Detection of Genomic Variation and Tepal Shape Related Genes Between Broad and Narrow Tepal Lotus (Nelumbo Adans.) by whole Genome Resequencing

Feng-Luan Liu
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Mi Qin
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Shuo Li
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Da-Sheng Zhang
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Qing-Qing Liu
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Meng-Xiao Yan
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Dai-Ke Tian (dktian@cemps.ac.cn)
Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden

Research Article

Keywords: Asian lotus, Genomic variation, Indel, Nelumbo, SNP, Tepal, Petal shape, Whole genome resequencing

Posted Date: February 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-208049/v1

Version of Record: A version of this preprint was published at Horticulturae on December 20th, 2021. See the published version at https://doi.org/10.3390/horticulturae7120593.

Abstract

Background: The shape, color, and size of petals or tepals are vital attributes of flowering plants, as they affect the pollinator attraction ability, environmental stress adaptation, and ornamental value of plants. Compared with rose, chrysanthemum, and water lily, the tepal shapes of lotus (*Nelumbo* Adans.) exhibit low variation, and the absence of short-broad and long-narrow tepals limits the commercial value of lotus flowers. Therefore, this study aimed to uncover the genomic variation underlying the broad and narrow tepal phenotypes of lotus flowers, and to screen candidate genes associated with different tepal shapes.

Results: Whole genome resequencing of two groups of lotus, NL (comprising three American lotus genotypes: broad tepal mutant, narrow tepal mutant, and wild type) and WSH (comprising two Asian lotus genotypes: narrow tepal mutant and wild type), generated 9.23–12.85 Gb clean data. A total of 716,656 single nucleotide polymorphisms (SNPs) and 221,688 insertion-deletion mutations (Indels) were obtained in the NL group, while 639,953 SNPs and 134,6118 Indels were obtained in the WSH group. Only a small proportion of these SNPs and Indels was mapped to exonic regions of genome: 1.92% and 0.47%, respectively, in the NL group, and 1.66% and 0.48%, respectively, in the WSH group. Gene Ontology (GO) analysis showed that out of 4,890 (NL group) and 1,272 (WSH group) annotated variant genes, 125 and 62 genes were enriched ($Q < 0.05$), respectively. Additionally, in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, 104 genes (NL group) and 35 genes (WSH group) were selected ($P < 0.05$). Finally, 41 genes were filtered and predicted to be associated with tepal shape in lotus.

Conclusions: This is the first comprehensive report of genomic variation controlling tepal shape in lotus. Significant genetic variation was detected between the wild-type and mutant lotus genotypes, and varying levels of differentiation between groups NL and WSH. Candidate genes containing epidermal growth factor (EGF) and wall-associated receptor kinase (WAK) domains are considered important targets for further research on tepal development. Overall, the genomic data and candidate genes obtained in this study are essential references for future identification of tepal-shaped control genes in lotus combined with transcriptome analysis and quantitative trait loci (QTL) mapping.

Background

Lotus (*Nelumbo* Adans.) is one of the most economically important plants in the world. The genus *Nelumbo* has two species: *Nelumbo nucifera* Gaertn., commonly known as Asian lotus, and *Nelumbo lutea* Willd., commonly known as American lotus [1, 2]. Asian lotus is one of the top 10 traditional flowers in China and the national flower of India. Lotus, particularly Asian lotus, is usually divided into three types, depending on its agricultural application: ornamental lotus, rhizome lotus, and seed lotus [3, 4]. Approximately 2,000 commercial cultivars of ornamental lotus have been developed to date, accounting for 96% of the total number of the three types [5]. However, as an ornamental character of importance [6], the shape of petals (referred to as tepals in lotus) exhibits little phenotypic variation in these cultivars. Few of lotus germplasms exhibit short-broad tepals (similar to rose) and long-narrow tepals (as in chrysanthemum), which limits the horticultural value of lotus. Therefore, to develop lotus cultivars with unique tepal shapes, understanding the genetic basis of tepal shape determination is critical.

Genetic mapping of quantitative trait loci (QTL), genome-wide association studies (GWAS), and transcriptome analysis using modern high-throughput sequencing methods are powerful approaches that facilitate the identification of novel genes and pathways. The completion of *de novo* sequencing of lotus genome [7, 8] and the rapid development of high-throughput sequencing technology have made it possible to analyze nucleotide sequence variation among various lotus cultivars. Whole genome resequencing is used to detect single nucleotide polymorphisms (SNPs) and insertion-deletion mutations (Indels) in the whole genome [9, 10] and to mine structural variation and copy number variation by high-depth sequencing [11, 12]. Using this technique in lotus, Hu et. al [13] identified a large number of nucleotide sequence variation and 103,656 simple sequence repeats (SSRs) between the genomes of ‘Chiang Mai Wild’ and ‘Middle Lake Wild’, which

could be used for QTL mapping and molecular-assisted breeding. In another study, whole genome resequencing of 19 accessions belonging to three subgroups of cultivated temperate lotus revealed that rhizome lotus showed the lowest genomic diversity, and the candidate genes related to temperate lotus and tropical lotus always exhibited divergent expression patterns [14].

Petal growth depends on the precise orchestration of the frequency of cell division, orientation of cell division planes, and extent of cell elongation [15–17]. For example, non-uniform cell expansion within a petal resulted in distorted petals in *Eustoma grandiflorum* [18]. Although the molecular basis of petal shape is largely unclear, it has been shown that several genes, such as *LEUNIG* [19, 20], *SEUSS* and *APETALA2* [21, 22], *JAGGED* [23–25], *CYCLOIDEA* [26], INDOLE-3-BUTYRIC ACID-RESPONSE5 [27], and SPIKE1 [28], regulate petal shape by affecting cell proliferation and or cell expansion. The roles of these genes in petal shape regulation were mainly verified in model plants including *Arabidopsis* and *Antirrhinum*.

In this study, two groups of lotus, with wide-short and narrow-long tepals, were analyzed by whole genome resequencing. We explored the differences in the number and distribution of SNPs and Indels within each group and between the two lotus groups and screened functional genes and metabolic pathways related to the development of broad and narrow tepals. The results of this study provide important clues for the exploration of genes affecting tepal shape in lotus.

Results

Whole genome resequencing data of the six genotypes

A total of six samples belonging to groups NL and WSH were subjected the whole genome resequencing, and 9.23–12.85 Gb clean data, with ‘Q20 rate’ > 97% and ‘Q30 rate’ > 92%, were generated (Additional file 1). Among the six samples, M652 produced the lowest amount of clean data (9.23 Gb), whereas NL-CK produced the highest amount (12.85 Gb). The GC content of all samples varied from 39.28% to 40.14%, and no significant difference was detected in GC content between and within the two groups (Additional file 1). The mapping rate of NL group genotypes (93.63%–94.43%) was lower than that of WSH group accessions (98.22%–98.43%). Additionally, the 1× and 4 × coverage of NL group samples was also lower than those of WSH group accessions; the 4 × coverage of NL group ranged from 80.41% to 83.54%, whereas that of the WSH group varied from 94.65% to 98.67% (Additional file 1).

SNPs and Indels in the six genotypes

The sequence reads of all six accessions were mapped onto the reference genome sequence of lotus, and SNPs and Indels were called for each accession. The total number of SNPs identified in each sample of NL group was about 7–10-fold higher than that of WSH group, and it was about 8–10-fold higher in Indel (Additional file 2). Of all the SNPs identified in the NL group, 6.15% (M512), 5.90% (DP23), and 5.73% (NL-CK) were mapped to exonic regions, which were approximately 3.5 times greater than the number of exonic SNPs in the WSH group. Of all the Indels identified in the NL group, 0.84% (M512), 0.81% (DP23), and 0.79% (NL-CK) were mapped to exonic regions, which was approximately 1.5 times those of group WSH (Additional file 2).

Within each group, the controls contained more SNPs and Indels than their mutants (Additional file 2), which was probably associated with the increase in variant sites caused by the pooling of DNA extracted from three seedlings per control. Additionally, the ratio of nonsynonymous to synonymous SNPs (Nonsyn/Syn), transitions to transversions (Ts/Tv), and exonic sequence variation to total variation (exonic/total) varied slightly among the three accessions within each group (Additional file 2), to some extent, suggesting that the three genotypes within each group exhibit very low genetic diversity.

Analysis of SNPs and Indels between mutant accessions and their controls

Within each group, the SNPs and Indels were filtered between any two samples (except between WSH-CK1 and WSH-CK2). A total of 716,656 SNPs and 221,688 Indels were obtained from the NL group, while 639,953 SNPs and 134,6118 Indels were obtained from the WSH group (Fig. 1). Compared with the controls, nucleotide polymorphisms in the mutant samples of the two groups occurred mainly in the intergenic regions: intergenic SNPs accounted for 61.3% and 72.6% of all SNPs identified in NL and WSH groups, respectively, and intergenic Indels accounted for 58.4% and 67.4% of all Indels identified in NL and WSH groups, respectively (Fig. 1). By contrast, the proportion of SNPs and Indels that mapped to exonic regions was very small in both groups (1.9% SNPs and 0.47% Indels in the NL group; 1.7% SNPs and 0.48% Indels in the WSH group) (Fig. 1). This showed that the ratio of exonic SNPs to total SNPs was slightly higher in the NL group than in the WSH group.

Annotation and analysis of genes harboring SNPs and Indels in the exonic regions

The predicted amino acid sequences of the variant genes with exonic SNPs and Indels were blasted in the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases to annotate their functions. In the GO database, 4,890 (72.8%) of the 6,715 variant genes in the NL group were selected for annotation and assigned to three functional categories for a total of 19,785 times (Table 1). The most times of these variant genes were assigned to the biological process category (99,311 times [50.2%]), followed by the cellular component and molecular function categories (Table 1). In the entire genetic background and variant genes, GO terms including 'ADP binding', 'pattern binding', and 'polysaccharide binding' were significantly ($Q < 0.05$) enriched in the molecular function category and contained 88, 37, and 37 candidate genes, respectively (Fig. 2A). After excluding duplicates, a total of 125 candidate genes were selected from the above three GO items (Additional file 3).

Table 1
Annotation and distribution of variant genes in GO and KEGG databases

Database	Sample group	No. of variant genes	No. of annotated genes	Percentage of annotated genes	Distributed times of annotated genes			
					Total times	Biological process	Cellular component	Molecular function
GO	NL	6,715	4,890	72.8	197,885	99,311	72,072	26,502
	WSH	5,353	1,272	23.8	52,425	25,824	20,751	5,850
Database	Sample group	No. of variant genes	No. of genes annotated	Percentage of annotated genes	No. of pathways	No. of annotated genes in the top three pathways		
						Metabolic pathways	Biosynthesis of secondary metabolites	Starch and sucrose metabolism
KEGG	NL	6,715	2,286	34.0	122	430	256	67
	WSH	5,353	645	12.0	68	128	84	39

In the WSH group, 1,272 genes were annotated in the GO database, accounting for 23.8% of the 5,353 variant genes, which was significantly lower than the proportion of annotated genes in the NL group (Table 1). These annotated genes of WSH group received a total of 52,425 assignments in the three functional categories (in the same order as that

observed in the NL group): 25,824 times (49.3%) in the biological process category, followed by 20,751 times in the cellular component category, and 5,850 times in the molecular function category (Table 1). Seven GO terms, such as 'ADP binding', 'chitin metabolic process', 'chitin catabolic process', showed significant differences ($Q < 0.05$). Among them, 'ADP binding' contained 55 genes, while each of the remaining six GO terms contained 12 identical genes (Fig. 2B). Finally, A total of 62 candidate genes were identified in the seven GO terms, after excluding gene duplicates (Additional file 3).

In the KEGG database, 2,286 variant genes of the NL group were assigned into 122 pathways, of which the top three pathways were 'metabolic pathways' (430 genes), 'biosynthesis of secondary metabolites' (256 genes), and 'starch and sucrose metabolism' (67 genes) (Table 1). By contrast, only 645 genes of the WSH group were annotated and distributed to 68 pathways, and the top three pathways with the most annotated genes and their order were the same as those in the NL group, each with 128, 84, and 39 genes, respectively (Table 1). However, no pathway that showed a significant difference, based on the Q-value ($Q < 0.05$), was found in the two groups. While four pathways (comprising 104 genes, Additional file 3) with significant *P*-value ($P < 0.05$) were detected in the NL group, including 'ribosome biogenesis', 'ABC transporters', 'propanoate metabolism', and 'phenylpropanoid biosynthesis'. And five pathways (comprising 35 genes, Additional file 3) were detected in the WSH group ($P < 0.05$), including 'ABC transporters', 'starch and sucrose metabolism', 'plant-pathogen interaction', 'ether lipid metabolism', and 'amino sugar and nucleotide sugar metabolism'.

In summary, 227 (125 from GO database and 104 from KEGG database) and 89 (62+35) enriched genes (based on significant Q- or *P*- value) were collected from the NL and WSH groups, respectively, after GO and KEGG annotations (Additional file 3), which are an essential reference for locating and screening tepal shape-regulating genes of lotus in the future.

Candidate genes related to the development of tepal shape in lotus

According to the 227 (NL group) and 89 (WSH group) genes enriched in the GO and KEGG databases, those shared by both groups and with mutation rate > 5‰ in the exonic regions were selected. Thus, a total of 41 crucial candidate genes were obtained that potentially regulate the tepal shape in lotus, including eight genes shared by the two groups and 24 and nine genes exclusive to the NL and WSH groups, respectively (Additional file 4). Among the eight shared genes, three (LOC104590721, LOC104605275, and LOC104607443) were enriched in the GO but not the KEGG database, and the remaining five were enriched in both GO and KEGG databases (Additional file 4).

Of the 41 candidate genes, 35 were annotated with a clear function, including 17 disease resistance related genes, six (like-) receptor protein kinases, and 12 related to chitinase, transferase, synthase, and other functions (Additional file 4). The lengths of the 41 genes varied from 891 to 17,163 bp, in which two candidate genes (LOC104607832 and LOC104609114) produced four transcript variants, and 10 genes produced two to three transcript variants (Additional file 4).

In terms of 'family and domain', the genes harboring nucleotide-binding adaptor shared by APAF-1, R proteins and CED-4 (NB-ARC) and leucine-rich repeat (LRR) domains were the most (18, 44%) and mainly related to the function of 'ADP binding' (Additional file 4). While the domains of the six (like-) receptor protein kinases were epidermal growth factor (EGF)-like, Pkinase, and wall-associated receptor-kinase galacturonan-binding (GUB_WAK), which are associated with the morphological development of plants due to their functions such as 'ATP binding', 'calcium ion binding', 'polysaccharide binding', and 'protein serine/threonine kinase activity' (Additional file 4). In the KEGG database, only 18 of the 41 candidate genes were assigned to specific pathways, while the remaining 23 genes were unknown (Additional file 4).

Discussion

Differences in genome mapping between the NL and WSH groups

The two groups of samples investigated in this study were representative wild-type germplasms of the two species of *Nelumbo*, and whole genome resequencing results revealed many differences at the DNA level. For example, the mapping rate, 1× coverage, and 4× coverage of the three lotus accessions in the NL group were lower than those in the WSH group (Additional file 1). These differences occurred probably because 'China Antique' (source of the reference genome) exhibited a closer genetic relationship with each of the three accessions within the WSH group (native to China) than with the geographically isolated American lotus accessions in the NL group. This reason may also explain why more SNPs and Indels were detected in the NL group than in the WSH group (Table 1). Previously, we showed that many SNPs and Indels among M512, DP23, and 'China Antique' could be detected using expressed sequence tags (ESTs) and SSRs [33], which indirectly supported the above inference that the reference genome exhibited lower sequence similarity with the American lotus genome than with 'Weishan Hong' genome.

In addition, the three accessions in the WSH group exhibited higher Nonsyn/Syn ratio and Ts/Tv ratio than those in the NL group (Additional file 2). In other words, less non-synonymous mutations and transitions occurred in the genomes of NL group accessions, probably because of the lack of gene transfer between Asian lotus and American lotus during the long period of geographical isolation [14, 34].

Distribution of tepal shape related SNPs and Indels on lotus chromosomes

To visualize the genomic diversity among samples, we drew 12 separate maps of SNP and Indel densities on the chromosomes for the six accessions (Additional file 5–8), with the parameters of "window = 100k" and "step = 100k". On the SNP density map of the WSH group, three color-blocks showed visible differences between M652 and its two controls (WSH-CK1 and WSH-CK2) in the 20 longest scaffolds of the lotus genome (Additional file 5), which represent key regions for QTL mapping and gene scanning for the tepal shapes trait in future studies. For instance, the SNP density of M652 in the 6.0–6.3 M interval of NW010729074.1 (located on lotus chromosome 3) was higher than those of the two controls (Additional file 5). In this region, there were 12 variant genes, among which the LOC104586031 locus in M652 genome contained the highest number of exonic SNPs (five) compared with the two controls and was also significantly enriched in the GO database (Additional file 5). Besides, LOC104586031 was known to encode a 2-OG-Fe(II) oxygenase superfamily protein (ISP7), but its function has not been confirmed. On the Indel density map, the narrow tepal accession, M652, exhibited four sites with lower Indel density than the WSH-CK1 and WSH-CH2 controls (Additional file 6), and these sites may represent the chromosome regions associated with tepal shape in lotus.

As described above, the mapping rate in the NL group was lower than that in the WSH group, probably because of the distant genetic relationship between NL accessions and 'Chinese Antique' (the reference genome). Therefore, no visual distinction was evident on the maps of SNP or Indel density among the three NL group accessions (Additional file 7–8).

Genes associated with tepal shape in lotus

Tepal shape and size are remarkably consistent within a given ecotype [35, 36] but are influenced by the environment [37-40]. Among the 41 tepal shape related candidate genes identified in this study, 17 genes encoded disease resistance proteins belonging to the NB-LRR family, of which 12 were exclusive to the NL group, four were common to both groups, and only one gene was unique to the WSH group (Additional file 4), implying that M512 and or DP23 in the NL group may have undergone natural selection for resistance to disease and survived. This hypothesis was supported, to some extent, by the explicit functions of 21 genes detected by the polymorphic bands between M512 and DP23, which were not only

involved in plant organogenesis, plant hormone synthesis, and signal transduction but reportedly are also involved in plant physiology and metabolism under stress [33]. Additionally, the contrasting number of disease resistance genes between the two groups was consistent with, at least to some degree, the fact that M512 and DP23 were derived from natural mutation or selection of lotus seeds, whereas M652 was induced by artificial radiation.

The development of tepal shape mainly regulated by cellular proliferation and expansion and the direction of cell division [15, 17]. In this study, six genes predicted to encode EGF-like or WAK domain-containing proteins (Additional file 4), required for cell expansion [41–43], may play an essential role in the morphogenesis of lotus tepals. *ERECTA*, which encodes a LRR receptor-like serine-threonine kinase (STK), was verified to be a major effect locus for shaping petals in *Arabidopsis thaliana* [44]. A BLAST search of this *Arabidopsis* gene sequence in the lotus genome identified its homolog, LOC104600563, which was also annotated as *ERECTA*. However, this gene was not one of the 366 genes (NL 277 + WSH 89) screened in this study. Nonetheless, the LOC104604923 gene identified in the WSH group as well as the previously reported WAK domain-containing genes LOC104590721, LOC104596880, LOC104608917, and LOC109115528 were all STK-type genes [41] like *ERECTA*, they will be the focus of subsequent research.

Conclusions

The shape and size of petals are essential for the survival and reproduction of plants, and both these morphological features are two of the most important targets in ornamental crop breeding. Lotus germplasms with broad and narrow tepals generated by natural mutation and artificial irradiation, respectively, represent ideal materials for exploring the mechanism of tepal shape development. The mutant samples and their respective controls within each group showed a large number of SNPs, Indels, and candidate genes, which also exhibited differences between group NL and WSH. Candidate EGF and WAK domain-containing genes in this study related to cell expansion and division must be investigated in future studies for understanding tepal development in lotus. Overall, the results of this study provide a valuable resource for future identification of petal-shaped control genes combined with transcriptome technology and QTL mapping and for understanding the mechanism of flower morphogenesis in *Nelumbo*.

Methods

Plant material

In 2015, two accessions of American lotus with significant differences in tepal shape were obtained from the ‘International Nelumbo Collection’ (certified by International Waterlily & Water Gardening Society in 2016) of Shanghai Chenshan Botanical Garden, Shanghai, China. Both these accessions were germinated from wild lotus seeds collected from Florida, USA. Compared with the wild American lotus, one genotype (M512) showed broad and short tepals, while the other genotype (DP23) exhibited narrow and long tepals (Fig. 3), and the tepal shapes of both accessions were stable after years of vegetative propagation via rhizomes. In 2016, 1,024 seeds of ‘Weishan Hong’, a wild Asian lotus, were irradiated with cobalt-60, which emits gamma rays, and a mutant (M652) with narrow and long tepals was obtained (Fig. 3), and the tepal shape was stable after three years of clonal propagation.

The NL group accessions include M512 (wide-short tepals), DP23 (narrow-long tepals), and NL-CK (wild American lotus; the control). The WSH group accessions include M652 (derived from ‘Weishan Hong’ by gamma irradiation; narrow-long tepals) and WSH-CK (‘Weishan Hong’ belonging to wild Asian lotus; the control). Scale bars = 1 cm.

Grouping of lotus accessions

The lotus genotypes investigated in this study were divided into two groups: NL (short for *N. lutea*), which comprised American lotus genotypes, M512, DP23, and their wild-type control (NL-CK), and WSH (short for 'Weishan Hong'), which comprised M652 and its two wild-type controls, WSH-CK1 and WSH-CK2. Each control contained three wild seedlings.

DNA extraction

Three young leaves of each mutant (M512, DP23, and M652) were harvested separately from three pots of their clones, and one young leaf of three seedlings of each control (NL-CK, WSH-CK1, and WSH-CK2) were collected. DNA was extracted separately from the three leaves of the same genotype using the modified cetyl triethylammonium bromide (CTAB) method [29], in which the sugar removal was performed three times, then they were pooled together at equimolar concentrations. The integrity of the pooled DNA samples was detected by electrophoresis on 1% agarose gels, and the DNA quality and concentration were measured by NanoDrop 2000c and Qubit 2.0 Fluorometer (Thermo Fisher Scientific). DNA samples with OD260/OD280 (optical density, OD) values of 1.8–2.0 and with contents greater than 1.5 μg were used for library construction.

Library construction and whole genome resequencing

The purified genomic DNA was randomly sheared into 350 bp fragments using a COVARIS M220 Focused-ultrasonicator^{TM} (Covaris, Woburn, MA, USA). Genomic DNA library was constructed using the TruSeq Nano DNA HT Sample Preparation kit (Illumina Inc., San Diego, CA, USA), and its quality was assessed on the Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA). The size of inserts was determined using Agilent Technologies 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Finally, whole genome resequencing was conducted by Beijing Novogene Bioinformatics Technology Co. Ltd. (Beijing, China) at a depth of 10× using the Illumina HiSeq 2500 platform.

Date filtering

The original image data generated by the sequencing machine were converted into sequence data via base calling (Illumina pipeline CASAVA v1.8.2). The sequence data were then subjected to the quality control (QC) procedure to remove unusable reads: 1) The reads with a Phred quality (Q) score <20; 2) The reads contain the Illumina library construction adapters; 3) The reads contain more than 10% unknown bases (N bases); 4) One end of the read contains more than 50% of low-quality bases (sequencing quality value ≤5).

Read mapping, variant detection, and annotation

The clean reads were mapped to the reference genomes of 'China Antique' (accession number: GCA_000365185.2) using the Burrows-Wheeler Aligner (BWA) program [30] using the following parameters: 'mem -t 4 -k 32 -M'. Subsequent processing, including duplicate removal, was performed using SAM tools [31] with the parameter 'rmdup' and PICARD (http://picard.sourceforge.net). Raw SNP and Indel sets were called with SAM tools using the following parameters: 'mpileup -m 2 -F 0.002 -d 1000'. Then, these sets were filtered using the following criteria: mapping quality > 20; depth of the variate position > 4. ANNOVAR [32] was used for the functional annotation of variants.

Screening and enrichment analysis of variant genes

Genes showing SNPs and Indels between any two samples (except between WSH-CK1 and WSH-CK2) within each group were selected. Then, among these genes, those showing SNPs and Indels in the exonic regions were manually extracted and searched in the NCBI non-redundant protein (Nr) database for annotation and in the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases for enrichment analyses.

Screening of genes controlling tepal shape

Based on the blasting in the GO and KEGG databases, genes from group NL and WSH showing significant enrichment ($Q < 0.05$) were identified. Among these genes, two types of genes, common to both groups and showing a mutation rate (SNP and Indel density per kb in the coding region) greater than 5‰, were extracted and searched in NCBI and UniProtKB databases to clarified their biological information. Both types of genes were considered as candidates that regulate tepal shape in lotus.

Abbreviations

ABC: ATP-binding cassette; ADP: Adenosine 5'-diphosphate; ATP: Adenosine triphosphate; CTAB: Cetyltriethylammnonium bromide; EGF: Epidermal growth factor; EST-SSR: Simple sequence repeat markers based on express sequence tags; GO: Gene Ontology database; GUB: Galacturonan-binding; KEGG: Kyoto Encyclopedia of Genes and Genomes database; LRR: Leucine rich repeat; NB-ARC: *Nucleotide binding* adaptor shared by APAF-1, R proteins, and CED-4; NL: The group of American lotus; NL-CK: The control in the NL group; Nonsyn: Non-synonymous; OD: Optical density; QTL: Quantitative trait loci; QC: Quality control; Syn: Synonymous; Ts: Transitions; Tv: Transversion; WAK: Wall-associated receptor kinase; WSH: The group of Asian lotus; WSH-CK; The control in the WSH group.

Declarations

Acknowledgements

We thank Dr. Ken Tilt (Auburn University, Alabama, USA) and Dr. Lyn Gettys (University of Florida, Florida, USA) for facilitating the collection of American lotus seeds. We also thank Dr. Xiao-Li Chen (Beijing Genomics Institution) for providing suggestions on data processing.

Authors’ contributions

DT and FL designed the study. FL, MQ, and SL carried out comparative genomic analyses. DZ, QL, and MY cultivated lotus plants, performed DNA extractions and carried out bioinformatics analysis of candidate genes. FL, MQ, and SL drafted the manuscript. DT revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by the National Natural Science Foundation of China (NSFC) (Grant No. 31601791), the Shanghai Landscaping Administration Bureau (Grant No. G162402).

Availability of data and materials

All supporting data can be found within the manuscript and its additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Plant Science Research Center, the Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden, Shanghai 201602, China.

References

1. Li HL. Classification and phylogeny of Nymphaeaceae and allied families. Am Midl Nat. 1955;54:33–41.
2. Li Y, Svetlana P, Yao J X, Li CS. A review on the taxonomic, evolutionary and phytogeographic studies of the lotus plant (Nelumbonaceae: *Nelumbo*). Acta Geol Sin. 2014;88:1252–61.
3. Wang QC, Zhang XY. Lotus flower cultivars in China. Bao MZ trans. Forestry Publishing House Press. 2005:59.
4. Guo HB. Cultivation of lotus (*Nelumbo nucifera* Gaertn. ssp. *nucifera*) and its utilization in China. Genet Resour Crop Evol. 2009;56:323–30.
5. Liu L Li YC Min J Tian DK. Analysis of the cultivar names and characteristics of global lotus (*Nelumbo*). Agric Sci (Hans, China). 2019;9:163–81.
6. Lin ZY, Zhang CH, Cao DD, Damaris RN, Yang PF. The latest studies on lotus (*Nelumbo nucifera*)-an emerging horticultural model plant. Int J Mol Sci. 2019;20:3680.
7. Ming R, Van Buren R, Liu YL, Yang M, Han YP, Li LT et al. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol. 2013;14:R41.
8. Gui ST, Peng J, Wang XL, Wu ZHH, Cao R, Salse J, et al. Improving *Nelumbo nucifera* genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. Plant J. 2018;94:721–34.
9. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
10. Say YH. The association of insertions/deletions (INDELs) and variable number tandem repeats (VNTRs) with obesity and its related traits and complications. J Physiol Anthrop. 2017;36:25.
11. Neerman N, Faust G, Meeks N, Modai S, Kalfon L, Falik-Zaccai T, et al. A clinically validated whole genome pipeline for structural variant detection and analysis. BMC Genomics, 2019;20:545.
12. Magi A, Pippucci T, Sidore C. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. BMC Genomics. 2017;18:1–15.

13. Hu JH, Gui ST, Zhu ZX, Wang XL, Ke WD, Ding Y. Genome-wide identification of SSR and SNP markers based on whole-genome re-sequencing of a Thailand wild sacred lotus (*Nelumbo nucifera*). PLoS ONE. 2015;10:e0143765.
14. Huang LY, Yang M, Li L, Li H, Yang D, Shi T, et al. Whole genome re-sequencing reveals evolutionary patterns of sacred lotus (*Nelumbo nucifera*). J Integr Plant Biol. 2018;60:2–15.
15. Hill JP, Lord EM. Floral development in *Arabidopsis thaliana*: A comparison of the wild type and the homeotic pistillata mutant. Can J Bot. 1989;67:2922–36.
16. Hase Y, Fujioka S, Yoshida S, Sun G-Q, Umeda M, Tanaka, A. Ectopic endoreduplication caused by sterol alteration results in serrated petals in Arabidopsis. J Exp Bot. 2005;56:1263–8.
17. Peaucelle A, Wightman R, Höfte H. The control of growth symmetry breaking in the Arabidopsis hypocotyl. Curr Biol. 2015;25:1746–52.
18. Kawabata S, Nii K, Yokoo M. Three-dimensional formation of corolla shapes in relation to the developmental distortion of petals in *Eustoma grandiflorum*. Sci Hortic. 2011;132:66–70.
19. Liu ZC, Meyerowitz EM. *LEUNIG* regulates *AGAMOUS* expression in *Arabidopsis* flowers. Development. 1995;121:975–91.
20. Conner J, Liu ZH-CH. *LEUNIG*, a putative transcriptional corepressor that regulates *AGAMOUS* expression during flower development. Proc Natl Acad Sci. 2000;97:12902–7.
21. Franks RG, Wang CX, Levin JZ, Liu ZC. *SEUSS*, a member of a novel family of plant regulatory proteins, represses floral homeotic gene expression with *LEUNIG*. Development. 2002;129:253–63.
22. Franks RG, Liu ZC, Fischer RL. *SEUSS* and *LEUNIG* regulate cell proliferation, vascular development and organ polarity in *Arabidopsis* petals. Planta. 2006;224:801–11.
23. Dinneny JR, Yadegari R, Fischer RL, Yanofsky MF, Weigel D. The role of *JAGGED* in shaping lateral organs. Development. 2004;131:1101–10.
24. Sauret-Güeto S, Schiessl K, Bangham A, Sablowski R, Coen E. JAGGED controls Arabidopsis petal growth and shape by interacting with a divergent polarity field. PLoS Biol. 2013;11:e1001550.
25. Schiessl K, Muiño JM, Sablowski R. Arabidopsis JAGGED links floral organ patterning to tissue growth by repressing Kiprelated cell cycle inhibitors. Proc Natl Acad Sci. 2014;111:2830–35.
26. Juntheikki-Palovaara I, Tähtiharju S, Lan TY, Broholm SK, Rijpkema AS, Ruonala R, et al. Functional diversification of duplicated CYC2 clade genes in regulation of inflorescence development in *Gerbera hybrida* (Asteraceae). Plant J. 2014;79:783–96.
27. Johnson KL, Ramm S, Kappel C, Ward S, Leyser O, Sakamoto T, et al. The *Tinkerbell* (*Tink*) Mutation Identifies the dual-specificity MAPK phosphatase INDOLE-3-BUTYRIC ACID-RESPONSE5 (IBR5) as a novel regulator of organ size in Arabidopsis. PLoS ONE. 2015;10:e0131103.
28. Ren HB, Dang X, Yang YQ, Huang DQ, Liu MT, Gao XW, et al. SPIKE1 activates ROP GTPase to modulate petal growth and shape. Plant Physiol. 2016;172:358–71.
29. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5.
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.

33. Liu FL, Qin M, Liu QQ, Zhang DS, Tian DK. Detection of genetic variation between broad and narrow-tepalled American lotus (*Nelumbo lutea* Willd.) by EST-SSR markers. Acta Agric Boreali-occident Sin. 2020;29:306–14.
34. Li Z, Liu XQ, Gituru RW, Juntawong N, Zhou MQ, Chen LQ. Genetic diversity and classification of *Nelumbo* germplasm of different origins by RAPD and ISSR analysis. Sci Hortic. 2010;125:724–32.
35. Pyke KA, Page AM. Plastid ontogeny during petal development in Arabidopsis. Plant Physiol. 1998;116:797–803.
36. Brock MT, Weinig C. Plasticity and environment-specific covariances: an investigation of floral-vegetative and within flower correlations. Evolution. 2007;61:2913–24.
37. Marsden-Jones EM, Turrill WB. A quantitative study of petal size in *Saxifraga granulata* L. J Genet. 1947;48:206–18.
38. Yoshioka Y, Iwata H, Ohsawa R, Ninomiya S. Analysis of petal shape variation of *Primula sieboldii* by elliptic fourier descriptors and principal component analysis. Ann Bot. 2004;94:657–64.
39. Yoshioka Y, Iwata H, Ohsawa R, Ninomiya S. Quantitative evaluation of the petal shape variation in *Primula sieboldii* caused by breeding process in the last 300 years. Heredity (Edinb.). 2005;94:657–63.
40. Chacón B, Ballester R, Birlanga V, Rolland-Lagan AG, Pérez-Pérez JM. A quantitative framework for flower phenotyping in cultivated carnation (*Dianthus caryophyllus* L.). PLoS ONE. 2013;8:e82165.
41. He ZH, Cheeseman I, He DZ, Kohorn BD. A cluster of five cell wall-associated receptor kinase genes, *Wak1–5*, are expressed in specific organs of *Arabidopsis.* Plant Mol Biol. 1999;39:1189–96.
42. Wagner TA, Kohorn BD. Wall-associated kinases are expressed throughout plant development and are required for cell expansion. Plant Cell. 2001;13:303–18.
43. Kohorn BD, Kohorn SL. The cell wall associated kinases, WAKs, as pectin receptors. Front Plant Sci. 2012;3:88.
44. Abraham MC, Metheetrairut C, Irish VF. Natural variation identifies multiple loci controlling petal shape and size in *Arabidopsis thaliana*. PLoS ONE. 2013;8(2):e56743.

Figures

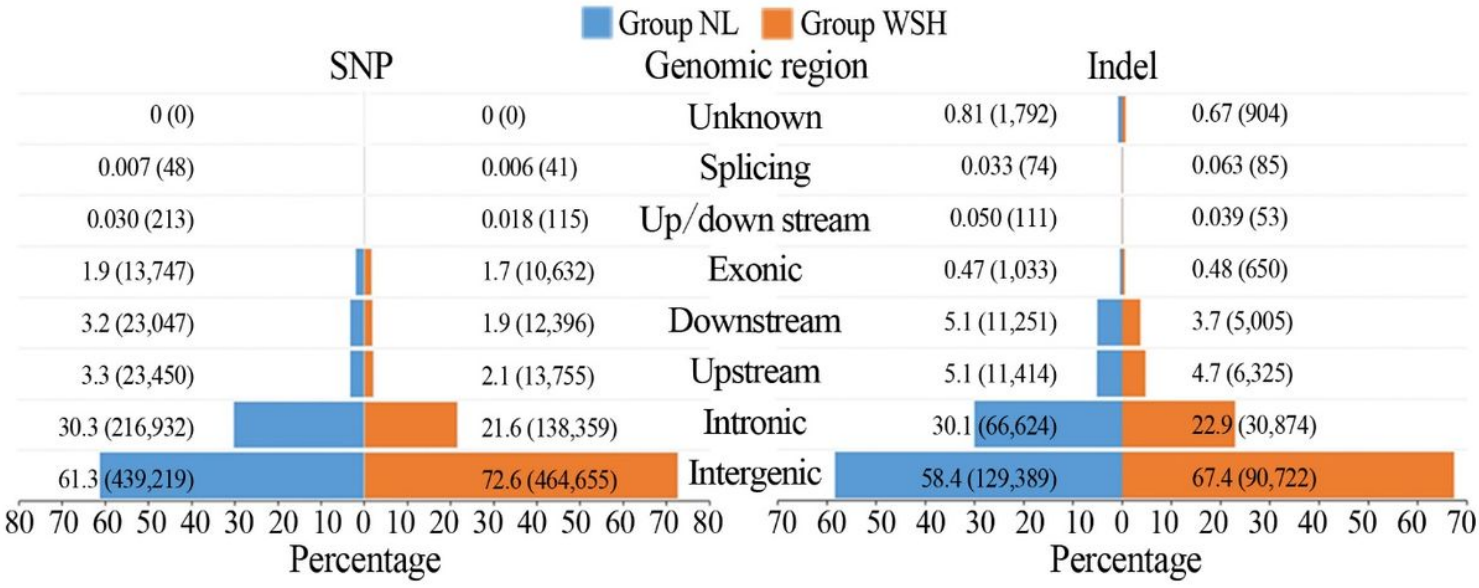

Figure 1

Distribution, proportion, and number of SNPs and Indels identified within each group. Each pair of data represented the percentage (numbers out the brackets) and the amount (numbers in the brackets), respectively.

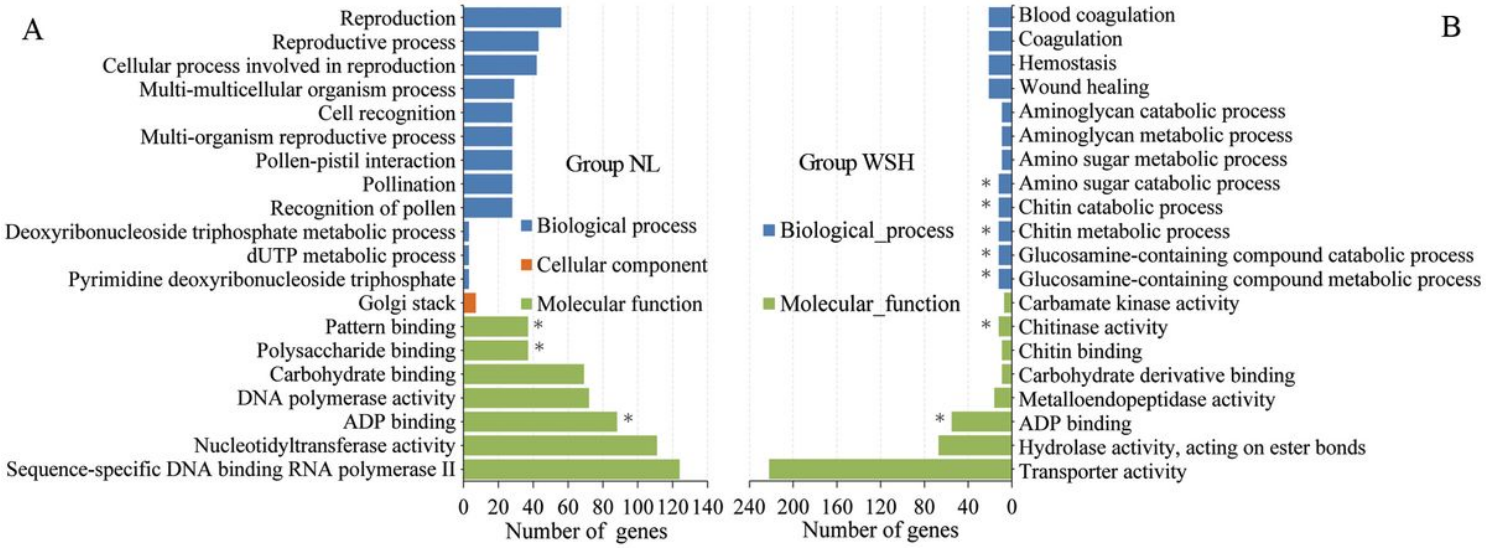

Figure 2

Top 20 of most enriched Gene Ontology (GO) terms among variant genes. GO terms showing significant differences (Q < 0.05) within each group were indicated with an asterisk (*).

Figure 3

Buds and second-day flowers of NL group (top panel) and WSH group (bottom panel). The NL group accessions include M512 (wide-short tepals), DP23 (narrow-long tepals), and NL-CK (wild American lotus; the control). The WSH group accessions include M652 (derived from ‘Weishan Hong’ by gamma irradiation; narrow-long tepals) and WSH-CK (‘Weishan Hong’ belonging to wild Asian lotus; the control). Scale bars = 1 cm.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile2.jpg
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.jpg
- Additionalfile6.jpg
- Additionalfile7.jpg
- Additionalfile8.jpg