embryos and partial loss of the aleurone layer on the ventral side. It is observed from

previous studies that most of the genes affecting the aleurone layer phenotype also result in defects in grain shape, rendering them unsuitable for commercial purposes. Furthermore, most studies have utilized mutant lines, which may not accurately reflect natural variation in this trait. As such, further investigation of the aleurone layer in naturally occurring genetic variation may provide valuable insights for rice breeding efforts aimed at improving grain quality and nutritional content.

The embryo, or germ, of a rice grain contains essential fatty acids and vitamins. Further understanding of the structure and composition of the rice embryo is crucial in developing strategies aimed at increasing its nutritional content and ensuring seed viability. Genes such as *Giant Embryo* (*GE*), *GOLIATH* (*GO*) and *Large Embryo*(*LE*) is known to affect the embryo size. The *GE* gene, encoding the cytochrome P450 protein CYP78A13, is a crucial determinant of the size and development of rice embryos and endosperm. The loss of function mutation of the *GE* gene results in an increase in the size of rice embryos and a reduction in the endosperm size (Nagasawa, Hibara et al. 2013). The *LE*, which encodes a C3HC4-type RING finger protein and is highly expressed in seeds at a late developmental stage, was confirmed to play a role in determining the size of the rice embryo. The size of the rice embryo was increased

because of the suppression of the expression of the *LE* gene through the utilization of RNA interference (Lee, Piao et al. 2019). The study revealed that the size of the embryos in three mutants, *le*, *ge*, and *ge^s*, was found to be in the order of *ge^s*, *ge*, *le*. The study also found that these mutants had higher nutritional content compared to wild rice, including higher levels of protein, lipid, vitamin B1, vitamin E, and GABA. Especially the GABA content in mutants of *le*, *ge*, and *ge^s* was found to be significantly increased by 150%, 400%, and 850%, respectively. In conclusion, understanding the internal structure of rice grain is crucial for improving rice quality, nutritional value, and processing efficiency. Further research on the internal structure of rice grain will aid in the development of new rice varieties that are better suited to the needs of consumers and farmers.

Regardless of its importance, phenotyping of the grain internal structure is challenging due to its small size and opaque nature. The internal structures are not visible to the naked eye and require specialized techniques such as microscopy to be visualized. However, the whole mount or sectioning methods can be time consuming and the detail that can be obtained from it is limited. Additionally, many of the internal structures within a seed are very delicate and are easily damaged during the phenotyping process. To overcome this issue, techniques like X-ray imaging, CT scanning, or MRI have been

recently emerged to visualize the internal structure of seeds.

Micro-Computed tomography (micro CT) is a non-destructive imaging technique that can create detailed 3D images of plants. Micro CT uses a micro-focus X-ray tube to irradiate a conical X-ray beam that projects the sample onto a 2-D image detector. A 3-D matrix is reconstructed to represent the scanned object by transforming these projections acquired at different angles. Each element in the matrix holds a grey value corresponding to the material's density and atomic number (Dhondt, Vanhaeren et al. 2010).

These images can be used to phenotype traits, such as their shape, size, and internal structure. The use of micro-CT is becoming more popular for accurate characterization of internal seed morphology, as it provides higher spatial and temporal resolution compared to the traditional destructive sampling method (Gargiulo, Leonarduzzi et al. 2020). The use of micro-CT has enabled the accurate phenotyping of previously difficult to assess traits, such as chalkiness, and has been demonstrated as a powerful new phenotyping method (Su and Xiao 2020).

Overall, understanding the internal structure of rice and the genetic factors that control its development can help to improve the nutritional value, cooking properties and appearance of the grain, making it more desirable for consumption. Additionally, breeding for

desired internal structure can also improve the functional properties of rice like its milling, cooking, and eating quality. Therefore, micro-CT was used to obtain detailed information about size and shape of the various structures within a seed, including the embryo, endosperm, and bran layer. Then these phenotype data was combined with genotype data to conduct genome-wide association study to find QTLs controlling the internal and outer structure of rice grain.

MATERIALS AND METHODS

- **Plant material and sample preparation**

A total of 215 rice accessions were used. This population has been maintained in Agricultural Genetic Resource Center, Seoul National University(SNU), Suwon, South Korea. All rice accessions were cultivated in SNU, Suwon, South Korea using conventional lowland cultivation methods and natural long-day conditions. The cultivars were harvested 45 to 50 days after heading and air-dried to reach a moisture content of 13-14%. The plants were then threshed, and the resulting grains were stored in a cold environment at 12° C until the experiment was conducted.

- **Sample Preparation and micro-CT imaging**

Before the phenotyping, obtained grains were dehulled using a dehull machine to yield brown rice. Then at least 10 intact kernel per accession were wrapped and held securely to the sample holder using Parafilm M®. The sample holder was placed inside the micro-CT (Skyscan 1272®, Bruker). It was then scanned at 60 kV, 166 μA, with image resolution of 7μm with 360° rotation around the vertical axis with rotation step of 0.4° , and frame averaging of 2. Images were

reconstructed using NRecon software v. 1.7.4.2 (Bruker) using 40% beam hardening correction and smoothing of 2.

- **Segmentation and phenotype analysis**

Separation of the reconstructed 3D images were required before measuring the structural characteristics. 3D Slicer v.5.1.0 software was used for the segmentation process, and MONAILabel module was implemented to aid labeling embryo, endosperm, and bran (Kikinis, Pieper et al. 2014). First, the grains were roughly segmented by using pretrained model with 95% validation accuracy. Then the segments were further modified to increase the accuracy.

Then, segments of whole kernel were repositioned with principal axes to analyze the internal grain structure using the SegmentGeometry module (Huie, Summers et al. 2022)(Figure 2). Realigning the rice grain's position makes it possible to ensure that the 2D sectioning image is taken at the exact position for all samples. The imaging process was carried out by obtaining cross-sectional, sagittal, and longitudinal sectioning images of the grains after repositioning. The area of each structure were measured at the largest sectioning area for the Grain. Also, the whole grain's length, width, thickness, and volume were calculated for each structure. The segmented embryo volume was separated and repositioned using the

principal axes to measure the width, height and length for embryo structure. In total, 17 structural traits of grain were measured.
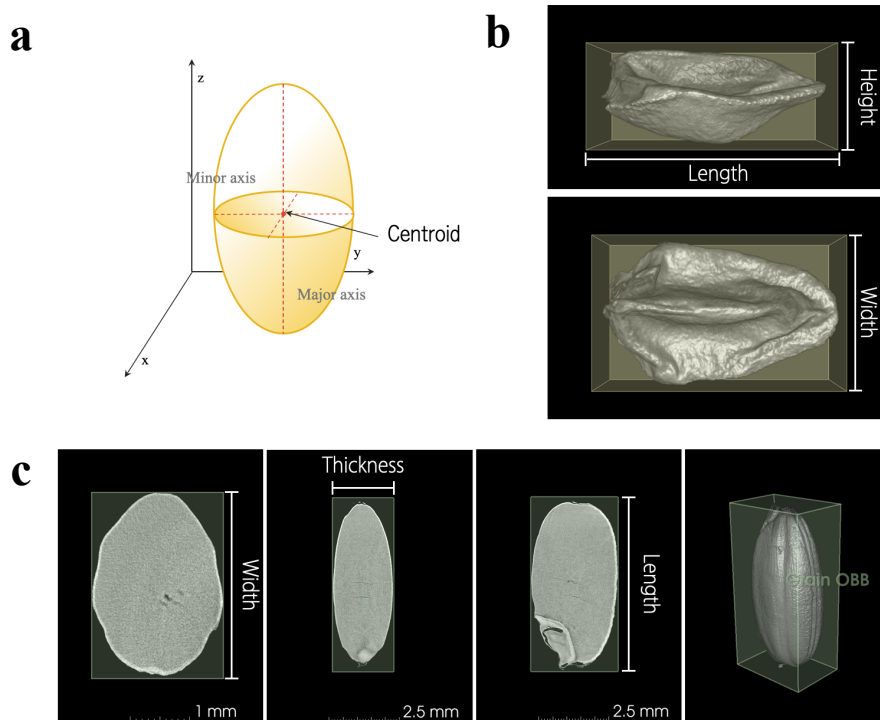
Figure 2. Repositioning and measurement method of of grain kernel and internal structures. (a) represents how principal axes of elliptical grain is measured then aligned along the plane. (b) Embryo measurements. Embryo segment is repositioned then the height, length, and width is measured. (c) Grain measurements. The grain is repositioned then the length, width, thickness is measured. For the cross-sectional, sagittal, longitudinal sectioning images, the slice with largest area were chosen for each and the area of grain parts were measured.

- NGS Analysis and Genotyping

The CTAB method by Murray and Thompson was used to extract genomic DNA from the leaves of 90-day-old seedlings. Using the TruSeq Nano DNA Library Prep kits (Illumina) and the manufacturer's instructions, total DNA was sheared into 450-500 bp fragments and used for library assembly. The library size distribution was verified with the Agilent Technologies 2100 Bioanalyzer and a DNA 1000 chip. Whole genome sequencing data were generated on the Illumina HiSeq X system, producing 2×150 bp paired-end reads with a sequencing depth of $>10\times$ per sample. Raw reads were processed and aligned to the rice reference genome (Nipponbare, IRGSP v1.0) using the BWA v0.7.17 MEM algorithm with default parameters. Nucleotide variants were called by the HaplotypeCaller function of GATK v4.1.2 with the parameters: max-missing 0.95, minQ 30, minDP 5. Furthermore, vc.getHetcount command in GATK v4.12 was used to filter nucleotide variants with proportions of heterozygous genotypes of >0.05. GWAS analysis was conducted using the remaining set of 1,254,682 high-quality SNP markers.

- Statistical analysis

All statistical analyses were performed using With R statistical software (version 4.1.2). Correlation between all traits was assessed using Pearson's correlation method from the corrplot package (Wei and Simko 2017). The significance of correlation was determined by correlation significance of 0.05. The frequency distribution of grain structure traits were visualized along with correlation.

The statistical analysis of haplotype and allelic variation was performed through Anova, t-test, and Duncan test to assess the significance of the relationships between phenotypic traits and genotypic variations.

- Population structure and linkage decay analysis

In order to analyze the molecular diversity of 215 rice accession, principal component analysis(PCA) was conducted using the 25 rice accessions from the 3K rice genome as reference panel. Principal component(PC) of the accessions were analyzed using PLINK 1.9.0 then the first two PC were visualized using ggplot in R.

For further analyses of genetic relationships, a total of 33,319 SNPs (MAF > 0.05) were obtained by a linkage disequilibrium (LD)-based SNP pruning method using PLINK v1.9. Then the pairwise LD distance between the pruned SNPs were analyzed by PopLDdecay(Wang,

Shapey et al. 2019) software within a sliding window of 10000kb with a step size of 500bp. The mean linked LD was calculated by dividing the total r^2 value (a measure of LD between markers) by the number of corresponding loci pairs. We imported the r^2 data into R and the LD decay was drawn using ggplot.

- **Association mapping**

For the genome-wide association studies (GWAS), previously mentioned 1,254,682 high-quality SNP markers were used. The association study was conducted using linear mixed-models (LMM) implemented in FaST-LMM v2.07 (Lippert, Listgarten et al. 2011) . The threshold values for associated SNPs were obtained by dividing the significance level 0.05 by the effective number of independent SNPs, $Log10(P) \geq -Log10(0.05/33,319) \approx 5.82$. Based on the pattern of LD in the population, the distance at which LD decay drops in half was used as tagging threshold to select lead SNPs and determine each QTL's chromosomal region. The haplotype analysis was conducted utilizing all types of genetic variants, including SNPs and InDels without consideration of MAF. The accessions with any missing or heterozygous genotype were removed from the haplotype analysis.

–

Result

1. Population structure

The population structure of 215 accessions was studied using a principal component analysis (PCA) based on linkage disequilibrium (LD)-pruned single nucleotide polymorphisms (SNPs). The results showed that the first two PC dimensions explained 77.7% of the total genetic variation, with 59.3% and 18.4% accounted for by the first and second PC dimensions, respectively. The first two PCs allowed for the clear identification of the subpopulation groups within the panel. Using the PCA results, the population genetic structure of the 215 accessions was divided into four groups: 197 temperate japonica, 10 tropical japonica, 5 indica, and 1 aromatic.
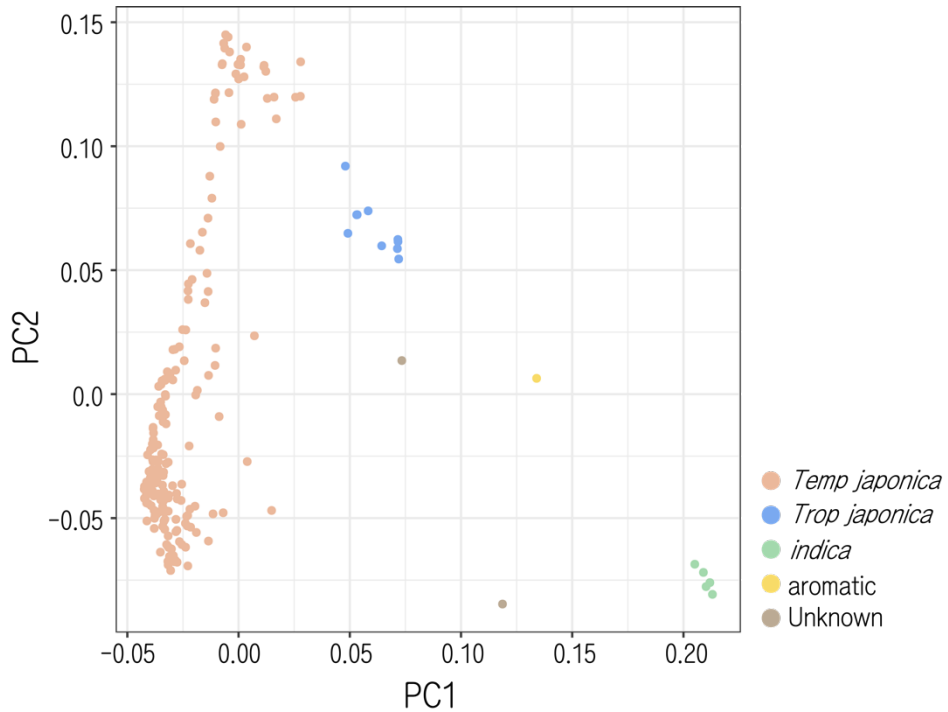
Figure 3. The population structure of 215 rice accessions determined by principal component analysis. The first and second principal components (PC1 and PC2) divided the population into five subgroups: temperate japonica (Temp japonica), tropical japonica (Trop japonica), indica, aromatic, and unknown. These subgroups are represented visually by different colors.

2. Phenotypic Variation and Correlation

In the study, a total of 17 morphological traits were assessed using the Slicer instrument. Additionally, 5 calculated traits were determined, which included grain length-width ratio, bran to grain longitudinal area ratio, bran to grain volume ratio, embryo to grain width ratio, and embryo to grain volume ratio (Table 1). The average volume for each of the internal components were as follows: 14.852 ±1.564 mm^3 for grain, 12.861 ±1.421 mm^3 for endosperm, 0.569 ±0.073 mm^3 for embryo, and 1.412 ±0.138 mm^3 for bran layer.

The phenotype data collected allows quantification of the volume occupied by each morphological component, which was traditionally assessed by measuring the weight of milling byproducts. The endosperm takes up about 86.6 % of the grain and the embryo and bran layer takes up 3.8% and 9.5% of the whole grain. These data provide insights into the internal structure of the sample and can be used to inform further studies.

The distribution analysis revealed that most of the traits showed a normal distribution. However, for embryo width, the data points were highly concentrated around the range of 1.09 to 1.38 mm, deviating from a normal distribution (Fig. 4). In addition to the distribution analysis, a correlation analysis was also conducted on the

traits. The results of the correlation analysis revealed that most of the traits showed positive correlation, with correlation coefficients ranging from 0.017 to 1. This could be because as the size of the grain increases, so does the size of the other components. However, the grain thickness, width, cross-sectional area had weak negative correlation with the grain length. Additionally, a strong correlation was observed between endosperm volume and variable grain volume ($r^2 =$ 1, $p < 0.05$). This correlation could be related to the fact that endosperm takes up a significant portion of the grain volume.

Table 1. Phenotype variation and distribution pattern of 22 grain structure related traits.

	Trait	Min	Max	Average±SD	Coefficient of variation (%)
Grain	GL (mm)	4.460	6.538	5.110 ±0.309	6.047
	GW (mm)	2.085	3.326	2.952 ±0.158	5.344
	GT (mm)	1.686	2.361	2.065 ±0.098	4.767
	GCSA (mm^2)	2.612	5.630	4.417 ±0.400	9.056
	GSA (mm^2)	6.287	9.860	7.796 ±0.577	7.396
	GLA (mm^2)	9.976	14.966	11.728 ±0.925	7.888
	GV (mm^3)	11.093	19.726	14.852 ±1.564	10.533
	GLWR	1.435	3.069	1.737 ±0.170	9.788
Endosperm	EV (mm^3)	9.285	17.172	12.861 ±1.421	11.049
Embryo	EL (mm)	1.655	2.509	2.066 ±0.109	5.286
	EW (mm)	0.795	2.279	1.266 ±0.119	9.420
	EH (mm)	0.763	1.301	0.946 ±0.066	7.025
	ELA (mm^2)	0.794	1.686	1.128 ±0.122	10.816
	EV (mm^3)	0.299	0.889	0.569 ±0.073	12.889

	EGWR	0.612	0.900	0.701 ± 0.041	5.795
	EGVR	0.021	0.066	0.038 ± 0.004	11.525
Bran	BCSA (mm^2)	0.227	0.423	0.335 ±0.028	8.245
	BSA (mm^2)	0.366	0.672	0.490 ±0.048	9.857
	BLA (mm^2)	0.485	0.983	0.705 ±0.079	11.192
	BV (mm^3)	1.089	1.831	1.412 ± 0.138	9.750
	BGLAR	0.046	0.081	0.060 ± 0.006	10.05
	BGVR	0.081	0.114	0.097 ± 0.006	6.556

GL = grain length, GW = grain width, GT = grain thickness, GCSA = grain cross-sectional area, GSA = grain saggital area, GLA = grain longitudinal area, GV = grain volume, GLWR = grain length-width ratio, EnV = endosperm volume, EL = embryo length, EW = embryo width, EH = embryo height, ELA = embryo longitudinal area, EV = embryo volume, EGWR = embryo to grain width ratio, EGVR = embryo to grain volume ratio, BCSA = bran cross-sectional area, BSA = bran sagittal area, BLA = bran longitudinal area, BV = bran volume, BGLAR = bran to grain longitudinal area ratio, BGVR = bran to grain volume ratio
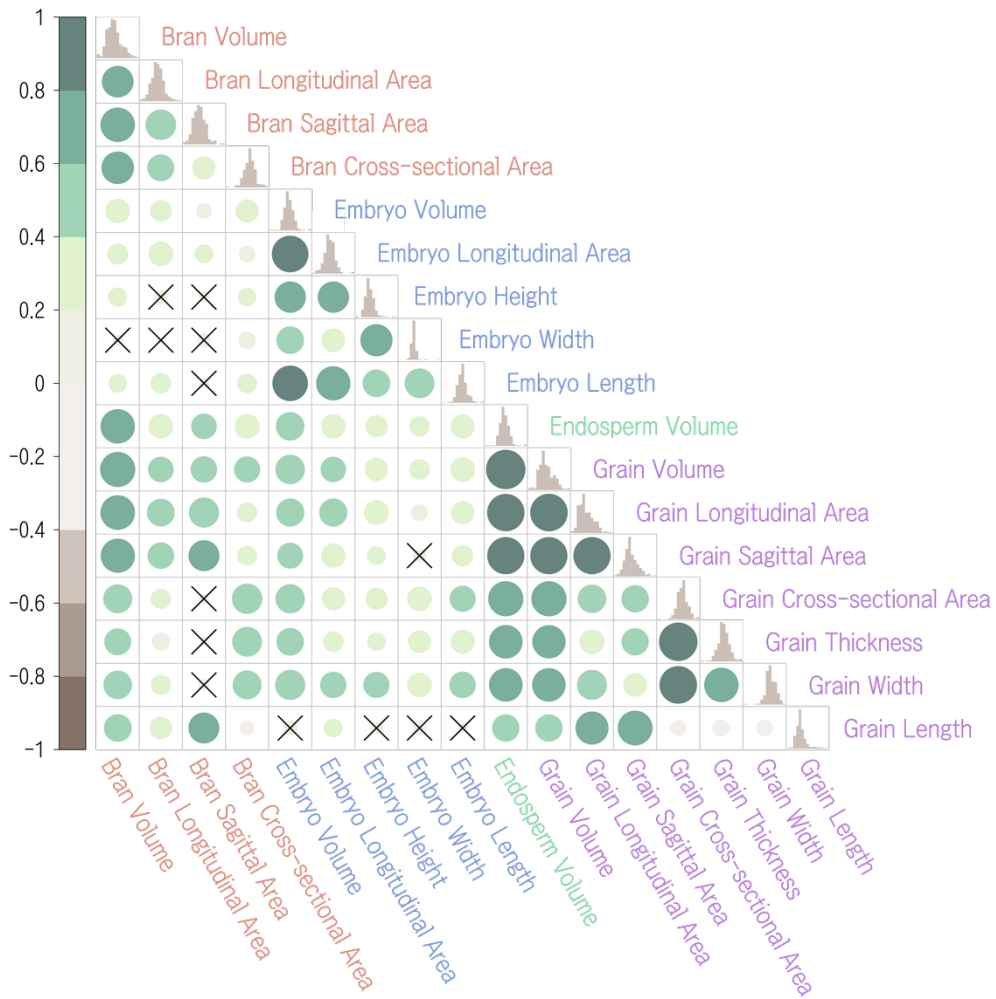
Figure 4. Distribution of grain morphological traits in rice and Pearson coefficient correlation analysis. The upper section of the figure illustrates the distribution of the 17 grain morphological traits, while the lower section presents the correlation coefficient between these traits. The color and size of the circles correspond to absolute values of corresponding r. The cross symbol in the matrix indicates weak correlation, with a correlation coefficient less than 0.05.

3. LD decay analysis

Linkage disequilibrium (LD) decay analysis was performed using the independent SNPs to understand the genetic structure of the population. To reduce the number of datasets and increase the efficiency of the GWAS, a pairwise pruning of SNPs was performed using the software PLINK with a window size of 50, a step size of 10, and an LD threshold of 0.2. This process removed SNPs with high correlation and 33,319 independent SNPs remained. Subsequently, these SNPs were utilized for conducting the LD decay analysis along the chromosomes. The average value of LD, represented as the correlation coefficient r^2, showed a pattern of decline as the distance between pairwise SNPs increased. The value of LD, represented by the correlation coefficient r^2, began with a maximum value of 0.84 and declined to half of that value at a physical distance of approximately 446 kb.

The half decay distance of 446 kb is relatively large in comparison to the LD intervals observed in a GWAS panel consisting of multiple subspecies. Previews research has shown that the LD size varies between temperate japonica (>500 kb), tropical japonica (~150 kb) and indica (~75 kb) (Mather, Caicedo et al. 2007). The LD distance is highly influenced by the population structure, and as demonstrated

by the PCA results, the panel primarily consisted of temperate japonica. Hence, the large half-decay distance observed in this study can be primarily attributed to the dominance of temperate japonica in the panel used.
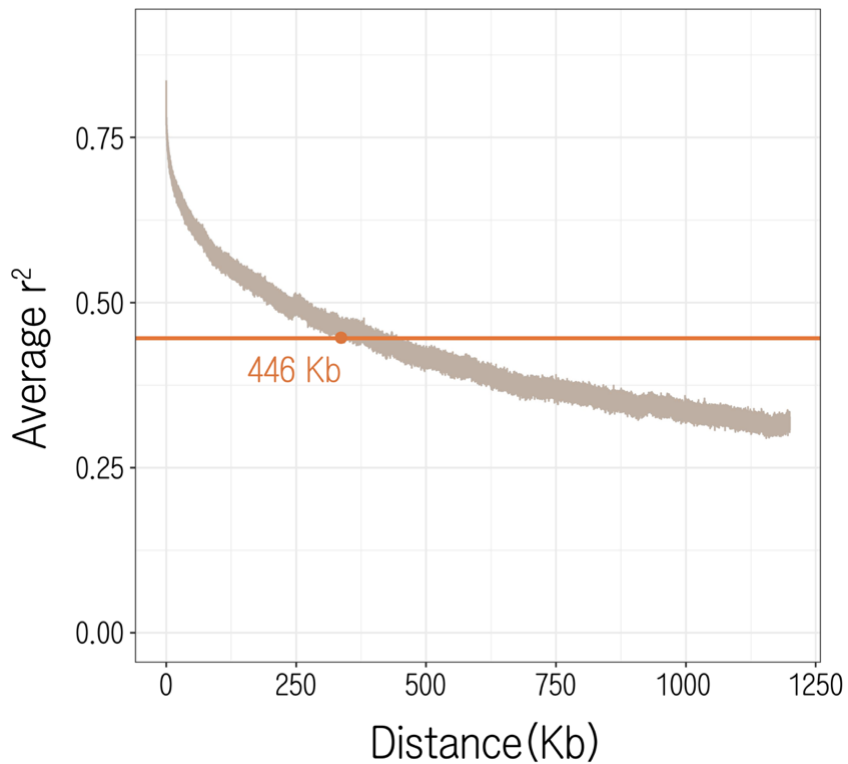

Figure 5. The genetic structure of population revealed by LD decay analysis. The x-axis indicates the distance between SNPs in Kb and the y-axis represents the average r^2. The half-decay point is marked on the graph.

4. Genome-wide association study

In an effort to identify QTLs related to grain morphology, 22 GWAS were performed using the FaST-LMM algorithm. The threshold for genome-wide significance was set at 5.82, corresponding to -log10(0.05/33319). For several traits such as GL, GLWR, GLA, EW, and EH, there were a high degree of linkage disequilibrium between neighboring allelic variations, leading to a significant number of SNP sites exhibiting redundant information. To address this, many studies have aimed to identify "representative" SNPs, referred to as tag SNPs, that can provide adequate information about nearby variants. By selecting an appropriate subset of SNPs, tag SNPs can accurately predict the rest of the SNPs with minimal error (Halperin, Kimmel et al. 2005). To summarize, the half-decay of LD was applied in the identification of tag SNPs through the PLINK software with a half decay distance of 446 kb, in order to reduce the number of SNPs required for association studies while still preserving considerable representation of genetic variation.

A total of 128 lead SNPs across all traits were identified, with 44 SNPs associated with grain-related traits, 79 SNPs associated with embryo-related traits, 3 SNPs associated with bran layer traits, and 2 SNPs associated with endosperm volume traits (Table S1 and Figure S1). We found that SNP 5:27943304, were associated with total of four

grain traits: GW, GV, GCSA, GT, GLWR. Additionally, a lead SNP detected at 5:28063233 for BCSA was within close proximity (119 kb) and exhibited high LD (r^2 = 0.92) with 5:27943304, suggesting that these two SNPs may represent a shared association signal. The detection of two lead SNPs at positions 3:26977492 and 5:27943304 for the endosperm phenotype were found to coincide with the lead SNPs identified for grain, implying the presence of a shared association signal between the two as well. Approximately 27 loci were detected for more than two traits among the 128 detected. The 892 kb genomic region surrounding each lead SNP (with 446 kb upstream and 446 kb downstream) was considered as a potential QTL and genes with relevant functions were evaluated.

5. Haplotype Analyses of the Candidate Gene

The gene *OsYUC9*, which is involved in the synthesis of indole acetic acid (IAA), was identified within a QTL region that was closely linked to a lead SNP (1:9237559) associated with embryo width (Figure. 6). Two variants were discovered on the gene, located at positions 1:9488879 and 1:9493610. The first variant, at 1:9488879, is a missense mutation with a change from T to G, while the second variant, at 1:9493610, is a possible splice region alteration from T to C. Two haplotypes were constructed and the

dominant haplotype 2 accounted for 92% of the total population. The significance of the differences between the two haplotypes was estimated using a t-test, with a p-value of 0.02, indicating significant differences. The two haplotypes were further confirmed by a Duncan multiple range test, which indicated significant differences between them. The mean and standard deviation for haplotype 1 was calculated to be 1.20 ± 0.16, while for haplotype 2 it was 1.27 ± 0.12. However, the haplotypes were not found to have a significant impact on other traits such as the length and height of the embryo, and the overall volume of the embryo was also not affected.
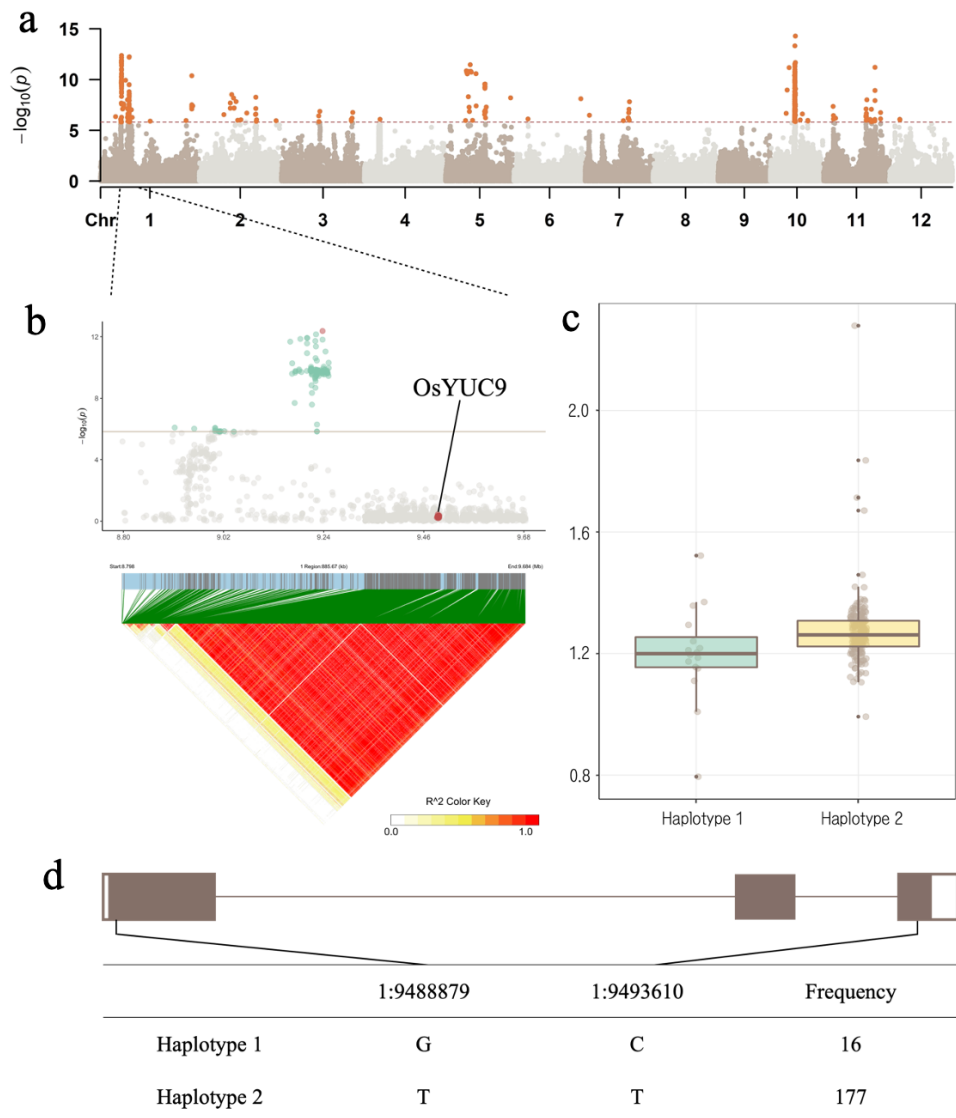
Figure 6. Candidate gene analysis of position 1:9237559. (a)
Manhattan plot for GWAS result of embryo width (b) LD heat map of
QTL region for the detected lead SNP. The lead SNP and the
candidate gene *OsYUC9* is marked in red. (c) The distribution of
embryo width for the two haplotypes of 193 accessions. (d) The
frequency of each haplotype among the population.

The gene *OsPUP7* was found near a SNP at 5:27943303, which affects various grain traits such as GW, GV, GCSA, GT, and GLWR. Previous research has shown that this gene is involved in the regulation of grain size and weight (Qi and Xiong 2013). The two missense mutations were identified at the chromosomal positions 5:27690359 and 5:27690860, respectively (Figure 7). These mutations resulted in the alteration of alleles from A to G and C to T, leading to the formation of two haplotypes, namely AC and GT. The haplotype AC dominated the population, with 92% frequency.

Statistical analysis, including t-tests, revealed that the haplotypes showed a p-value less than 0.01 for GT, GCSA, GLWR and less than 0.05 for GW and GV. Furthermore, the Duncan test confirmed that all haplotypes were significantly different from each other. The average and standard deviation for the two haplotypes of five traits were determined as follows: For the trait GW, haplotype 1 was found to have a mean of 2.96 ± 0.14, while haplotype 2 had a mean of 2.85 ± 0.31. For the trait GT, haplotype 1 had a mean of 2.07 ± 0.09 and haplotype 2 had a mean of 1.98 ± 0.11. For the trait GLWR, haplotype 1 had a mean of 1.68 ± 0.10 and haplotype 2 had a mean of 2.03 ± 0.39. For the trait GCSA, haplotype 1 had a mean of 4.45 ± 0.37 and haplotype 2 had a mean of 4.09 ± 0.58. Finally, for the trait GV, haplotype 1 had a mean of 14.80 ± 1.51 and haplotype 2

had a mean of 15.66 ± 2.15. Even though haplotype 1 had greater width, thickness, and cross-sectional area, the results indicated that haplotype 2 had a larger length-width ratio and volume.

The results of the analysis on the two haplotypes revealed a statistically significant difference in grain length as well. The p-value for the comparison of the two haplotypes was less than 0.01. The average grain length and standard deviation for haplotype 1 was 5.06 ± 0.23, while for haplotype 2, it was 5.7 ± 0.50. The results suggest that while haplotype 1 displays a wider grain type, haplotype 2 presents a longer grain type. This finding is consistent with previous research that showed grain length plays a more significant role in determining the shape and weight of the grain(Wu, Xu et al. 2018) .
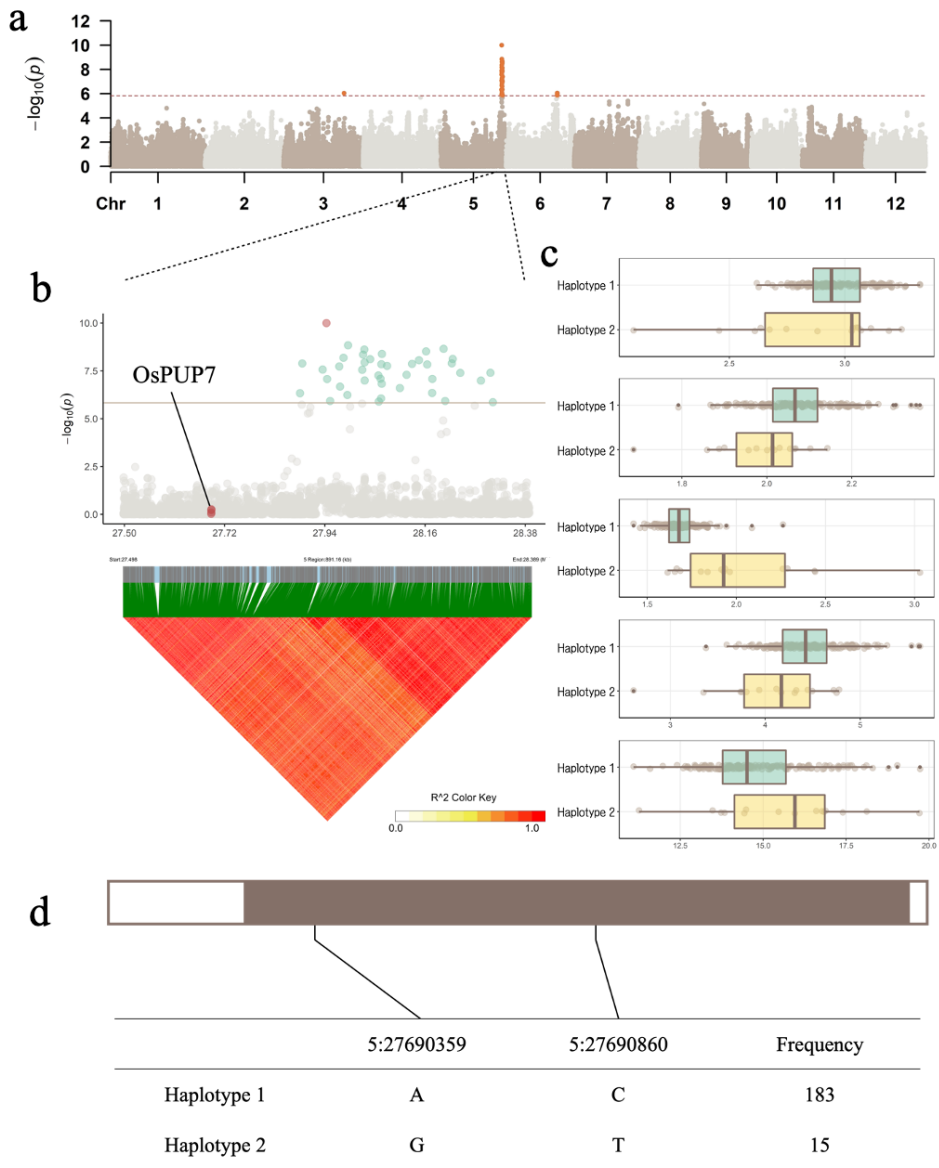
Figure 7. Candidate gene analysis of position 5:27943303. (a) Manhattan plot for GWAS result of grain width (b) LD heat map of QTL region. The lead SNP and the candidate gene *OsPUP7* is marked in red. (c) The distribution of GW, GT, GLWR, GCSA, and GV for the two haplotypes of 198 accessions.　(d) The frequency of each haplotype among the population.

6. Allele frequency analysis

Out of all the Single Nucleotide Polymorphisms (SNPs) detected for each morphological part of the grain, the most frequently occurring SNPs were selected to investigate their impact on the grain's morphology. Two lead SNPs sites (5:27943304 and 3:26977492) were chosen for the analysis of the allelic frequency on morphological traits of the grain(Figure 8). The alleles at each site differed from G to A and T to G, respectively, with a total of four allelic combinations. However, due to insufficient sample size, the allele frequency of allele 3 was excluded from further statistical analysis. The lead SNP site on chromosome 3 showed a large area of highly correlated SNPs ($r^2 > 0.8$). An ANOVA test showed significant differences between the alleles. The Duncan test revealed that for GL, GW, GT, GCSA, and GV, allele 1 was significantly different from alleles 2 and 4. Furthermore, all three alleles were found to be different from each other for GLA, GSA, and GLWR. The results suggest that the A allele from the 3:26977492 site had the most significant effect, leading to longer and larger grain volume while having a narrower grain width.

For the morphological traits of the bran layer, three SNPs (detected from bran-related traits) were analyzed due to the absence of reoccurring SNPs(Figure 9). Out of the eight allelic combinations

from the three SNP sites, six were sufficient for statistical testing. The results of the ANOVA test showed that BV was statistically significant with a p-value less than 0.05, while BCSA, BLA, BGCR, BGLR, and BGVR showed a p-value less than 0.01. Upon individual testing of each haplotype, allele T, T, and C were found to be positive alleles for BGCR, BGVR, and BGLR, respectively, while having a negative effect on BCSA. Although, there was no statistical difference between the groups for BV according to the Duncan test. A clear grouping was observed between alleles 1, 2, 4, and 5, 7, 8, which was caused by the SNP at 9:375565 for the bran-to-grain longitudinal area ratio. The C allele showed a larger bran area compared to the grain compared to the G allele. The allele 4 with positive alleles 3 showed the largest BGVR and BGCR, while the allele 5 with all negative values showed the lowest values for these traits.

Five repeating SNP sites (1:9223183, 1:9989728, 10:10946163, 11:22576941, 11:5330959) were selected for the allelic combination analysis of the embryo (Figure 10). Out of the ten allelic combinations, only two (allele 3 and allele 4) were found to have enough accessions to conduct a statistical analysis. The results showed that the alleles were different at the 1:9223183 site, with a transition from T to G. The statistical analysis using the t-test

revealed that the two alleles were significantly different, with a p-value less than 0.01, for traits such as EV, EH, and ELA. This finding suggests that the allele with G at the 1:9223183 site has a positive impact on the height and size of the embryo.

The results of the analysis provided valuable insights into the genetic basis of the grain structure and the relationship between the allelic variations and the phenotypic traits and the positive alleles for each grain trait can be further investigated in order to develop high-yielding varieties with improved grain quality.
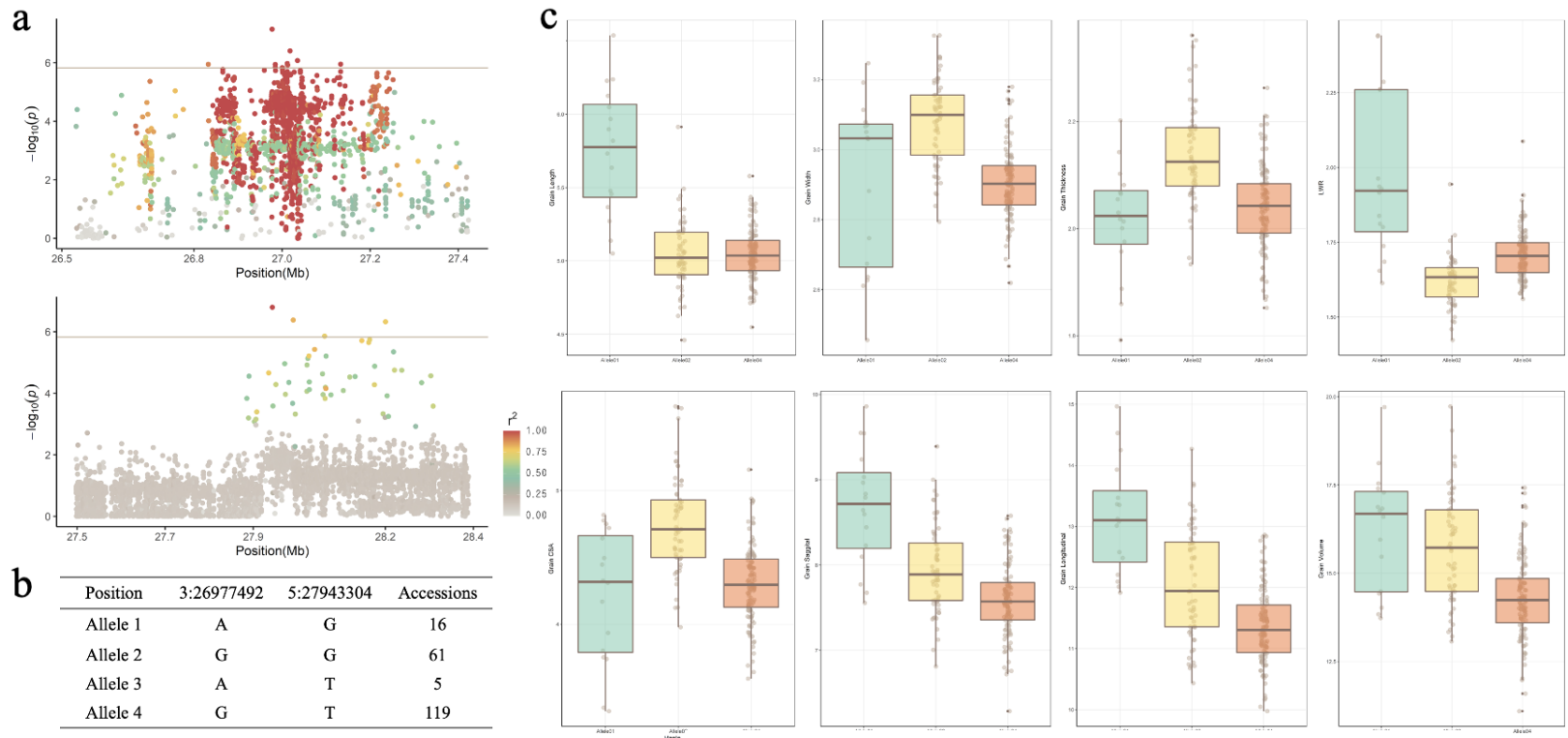
Figure 8. Allele frequency of lead SNPs related to grain. (a) Manhattan plot of the QTL region for SNPs at 3:26977492(top) and 5:27943304(bottom). The r^2 value between lead SNP and other SNPs are indicated in the color gradient. (b) the allele frequency (c) The phenotype difference between the allele combinations constructed by two lead SNPs

a

b

Position	2:19175513	5:28063233	9:375565	Accessions
Allele 1	G	C	C	15
Allele 2	T	C	C	22
Allele 3	G	T	C	7
Allele 4	T	T	C	39
Allele 5	G	C	G	16
Allele 6	T	C	G	3
Allele 7	G	T	G	11
Allele 8	T	T	G	36

c

Figure 9. Allele frequency of lead SNPs related to bran layer. (a) Manhattan plot of the QTL region for SNPs at 2:19175513(left), 5:28063233(center), and 9:375565(left). The r^2 value between lead SNP and other SNPs are indicated in the color gradient. (b) the allele frequency for 8 alleles. (c) The phenotype difference between the allele combinations constructed by 3 lead SNPs

b

Position	1:9223183	1:39989728	10:10946163	11:5330959	11:22576941	Accessions
Allele 1	T	A	A	C	C	3
Allele 2	T	G	A	C	C	1
Allele 3	G	A	G	C	C	11
Allele 4	T	A	G	C	C	138
Allele 5	T	G	G	C	C	2
Allele 6	T	A	G	T	C	4
Allele 7	T	A	A	C	T	1
Allele 8	G	A	G	C	T	1
Allele 9	T	A	G	C	T	6
Allele 10	G	G	A	T	T	2

Figure 10. Allele frequency of lead SNPs related to embryo. (a)
Manhattan plot of the QTL region for SNPs at 1:9223183, 1:9989728,
10:10946163, 11:22576941, 11:5330959(from left to right). The r^2
value between lead SNP and other SNPs are indicated in the color gradient.
(b) the allele frequency for 10 alleles (c) The phenotype difference between
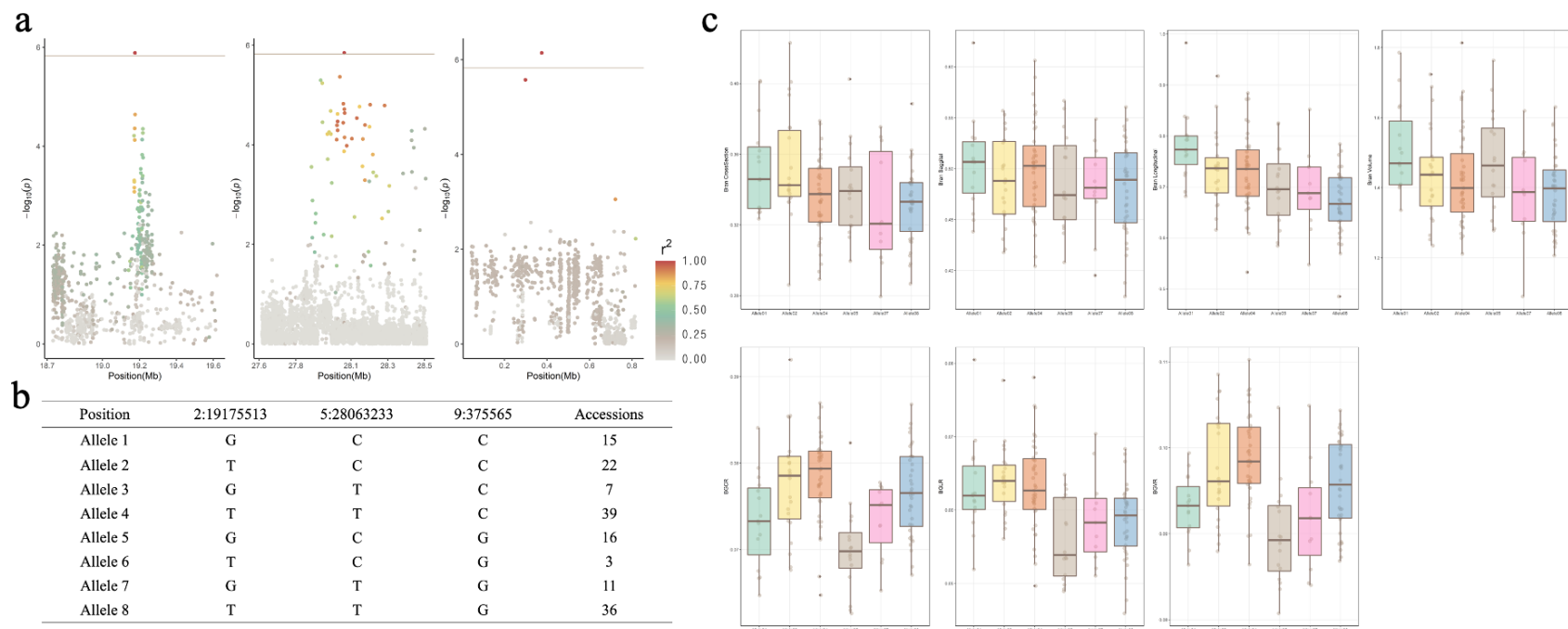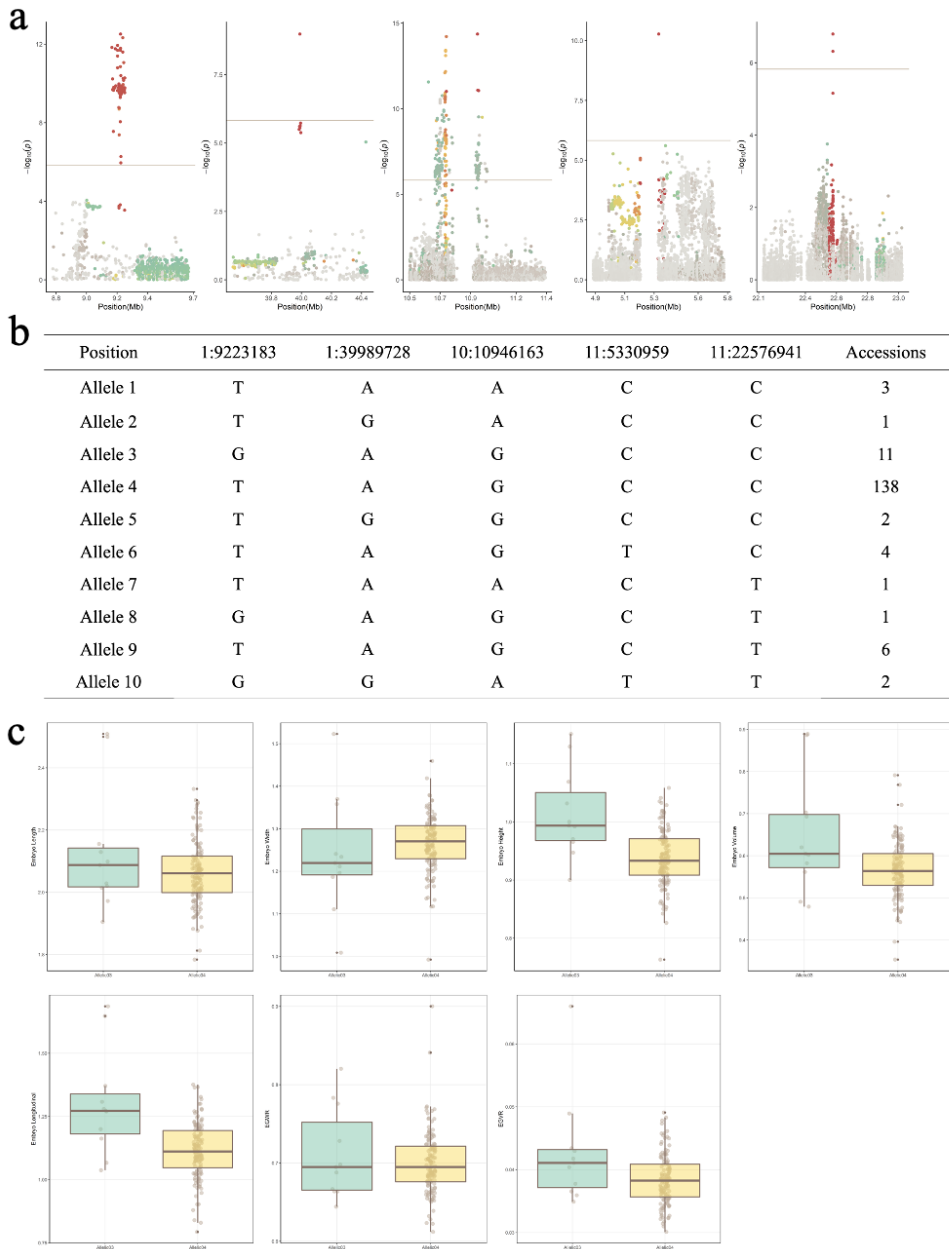the allele combinations constructed by lead SNPs.

3 7

Discussion

Brown rice, the whole grain with only the husk removed, is composed of three parts: the endosperm, the embryo, and the bran layer. The grain morphology is a crucial aspect to understand in rice breeding, as it is closely related to both yield and grain quality. Despite the importance of grain internal structure for plant breeding and crop improvement, research in this area has been limited as compared to the extensive studies that have been conducted on grain shape and size. This is predominantly due to the grain's delicate structure, compact size, and opaque characteristics.

To overcome this limitation, our study utilized a non-destructive method known as micro-CT imaging. This imaging technique was combined with genotype data to conduct a GWAS to identify genetic regions that control grain internal morphology. The results from this study offer valuable information on the genetics of grain internal structure and can be used in rice breeding and improvement programs.

The phenotype data achieved by micro-CT showed a positive correlation for most of the traits. This is probably because the inner structure becomes larger as the grain size increases. This suggests a strong relationship between the size of the grain and the size of the

internal structure, which may be an important factor to consider in rice breeding and crop improvement efforts. Few traits such as grain width, thickness showed negative correlation with the grain length which is consistent with previews research (Rasheed, Fiaz et al. 2022). Also, the correlation between the grain and endosperm showed a strong correlation with $r^2 = 1$. This can be understood as we look at the phenotype data. The endosperm takes up about 86.6% of the entire volume meaning that grain size is mainly controlled by the size of the endosperm. These results suggest that manipulation of endosperm size through breeding or agronomic practices may represent a viable strategy for improving grain size and ultimately, yield in rice.

Principal Component Analysis (PCA) showed four distinct clusters among the investigated accessions. These clusters were identified as representing indica, temperate japonica, tropical japonica, and aromatic subpopulations. The majority of rice accessions were found to belong to the temperate japonica subpopulation, which is in accordance with their morphological characteristics.

In the study, a GWAS was performed by combining the phenotyping data obtained from the non-destructive 3D imaging technique (micro-CT) and the genotyping data. The results showed that some genomic regions had an excessive number of highly correlated SNPs associated with them. To address this problem, tag

SNPs were selected to minimize the number of SNPs that needed to be genotyped. In total, 128 SNPs were detected, with the majority being found for traits related to embryo width and height, and grain length.

The gene *OsYUC9* (Os01g0273800) was detected within the QTL region near the lead SNP at 1:9237559 (p=4.21E-13) associated with embryo width. This gene is a putative biosynthesis gene for indole acetic acid (IAA) and has shown a strong correlation with the increase of IAA content in developing grains. Additionally, the expression of YUCCA genes (*OsYUC9* and *OsYUC11*) was found to be restricted to localized regions of the plant and highly expressed in developing grains, implying a primary role in endosperm development and potential involvement in the starch deposition in rice grains. These findings support the idea that the YUCCA genes play a crucial role in the regulation of IAA content (Abu-Zaitoon, Bennett et al. 2012).

Auxin is an important group of phytohormones involved in plant organ development through its effects on gene expression and cell expansion. Higher levels of auxin, in the form of IAA, are implicated in tropisms such as movement towards light and gravity signals, while decreased auxin levels can down-regulate the expression of cell cycle genes and result in enlarged organs

(Hentrich, Sánchez-Parra et al. 2013). Therefore, it can be hypothesized that the high expression of *OsYUC9* during the early stage of grain development may have an impact on the size of the embryo. Further studies are necessary to confirm this relationship and to fully understand the role of *OsYUC9* in the regulation of organ growth in rice.

The gene *OsPUP7* was also identified as a potential candidate in near the lead SNP located at 5:27943303, which was associated with GW, GV, GCSA, GT, GLWR, BCSA, and EnV. According to previous research, the *OsPUP7* is a PUP-type cytokinin transporter gene that plays a crucial role in the transportation of cytokinin. The mutant contained higher levels of cytokinin in the spikelet compared to the wild type, leading to multiple phenotypic changes such as increased plant height, larger seeds, and delayed flowering (Qi and Xiong 2013). According to the results of the study, it was observed that haplotype 2 with the GT allele showed increased grain volume and length, while width and thickness were comparatively narrower. The correlation analysis indicated a slight negative correlation between width and length. These observations agree with previous studies on the topic. This information can be taken into consideration during breeding efforts to enhance grain quality and size.

The allelic frequency of reoccurring single nucleotide polymorphisms (SNPs) was analyzed in three distinct morphological parts of the grain: the grain, bran layer, and embryo. For the grain morphological traits, two lead SNPs (5:27943304 and 3:26977492) were selected, and the allelic frequency of the four allelic combinations was investigated. The results showed significant differences between the alleles, with the A allele at the 3:26977492 site having the most pronounced effect on the grain morphology. This allele was associated with longer grain volume and larger grain size while also resulting in a narrower grain width.

For the bran layer, three SNPs were analyzed, and results showed statistical significance for 5 out of the 8 allelic combinations. Although no significant differences in single nucleotide polymorphisms (SNPs) were detected in the directly measured morphological traits of the bran layer, except for BCSA, which is a shared SNP with the Grain size, the comparison-based traits revealed significant differences. This may be due to the bran's thinness, which primarily affects grain length rather than its thickness. The comparison-based traits, such as bran-to-grain volume ratio (BGVR), bran-to-grain cross-sectional ratio (BGCSA), and bran-to-grain longitudinal area ratio (BGLR), showed an increase in the presence of alleles T, T, and C at SNP sites

2:19175513, 5:28063233, and 9:375565. This suggests that with these alleles, there is a higher proportion of bran within the grain than the grain size. These findings highlight the potential nutritional benefits of such traits, as well as the challenges that may arise in the milling process. Thus, these alleles should be considered when breeding rice with improved quality.

An analysis of five reoccurring SNP sites in the embryo was performed, and two allelic combinations were identified with sufficient sample size for statistical analysis. The results indicated that the presence of the G allele at the 1:9223183 site had a significant effect on the height and size of the embryo, as evidenced by statistically significant differences (p-value < 0.01) in traits such as embryo volume (EV), embryo height (EH), and embryo length (ELA). These findings suggest that the height of the embryo is a key factor in determining the volume of the embryo and that the identified haplotype could be utilized to breed rice with larger embryos for improved nutritional quality.

Supplementary

Table S1. SNPs associated with grain morphology. The information on candidate genes were collected from Gramene.

	Trait	Lead SNP	p-value	Candidate Gene
	GL	chr01:35189500	9.89E-07	Os01g0823500
	GT	chr01:40534600	1.14E-06	Os01g0921200, Os01g0921800, Os01g0927600
	GLWR	chr02:26050020	1.44E-06	Os02g0643200, Os02g0644000, Os02g0644000, Os02g0649300
	GL	chr02:31682969	1.16E-06	Os02g0747900, Os02g0749300, Os02g0755000, Os02g0756100, Os02g0759700, Os02g0761000, Os02g0762400, Os02g0762600
Grain	GLWR	chr03:13191097	7.44E-08	Os03g0346200, Os03g0352500
	GLWR	chr03:15294435	3.22E-08	Os03g0385100, Os03g0385400
	GLA, GL	chr03:16148641	1.15E-07	Os03g0402800
	GLWR	chr03:17688006	3.79E-09	Os03g0417700, Os03g0418600, Os03g0418700, Os03g0430500
	GL	chr03:19735124	6.32E-10	
	GLWR	chr03:23673281	8.01E-07	Os03g0626100

Trait	Position	P-value	Genes
GL, GLA	chr03:24530212	5.98E−09	Os03g0637800, Os03g0640100, Os03g0640400, Os03g0640800, Os03g0643300
GLA, GSA, GV, GL	chr03:26977492	1.6E−08	Os03g0672900, Os03g0685000, Os03g0687000
GW	chr03:27628447	9.27E−07	Os03g0685000, Os03g0687000, Os03g0691800, Os03g0693600, Os03g0699000
GLA, GSA	chr03:28719686	6.33E−08	Os03g0706900, Os03g0709100, Os03g0710800, Os03g0712800, Os03g0722400, Os03g0723000
GLA, GSA, GV, GLWR	chr04:20166831	1E−06	Os04g0409600, Os04g0413500
GLWR	chr04:26224222	6.95E−08	Os04g0517100, Os04g0518800, Os04g0526600
GLWR	chr04:28370718	6.31E−07	Os04g0559800, Os04g0566500, Os04g0566500, Os04g0568400
GT	chr05:1602676	4.04E−07	
GW, GV, GCSA, GT, GLWR	chr05:27943304	1.6E−07	Os05g0554000, Os05g0556800, Os05g0563500, Os05g0569500
GW	chr06:23461393	8.84E−07	Os06g0597500, Os06g0601500, Os06g0602700
GLWR	chr06:30882299	3.1E−07	Os06g0720900
CSA	chr07:16023071	3.76E−07	Os07g0463800

	GLWR	chr07:20639384	4.06E−07	
	GL, GLW R	chr07:28593722	2.72E−07	Os07g0668600, Os07g0669200, Os07g0669500, Os07g0669800
	GL, GLWR	chr08:21781094	4.14E−07	Os08g0440300, Os08g0441200, Os08g0441300, Os08g0452500
	GLWR	chr08:27032041	8.65E−07	Os08g0536800, Os08g0537800, Os08g0546100, Os08g0547300
	GLWR	chr09:4629202	3.02E−08	
	GLWR	chr11:17596430	4.28E−07	Os11g0490800, Os11g0497000, Os11g0497350
	GL	chr11:370937	2.96E−07	Os11g0117600
Endosperm	EnV	chr03:26977492	6.29E−07	Os03g0672900, Os03g0685000, Os03g0687000
	EnV	chr05:27943304	3.06E−07	Os05g0554000, Os05g0556800, Os05g0563500, Os05g0569500
Embryo	EH	chr01:12560209	1.23E−17	Os01g0322700 , Os01g0328500
	EH	chr01:13899196	5.23E−08	Os01g0343300, Os01g0346900
	EW	chr01:21767014	1.25E−06	Os01g0567500
	EW	chr01:37480054	1.09E−06	Os01g0864000, Os01g0867800
	EW, EH	chr01:39989728	4.24E−11	Os01g0911700, Os01g0914100, Os01g0914300, Os01g0921200, Os01g0921800

EW	chr01:6647283	4.51E−07	Os01g0218032, Os01g0218800
EW, EH	chr01:9237559	4.21E−13	Os01g0273800
EW	chr02:10800354	2.82E−07	Os02g0290900, Os02g0293400
EW	chr02:13576647	2.04E−08	
EW	chr02:14174996	3.03E−09	
EW	chr02:16087889	1.45E−08	Os02g0477700
EW	chr02:16880148	1.02E−06	Os02g0490000, Os02g0491300
EW	chr02:18203572	9.1E−07	Os02g0512400, Os02g0516400
EW	chr02:20795899	1.96E−07	Os02g0550000, Os02g0550600, Os02g0554000, Os02g0554300
EW	chr02:24787072	5.4E−09	Os02g0621300, Os02g0622100
EW	chr02:33637595	1.1E−06	Os02g0783700, Os02g0787300
EH	chr02:7697263	7.98E−11	Os02g0234200, Os02g0244100
EW , EH, EGVR	chr03:16619438	7.38E−07	Os03g0402800, Os03g0407400, Os03g0411500, Os03g0414400, Os03g0414900, Os03g0417700
EW, EH	chr03:30997208	2.47E−07	Os03g0747400, Os03g0747500, Os03g0753100, Os03g0757500, Os03g0760800
EL, EV	chr04:26351178	3.83E−07	Os04g0518800, Os04g0526600, Os04g0535600

EGVR	chr04:28862203	8.41E−08	Os04g0566500, Os04g0566500, Os04g0568400, Os04g0573100, Os04g0573900, Os04g0574200
EW, EH	chr04:6786926	5.04E−07	
EW, EH	chr05:10774446	3.4E−12	
EW	chr05:11823263	1.05E−06	
EW	chr05:13315994	2.6E−11	
EW, EH	chr05:17114036	5.48E−07	
EV	chr05:20232747	1.05E−06	Os05g0411200, Os05g0420800
EW, EGWR	chr05:28483453	6.17E−09	Os05g0563500, Os05g0569500, Os05g0578900
EH	chr05:6633350	6.29E−07	Os05g0199800, Os05g0203800
EL, EV, ECSA	chr06:15116015	1.27E−06	
EGWR	chr06:17625938	7.16E−07	Os06g0498150, Os06g0498400
EW	chr06:29330753	7.69E−09	Os06g0699400, Os06g0701900
EL, EV	chr06:30112006	5.09E−07	Os06g0706400, Os06g0710800, Os06g0711600, Os06g0711900, Os06g0712700, Os06g0712800
EW	chr06:6209808	7.54E−07	Os06g0213700, Os06g0225800
EW, EH	chr07:16686455	1.1E−06	Os07g0463800, Os07g0471300

EW	chr07:1795360	3.25E−07	Os07g0129700
EW, EH	chr07:19019270	4.92E−08	Os07g0497100, Os07g0500300, Os07g0505200
EL, EV	chr07:808643	7.47E−07	Os07g0109400, Os07g0110400, Os07g0111400
EV	chr08:16347822	1.16E−06	Os07g0463800
EV	chr08:17967598	1.34E−06	Os08g0378900, Os08g0385000
EW, EH	chr10:10681829	4.29E−16	
EH	chr10:13445949	8.64E−07	Os10g0397400, Os10g0400200, Os10g0402200
EW	chr10:14028348	2.43E−07	Os10g0402200
EH	chr10:7353469	1.33E−13	
EW	chr10:8208500	6.64E−12	
EV	chr11:16437583	1.2E−07	
EW, EH	chr11:18747673	1.26E−11	Os11g0512600, Os11g0514400, Os11g0523800, Os11g0523800, Os11g0525200, Os11g0525700, Os11g0533500
EW, EH	chr11:21005931	8.59E−08	
EW, EH	chr11:22576941	6.13E−12	Os11g0591100, Os11g0595400
EW	chr11:25084725	1.81E−07	Os11g0637700
EL	chr11:29009277	3.83E−07	Os11g0700500

	EW, EH	chr11:4286608	4.39E-08	Os11g0184800
	EW, EH	chr11:5330959	6.36E-07	Os11g0199700, Os11g0199700, Os11g0199700, Os11g0201450, Os11g0206700
	EL	chr11:6400495	1.73E-07	Os11g0216000, Os11g0221300
	EH	chr12:2467058	3.41E-07	Os12g0143200, Os12g0145700, Os12g0147800, Os12g0151600
	EW	chr12:4467292	8.08E-07	Os12g0189500, Os12g0190000, Os12g0194900
Bran	BGVR	chr02:19175513	1.3E-06	Os02g0517531, Os02g0521300, Os02g0523800, Os02g0530300
	BCSA	chr05:28063233	1.43E-06	Os05g0556800, Os05g0563500, Os05g0569500
	BGLR	chr09:375565	7.19E-07	Os09g0104200

GL = grain length, GW = grain width, GT = grain thickness, GCSA = grain cross-sectional area, GSA = grain saggital area,

GLA = grain longitudinal area, GV = grain volume, GLWR = grain length-width ratio, EnV = endosperm volume,

EL = embryo length, EW = embryo width, EH = embryo height, EGWR = embryo to grain width ratio, EGVR = embryo to

grain volume ratio, BCSA = bran cross-sectional area, BSA = bran sagittal area, BLA = bran longitudinal area, BV = bran

volume, BGLAR = bran to grain longitudinal area ratio, BGVR = bran to grain volume ratio
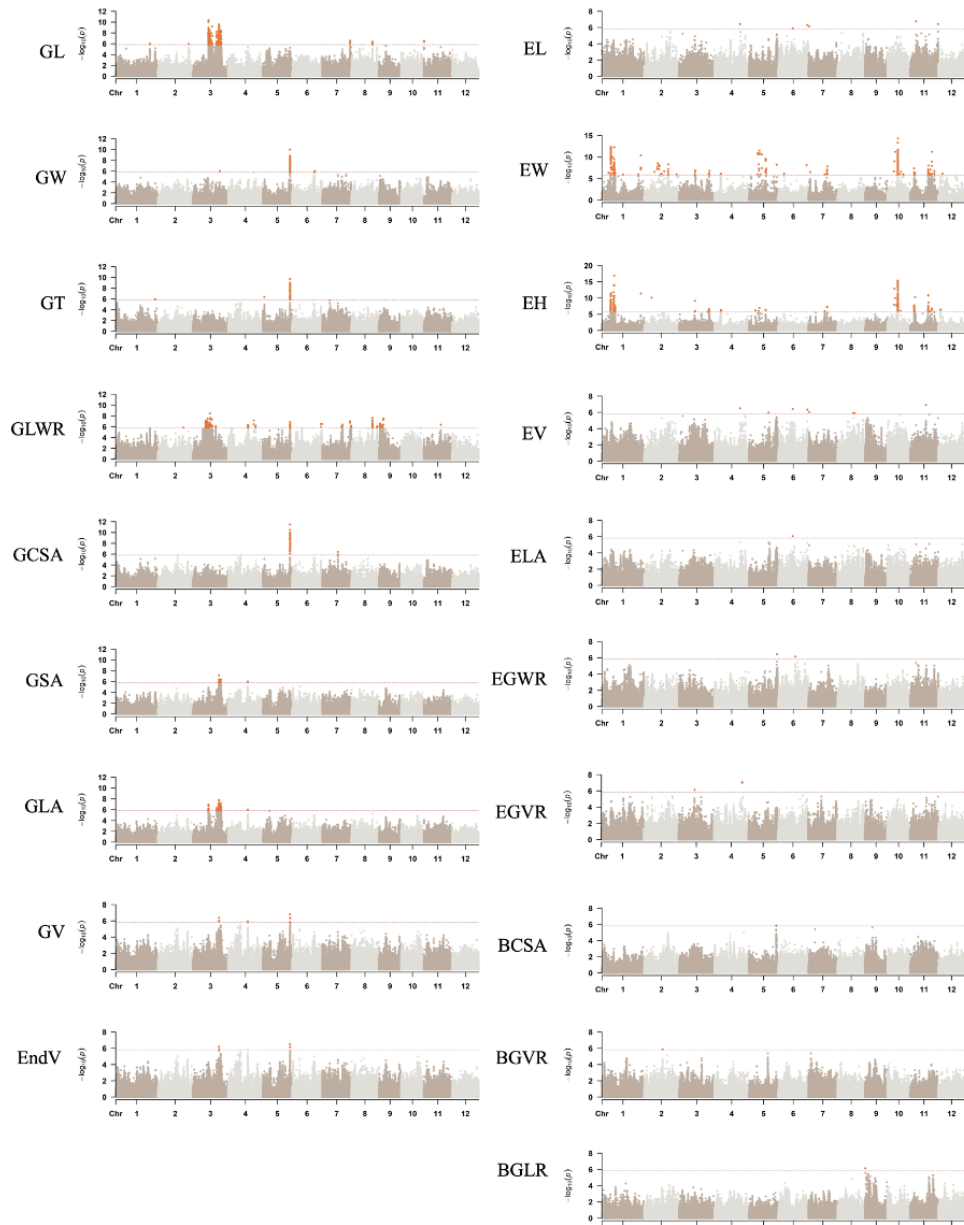
Figure S1. Manhattan plot of the grain morphology related traits.
Genome-wide association study was conducted using Factored
Spectrally Transformed Linear Mixed Models(FaST-LMM).

Bibliography

Abu-Zaitoon, Y. M., et al. (2012). "A large increase in IAA during development of rice grains correlates with the expression of tryptophan aminotransferase OsTAR1 and a grain-specific YUCCA." Physiol Plant **146**(4): 487-499.

Bodie, A. R., et al. (2019). "Current Trends of Rice Milling Byproducts for Agricultural Applications and Alternative Food Production Systems." Frontiers in sustainable food systems **3**.

Chang, T.-T. and E. A. Bardenas (1965). The morphology and varietal characteristics of the rice plant, Int. Rice Res. Inst.

Dhondt, S., et al. (2010). "Plant structure visualization by high-resolution X-ray computed tomography." Trends Plant Sci **15**(8): 419-422.

Diaz-Pinto, A., et al. (2022). "MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images."

Fan, C., 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." Theoretical and Applied Genetics **112**(6): 1164-1171.

Fischer, R. A. T. and G. O. Edmeades (2010). "Breeding and Cereal Yield Progress." Crop Science **50**(Supplement_1): S-85-S-98.

Gargiulo, L., et al. (2020). "Micro-CT imaging of tomato seeds: Predictive potential of 3D morphometry on germination." Biosystems Engineering **200**: 112-122.

Halperin, E., et al. (2005). "Tag SNP selection in genotype data for maximizing SNP prediction accuracy." Bioinformatics. **21**(Suppl 1): i195-i203.

Hentrich, M., et al. (2013). "YUCCA8andYUCCA9overexpression reveals a link between auxin signaling and lignification through the induction of ethylene biosynthesis." Plant Signaling & Behavior **8**(11): e26363.

Huang, R., et al. (2013). "Genetic bases of rice grain shape: so many genes, so little known." Trends in Plant Science **18**(4): 218-226.

Huie, J. M., et al. (2022). "SegmentGeometry: A Tool for Measuring Second Moment of Area in 3D Slicer." Integrative organismal biology. **4**(1).

Kikinis, R., et al. (2014). 3D Slicer: A Platform for Subject-Specific Image

Analysis, Visualization, and Clinical Support, Springer New York: 277-289.

Lee, G., et al. (2019). "Identification and Characterization of LARGE EMBRYO, a New Gene Controlling Embryo Size in Rice (Oryza sativa L.)." Rice 12(1).

Li, N., et al. (2018). "Control of grain size in rice." Plant Reprod 31(3): 237-251.

Li, P., et al. (2022). "Genes and Their Molecular Functions Determining Seed Structure, Components, and Quality of Rice." Rice 15(1).

Lippert, C., et al. (2011). "FaST linear mixed models for genome-wide association studies." Nature Methods 8(10): 833-835.

Liu, J., et al. (2018). "Mutations in the DNA demethylase OsROS1 result in a thickened aleurone and improved nutritional value in rice grains." Proceedings of the National Academy of Sciences 115(44): 11327-11332.

Mao, H., et al. (2010). "Linking differential domain functions of the GS3 protein to natural variation of grain size in rice." Proceedings of the National Academy of Sciences 107(45): 19579-19584.

Mather, K. A., et al. (2007). "Extent of Linkage Disequilibrium in Rice (Oryza sativa L.)." Genetics 177(4): 2223-2232.

Nagasawa, N., et al. (2013). "GIANT EMBRYO encodes CYP78A13, required for proper size balance between embryo and endosperm in rice." The Plant Journal 75(4): 592-605.

Pu, C.-X., et al. (2012). "Crinkly4 receptor-like kinase is required to maintain the interlocking of the palea and lemma, and fertility in rice, by promoting epidermal cell differentiation." The Plant Journal 70(6): 940-953.

Qi, Z. and L. Xiong (2013). "Characterization of a Purine Permease Family Gene OsPUP7 Involved in Growth and Development Control in Rice." Journal of Integrative Plant Biology 55(11): 1119-1135.

Rasheed, H., et al. (2022). "Characterization of functional genes GS3 and GW2 and their effect on the grain size of various landraces of rice (Oryza sativa)." Molecular Biology Reports 49(6): 5397-5403.

Su, Y. and L.-T. Xiao (2020). "3D Visualization and Volume-Based Quantification of Rice Chalkiness In Vivo by Using High Resolution Micro-CT." Rice 13(1).

Wang, G., et al. (2019). Automatic Segmentation of Vestibular Schwannoma from T2-Weighted MRI by Deep Spatial Attention with Hardness-Weighted Loss. Medical Image Computing and Computer Assisted Intervention -

MICCAI 2019.

Wei, T. and V. Simko (2017). "R package "corrplot": Visualization of a Correlation Matrix, R package version 0.84." Retrieved January **20**: 2021.

Wu, K., et al. (2018). "The rational design of multiple molecular module-based assemblies for simultaneously improving rice yield and grain quality." J Genet Genomics **45**(6): 337-341.

Wu, X., et al. (2016). "Rice caryopsis development II: Dynamic changes in the endosperm." Journal of Integrative Plant Biology **58**(9): 786-798.

초록

쌀의 품질을 결정하는 요소에는 식미, 도정, 외관, 영양 등이 있다. 현미상태의 종자는 배유, 배아, 속겨로 구성되며 각 구조는 서로 다른 영양성분을 가지고 있다. 종자의 대부분을 구성하는 배유는 전분과 단백질로 구성되어 있으며 배아와 속겨는 GABA, 비타민, 지질 등과 같은 기능성 물질이나 산업적 가치가 풍부한 물질들을 함유하는 것으로 보고되어 있다. 종자의 구조가 쌀 품질의 결정요소들에 직간접적으로 영향을 미치기 때문에 종자의 구조에 관련하는 양적형질유전자좌 (QTL)의 발굴은 더 나은 품질과 수량을 가지는 벼의 육성에 매우 중요한 의의를 갖는다. 따라서 본 연구는 현미의 내부 구조에 관여하는 QTL을 탐지하기 위해 215개의 벼 유전자원을 대상으로 micro-CT를 이용하여 종자 내부 구조 관련 형질들을 측정하였고, 전유전체연관분석 (Genome-wide association study, GWAS)을 실시하여 성공적으로 연관 SNP들을 탐지하였다.

그 결과 종자 입형에 관여하는 너비, 두께, 길이, 장폭비, 부피과 같은 형질들에서 총 128개의 연관 SNP가 탐지되었다. 탐지된 SNP들의 후보영역 내에서 배아 길이와 관련해 후보유전자로 생각되는 옥신관련 유전자 *OsYUC9*가 탐지되었으며 종자 크기에 관련하는 것으로 기존에 보고된 *OsPUP7*이 종자와 배유에 관련하는 QTL 영역에 존재했다. 이 유전자들이 해당 유전자원들의 표현형에 미치는 영향을 알아보기 위해 haplotype 분석을 실시한 결과 이들의 서열 변이로 인한 형질의 차이가 일부에서 나타났다.

또한 GWAS에서 탐지된 SNP들의 allelic combination에 따라 표현형에 유의미한 차이가 나타났다. 같은 종자 부위 내에서 여러번 중복되어 탐지된 연관 SNP들의 조합으로 분석을 실시한 결과 종자의

경우 총 4개의 조합으로 나타났으며 이중 통계분석이 가능한 allele 1이 종자의 부피가 가장 큰 조합으로 나타났다. 속겨의 경우 횡단면, 종자대비 종단면 면적비, 종자대비 부피비 에서 하나씩의 연관 SNP가 탐지되었으며 이들의 조합을 본 결과 allele 4가 가장 종자크기 대비 큰 부피의 속겨를 가지는 것으로 드러났다. 이는 영양 가치가 높은 종자의 육종에 활용될 수 있으나 백미를 목적으로 하는 경우 도정에 어려움을 줄 수 있기 때문에 이 조합을 활용하여 목적에 맞는 벼의 생산이 가능할 것으로 전망된다. 배아에서 역시 여러 형질에 대해 반복해서 나타나는 5개 연관 SNP의 allele combination을 분석한 결과, 총 10개의 조합으로 구별되었다. 이 중 통계 가능한 2개의 조합을 살펴 본 결과 allele 3의 조합이 배아의 높이와 부피가 가장 큰 것으로 나타났다. 이 조합을 가지는 벼를 육종 재료로 활용한다면 더 영양적 가치가 높은 쌀을 생산할 수 있을 것으로 판단된다. 본연구를 통해 밝혀진 결과들은 벼 종자 구조의 표현형에 대한 종합적 이해를 제공하며 영양가 높고 외관 품질과 수량이 우수한 벼의 육종에 기초자료로써 활용될 수 있다.