COMBINING IMAGE PROCESSING AND GENOME-WIDE ASSOCIATION STUDY FOR REVEALING QUANTITATIVE TRAIT LOCI RELATED TO GRAIN SIZE IN RICE

Tran Giang Son, Nguyen Quynh Hoa, Chu Bao Minh, Le Nhu Chu Hiep, To Thi Mai Huong*
University of Science and Technology of Hanoi - Vietnam Academy of Science and Technology

ARTICLE INFO

Received: 03/12/2024

Revised: 27/3/2025

Published: 28/3/2025

KEYWORDS

Rice grain traits
Candidate genes
Quantitative trait locus
Genome-wide Association Study
Image processing

ABSTRACT

Genome-wide Association Studies (GWAS) are known as the popular method to discover the quantitative trait loci (QTLs) related to the size traits of rice grains. However, most of the current GWAS analyses use manual measurement to create phenotypic data of rice grain traits which are usually labor-intensive, time-consuming, and prone to human error. In this study, a combination of low-cost image processing and GWAS analysis was investigated to reveal the candidate genes related to the size traits of rice grains. Image processing methods allow the automatic extraction of rice grain size (width and length) from the captured color digital images. Extracted size traits are then used as phenotypic data for further GWAS analyses. As results, both manually measured and estimated dataset provide reliable identification of significant QTLs such as *GS3*, *GW6* and *GL7*. Interestingly, the estimated dataset showed greater sensitivity, identifying additional significant QTLs such as *GS5* and *GW5*, which were not detected in the manually measured dataset. The results open a promising direction for applying advanced low-cost image processing techniques for automatic extraction of other quantitative traits of rice.

KẾT HỢP XỬ LÝ ẢNH VÀ NGHIÊN CỨU LIÊN KẾT TRÊN TOÀN HỆ GEN ĐỂ PHÁT HIỆN CÁC LÔ-CUT TÍNH TRẠNG SỐ LƯỢNG LIÊN QUAN ĐẾN KÍCH THƯỚC CỦA HẠT THÓC

Trần Giang Sơn, Nguyễn Quỳnh Hoa, Chu Bảo Minh, Lê Như Chu Hiệp, Tô Thị Mai Hương*
Trường Đại học Khoa học và Công nghệ Hà Nội - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

THÔNG TIN BÀI BÁO

Ngày nhận bài: 03/12/2024

Ngày hoàn thiện: 27/3/2025

Ngày đăng: 28/3/2025

TỪ KHÓA

Kiểu hình hạt thóc
Gen ứng viên
Lô-cut tính trạng số lượng
Nghiên cứu liên kết toàn hệ gen
Xử lý ảnh

TÓM TẮT

Nghiên cứu liên kết toàn hệ gen (GWAS) là phương pháp tiềm năng để tìm ra các lô-cut tính trạng số lượng (QTL) liên kết tới tính trạng kích thước hạt thóc. Tuy nhiên, phần lớn các phân tích kiểu hình hiện nay sử dụng các phép đo thủ công, đòi hỏi nhiều công sức, thời gian và dễ xảy ra lỗi. Nghiên cứu kết hợp giữa kỹ thuật xử lý ảnh chi phí thấp và phân tích GWAS để xác định các gen ứng viên liên quan đến kiểu hình kích thước của hạt thóc. Phương pháp xử lý ảnh cho phép tự động trích xuất kích thước hạt thóc từ ảnh kỹ thuật số màu. Kích thước sau đó được sử dụng làm dữ liệu đầu vào cho phân tích GWAS. Kết quả cho thấy sử dụng cả hai bộ dữ liệu được đo thủ công và được tính toán đều có khả năng xác định các QTL quan trọng như *GS3*, *GW6* và *GL7*. Đáng chú ý, bộ dữ liệu tính toán cho thấy độ nhạy cao hơn, phát hiện thêm được các QTL quan trọng như *GS5* và *GW5* mà bộ dữ liệu đo thủ công không phát hiện được. Kết quả mở ra hướng đi triển vọng trong việc ứng dụng các kỹ thuật xử lý ảnh tiên tiến có chi phí thấp để tự động trích xuất các kiểu hình định lượng khác của lúa.

DOI: https://doi.org/10.34238/tnu-jst.11646

* Corresponding author. *Email: to-thi-mai.huong@usth.edu.vn*

1. Introduction

Rice (*Oryza sativa* L.) is one of the most staple foods in the world population. The demand to increase the quality of rice holds the utmost importance [1]. To evaluate the rice yield and quality, grain features such as length, width, and thickness can be used. Usually, the big rice grains have a strong positive correlation with yield and contain more nutrition than the small ones. To obtain phenotypic data of the rice productivity-related traits (panicles, spikelet and grains), manual techniques are traditionally used to measure rice grain features such as width and length [2]. The manual measurement tasks are usually time-consuming, labor-intensive, and prone to human error [3]. Besides, the threshing process may affect measured accuracy due to its destructive nature [4].

In recent years, image processing and computer vision techniques have emerged as cost-effective methods for phenotypic trait extraction of rice grains [5]-[7]. For instance, Su and Ping [8] used X-ray computed tomography (CT) to characterize the panicle traits (spikelet number and the seed setting rate) of rice. A 3D image analysis pipeline using X-ray CT was also utilized for rice grain trait extraction [9]. In 2021, Yu and his team integrated X-ray and Red Green Blue (RGB) scanning for measuring spikelet and size traits of rice panicle [10]. Similarly, visible light scanning imaging and deep learning techniques were applied for high-throughput extraction of size traits in rice panicles by Lu and his colleagues in 2023 [11].

Nowadays, the length and width of rice grains can be improved by modifying the related genes which regulate their main features. Before the modification process starts, this genetic information, such as Single nucleotide polymorphism (SNP) variants, Quantitative trait loci (QTL), and associated genes, must be identified. Genome-Wide Association Studies (GWAS) is known as a powerful method for identifying genetic variants associated with phenotypic traits in rice and have successful applied with the Vietnamese rice collection [12]-[15]. However, most of these methods use manual measurements to create phenotypic data of rice grain traits for GWAS analysis, which are usually low-throughput, labor-intensive, and human-prone error.

Therefore, in the present study, a new method that combines low-cost image processing techniques and GWAS analysis for identifying related candidate genes to size trait of rice grains is proposed. The results from the study may be helpful for potential rice breeding programs and functional gene identification of rice varieties.

2. Materials and Methods

2.1. Materials

In this study, 88 rice varieties belonging to Vietnamese rice landrace provided by the Plant Resource Center (PRC) in Hanoi, Vietnam were employed in the image analysis. These varieties are scattered in different provinces of Vietnam to ensure the diversity of genetic background. Information about each rice landrace including original name, accession number on Genebank and origin is provided in Supplementary material S1.

The rice seed images were captured by the International joint laboratory LMI-RICE (USTH/AGI/IRD/UM) right after receiving the seeds. The images were taken using Canon digital cameras, ensuring uniformity in terms of distance from the rice grains, lens type, lighting conditions, and aperture settings. Each image has a label with the variety name of rice grains and a 13.5 cm long ruler. An excel file containing the length and width of 88 rice varieties measured manually by the biologists (10 grains per measure) was also provided by the PRC for comparative purposes.

The genome database of 88 Vietnamese rice varieties was also genotyped by sequencing. Approximately 21623 markers have been yielded by sequencing using DArTseq^{TM} and Illumina NGS technology [16].

2.2. Size trait extraction using image processing

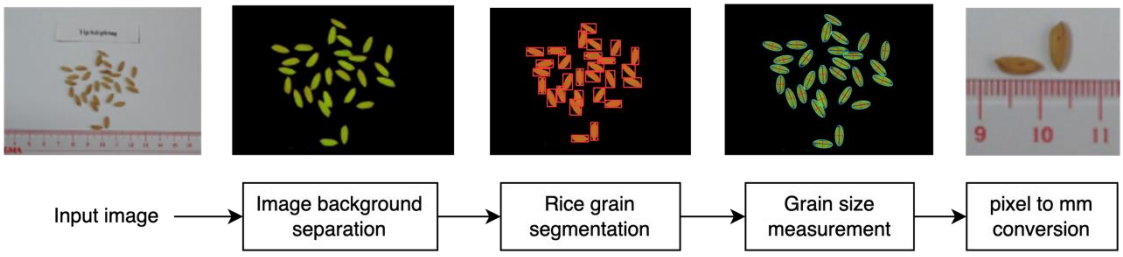

Figure 1. *Diagram of rice grain size extraction* [17].

In this study, size traits of rice grains are extracted using image processing pipeline developed by Chu and his colleagues in 2024 [17]. First, rice grain region is separated from image background using color space conversion (Figure 1). Then, rice grains are segmented based on YOLOv8 deep learning model. The detected rice grains are measured using fitEllipse() function in OpenCV library. Finally, pixel-based length and width are converted into physical units in mm for later use as phenotypic data.

From 88 images in the database, 176 estimated values of size trait (88 for width and 88 for length) of rice grains are extracted and stored in an excel file for later use in GWAS experiments.

2.3. Genome-wide association studies (GWAS)

GWAS analysis leverages large-scale single nucleotide polymorphism (SNP) datasets to examine the relationships between genetic polymorphisms and phenotypic traits using statistical frameworks such as linear mixed models. These models considered confounding factors like population structure and genetic relatedness by incorporating principal components from principal component analysis and kinship matrices. A suggestive threshold of -log(P-value) $\geq$ 3.5 was applied to reduce false discovery rates and identify lead SNPs which are most likely to influence traits of interest.

In this study, GWAS was performed using TASSEL 5.0 software to detect significant biomarkers related to grain size variation [18]. The analysis was conducted separately for manually measured and estimated phenotypic traits, incorporating kinship and principal components as covariates. Manhattan plots and QQ plots of manually measured and estimated data were compared to assessing which dataset produced better results. In addition, linkage disequilibrium (LD) heatmaps were also generated for the identification of significant correlation regions. LD heatmaps were produced using the package LDheatmap in R.4.4.1 [19].

2.4. Candidate gene identification

The gene database from the rice genome annotation project was used to screen the candidate genes [20]. The gene database is based on the genome of Nipponbare rice variety. The screening position was around 25 kb before and after the significant SNPs.

3. Results and Discussion

3.1. Comparison between the manually measured and estimated size traits of rice grains

Table 1 shows the comparison between the estimated size traits using image processing and the size traits manually measured by the biologists. Frequency distribution of each measurement is also provided in Figure 2. The mean difference between the two measuring methods did not exceed the control (0.005 mm and 0.055 mm difference in length and width, respectively).

Table 1. *Comparison between size traits of rice grains manually measured and estimated by the image analysis*

Values	Grain length (mm)		Grain width (mm)	
	Measured	Estimated	Measured	Estimated
Mean	8.728	8.732	2.918	2.862
Standard deviation	0.757	0.833	0.452	0.423
Min	7.370	6.860	2.090	2.090
Max	10.560	10.670	4.000	3.960
Mean difference	0.005		0.055	

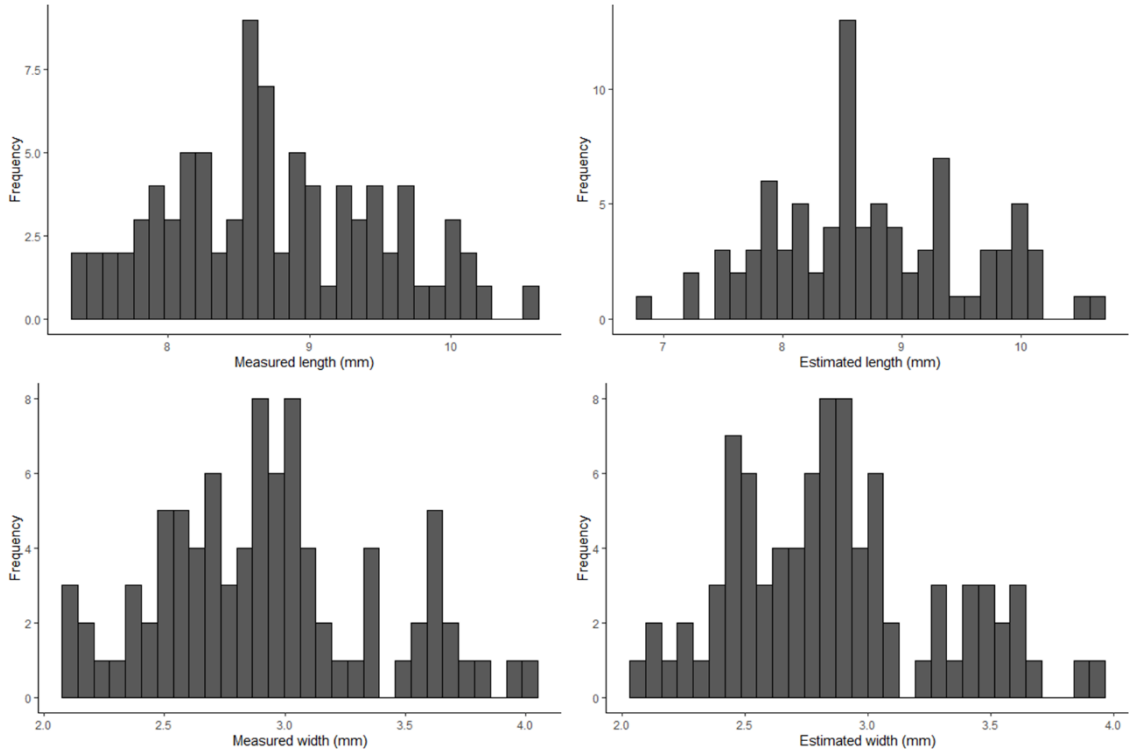

Figure 2. *Frequency distribution of each measurement of size trait*

Also, the association patterns between manually measured and estimated datasets demonstrated strong correlations, with Pearson's correlation coefficients of r = 0.91 for grain length and r = 0.86 for grain width (Figure 3).

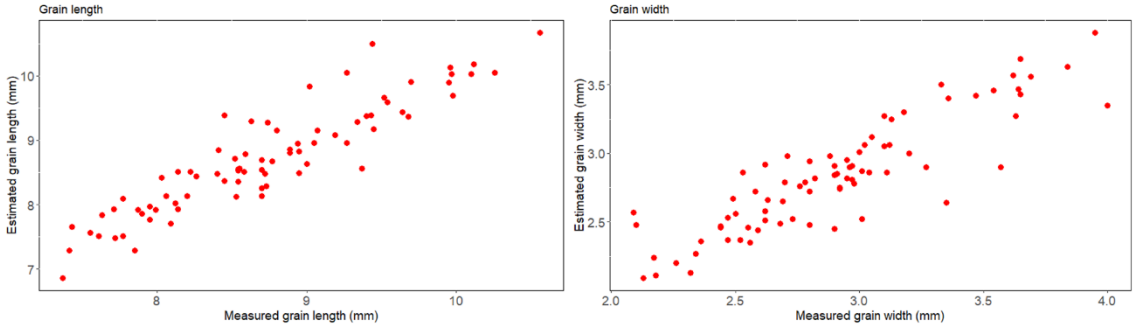

Figure 3. *Pairwise correlation between manually measured and estimated grain length (left) and grain width (right)*

3.2. Comparative GWAS between the manually measured and estimated phenotypic traits

In this study, genome-wide association studies (GWAS) were performed using two datasets: one derived from manual measurements and the other from algorithmically estimated grain size data. GWAS results confirmed that the linear mixed model effectively captured the association patterns for grain length and width, as illustrated in the QQ plots (Figure 4). Comparisons of Manhattan plots and QQ plots further highlighted the enhanced performance of the algorithmically derived dataset, as it displayed sharper signal peaks and better model fit. This underscores the utility of automated phenotyping tools in GWAS, offering higher throughput and precision for trait association studies.

Specifically, six significant overlapping SNPs were identified between the manually measured and estimated GWAS results for grain length (Table 2). Next, Linkage Disequilibrium analysis analysis were conducted using identified markers using the package LDheatmap in R.4.4.1 [19]. Among those, four LD heatmaps equivalent to four regions of significant association between the identified SNPs were also detected. Figure 5 shows the pairwise linkage disequilibrium (LD) patterns surrounding the significant SNPs associated with key grain-size QTLs, namely **GS3**, **GS5.1**, **GS5.2**, and **GW6**. Each triangular heatmap represents the extent to which nearby SNPs within a ~100–200 kb window is correlated with one another (with **red** indicating higher LD and **yellow** to **white** indicating lower LD). The presence of strong LD blocks (darker red clusters) suggests that these SNPs tend to be inherited together, reflecting limited recombination in those genomic regions. For instance, the tight LD blocks near **GS3** and **GW6** highlight the likelihood of a single major haplotype carrying the causative variant(s), whereas **GS5** is split into two regions (GS5.1 and GS5.2), each showing its own LD pattern. Together, these heatmaps provide a clearer picture of how genetic variation is structured around the candidate loci and help pinpoint the specific haplotype blocks most likely to harbor the functional polymorphisms influencing rice grain size. Based on the LD, the list of candidated genes was screened. Interestingly, these results revealed a consistent overlap between the two datasets for well-known QTLs regulating grain size, such as *GS3* (grain length regulator on chromosome 3), *GW6* (grain width regulator on chromosome 6), and *GL7* (grain length and slenderness regulator on chromosome 7). These findings validate the robustness of the data in identifying key genetic loci for grain size across different measurement methods.

Moreover, the estimated dataset showed greater sensitivity, identifying additional significant QTLs such as *GS5* and *GW5* on chromosome 5, which were not detected in the manually measured dataset. Both loci are well-established regulators of grain size and are pivotal for improving rice yield. This suggests that the estimation algorithm provides higher resolution and reduces phenotypic variability or measurement error, potentially uncovering subtle genetic effects that manual measurements might overlook.

These findings demonstrate the reliability and advantages of algorithmic estimation for phenotypic measurements. As GWAS serves as a screening tool for identifying potential genetic biomarkers, increasing the sample size and improving image processing techniques could enhance the accuracy and reliability of the estimated dataset, thereby improving GWAS outcomes. The enhanced ability to detect significant QTLs reinforces the value of integrating advanced computational tools into genetic research, paving the way for improved marker-assisted selection and molecular breeding strategies targeting grain size in rice.

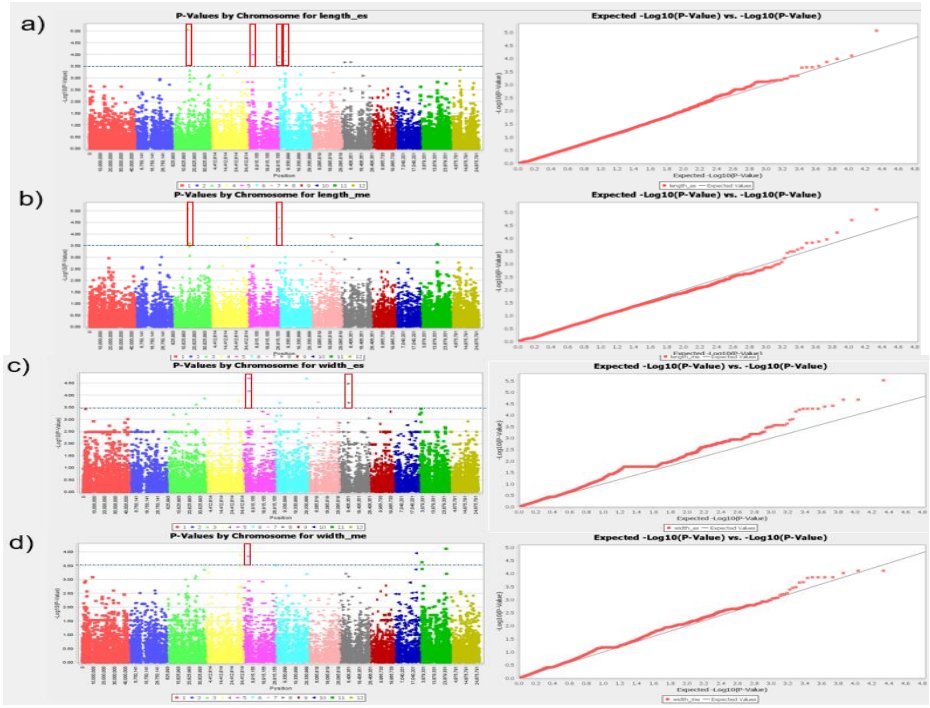

Figure 4. *GWAS results including Manhattan plots and corresponding QQ plots for the manually measured and estimated grain sizes. Dashed lines indicate similar peaks identified in GWAS results of both measurements. (a) Estimated grain length; (b) manually measured grain length; (c) Estimated grain width; (d) manually measured grain width*

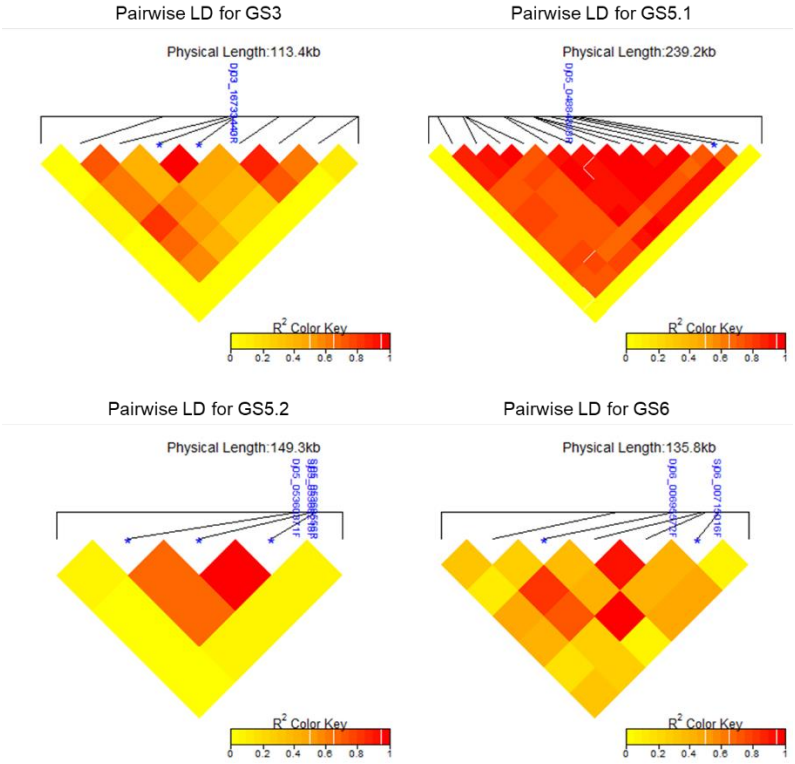

Figure 5. *Linkage disequilibrium heatmaps for regions of significant correlation between the identified single polymorphism loci.*

Table 2. *Overlapping single nucleotide polymorphisms (SNPs) between GWAS results of manually measured and estimated grain sizes. Chr: Chromosome; Pos: position of SNP on chromosome; length_es: estimated grain length; length_me: manually measured grain length; width_es: estimated grain width; width_me: manually measured grain width*

Trait	SNP	Chr	Pos	-logp	QTLs	candidate gene
length_est	Dj03_14863383F	3	14863383	13.51		
	Dj03_16733440F	3	16733440	4.31	*GS3*	*GS3* [21]
	Dj03_16733443R	3	16733443	4.31	*GS3*	
	Dj05_05360371F	**5**	**5360371**	**4.93**	*GS5.2*	*GS5* [22], *GW5* [23]
	Dj06_00695572F	6	695572	3.55	*GW6*	
	Sj06_00715016F	6	715016	3.77	*GW6*	*GW6* [24]
	Dj07_21410033R	7	21410033	3.67	*GL7*	*GL7* [25]
length_mes	Dj03_14863383F	3	14863383	5.21		
	Dj03_16733440F	3	16733440	4.21	*GS3*	*GS3* [21]
	Dj03_16733443R	3	16733443	4.21	*GS3*	
	Dj06_00695572F	6	695572	4.02	*GW6*	
	Sj06_00715016F	6	715016	4.57	*GW6*	*GW6* [24]
	Dj07_21410033R	7	21410033	4.20	*GL7*	*GL7* [25]
width_est	**Dj05_05360371F**	**5**	**5360371**	**5.52**	*GS5.2*	*GS5.2* [22] *GW5.2* [23]
	Sj05_05368218F	**5**	**5368218**	**4.67**	*GS5.2*	
	Sj05_05369556R	**5**	**5369556**	**4.67**	*GS5.2*	
width_mes	Dj05_04884858R	5	4884858	4.09	*GS5.1*	

4. Conclusions and future works

In this study, the presented results highlighted the combination of low-cost and high-throughput image processing techniques for phenotyping rice grain related to the size traits of rice grains. Through the GWAS analysis, both datasets are able to identify the significant QTLs such as *GS3*, *GW6* and *GL7*. Interestingly, the estimated dataset showed greater sensitivity, identifying additional significant QTLs such as *GS5* and *GW5* on chromosome 5, which were not detected in the manually measured dataset. These results open a promising direction for applying more advanced low-cost image processing and computer vision techniques for the automatic extraction of other quantitative traits of rice varieties.

Acknowledgments

This research was financially supported by the Vietnam Academy of Science and Technology through grant number CN4000.01/22-24.

REFERENCES

[1] T. Mark and P. Langridge. "Breeding technologies to increase crop production in a changing world," *Science*, vol. 327, no. 5967, pp. 818-822, 2010, doi: 10.1126/science.1183700.

[2] W. Michelle *et al.*, "Phenotyping: new windows into the plant for breeders," *Annual review of plant biology*, vol. 71, pp. 689-712, 2020, doi: 10.1146/annurev-arplant-042916-041124.

[3] S. Dawei *et al.*, "Advances in optical phenotyping of cereal crops," *Trends in plant science*, vol. 27, no. 2, pp. 191-208, 2022, doi: 10.1016/j.tplants.2021.07.015.

[4] H. Chenglong *et al.*, "Development of a whole-feeding and automatic rice thresher for single plant," *Mathematical and Computer Modelling*, vol. 58, no. 3-4, pp. 684-690, 2013, doi: 10.1016/j.mcm.2011.10.033.

[5] W. P. Alex *et al.*, "GrainScan: a low cost, fast method for grain size and colour measurements," *Plant methods*, vol. 10, pp. 1-10, 2014, doi: 10.1186/1746-4811-10-23.

[6] Y. Wanneng *et al.*, "Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives," *Molecular plant*, vol. 13, no. 2, pp. 187-214, 2020, doi: 10.1016/j.molp.2020.01.008.

[7] W. Dong *et al.*, "The development of a vision-based phenotypic analysis algorithm for measuring spikelet-related traits in rice," *Plant Physiology Journal*, vol. 58, no. 5, pp. 957-971, 2022, doi: 10.13592/j.cnki.ppj.200009.

[8] S. Lin and P. Chen, "A method for characterizing the panicle traits in rice based on 3D micro-focus X-ray computed tomography," *Computers and Electronics in Agriculture*, vol. 166, 2019, Art. no. 104984, doi: 10.1016/j.compag.2019.104984.

[9] H. Weijuan *et al.*, "Nondestructive 3D image analysis pipeline to extract rice grain traits using X-ray computed tomography," *Plant Phenomics*, 2020, doi: 10.34133/2020/3414926.

[10] Y. Lejun *et al.*, "An integrated rice panicle phenotyping method based on X-ray and RGB scanning and deep learning," *The Crop Journal*, vol. 9, no. 1, pp. 42-56, 2021, doi: 10.1016/j.cj.2020.06.009.

[11] L. Yuwei *et al.*, "High-throughput and separating-free phenotyping method for on-panicle rice grains based on deep learning," *Frontiers in Plant Science*, vol. 14, 2023, Art. no. 1219584, doi: 10.3389/fpls.2023.1219584.

[12] N. T. P. Mai, L. T. T. Nguyen, and H. T. M. To, "Genome-wide association studies for identification of genes and QTLs controlling the palmitic acid content in rice bran oil," *Vietnam Journal of Biotechnology*, vol. 21, no. 2, pp. 337-345, 2023, doi: 10.15625/1811-4989/18607.

[13] N. T. P. Mai, Q. K. Le, T. Q. A. Chu, and H. T. M. To, "Genome wide association studies analysis of the natural ability of uptaking the phosphate in Vietnamese rice landraces," *Vietnam Journal of Biotechnology*, vol. 19, no. 4, pp. 677-686, 2021, doi: 10.15625/1811-4989/15374.

[14] N. T. P. Mai, L. T. T. Nguyen, S. G. Tran, *et al.*, "Genome-wide association study reveals useful QTL and genes controlling the fatty acid composition in rice bran oil using Vietnamese rice landraces," *Funct. Integr. Genomics*, vol. 23, article no. 150, 2023, doi: 10.1007/s10142-023-01080-6.

[15] G. S. Tran, Q. H. Nguyen, L. T. T. Nguyen, *et al.*, "Unraveling New Genetic Elements Associated with the Morphological Changes and Relative Silicon Content in Rice Using Genome-Wide Association Studies (GWAS)," *J. Plant Biol.*, vol. 67, pp. 467-480, 2024, doi: 10.1007/s12374-024-09448-2.

[16] N. T. P. Phung, C. D. Mai, P. Mournet, *et al.*, "Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes," *BMC Plant Biol*., vol. 14, article no. 371, 2014, doi: 10.1186/s12870-014-0371-7.

[17] B. M. Chu, H. T. M. To, and G. S. Tran, "Rice Grain Trait Estimation Using Color Space Conversion and Deep Learning-based Image Segmentation," *TNU Journal of Science and Technology*, vol. 229, pp. 133-140, 2024, doi: 10.34238/tnu-jst.10191.

[18] P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler, "TASSEL: Software for association mapping of complex traits in diverse samples," *Bioinformatics*, vol. 23, pp. 2633-2635, 2007, doi: 10.1093/bioinformatics/btm308.

[19] R Core Team, "R: A language and environment for statistical computing," 2024. [Online]. Available: https://www.R-project.org/. [Accessed December 4, 2024].

[20] K. Yoshihiro *et al.*, "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, pp. 1-10, 2013, doi: 10.1186/1939-8433-6-4.

[21] C. Fan, Y. Xing, H. Mao, T. Lu, B. Han, C. Xu, X. Li, and Q. Zhang, "*GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein," *Theor. Appl. Genet.*, vol. 112, pp. 1164-1171, 2006.

[22] C. Xu, Y. Liu, Y. Li, X. Xu, C. Xu, X. Li, J. Xiao, and Q. Zhang, "Differential expression of *GS5* regulates grain size in rice," *J. Exp. Bot.*, vol. 66, pp. 2611-2623, 2015.

[23] J. Weng, S. Gu, X. Wan, H. Gao, T. Guo, N. Su, C. Lei, X. Zhang, Z. Cheng, X. Guo, *et al.*, "Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight," *Cell Res.*, vol. 18, pp. 1199-1209, 2008.

[24] C. L. Shi, N. Q. Dong, T. Guo, W. W. Ye, J. X. Shan, and H. X. Lin, "A quantitative trait locus GW6 controls rice grain size and yield through the gibberellin pathway," *Plant J.*, vol. 103, no. 3, pp. 1174-1188, 2020, doi: 10.1111/tpj.14793.

[25] Y. Wang, G. Xiong, J. Hu, L. Jiang, H. Yu, J. Xu, Y. Fang, L. Zeng, E. Xu, J. Xu, *et al.*, "Copy number variation at the *GL7* locus contributes to grain size diversity in rice," *Nat. Genet.*, vol. 47, pp. 944-948, 2015.