Genome-wide association analysis provides molecular insights into the natural variation of watermelon seed size

Running title: Molecular mechanism of watermelon seed size variation

Chengsheng Gong 1, Shengjie Zhao 1, Dongdong Yang 1, Xuqiang Lu 1, Muhammad Anees 1, Nan He 1, Hongju Zhu 1, Yong Zhao 1, Wenge Liu 1 *

1Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou 450009, China;
E-mail: Chengsheng Gong: 15738391652@163.com
Shengjie Zhao: zhaoshengjie@caas.cn
Dongdong Yang: 2625478799@qq.com
Xuqiang Lu: luxuqiang@caas.cn
Muhammad Anees: aneesgcaas@outlook.com
Nan He: henan@caas.cn
Hongju Zhu: zhuangju@caas.cn
Yong Zhao: 1690539046@qq.com
Wenge Liu: liuwenge@caas.cn
*Corresponding author: Wenge Liu; Tel.: +86-0371-6533-0936

© The Author(s) 2022. Published by Oxford University Press. All rights reserved.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract:

Seed-consumption watermelon tend to have larger-sized seeds, while flesh-consumed watermelon often require relatively smaller seed. Therefore, the seed size of watermelon has received extensive attention from consumers and breeders. However, the study on the natural variation and genetic mechanism of watermelon seed size is not clear enough. In the present study, 100 seed weight, seed hilum length, seed length, seed width, and seed thickness in 197 watermelon accessions were examined. Furthermore, association analysis was conducted between seed size traits and high-quality SNP data. The results revealed that there was a strong correlation between the five seed traits. And seed enlargement was an important feature during watermelon seed size domestication. Meanwhile, the seed consumption biological species *C. mucusospermu* and *C. lanatus* edible seed watermelon had a significantly bigger seed size than other species's. Eleven non-repeating significant SNPs above the threshold line were obtained by GWAS analysis. Four of them on chromosome 5 were considered to be closely associated with seed size traits, i.e., S5: 32250307, S5: 32250454, S5: 32256177, S5: 32260870, which could be used as potential molecular markers for the breeding of watermelon cultivars with target seed size. In addition, combined with gene annotation information and previous reports, five genes near the four significant SNPs may regulate seed size. And qRT-PCR analysis showed that two genes *Cla97C05G104360* and *Cla97C05G104380*, which may be involved in abscisic acid metabolism, may play an important role in regulating the seed size of watermelon. Our findings provide molecular insights into natural variation in watermelon seed size, and gives valuable information of molecular marker-assisted breeding.

**Keywords:** watermelon, GWAS, seed size
Introduction

Watermelon is an important Cucurbitaceae crop, widely grown all around the world. And classified as seven biological species, i.e., CA (C. amarus), CC (C. colocynthis) CE (C. ecirrhous), CL_CUL (C. lanatus cultivar), CL_LR, (C. lanatus landrace), CM (C. mucosopermus), CN (C. naudinianus). Notably, natural and human selections have caused significant phenotype differences such as fruit size and sugar content during long-term domestication of watermelon. And seed size of watermelon may also have been domesticated. Flesh-consumed watermelon often belong to C. lanatus, and relatively smaller seed can increase the edible part of the watermelon flesh, so the cultivated seedless or small seeded watermelon were more popular in the market; while CM and C. lanatus edible seed watermelon (CL_ES), which were used as seed consumption, often have larger seed. A more comprehensive understanding of the natural variation of watermelon seed is of great significance for breeding cultivars with target seed size, and may provide insights into the domestication process of different species. However, there is currently a lack of complete data support related to the seed size of watermelon in the natural population.

In the plant kingdom, seed size was usually regulated by multiple genes. Previous studies have shown that the ubiquitin-proteasome pathway, G-protein signaling, signaling pathways such as mitogen-activated protein kinase (MAPK) signaling were identified as three main pathways that affect seed size by regulating the size of maternal tissue. In addition, transcription factors such as BS1, AP2, WRKY, the genes in HAIKU (IKU) pathway, and plant hormones, such as abscisic acid, brassinosteroids, auxin, were also thought to play an essential role in regulating plant seed size. At present, some progress has been made in watermelon seed traits. Li, et al. found that the black seed coat of watermelon was controlled by a single dominant gene and identified a candidate gene CICS1 regulating seed coat color by fine-mapping; Li, et al. identified three candidate genes by fine mapping based on the main QTL qSS6 for thousand-seed weight, seed length and seed width, including homologous genes of the seed size regulator SRS3 in rice, Wang, et al. obtained the key gene ClBG1 regulating seed size of watermelon, and after knockout of this gene, the content of abscisic acid (ABA) was reduced, leading to smaller watermelon seed and increased germination. It is worth noting that abscisic acid is a key hormone regulating seed germination. Previous studies have shown that the content of ABA plays an important role in regulating watermelon seed size. Therefore, those genes affect the accumulation of abscisic acid, such as PP2C, may also affect seed size through regulating ABA content.

Genome-Wide Association Study (GWAS) is a method to analyzing genetic variation polymorphisms of multiple individuals by combining with the phenotypic traits in the natural population. Currently, association analysis is often carried out by Gemma, Fast-LMM, Emmax genome-wide association methods. And significant progress has been made in the seed size of rice, Brassica napus and other plants by GWAS. Important progress have also been made in the genetic basis of watermelon agronomic traits by GWAS. Guo, et al. resequenced 414 watermelon accessions, and obtained key regulation genes of sugar content, fruit color, seed coat color and other important traits through GWAS analysis. Dou, et al. combined with the results of GWAS and BSA, obtained the key gene ClFIS1 regulating watermelon fruit shape. GWAS analysis provides a useful tool for the genetic study of important agronomic traits. However, there is no research report on obtaining SNPs and candidate genes closely related to seed size traits through this method.

Understanding the natural variation of seed size may increase the cognition of watermelon domestication and improvement, and genetic research will provide molecular insights for the selection of new cultivars with target seed traits. In the present study, five seed size related traits, i.e., 100 seed...
weights (100 SWT), seed hilum length (SHL), seed length (SL), seed width (SWD), and seed thickness (STH), were measured and analyzed of 197 watermelon accessions in 2019 and 2020. The significant SNP loci linked to seed traits obtained by GWAS could be used as potential molecular markers for molecular marker assisted breeding. Furthermore, candidate genes near SNP loci were analyzed by qRT-PCR (quantitative real-time PCR). Taken together, our findings provides data support for the genetic basis of watermelon seed size.

Results

Natural variation of seed size traits in 197 watermelon accessions

The natural variation of five seed size traits in 2019 and 2020 was evaluated by the statistics of seven indicators: maximum, minimum, mean, standard deviation, coefficient of variation, kurtosis and skewness. The results showed that there was no significant difference in the values of these indexes by comparing the data of two years (Table 1). The correlation of the five seed traits between 2019 and 2020 was more than 94%, indicating that the phenotypic data between two years had a good correlation. In addition, phenotypic coefficient of variation analysis showed that 100 seed weight had the highest variability, i.e., 53.97% and 53.59% in 2019 and 2020, respectively, which was much higher than that of seed hilum length (23.97%, 24.46%), seed length (25.23%, 25.01%), seed width (24.56%, 24.30%), and seed thickness (13.76%, 14.01%). The skewness values of all traits were greater than 0, and kurtosis values of all traits were greater than 3 except for seed length. So it showed that these five traits were quantitative traits following normal distribution in 197 watermelon accessions. In short, the five seed traits were quantitative traits with significant phenotypic variation and strong correlations between two years.

Table 1 Descriptive statistics of five seed traits in watermelon

| Trait   | Mean   | Maximum | Minimum | SD     | CV/%   | Kurtosis | Skewness |
|---------|--------|---------|---------|--------|--------|----------|----------|
| 100SWT-2019 | 7.93   | 23.50   | 0.90    | 4.28   | 53.97  | 3.49     | 1.03     |
| 100SWT-2020 | 7.51   | 21.44   | 0.84    | 4.02   | 53.59  | 3.62     | 1.09     |
| SHL-2019  | 3.34   | 6.23    | 1.35    | 0.80   | 23.97  | 3.23     | 0.18     |
| SHL-2020  | 3.39   | 6.46    | 1.30    | 0.83   | 24.46  | 3.39     | 0.23     |
| SL-2019   | 10.69  | 17.98   | 4.56    | 2.70   | 25.23  | 2.37     | 0.45     |
| SL-2020   | 10.53  | 18.51   | 4.13    | 2.63   | 25.01  | 2.73     | 0.57     |
| SWD-2019  | 6.59   | 12.47   | 3.20    | 1.62   | 24.56  | 3.40     | 0.74     |
| SWD-2020  | 6.47   | 11.06   | 2.63    | 1.57   | 24.30  | 3.26     | 0.74     |
| STH-2019  | 2.26   | 3.78    | 1.64    | 0.31   | 13.76  | 5.58     | 1.02     |
| STH-2020  | 2.22   | 3.91    | 1.60    | 0.31   | 14.01  | 6.79     | 1.24     |

Note: SD, standard deviation. CV, coefficient of variation.

The 100 seed weight is an important indicator of seed size, because it is usually a combination of seed length, seed width, seed hilum length and seed thickness. To better understand the relationship between five seed size traits and their contribution to seed size formation, correlation analysis on two years data were conducted (Fig. 1). In 2019, the correlation between 100 SWT and seed length/seed width was exceeded 90%, while the correlation with seed hilum length and seed thickness was relatively
low, at 72% and 67%, respectively. In addition, the correlation between seed length and seed width was the highest (95%), the correlation between seed thickness and other four traits was less than 70%, and the correlation between the umbilical length and other four traits was less than 80%. At the same time, there was no significant difference between the correlation value in 2020 and 2019 (Fig. 1). These data indicated that seed length and seed width were more likely to determine seed size variability than seed thickness and hilum length.

Fig. 1 Correlation analysis of five seed traits of 197 watermelon accessions in 2019 and 2020. The correlation value and correlation pie chart were displayed in the upper triangle, and the scatter plot between two seed traits were showed in the lower triangular. 100 seed weights (100 SWT), seed hilum length (SHL), seed length (SL), seed width (SWD), and seed thickness (STH).

Obvious phenotypic differences of seed size traits among different biological species

The statistics of the ten accessions with the largest and smallest 100 SWT in two years showed that, the seed weight of the largest ten accessions exceeded 15.5 g, NingXianXiGua, PI179240, PI532722, PI560014, DaBanGuZi, DaBanHongZiGua, DengKouZiGua, GaolanZiGua were common accessions in 2019 and 2020. Of these eight accessions, GaolanZiGua had the largest 100 SWT, exceeding 20 g in 2019 and 2020. And it is worth noting that NingXianXiGua belongs to CL, PI179240, PI532722, and PI560014 belong to CM, and the other four accessions belong to CL_ES. At the same time, among the ten accessions with the smallest 100 SWT in 2019 and 2020, TOMATOSEED, SuXianXiaoZi, 9904,
ZhuXiaoHeiXiaoZi, XiaoZiMiBao, ZXG1702, and PI386015 were co-existed in the two years. Among the seven accessions, TOMATOSEEDD has the smallest seed, and the seed weight was less than 1 g. Interestingly, the first five accessions belong to CL, and the other two accessions belongs to CC. Through the analysis of the extreme watermelon seed size, we found that there may be significant differences between different species of watermelon.

Of the 197 watermelon accessions, 18 accessions were from North America, 20 accessions were from Africa, 5 accessions were from Europe, and 152 accessions were from Asia. The result showed no significant difference of 100 SWT from different regions (Fig. 2A). These accessions including 8 CC, 5 CA, 17 CM, 11 CL_ES, and 156 CL (Table S1). Analysis of 100 SWT in 2020 of different species of watermelon showed that CC has the lowest seed weight, with an average 100 SWT of 3.52 g; CL_ES has the highest seed weight, with an average weight of 15.28 g, and CA, CM, and CL have an average weight of 7.54 g, 12.60 g and 6.67 g (Fig. 2B). Similarly, seed hilum length, seed length, and seed width showed a similar trend to 100 SWT among different watermelon species (Fig. 2C-E). However, for seed thickness, CA and CL_ES had thicker seed, while CC, CM, and CL had thinner seed (Fig. 2F). In addition, there was no significant difference in the performance of these five seed traits between 2019 and 2020 (Fig. 2, Fig. S1). These results showed that the five seed traits of CC were the smallest, while the seed of CA were relatively smaller but thicker seed, which was contrary to CM, while CL_ES has the largest seed weight and was not significantly different from CM, which was consistent with the characteristics that these two types of watermelon were mainly used for seed consumption. And CL had a large degree of seed variation in the natural population. To sum up, The significant differences of seed size between different species might be related to the long-term domestication and artificial selection of watermelon.

Fig. 2 Statistical analysis of five seed size traits in different types of watermelon in 2020. (A) The box plot of 100 SWT in different geographic origin watermelon. (B-F) The box plot of 100 seed weight (g), seed hilum width, seed length (mm), seed width (mm), seed thickness (mm) in different types watermelon. CC: C. colocynthis; CA: C. amarus; CM: C. mucosospermus; CL_ES: C. lanatus edible seed watermelon; CL: C. lanatus.
GWAS analysis of watermelon seed size traits

Principal component analysis was performed on the high-quality SNP data of 197 watermelon accessions. The results showed that the first two principal components PC1 (61.86%) and PC2 (6.35%) could divide these watermelon into five groups, and the group structure was consistent in the phenotypic data analysis of seed size traits, which indicated that accessions we selected were reasonable (Fig. S2).

To further establish the association between SNP and seed size traits in 197 watermelon accessions, we performed two algorithms of EMMAX and Fast-LMM for association analysis. The results showed that no significant SNP loci were associated with seed hilum width and seed thickness; while SNP loci were significantly correlated with the three seed traits, i.e., 100 seed weight, seed length, and seed width. Moreover, the consistency between observed and expected $P$ values were evaluated by the QQ plot, which effectively controlled the generation of false positives (Fig. S3). In 2020 and 2019, 5 and 7 significant SNP loci has obtained of 100 SWT by EMMAX association analysis, respectively; 5 and 10 significant SNPs were obtained of SL; while 3 and 9 significant SNP loci were found of SWD, respectively (Fig. 3, Table S2). A total of 11 non-repeated significant SNPs were obtained on chromosome 5 and chromosome 10. Specifically, S5:32250307 (SNP locus 32250307 on chromosome 5), S5:32256332, S5:32252937 were all associated with three traits in two years; S5:32250454, S5:32254669 and the two SNPs were all significantly associated with traits in association analysis except for SWD in 2020; S5:32251321, S5:32256177 were significant associated with three seed traits in 2019; S5:32180774 was only associated with seed length and seed width in 2019. In addition, S10:28482105, S5:6272374 were the additional SNPs obtained of seed length in 2019, and S5:32260870 was only associated with SWD in 2019. In addition, Fast-LMM algorithm also has similar location results on chromosome 5, indicating the reliability of our analysis results (Fig. S4). These 11 non-repeating SNPs exceeded the threshold that were significantly associated with three seed size traits will be further analyzed to explore whether they can be used as potential molecular markers.
Fig. 3 Manhattan plots of GWAS for 100 seed weight (A), seed length (B), seed width (C) in 2020 and 2019. The abscissa represents the position of a chromosome, the ordinate represents the negative logarithm of $P$ value with the base of 10 ($-\log_{10}(P)$), and the scatter on the graph represents the corresponding $-\log_{10}(P)$ value of each SNP locus. The blue and red dotted lines parallel to the X-axis represent the two significance thresholds.

Analysis of significant SNPs associated with seed size traits

The analysis of eleven significant SNPs correlated with seed traits showed that, for the SNP locus S10:28482105 on chromosome 10, the average 100 SWT was 7.47 g and 7.58 g under the base T (173 accessions) or C (19 accessions), and there was no significant difference; For the locus S5:6272374 on chromosome 5, there was also no significant difference in 100 SWT under different bases. In addition, for the other nine significant SNPs on chromosome 5, the seed size traits have significant differences under different bases (significance level: 0.05, Table S3). For the most significant SNP locus S5:32250307 in 100 SWT and SL, 112 accessions correspond to SNP base T, with an average 100 SWT of 5.66 g, while 76 accessions correspond to base C, with the average 100 SWT was 10.17 g (Fig. 4A). The other eight significant SNPs on chromosome 5 also showed significant differences at different bases of 100 SWT (Fig. 4B-I). In addition, these significant difference in different SNP loci also existed in seed length and seed width (Fig. S5 and Fig. S6). It can be clearly observed that seeds under mutation SNP bases were significantly larger than the seed size with reference bases. The previous analysis shows that CC and CA have smaller seed, so combined with the existing data results, we performed two criteria to screen for SNPs that were more strongly associated with seed size. First, the 100 SWT exceeded 15.4 g in 2019 and 17 g in 2020 under the mutant base; second, CC and CA with smaller seed should match with the reference bases. According to the above criteria, we further screened out four SNP loci
significantly related to seed traits, i.e., S5: 32250307, S5: 32250454, S5: 32256177, S5: 32260870, which
could be used as potential molecular markers for watermelon seed size screening.

**Fig. 4** Violin diagram of significant SNP loci at different bases of 100SWT in 2020. (A-I) represent 9
significant SNP loci, i.e., S5: 32250307, S5: 32256332, S5: 32250454, S5: 32252937, S5: 32254669, S5:
32251321, S5: 32256177, S5: 32180774, S5: 32260870, respectively. The black parts of the boxplot in
the graph represent the range of 25%-75%, and the white circles represent the median.

**Prediction of candidate genes near significant SNPs**

Genes closer to significant SNPs are more likely to be candidate genes for regulating target traits.
Therefore, combining the analysis results of LD decay in the previous report, we further analyzed the
upstream and downstream 100 kb intervals (S5:32150307 - S5:32360870) of the four significant SNPs.
Combined with LD block analysis result, candidate genes were predicted to be in this interval (Fig. S7).
In this interval, there were 31 genes (Table 2). Interestingly, according to previous gene annotations and
reports on regulating seed size, 5 genes in the upstream and downstream 30 kb interval of the most
significant SNP locus S5: 32250307 may be key genes in regulating watermelon seed size. Specifically,
according to gene annotation information and previous research reports, F-box protein is related to organ
size, protein kinase and zinc finger protein have a regulatory effect on rice seed size; in addition,
PP2C and chaperone protein play an important role in the formation of abscisic acid, and abscisic
acid was involved in the development of watermelon seed. Therefore, We suspect that these five genes,
i.e., *Clavata1* (*Cla*) *97C05G104340*, *Cla*97C05G104350, *Cla*97C05G104360, *Cla*97C05G104380,
*Cla*97C05G104390 may be related to watermelon seed size. And the two SNPs S5: 32250307 and S5:
32250454 were located in promoter region of gene *Cla*97C05G104380.

**Table 2** Annotation information of genes that may regulate watermelon seed size in the candidate interval
The RNA of seed in six watermelon accessions were extracted for qRT-PCR analysis to verify the expression level of candidate genes. Among them, the base at the most significant SNP locus S5:32250307 was T in three accessions XiangXiaoGua, SuXianXiaoZi, and XiaoHongYu, and the average 100 SWT was 3.71 g, 1.66 g, and 5.18 g, respectively; HeTaoPi, JiZhuaGua, NingXiaHongZiGua correspond to the mutant base C at this locus was, and the average seed weight was 10.19 g, 14.33 g, and 15.31 g, respectively (Fig. 5A). Further combined with the results of qRT-PCR, we found that the expression levels of Cla97C05G104340, Cla97C05G104350, and Cla97C05G104390 in watermelon seed did not show obvious changes or differences under different bases in 34 DAP (Figs. 5B, 5C, 5F). In contrast, Cla97C05G104360 and Cla97C05G104380 have obvious differences under different base backgrounds in 34 DAP (days after pollination) (Figs. 5D, 5E). In addition, Cla97C05G104360 and Cla97C05G104380 had relatively higher gene expression levels in the three smaller seed materials at 20 DAP (Fig. S8), and similar gene expression results were also found during watermelon flesh development (Fig. S9). As in other plants, members of these two gene families were thought to negatively regulate ABA levels. It also has been confirmed that when CLBG1, a gene regulating seed size, was knocked out, the seeds became smaller and ABA content decreased. Therefore, we hypothesized that Cla97C05G104360 and Cla97C05G104380 may negatively regulate ABA content in watermelon seed and affect watermelon seed size. In addition, the ABA content of the six cultivars was determined and the larger seed had a relatively higher ABA content (Fig. 5A), which was consistent with the hypothesis.
that the two genes may negatively regulate ABA content.

Fig. 5 Seed weight (g), abscisic acid content (μg/L) and the expression levels of five candidate genes. (A) The bar graph represents the weight of the seed, and the dotted graph represents the content of abscisic acid. Gene expression levels of Cla97C05G104340 (B), Cla97C05G104350 (C), Cla97C05G104360 (D), and Cla97C05G104380 (E) Cla97C05G104390 (F) in watermelon seed at 34 days after pollination. L1, L2, L3, H1, H2, H3 represent accessions XiangXiaoGua, SuXianXiaoZi, XiaoHongYu, HeTaoPi, JiZhuaGua and NingXiaHonGZiGua, respectively. T and C represent the reference and mutant bases at SNP locus S5: 32250307 on chromosome 5.

Discussion

Seed enlargement was an important characteristic of watermelon domestication

Natural variation of crops is generated by wild ancestral plants under natural and human selection, and a full understanding of crop phenotypic variation and domestication process can help us to improve the diversity of genetic resources more effectively. For watermelon, the wild watermelon has bitter, pale-coloured, and non-sweet flesh, while the cultivated watermelon has non-bitter, well-coloured and sweet flesh. These differences in traits were the result of long-term selection. For crops, some of the most common domesticated traits in the natural population include decreased seed dormancy and seed enlargement, which has also been confirmed in our research. Compared with other types of watermelon, CC and CA had significantly smaller seed. Specifically, as the ancestor, CC had the smallest seed; and CA had larger seed than CA, but the seed thickness was biggest than other types, which may be an important reason why wild-type watermelon are not easy to sprout. CM was distributed in Western Africa, and CL_ES (belong to CL-LR) was mainly distributed in China. These two types of watermelon were mainly used for seed consumption. The average seed length and width of CM were the largest, while
with smaller seed thickness, the seed weight of CM was relatively smaller than CL_ES. The seed of CL (C. lanatus cultivar and C. lanatus landrace) have a great degree of variation, CL seeds have a greater degree of variation, but the average seed weight was similar to CA, which may be determined by the characteristics of this type species, that is, flesh-consumed watermelon should have smaller seeds. It is worth noting that CM and CL_LR (including CL-ES) with larger seeds linkage disequilibrium decayed to the maximum value within the corresponding distance of 100 kb, while CC and CA were very small. In general, seed expansion is an important feature of watermelon in the process of domestication, while the seed size characteristics of CM, CL_ES and CL were mainly determined by seed consumption or fruit consumption.

GWAS analysis provided potential SNP molecular markers for watermelon seed size traits

Genes are the root cause of differences in traits and provide strong support for research to improve crop yield and quality. For example, overexpression of zmm28 in maize significantly increases the yield of maize, and the monoterpen synthase gene cluster is involved in the formation of carotene flavor. BSA and genetic map were common methods of gene fine mapping. Xue et al. obtained the candidate interval of pear red skin by BSA and developed available molecular markers; important QTLs regulating watermelon seed size were found on chromosomes 2, 6, 8, et. al. However, BSA and genetic map tend to only be able to take a long time to build separation population. GWAS provides an effective tool for fast and accurate gene mapping of target traits. Guo et al. identified significantly SNP related to grape fruit traits through GWAS analysis, which may be applied to molecular marker-assisted breeding. The development and utilization of molecular markers have greatly improved the efficiency of breeding. In the current study, four SNP loci significantly correlated with 100 SWT, SL and SWD were obtained, i.e., S5: 32250307, S5: 32250454, S5: 32256177, S5: 32260870. These four significant SNPs providing potential molecular markers for seed traits in natural populations for the first time.

Two candidate genes may be involved in the formation of watermelon seed size by regulating ABA content

Based on gene annotation information and previous reports, 5 genes may regulate watermelon seed size. Specifically, the gene annotation information of Cla97C05G104340 is F-box protein, in Arabidopsis, F-box protein STERILE APETALA (SAP)/SUPPRESSOR OF DA1 (SOD3) affects the stability of transcriptional regulatory factors Peapods (PPDs), and in turn to regulate organ size. Cla97C05G104350 has a conserved domain of protein kinase, STKC_IRAK. In previous study, non-synonymous mutations in the exon of BAK1, an important gene with this conserved domain, resulted changes in rice grain size. Cla97C05G104390 is a zinc finger protein, and it has a C2H2 conserved domain. C2H2 zinc finger protein was thought to affect the development of spikelets in rice, thus affecting seed size and yield. The content of ABA was an important hormone affecting the seed size of plants such as watermelon. The gene annotation information of Cla97C05G104360 is a PP2C family protein. PP2C can negatively regulate the content of ABA in previous studies, and also had a study found that PP2C regulation of seed size in soybean. Cla97C05G104380 encodes a chaperone protein dnaJ gene. In Arabidopsis, ATJ3 (chaperone, dnaJ homolog 3) inhibits PKS5 (Protein Kinase5) activity, and PKS5 (SOS2-LIKE PROTEIN KINASE5) was an important
gene involved in ABA responses \(^{36}\), suggesting that dnaJ may affect the content of ABA through negative regulation in plants. The seed of six watermelon accessions with different base backgrounds were used for qRT-PCR analysis. The results showed that genes \(Cla97C05G104360\) and \(Cla97C05G104380\) have obvious differences in gene expression levels in different types of watermelon. And these two genes may be closely related to the accumulation of ABA. Combined with previous studies on watermelon seed, after the knockout of the gene \(CLBG1\), ABA content decreased and watermelon seed became smaller \(^{5}\), so we speculate that these two genes may be candidate genes for regulating the size of watermelon seed. It is worth noting that seed size is a complex quantitative trait, and we need to combine more experiments with candidate genes and retrograde validation. Specifically, in the next step, we will conduct further laboratory work to verify the function of candidate genes and construct isolated population validation experiments to further verify the possible key role of candidate genes in regulating watermelon seed size. Overall, our study provides possible molecular markers for the gene mapping of watermelon seed and clarifies the variation of watermelon seeds in natural population.

Materials and Methods

Plant materials and cultivation management

A total of 197 watermelon accessions have been sequenced and used in this experiment. These materials were collected from the National Mid-term Genebank for Watermelon and Melon, Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences (Zhengzhou, Henan, China). The 197 watermelon accessions were planted in Xinxiang (Henan, China) in 2019 and Zhongmu (Zhengzhou, Henan, China) in 2020. Divided by geographical location, these germplasm resources were mainly from North America, Africa, Europe, and Asia; The classification of different species of accessions was mainly based on the previous report of Guo, et al. \(^{4}\). These accessions were divided into 5 types, CC and CA were wild-type watermelon; CM and CL_ES were often used for seed consumption and were separately classified; \(C. lanatus\) cultivar and \(C. lanatus\) landrace were closely related, so classified into CL.

In addition, six accessions used for qRT-PCR analysis were planted in Xinxiang in the spring of 2021 and sampled watermelon seed at 20 DAP and 34 DAP. The watermelon plants were planted with double pruning, the time of sowing and transplanting to the field according to local climatic conditions, and random block design (RCBD) was adopted and maintained at 0.8 m of plant spacing and 1.5 m of row spacing. Agronomic practices remained consistent in both environments, including fertilization, irrigation, pest control, etc. The pollination was carried out by self-crossing, and only one fruit was grown on each plant; the harvest out of watermelon according to the characteristics of different accessions. Fresh seed were washed and placed in breathable gauze bags. After drying, the dried seed were obtained for further phenotypic determination.

Measurement of seed size traits

Five seed size traits were manually determined, including 100 seed weights (100 SWT), seed hilum length (SHL), seed length (SL), seed width (SWD), and seed thickness (STH). The seed with good growth were selected from each independent watermelon to determine the 100-seed weight. Three watermelon was measured as three biological replicates, which was measured with an electronic balance (JA2003, Soptop). Meanwhile, for SHL, SL, SWD and STH, the seed with relatively uniform appearance
and good growth were selected for the determination. After seed were selected from each watermelon, a
digital vernier caliper (AIRAJ) was used to measure and record the four agronomic traits. The results of
the same cultivars from different watermelon were averaged for further analysis.

Statistics and analysis of phenotypic data

The maximum values, minimum values, average values, variance values, standard deviation, and
variation values of the five seed traits in 197 watermelon accessions were counted and calculated by
Excel 2020. In addition, to further clarify the overall distribution of the five traits in the natural population,
two important indicators, including kurtosis and skewness, were calculated. The calculation formulas for
the kurtosis and skewness of the population were as follows:

\[ K = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n - 1)s^4} \]

\[ g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n - 1)s^3} \]

K: kurtosis; g: skewness; \( \bar{x} \): mean; \( x_i \): the ith sample; n: total number of samples; s: standard
deviation.

Correlation analysis between every two seed traits was calculated by the default statistical method
of Origin software (Origin2021). The correlation analysis different seed size trait was obtained by the
Corrplot package of R software.

SNP acquisition and GWAS analysis

Of the 197 watermelon accessions used for GWAS analysis, 163 of them were sequenced from
previous research reports (Number WM in Table S1), and the genome library construction, sequencing
and analysis methods of the other 34 accessions were based on previous research methods (Number R in
Table S1). SNP data with variation were obtained also based on the previous analysis method, and
filtered according to secondary allele frequency (MAF: 0.05) and site integrity (INT: 0.8) to obtain highly
consistent SNP for further analysis. To evaluate whether the selected population was suitable for GWAS
analysis, Principal Component Analysis (PCA) was performed based on SNP data of different types
species, and EIGENSOFT \( ^{37} \) software was used to obtain the result of sample clustering. EMMAX \( ^{15} \)
(efficient mixed model association eXpedited) and Fast-LMM (Factored Spectrally Transformed Linear
Mixed Models) \( ^{16} \) were used to analyze five seed size traits for GWAS analysis. QQ plot and manhattan
plot were drawn by R software. The significant threshold was calculated by Bonferroni correction, and
the \(-\log_{10}(P) 0.1/ Ne (Ne = effective SNP number)\) and \(-\log_{10}(P) 0.01/Ne\) were set as two threshold
lines for screening the significant SNPs. According to the characteristics of seed size traits and previous
LD decay results, we selected the upstream and downstream 100 kb interval near the significance SNP
as the candidate interval, and conducted LD block analysis to increase the possibility of candidate interval.
97103 V2 (http://cucurbitgenomics.org/organism/21) was used as a reference genome for related analysis.

RNA extraction and quantitative real-time PCR

The seed of six watermelon accessions were fully developed and had the largest seed size at 34 days
after pollination, so we sampled watermelon seed from this period. Fresh seed from the same watermelon
were sampled and quickly quick-frozen in liquid nitrogen and stored at -80°C refrigerator. The seed were
ground into powder in liquid nitrogen for further RNA extraction. Total RNA extraction was carried out
according to the instructions of one RNA isolation kit (Huayueyang Biotechnologies, China). And cDNA
was synthesized from 1 μg RNA as a template according to the instructions of NovoScript plus all-in-one 1st stand cDNA synthesis supermix (Novoprotein, China). Gene sequences were obtained based on Cucurbitaceae genome database (http://www.icugi.org), and ClCAC (Gene ID: Cla97C09G174930) was used as the reference gene for qRT-PCR. Primers of candidate genes were designed shown in Table S4.

Expression levels of a candidate gene was measured by a LightCycler480 RT-PCR system (Roche, Swiss). And the 20 μL reaction system was constructed according to the SYBR Green real-time PCR mix's instructions. The main parameters of the program were: after preheating at 95℃ for 5 min, 45 cycles were executed at 95℃ for 10 s, 56℃ for 30 s, and 72℃ for 30 s. Finally, to obtain the relative expression level of genes, the obtained original data were analyzed by the 2−ΔΔCT method.

**Determination of ABA content in watermelon seed**

The seed of six watermelon accessions at 34 DAP were fully powdered in liquid nitrogen. Accurately weight 0.2 g-0.5 g of seed powder into the collection tube, and add PBS solution (Ph=7.3) at the ratio of weight (g): volume (ml) = 1:9, and place the seed in a refrigerator at 4℃ for 2 h after full shock. Next, centrifuge in a centrifuge (Centrifuge 5810R) at 4℃ for 20 minutes and collect the supernatant as the sample to be tested. Standards were diluted gradually in accordance with 300 μg/L, 150 μg/L, 75 μg/L, 37.5 μg/L, 18.75 μg/L, and the content of abscisic acid was determined by ELISA kit (www.mmbio.cn). Follow the instructions of the PLANT Abscisic acid (ABA) ELISA Kit. The standard curve formula for ABA content determination was calculated, and the abscisic acid content of samples was obtained by substituting the spectrophotometric values of each sample into the formula.

**Acknowledgments**

This work was financially supported by Agricultural Science and Technology Innovation Program (CAAS-ASTIP-2021-ZFRI), the China Agriculture Research System of MOF and MARA (CARS-25-03), National Key R&D Program of China (2018YFD0100704), the National Nature Science Foundation of China (31672178 and 31471893) and Natural Science Foundation of Henan Province (212300410312).

**Author contributions**

L.W., G.C. and Z.S. designed this research; L.W., Z.S., L.X., H.N. and Z.H. were mainly responsible for the collection of germplasm resources and field management; G.C., Y.D., Z.Y. carried out the research, and mainly responsible for the investigation of traits and data sorting and analysis. G.C. wrote the manuscript, while L.W. and M.A. were responsible for the verification of the manuscript.

**Conflict of interest**

The authors declare no competing interests.

**Data availability**

All data included in this study is publicly available. All experimental data are shown in the attachment. SNP data are mainly from previous studies and uploaded to public databases or obtained data by contacting authors.

**Reference**

1. Guo, S. et al. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. Nat Genet 45, 51-58 2470 (2013).
2. Renner, S.S., Sousa, A. & Chomicki, G. Chromosome numbers, sudanese wild forms, and classification
of the watermelon genus Citrullus, with 50 names allocated to seven biological species. *TAXON* 66, 1393-1405 (2017).

3. Chomick, G. & Renner, S.S. Watermelon origin solved with molecular phylogenetics including Linnaean material: another example of museomics. *New Phytol* 205, 526-532 (2015).

4. Guo, S. *et al.* Resequeing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat Genet* 51, (2019).

5. Wang, Y. *et al.* CRISPR/Cas9-mediated mutagenesis of CIBG1 decreased seed size and promoted seed germination in watermelon. *Hortic Res* 8, 70 (2021).

6. Ge, L. *et al.* Increasing seed size and quality by manipulating BIG SEEDS1 in legume species. *Proc Natl Acad Sci USA* 113, 12414 (2016).

7. Jofuku, K.D., Omidyar, P.K., Gee, Z. & Okamuro, J.K. Control of seed mass and seed yield by the floral homeotic gene APETALA2. *Proc Natl Acad Sci USA* 102, 3117 (2005).

8. Cheng, Z.J. *et al.* Abscisic acid regulates early seed development in Arabidopsis by ABI5-mediated transcription of SHORT HYPOCOTYL UNDER BLUE1. *Plant Cell* 26, 1053-1068 (2014).

9. Ishimaru, K. *et al.* Loss of function of the IAA-glucose hydrolase gene TGW6 enhances rice grain weight and increases yield. *Nat Genet* 45, 707-711 (2013).

10. Feng, Z. *et al.* SLG controls grain size and leaf angle by modulating brassinosteroid homeostasis in rice. *J Exp Bot* 67, 4241-4253 (2016).

11. Li, B. *et al.* genetic mapping and discovery of the candidate gene for black seed coat color in watermelon (Citrullus lanatus). *Front Plant Sci* 10, 1689-1689 (2020).

12. Li, N. *et al.* Fine mapping and discovery of candidate genes for seed size in watermelon by genome survey sequencing. *Sci Rep* 8, 17843-17843 (2018).

13. Wang, K. *et al.* EAR1 negatively regulates ABA signaling by enhancing 2C protein phosphatase activity. *Plant Cell* 30, 815-834 (2018).

14. Park, S.-Y. *et al.* Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* 324, 1068-1071 (2009).

15. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44, 821-824 (2012).

16. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* 8, 833-835 (2011).

17. Sul, J.H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat Rev Genet* 14, 300 (2013).

18. Dunn, P. *et al.* Natural variation in the promoter of GSE5 contributes to grain size diversity in rice. *Mol Plant* 10, 685-694 (2017).

19. Miller, C. *et al.* Variation in expression of the HECT E3 ligase UPL3 modulates LEC2 levels, seed size, and crop yields in Brassica napus. *Plant Cell* 31, (2019).

20. Ren, D. *et al.* A new regulator of seed size control in Arabidopsis identified by a genome-wide association study. *Mol Plant* 222, 895-906 (2019).

21. Dou, J. *et al.* Genetic mapping reveals a candidate gene (CIFS1) for fruit shape in watermelon (Citrullus lanatus L.). *Theor Appl Genet* 131, 947-958 http://dx.doi.org/10.1007/s00122-018-3050-5 (2018).

22. Li, N. *et al.* STERILE APETALA modulates the stability of a repressor protein complex to control organ size in Arabidopsis thaliana. *Plas Genet* 14, c1007218 (2018).

23. Yuan, H. *et al.* 08SG2/OsBAK1 regulates grain size and number, and functions differently in Indica and Japonica backgrounds in rice. *Rice* 10, 25 (2017).
24. Xu, Q. et al. The C2H2 zinc-finger protein LACKING RUDIMENTARY GLUME 1 regulates spikelet development in rice. *Science Bulletin* **65**, 753-764 (2020).

25. Zhuang, H. et al. NONSTOP GLUMES1 encodes a C2H2 zinc finger protein that regulates spikelet development in rice. *Plant Cell* **32**, 392-413 (2020).

26. Yang, Y. et al. The Arabidopsis Chaperone J3 regulates the plasma membrane H+-ATPase through interaction with the PKS5 kinase. *Plant Cell* **22**, 1313-1332 (2010).

27. Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol* **65**, 531-51 (2014).

28. Paris, H.S. Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*. *Ann Bot* **116**, 133-48 (2015).

29. Olsen, K.M. & Wendel, J.F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu Rev Plant Biol* **64**, 47-70 (2013).

30. Wu, J. et al. Overexpression of zmm28 increases maize grain yield in the field. *Proc Natl Acad Sci U S A* **116**, 23850 (2019).

31. Reichardt, S. et al. The carrot monoterpene synthase gene cluster on chromosome 4 harbours genes encoding flavour-associated sabinene synthases. *Hortic Res* **7**, 190 (2020).

32. Xue, H. et al. Interval mapping for red/green skin color in Asian pears using a modified QTL-seq method. *Hortic Res* **4**, 17053 (2017).

33. Ren, Y. et al. An integrated genetic map based on four mapping populations and quantitative trait loci associated with economically important traits in watermelon (*Citrullus lanatus*). *BMC Plant Biol* **14**, 33 (2014).

34. Prothro et al. Main and epistatic quantitative trait loci associated with seed size in watermelon. *J AM SOC HORTIC SCI* (2012).

35. Guo, D.-L. et al. Genome-wide association study of berry-related traits in grape (*Vitis vinifera L.*) based on genotyping-by-sequencing markers. *Hortic Res* **6**, 11 (2019).

36. Zhou, X. et al. SOS2-LIKE PROTEIN KINASE5, an SNF1-RELATED PROTEIN KINASE3-type protein kinase, is important for abscisic acid responses in arabidopsis through phosphorylation of ABSCISIC ACID-INSENSITIVE5. *Plant Physiol* **168**, 659-676 (2015).

37. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).

38. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).