genes

MDPI

Article

Genome-Wide Survey and Analysis of Microsatellites in Waterlily, and Potential for Polymorphic Marker Development

Xiang Huang [1], Meihua Yang [1,*], Jiaxing Guo [1], Jiachen Liu [1], Guangming Chu [1] and Yingchun Xu [2]

[1] College of Agriculture, Shihezi University, Shihezi 832003, China
[2] College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China
* Correspondence: ymh_agr@shzu.edu.cn

Abstract: Waterlily (Nymphaeaceae), a diploid dicotyledon, is an ornamental aquatic plant. In 2020, the complete draft genome for the blue-petal waterlily (*Nymphaea colorata*) was made available in GenBank. To date, the genome-wide mining of microsatellites or simple sequence repeats (SSRs) in waterlily is still absent. In the present study, we investigated the characteristics of genome-wide microsatellites for *N. colorata* and developed polymorphic SSR markers across tropical and hardy waterlilies. A total of 238,816 SSRs were identified in 14 *N. colorata* chromosomes with an average density of 662.60 SSRs per Mb, and the largest number of SSRs were present on chromosome 1 (n = 30,426, 705.94 SSRs per Mb). The dinucleotide was the most common type, and AT-rich repeats prevail in the *N. colorata* genome. The SSR occurrence frequencies decreased as the number of motif repeats increased. Among 2442 protein-coding region SSRs, trinucleotides, accounting for 63.84%, were the most abundant. Gene ontology terms for signal transduction (e.g., GO: 0045859 and GO: 0019887) and the lipoic acid metabolism (ko00785,) were overrepresented in GO and KEGG enrichment analysis, respectively. In addition, 107,152 primer pairs were identified, and 13 novel polymorphism SSR markers were employed to distinguish among nine waterlily cultivars, of which Ny-5.2 and Ny-10.1 were the most informative SSR loci. This study contributes the first detailed characterization of SSRs in *N. colorata* genomes and delivers 13 novel polymorphism markers, which are useful for the molecular breeding strategies, genetic diversity and population structure analysis of waterlily.

Keywords: *Nymphaea colorata*; genome-wide mining; microsatellites; polymorphism markers

Citation: Huang, X.; Yang, M.; Guo, J.; Liu, J.; Chu, G.; Xu, Y. Genome-Wide Survey and Analysis of Microsatellites in Waterlily, and Potential for Polymorphic Marker Development. *Genes* **2022**, *13*, 1782. https://doi.org/10.3390/genes13101782

Academic Editor: Serena Aceto

Received: 9 September 2022
Accepted: 28 September 2022
Published: 2 October 2022

1. Introduction

Waterlily, belonging to the Nymphaeaceae family, is an ornamental aquatic plant. There are at least 50 species in this genus and more than 1000 horticultural cultivars, with a worldwide distribution in tropical to temperate regions [1,2]. According to its eco-physiological characteristics, it can be divided into tropical and hardy waterlily [3]. Among these, blue-petal waterlily (*Nymphaea colorata*) is a typical tropical waterlily possessing gorgeous flowers and a high ornamental value [4]. Meanwhile, *Nymphaea candida*, a species native to Xinjiang Province (northwestern China), is a typically hardy waterlily with strong cold resistance and high biological significance [3].

In previous studies, various methods of molecular markers, such as amplified fragment length polymorphism (AFLP), inter-simple sequence repeat (ISSR), random amplified polymorphic DNA (RAPD) and simple sequence repeat (SSR), have been applied to the study of aquatic plant genomics, such as *Nelumbo nucifera*, *Euryale ferox* and *Ranunculus nipponicus* [5–7]. Among them, SSR markers have the advantages of high polymorphism, genome richness and co-dominance in genetic diversity and phylogenetic analysis. Meanwhile, genome-wide microsatellite mining has been carried out on many species, such as *Brassica napus* [8], *Punica granatum* [9], *Brassica oleracea* [10], *Fagopyrum tataricum* [11]

and *Anemone coronaria* [12], which offered novel molecular markers for their genetic improvement and genetic diversity studies. Moreover, in aquatic plants, genome-wide SSR markers of *N. nucifera* [6] were mined and polymorphic primers were successfully developed to analyze the inner-species difference and genetic similarities between cultivated and wild lotus.

N. colorata, a diploid dicotyledon, lies at the base of the angiosperm linage and has a strong genetic significance. At present, there are many varieties of waterlily species and commercial species, but there are few studies on the specific molecular markers of *Nymphaea* [13–15] or the molecular-level identification of different varieties. Therefore, it is crucial to analyze the characteristics of the simple repeat sequence (SSR) loci of the whole-genome sequence of *N. colorata* to identify germplasm resources, analyze genetic diversity and construct a genetic linkage map of *Nymphaea*. In 2020, the *N. colorata* genome was discovered, which has a relatively small size of 409.472 Mb [16]. This finding, a research milestone, provides a chance to develop research on molecular markers, comparative genomics and functional genomics in waterlily.

In this study, we aimed to investigate genome-wide SSRs in *N. colorata* by an analysis of SSR density, occurrence frequency and the possible functions of SSRs in CDS regions. Furthermore, polymorphic SSR markers were developed to distinguish multiple waterlily cultivars. Our study provides a useful tool for molecular breeding strategies, genetic diversity and the population structure analysis of waterlily.

2. Materials and Methods

2.1. Source of Genomic Sequences

N. colorata genome assemble sequences were obtained from the National Center for Biotechnology Information (NCBI), and its accession number is GCF_008831285.1. In addition, the annotation file (7.6 Mb) of the *N. colorata* genome was downloaded from https://ftp.ncbi.nlm.nih.gov/genomes, accessed on 15 October 2020.

2.2. Identification of Microsatellite and Primer Design

The microsatellite identification software MISA (http://pgrc.ipk-gatersleben.de/misa, accessed on 20 November 2020) was used to identify genome-wide and coding region microsatellites in the *N. colorata* genome sequences. To identify perfect microsatellites, the minimum repeat number was defined as 10 for mono-, 6 for di- and 5 for tri-, tetra-, penta- and hexa-nucleotide SSRs [17,18]. The flanking sequences of SSRs were used as targets for the primer design by the Primer3 program (http://frodo.wi.mit.edu/, accessed on 6 May 2021) with a MISA-generated Primer3 input file [19], and the major selecting parameters for primer design were as follows: primer length, 18–23 bp, with 20 bp being optimal; PCR product size, 100–500 bp, with 200 bp being optimal; an annealing temperature of 50–65 °C; and an optimal GC content of 40–60%.

2.3. SSR Location, GO and KEEG Enrichment

Based on the *N. colorata* genome annotation file (.gff3 Data), the intergenic and gene-intergenic regions were calculated according to the length from the starting to ending position information of gene, CDS and exon regions. The gene ontology (GO) enrichment analysis of CDSs with microsatellites, including GO term mapping, classification and enrichment, was conducted using Blast2GO software [20], WEGO software [21] and the GoSeq program [22], respectively. In order to further study the functions of these genes in terms of the networks of genes and molecules, the CDSs containing microsatellites were aligned to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg, accessed on 10 December 2021) using BLASTx with an e-value of $<10^{-5}$ [23].

2.4. Plant Materials and DNA Extraction

A total of nine waterlily cultivars were collected from four localities (originating from Hainan, Guangdong, Xinjiang provinces and Beijing City) (Table S1). Young leaf materials were immediately preserved with liquid nitrogen, transported to the lab and stored at −80 °C until the DNA was extracted. DNA was then extracted from the samples using a Plant Genomic DNA Kit (Tiangen Biotech, Beijing, China) according to the manufacturer's instructions. DNA quality and quantity were checked using the Nanodrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA).

2.5. Validation of SSR Primer Pairs by PCR Amplification

The 50 primer pairs randomly selected from the 107,152 primers were synthesized by Sangon Biological Engineering Technology & Service Co. (Shanghai, China) for polymorphism validation. SSR-primed polymerase chain reactions (PCRs) were performed in 10 μL reaction volumes, which contained 5 μL PCR mixed solution (TransGen Biotec, Beijing, China), 0.3 μL forward primer (10 nmol/L), 0.3 μL reverse primer (10 nmol/L), 3.4 μL double distilled water and 1 μL DNA template. PCR amplification was carried out with the following cycling conditions: an initial denaturation at 95 °C for 5 min followed by 30 cycles of denaturation at 94 °C for 30 s, annealing at 50–60 °C (dependent on the different primers) for 25 s and extension at 72 °C for 40 s. The PCR reaction was ended with a 5 min incubation step at 72 °C. The amplified products were separated on 8% polyacrylamide gels with 1 × TBE buffer at a constant voltage of 220 V for 50–60 min and visualized by silver staining.

2.6. Data Analysis

After the silver staining of the PCR products, the alleles with the maximum molecular weight were manually recorded in binary format as 'A', followed by 'B', 'C', etc., in decreasing order by molecular weight. The polymorphism information content (PIC) was used to calculate the discriminatory power of primer pairs targeting each SSR locus using Power Marker Software (version 3.25) [24]. Meanwhile, the number of alleles (*Na*), the effective number of alleles (*Ne*), Shannon's information index (*I*), observed heterozygosity (*Ho*), expected heterozygosity (*He*) and inbreeding coefficient (Wright's fixation index, *Fis*) were calculated using POPGENE (version 1.3.1) [25]. The phylogenetic relationship among nine waterlily cultivars was constructed in a dendrogram based on similarity coefficients using the program NTSYS-pc (version 2.10) [26]. The clustering map was created based on genetic distances and the unweighted pair group method with arithmetic mean (UPGMA), and the simple matching (SM) coefficient was used to construct a tree.

3. Results

3.1. SSR Motifs Content in the N. colorata Genome

The *N. colorata* genome was assembled into 1429 contigs (with a contig N50 of 2.1 Mb), and the GC content was 38.59%. The total length of the genome was 409.5 Mb with 804 scaffolds, and it was anchored onto 14 chromosomes. A total of 238,816 SSRs were identified in 14 *N. colorata* chromosomes. Out of these, dinucleotide was the most common type, accounting for 51.67% (n = 123,399), followed by mononucleotide (43.56%, n = 104,031), trinucleotide (3.92%, n = 9354), tetranucleotide (0.65%, n = 1549), pentanucleotide (0.12%, n = 286) and hexanucleotide (0.08%, n = 197) (Figure 1A). The analysis of SSR distribution in 14 chromosomes revealed that (i) the average density was 662.60 SSRs per Mb; (ii) the largest number of SSRs were present on chromosome 1 (n = 30,426, 705.94 SSRs per Mb), followed by chromosome 2 (n = 24,996, 714.17 SSRs per Mb); and (iii) the largest numbers of SSR types were dinucleotide and mononucleotide on chromosome 1 (Figure 1B). In addition, Pearson correlation analysis revealed that chromosome length was significantly positively associated with the number of SSRs in each chromosome ($r = 0.982$, $p < 0.01$) (Figure 1C).

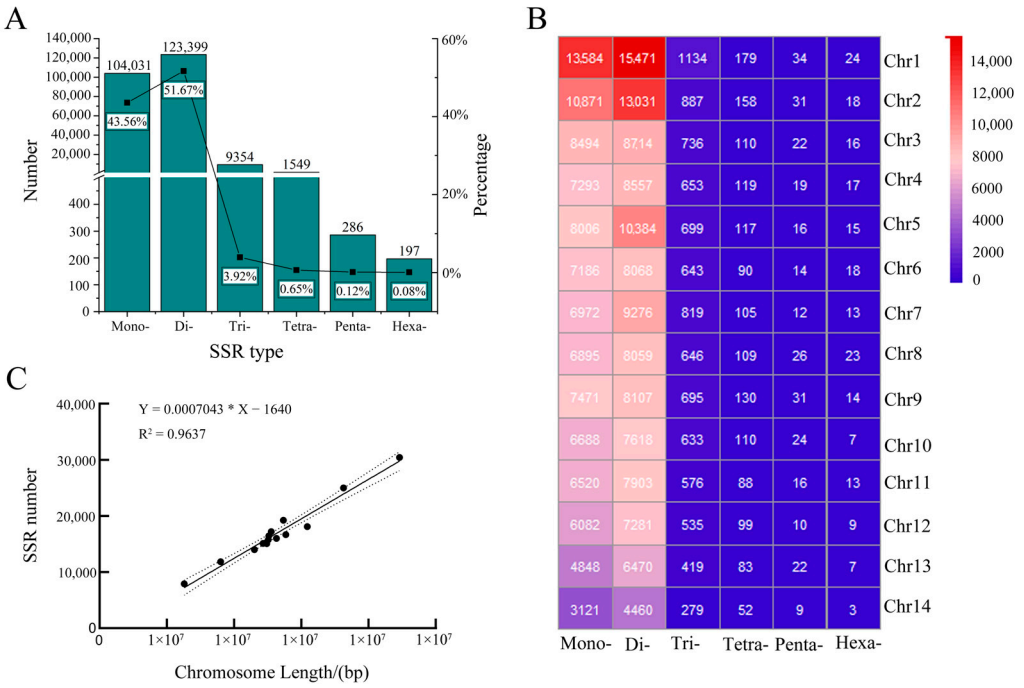

Figure 1. Number of SSRs identified on each chromosome in the *N. colorata* genome. Number and percentage of SSR markers ranging from mononucleotide to hexanucleotide repeats in the whole genome (**A**). SSR numbers ranging from mononucleotide to hexanucleotide repeats in each chromosome (**B**). Pearson correlation analysis between SSR number and each chromosome length (**C**).

3.2. Characterization of SSR Motifs in the N. colorata Genome

The analysis of the base composition for SSR motifs indicated that (i) A/T were dominant, accounting for 96.76% among the mononucleotide repeats; (ii) AT/TA were the most frequent (43.47%), followed by AG/CT (42.42%), AC/GT (14.04%) and CG/GC (0.08%) among dinucleotide repeats; (iii) AAG/CTT (40.53%) were the most abundant, followed by AAT/ATT (25.34%) and AGG/CCT (12.40%) among the trinucleotide repeats; and (iv) the most abundant repeats among the tetranucleotide, pentanucleotide and hexanucleotide repeats were AAAT/ATTT (42.39%), AAGGG/CCCTT (21.77%) and AACCCC/GGGGTT (71.86%), respectively (Figure 2).

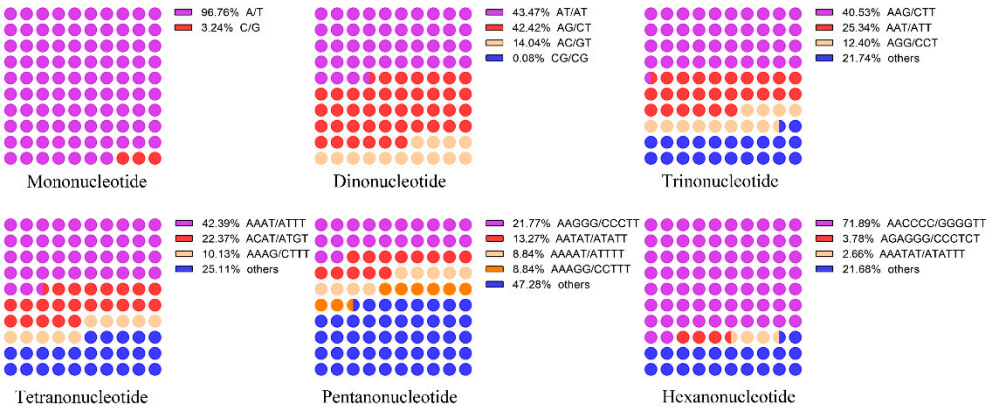

Figure 2. The percentage of mononucleotide to hexanucleotide SSR motifs in the *N. colorata* genome.

Regarding SSR repeat numbers, repeat times ranging from 5 to 30 times were the most common. Of these, 10 was the most common number of repetitions, accounting for 21.83% (52,145/238,816) of the total SSRs. Interestingly, mononucleotides with 10 repetitions; dinucleotides and hexanucleotides with 6 repetitions; and trinucleotide, tetranucleotide

and pentanucleotide with 5 repetitions were dominant among SSR repeat categories. In addition, the SSR number decreased with the increase in repeat times from mononucleotide to hexanucleotide repetitions (Figure 3).

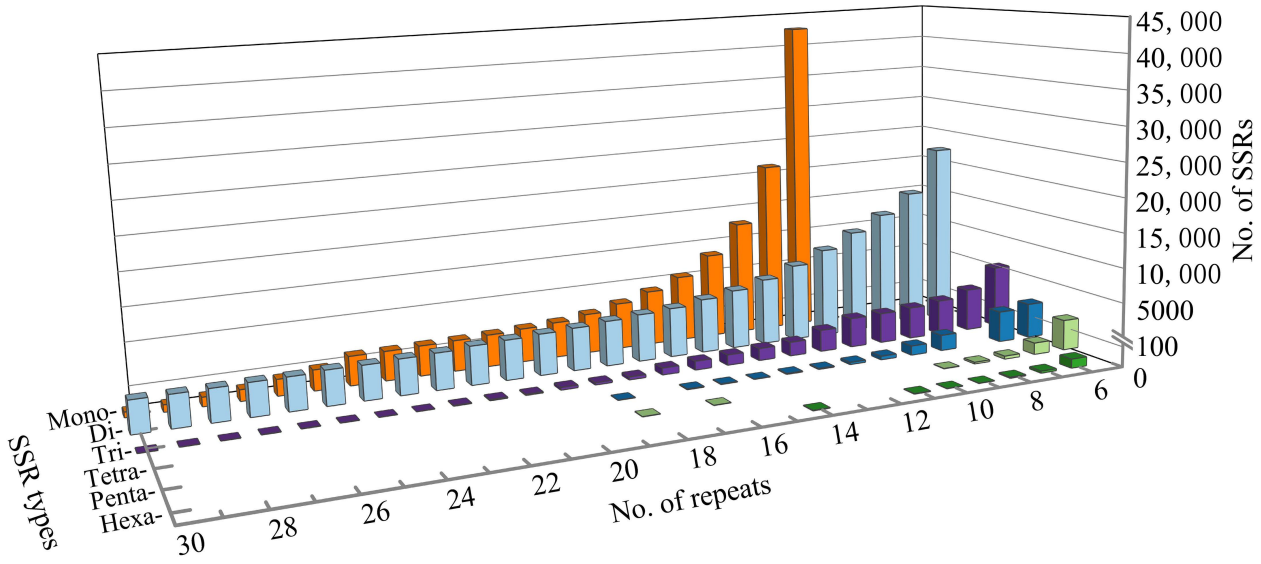

Figure 3. The number of repeat types with respect to the number of repeat motifs of SSRs in the *N. colorata* genome.

3.3. GO and KEGG Enrichment Analysis of CDSs Containing Microsatellite in the N. colorata Genome

A total of 82.20% (196,305/238,816) of SSRs were mainly found in the intergenic regions, followed by 15.76% (37,645/238,816) of SSRs that were found within the genes. Only 2.04% (4866/238,816) in the gene-intergenic regions were distributed in the *N. colorata* genome. The detailed information for this is shown in Table S2.

Among 2038 CDSs with SSRs, 939 CDSs were annotated and further subjected to a gene ontology (GO) analysis. A total of 793 GO functional terms were assigned, in which 183 GO terms were significantly overrepresented in the sets of genes with SSRs (p value < 0.05). In the three categories, protein binding, membranes and the regulation of transcription were the most abundant in the biological process, cellular component and molecular function ontology, respectively (Figure 4A). In addition, with respect to membrane, integral component of membrane, protein binding and regulation of transcription, DNA-templated dominated both in absolute number and rich factor. Meanwhile, the regulation of protein kinase activity and protein kinase regulator activity had the highest rich factor in the top 20 GO terms enrichment analysis results (Figure S1A).

A Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed that 191 CDSs containing microsatellites were annotated to 72 KEGG pathways, which were further categorized into four major categories (environmental information processing (EIP), genetic information processing (GIP), organismal systems (OS) and metabolism). In brief, pentose and glucuronate interconversions, thermogenesis, the PI3K-Akt signaling pathway and aminoacyl-tRNA biosynthesis were the most abundant in the metabolism, OS, EIP and GIP category, respectively (Figure 4B). In addition, the PI3K-Akt signaling pathway, the mTOR signaling pathway, thermogenesis, the relaxin signaling pathway and pentose and glucuronate interconversions were dominant both in absolute number and rich factor. Meanwhile, the lipoic acid metabolism pathway had the highest rich factor in the top 20 KEGG pathways enrichment analysis results (Figure S1B).

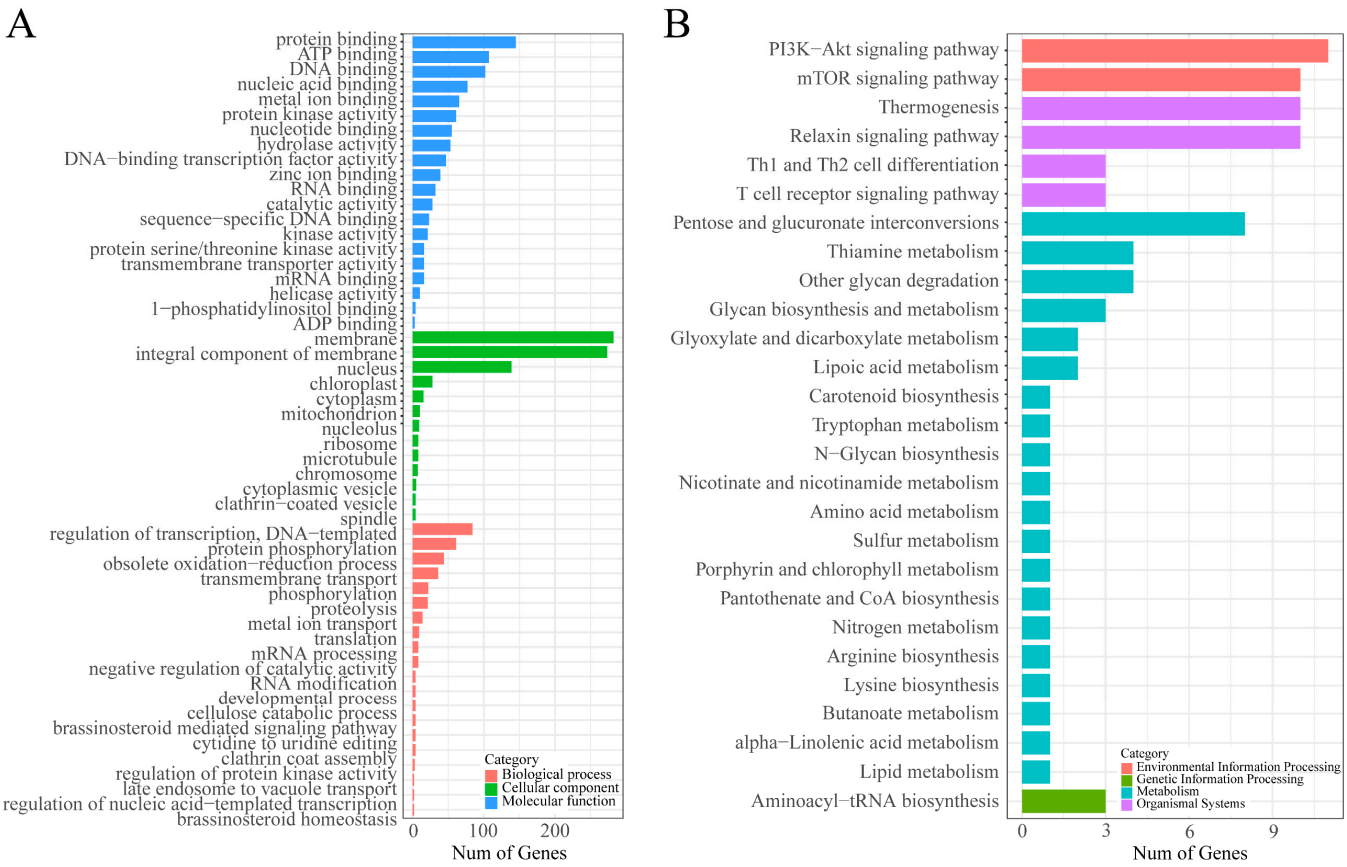

Figure 4. GO and KEGG enrichment analysis. GO classifications of CDSs with SSRs (**A**). KEGG pathway classifications of CDSs with SSRs (**B**).

3.4. Genome-Wide SSR Marker Development

Based on the start positions of the SSR markers, the reference genome physical map, which included 107,152 primer pairs, is described in Table S3, with results ranging from 540.68 per Mb on chromosome 1 to 198.66 per Mb on chromosome 10, with an average density of 318.67 per Mb (Figure 5). In addition, the distribution of each chromosome showed that these primer pairs were averagely distributed on each chromosome except chromosome 1. Interestingly, there were some regions with a higher density distribution of primers, such as 0–8 Mb, 14–21 Mb, 26–33 Mb and 36–41 Mb on chromosome 1.

3.5. Validation Analysis of SSR Primers

Among the 50 primer pairs, 13 primers were practically screened out and used for a polymorphic analysis of nine waterlily species. In total, 30 alleles were found (average 2.308 alleles/primer pair or locus). The *PIC* values of each SSR primer pair ranged from 0.290 to 0.624 (average of 0.475). Among them, Ny-5.2 and Ny-10.1 were the most informative SSR primer pairs, as they had the highest *PIC* values. Meanwhile, Ny-10.1 had the highest value of *Ho*, and Ny-5.2 had the highest values of *Ne*, *I* and *He*. The detailed information for this is shown in Table 1.

Figure 5. Overview of the high-density SSR primer pairs physical map in *N. colorata*. The bar represents the number of SSR primer pairs within a 1 Mb window.

Table 1. Characteristics of the 13 polymorphic microsatellite loci and primer sets.

Locus	Core Motif	Chr.	Primer Sequences (5′–3′)	*Na*	*Ne*	*I*	*Ho*	*He*	*Fis*	*PIC*
Ny-2.1	(T)13	2	F: AGAGCTGAGATTGGTTTGAAGC R: TCAGCGATTTCTTCTTGGGAT	2.000	1.800	0.637	0.000	0.471	1.000	0.444
Ny-2.2	(CT)10	2	F: TTTGTGGTGCGACTTCTTGC R: GTTAACAGCAGCCTTCACCG	3.000	1.409	0.557	0.333	0.307	−0.149	0.290
Ny-3.2	(A)12	3	F: CTATGGACAACACATGCCGC R: CACGAGCAACAAGACCAGTA	2.000	1.800	0.637	0.000	0.471	1.000	0.444
Ny-4.2	(GA)33	4	F: CACGGCGAGGGGACAATATA R: CCGCCAATTCCACCATTCAT	3.000	1.906	0.787	0.333	0.503	0.299	0.475
Ny-5.1	(A)13	5	F: AAAGAACTATGCGGACCTGC R: CAAGCTCAGGACATGGTTCG	2.000	2.000	0.693	0.111	0.529	0.778	0.500
Ny-5.2	(TC)20(TA)25	5	F: CCGCAGTTAGTGTCACATGG R: CGCGTTCTCCTTTGCCAATA	3.000	2.656	1.037	0.222	0.660	0.644	0.624
Ny-6.2	(T)10 . . . (AT)10	6	F: AATCAATGCTTCCATGGCCG R: TCATGTCCGGGATTCTAGGC	2.000	1.976	0.687	0.000	0.523	1.000	0.494
Ny-9.1	(TA)17(GA)14	9	F: ACCAAGGACTGCGAGTGTAT R: ATTTGAGTTGAGGGTTGCCG	2.000	1.670	0.591	0.556	0.425	−0.385	0.401
Ny-10.1	(AG)8 . . . (AG)10	10	F: CCCAGCATCGTAAATGACCG R: TTGGAGGAGGAGGAGATTGC	3.000	2.418	0.981	0.667	0.621	−0.137	0.586
Ny-11.1	(AG)6 . . . (T)11	11	F: AGCGTCACAACACTCCACTA R: AGGATTAGATGGGGCTCTGC	2.000	1.906	0.668	0.333	0.503	0.299	0.475
Ny-12.1	(TC)10(TA)13	12	F: AGGAGAAAACAGAGTGGGGC R: AGCATGCATGTATTCCCCAT	2.000	1.800	0.637	0.222	0.471	0.500	0.444
Ny-13.1	(CT)6(CA)9	13	F: CAGATGCAAGGATGGGAAGC R: GCAATGGGGATGATGAAGGC	2.000	2.000	0.693	0.556	0.529	−0.111	0.500
Ny-13.2	(TC)20(TA)12	13	F: GCCTACCCATGTCCTCTGAT R: CCCTGTTCTGTTTGTGTTGC	2.000	1.976	0.687	0.000	0.523	1.000	0.494
Mean	-	-	-	2.308	1.947	0.715	0.256	0.503	0.441	0.475

3.6. Cluster Analysis

The phylogenetic tree indicated nine waterlily cultivars that were divided into two clusters; Cluster I and Cluster II. *Nymphaea 'Black Beauty'*, *N. colorata* and *Nymphaea lotus*, traditionally classified as tropical waterlilies, were clustered into two subgenera (*Brachyceras* and *Lotos*) compared to *Nymphaea 'Islamorada'* and *Nymphaea 'Tina'*. In addition, *Nymphaea*

'Princess Elizabeth', *Nymphaea candida*, *Nymphaea 'Colorado'* and *Nymphaea Mexicana*, traditionally classified as hardy waterlilies, were divided into two clades (Figure 6).

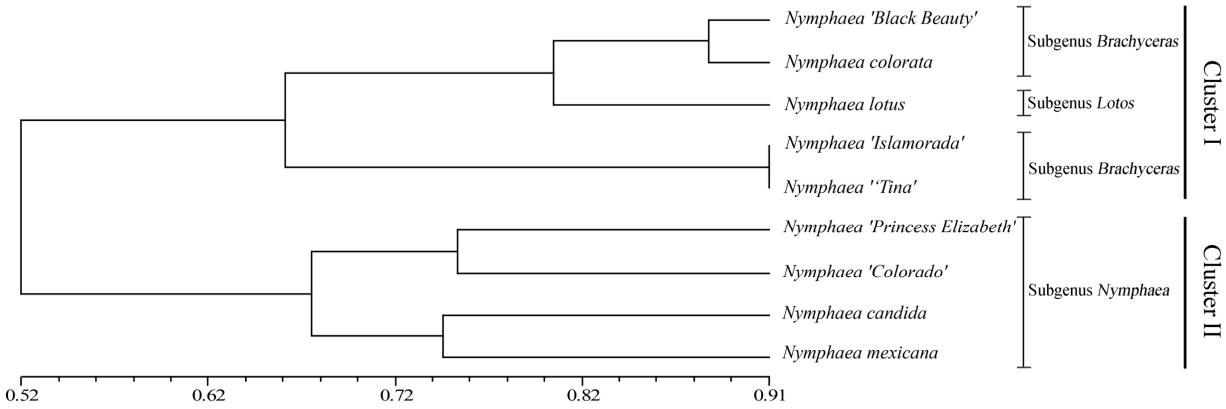

Figure 6. Cluster dendrogram of SSR markers of nine waterlily cultivars.

4. Discussion

Prior to this study, due to the lack of genomic information on waterlily, the mining and development of molecular markers were limited, and only a limited number of molecular markers have been used in waterlily [13,27,28]. In this study, we were the first to scan the genome-wide SSRs of *N. colorata*. A total of 238,816 SSRs were identified, presenting at a density of 583.75 SSRs per Mb. This density was extremely low compared with those reported in previous studies on the dicotyledon species *Prunus mume* (234.03 Mb, 794 SSRs/Mb) [29] and pomegranate (308.438 Mb, 1,230.6 SSRs/Mb) [30] but high in comparison to those of *N. nucifera* (804.648 Mb, 236.41 SSRs/Mb) [6] and *Malus domestica* (703.358 Mb, 40.8 SSRs/Mb) [31]. Our study supports the idea that SSR density is negatively correlated with genome size in plants [32].

Moreover, these motif types and their proportions in the *N. colorata* genome are in close agreement with the patterns observed in aquatic plants, such as Asian lotus (*N. nucifera*) and American lotus (*Nelumbo lutea*) [33], in which di- and tri-nucleotide repeat motifs were the most abundant. Moreover, the AT/TA and A/T are the most abundant types among the dinucleotide and mononucleotide repeat motifs, accounting for 43.47% and 96.76%, respectively. This result agrees with that of the dicot species (e.g., pomegranate, peanut and cucumber) [34–36]. In addition, the SSR number decreased with the increase in repeat times from mononucleotide to hexanucleotide repetitions based on a genome-wide scan of microsatellites in *N. colorata*. This finding presents similar characteristics to peanut (*Arachis hypogaea*) and asparagus (*Asparagus officinalis*) [35,37].

In previous studies, the distribution of SSRs varied in different regions across a genome and had a higher degree of abundance in noncoding than in coding regions in various animals and plants [38–40]. A similar feature was also observed in *N. colorata*. In our studies, 82.20% (196,305/238,816) of the SSRs were distributed in intergenic regions, and only 1.02% (2442/238,816) of SSRs were abundant in exon and CDS regions. Furthermore, we calculated that SSR repeat types ranged from mononucleotide to hexanucleotide. Interestingly, trinucleotides were much more abundant than other nucleotides and accounted for 63.84% (1559/2442) (Table S4). This finding conveyed to us that trinucleotide SSRs in CDS regions were significantly abundant in *N. colorata*, which did not cause frameshifts and were not notably influenced by coding status [41]. This phenomenon is similar to king cobra (*Ophiophagus hannah*) [42].

A GO analysis revealed that CDSs containing SSRs were mainly associated with signal identification, such as protein kinase activity (e.g., GO: 0045859, the regulation of protein kinase activity; GO: 0019887, protein kinase regulator activity), which might play a major role in response to a variety of stimuli such as phytohormone treatment and temperature stress [43]. A KEGG pathway analysis showed that CDSs within SSRs play a major role

in metabolism (ko00785, lipoic acid metabolism), which might strengthen the antioxidant network of the cells [44]. In the future, it will be necessary to explore SSR diversity in CDS regions when comparing the special function (e.g., freezing resistance and disease resistance) of different waterlily species originating from different geographical habitats.

In previous studies, SSR markers were effectively applied to the diversity of local and wild lotus varieties [45–47]. To analyze the genetic relationships between nine waterlily cultivars, a total of 13 pairs of primers targeting polymorphic SSR markers were screened out. In total, 30 alleles (average 2.308 alleles/primer pair or locus) were found. The UPGMA dendrograms show that the nine waterlily cultivars were divided into two clusters, which is consistent with the eco-physiological classification (tropical waterlily and hardy waterlily) [3]. This finding indicated that 13 pairs of primers targeting polymorphic SSRs are useful for identifying different waterlily species due to the transferability of SSR markers. Many studies have demonstrated the utility of the transferability of SSRs for the analysis of intra- and inter-specific genetic diversity and species identification [48–50]. These SSR markers in waterlily might also be applicable to cross-genera genotyping or to genotyping in other closely related plant species. In addition, the values for *PIC*, *Ho*, *He*, *Ne* and *I* indicated that Ny-5.2 and Ny-10.1 are highly polymorphic and could potentially be used in genetic diversity. In the future, the most effective SSR primers could be developed as more genomes of waterlily species are sequenced.

5. Conclusions

Here, we report the first comprehensive study of SSR density, occurrence frequency and GO and KEGG enrichment analysis based on the *N. colorata* genome. A total of 238,816 SSRs were identified in 14 *N. colorata* chromosomes with an average density of 662.60 SSRs per Mb. The dinucleotide was the most common type, and AT-rich repeats prevail in the *N. colorata* genome. In GO and KEGG enrichment analysis of CDSs containing microsatellites, signal recognition (e.g., GO: 0045859 and GO: 0019887) and metabolism (ko00785, lipoic acid metabolism) were significantly enriched, respectively. In addition, the large amount of SSR marks (n = 107,152) enriches molecular markers in waterlily. Among these, the 13 novel candidate SSR markers used in this study will be useful for genetic diversity and phylogenetic analysis to differentiate between waterlily species, hybrids and even lineages.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13101782/s1. Figure S1: The top 20 significantly enriched analysis of GO terms and KEGG pathways. GO term enrichment of CDSs with SSR(A). KEGG pathway enrichment of CDSs with SSR(B); Table S1: The localities and eco-physiological characteristics of nine waterlily cultivars; Table S2: The number of SSRs in different regions of the *N. colorata* genome; Table S3: List of genome-wide SSR marks in *N. colorata*; Table S4: The number and percentage of mononucleotide to hexanucleotide SSR motifs in CDS regions of the *N. colorata* genome.

Author Contributions: Conceptualization, investigation, methodology, writing—original draft preparation, X.H.; writing—review and editing, project administration, funding acquisition, M.Y.; investigation, software, formal analysis, J.G. and J.L.; investigation, formal analysis, supervision, G.C., Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by New Variety Cultivation Project of Shihezi University (YZZX202105) and High-Level Talent Initiative Foundation of Shihezi University (RCZK2018C04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data presented in this study are available in the article itself or in the provided Supplementary Materials files.

Acknowledgments: We are grateful to Professor Yuanzhi Wang of Shihezi University for assistance with the experiments and valuable discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Slocum, P.D. *Waterlilies and Lotuses: Species, Cultivars, and New Hybrids*; Timber Press: Portland, OR, USA, 2005; pp. 1–7. ISBN 9780881926842.
2. Yoo, M.J.; Bell, C.D.; Soltis, P.S.; Soltis, D.E. Divergence times and historical biogeography of Nymphaeales. *Syst. Bot.* **2005**, *30*, 693–704. [CrossRef]
3. Huang, G.Z.; Deng, H.Q.; Li, Z.X.; Li, G. *The Waterlilies*; Chinese Forestry Publishing House: Beijing, China, 2008; pp. 1–15. ISBN 9787503853326.
4. Tang, H.B.; Zhang, L.S.; Chen, F.; Zhang, X.T.; Chen, F.; Ma, H.; Peer, Y.V. *Nymphaea colorata* (Blue-Petal Water Lily). *Trends Genet.* **2020**, *36*, 718–719. [CrossRef] [PubMed]
5. Koga, K.; Kadono, K.; Setoguchi, H. The genetic structure of populations of the vulnerable aquatic macrophyte *Ranunculus nipponicus* (Ranunculaceae). *J. Plant Res.* **2007**, *120*, 167–174. [CrossRef]
6. Hu, J.H.; Pan, L.; Liu, H.G.; Wang, S.Z.; Wu, Z.H.; Ke, W.D.; Ding, Y. Comparative analysis of genetic diversity in sacred lotus (*Nelumbo nucifera* Gaertn.) using AFLP and SSR markers. *Mol. Biol. Rep.* **2012**, *39*, 3637–3647. [CrossRef]
7. Kumar, H.; Priya, P.; Singh, N.; Kumar, M.; Choudhary, B.K.; Kumar, L.; Singh, I.S.; Kumar, N. RAPD and ISSR marker-based comparative evaluation of genetic diversity among Indian germplasms of *Euryale ferox*: An aquatic food plant. *Appl. Biochem. Biotechnol.* **2016**, *180*, 1345–1360. [CrossRef]
8. Zhu, J.; Zhang, J.; Jiang, M.; Wang, W.; Jiang, J.; Li, Y.; Yang, L.; Zhou, X. Development of genome-wide SSR markers in rapeseed by next generation sequencing. *Gene* **2021**, *798*, 145798. [CrossRef] [PubMed]
9. Patil, P.G.; Singh, N.V.; Parashuram, S.; Bohra, A.; Mundewadikar, D.M.; Sangnure, V.R.; Babu, K.D.; Sharma, J. Genome wide identification, characterization and validation of novel miRNA-based SSR markers in pomegranate (*Punica granatum* L.). *Physiol. Mol. Biol. Plants* **2020**, *26*, 683–696. [CrossRef] [PubMed]
10. Xu, Y.; Xing, M.; Song, L.; Yan, J.; Lu, W.; Zeng, A. Genome-wide analysis of simple sequence repeats in Cabbage (*Brassica oleracea* L.). *Front. Plant. Sci.* **2021**, *12*, 726084. [CrossRef]
11. Hou, S.; Ren, X.; Yang, Y.; Wang, D.; Du, W.; Wang, X.; Li, H.; Han, Y.; Liu, L.; Sun, Z. Genome-wide development of polymorphic microsatellite markers and association analysis of major agronomic traits in core germplasm resources of *Tartary Buckwheat*. *Front. Plant Sci.* **2022**, *13*, 819008. [CrossRef]
12. Martina, M.; Acquadro, A.; Barchi, L.; Gulino, D.; Brusco, F.; Rabaglio, M.; Portis, F.; Portis, E.; Lanteri, S. Genome-wide survey and development of the first microsatellite markers database (AnCorDB) in *Anemone coronaria* L. *Int. J. Mol. Sci.* **2022**, *23*, 3126. [CrossRef]
13. Chaveerach, A.; Tanee, T.; Sudmoon, R. Molecular identification and barcodes for the genus Nymphaea. *Acta Biol. Hung.* **2011**, *62*, 328–340. [CrossRef] [PubMed]
14. Jeremy, D.; Suman, K.; Sayawada, R.R.; Pramod, T. Sequence characteristics and phylogenetic implications of the nrDNA internal transcribed spacers (ITS) in the genus Nymphaea with focus on some Indian representatives. *Plant Syst. Evol.* **2012**, *298*, 93–108. [CrossRef]
15. Parveen, S.; Singh, N.; Adit, A.; Kumaria, S.; Tandon, R.; Agarwal, M.; Jagannath, A.; Goel, S. Contrasting reproductive strategies of two *Nymphaea* species affect existing natural genetic diversity as assessed by microsatellite markers: Implications for conservation and wetlands restoration. *Front. Plant Sci.* **2022**, *13*, 773572. [CrossRef] [PubMed]
16. Zhang, L.S.; Chen, F.; Zhang, X.T.; Li, Z.; Tang, H.B. The water lily genome and the early evolution of flowering plants. *Nature* **2020**, *577*, 79–84. [CrossRef]
17. Zuo, L.H.; Zhang, S.; Zhang, J.; Liu, Y.; Yu, X.; Yang, M.; Wang, J. Primer development and functional classification of EST-SSR markers in Ulmus species. *Tree Genet. Genomes* **2020**, *16*, 74–87. [CrossRef]
18. Li, J.M.; Li, S.Q.; Kong, L.J.; Wang, L.H.; Wei, A.Z.; Liu, Y.L. Genome survey of *Zanthoxylum bungeanum* and development of genomic-SSR markers in congeneric species. *Biosci. Rep.* **2020**, *40*, BSR20201101. [CrossRef] [PubMed]
19. Rozen, S.; Skaletsky, H. Primer 3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **2000**, *132*, 365–386. [CrossRef]
20. Conesa, A.; Gotz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef]
21. Ye, J.; Fang, L.; Zheng, H.K.; Zhang, Y.; Chen, J.; Zhang, Z.J.; Wang, J.; Li, S.T.; Li, R.Q.; Bolund, L.; et al. WEGO: A web tool for plotting GO annotations. *Nuclei. Acids Res.* **2006**, *34*, 293–297. [CrossRef]
22. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef]
23. Altschul, S.F.; Madden, T.L.; Schffer, A.A.; Zhang, J.; Zhang, Z.; Webb, M.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
24. Pan, G.; Chen, A.; Li, J.J.; Huang, S.Q.; Tang, H.J.; Chang, L.; Zhao, L.N.; Li, D.F. Genome-wide development of simple sequence repeats database for flax (*Linum usitatissimum* L.) and its use for genetic diversity assessment. *Genet. Resour. Crop Evol.* **2020**, *67*, 865–874. [CrossRef]
25. Yeh, F.C.; Yang, R.C.; Boyle, T.B.J.; Yeh, Z.H.; Mao, J.X. *POPGENE Version 1.31: Microsoft Window-Based Freeware for Population Genetic Analysis*; University of Alberta: Edmonton, AB, USA, 1997.

26. Han, B.; Wang, C.B.; Tang, Z.H.; Ren, Y.K.; Li, Y.; Zhang, D.Y.; Dong, Y.H.; Zhao, X.H. Genome-wide analysis of microsatellite markers based on sequenced database in Chinese spring wheat (*Triticum aestivum* L.). *PLoS ONE* **2015**, *10*, e0141540. [CrossRef] [PubMed]
27. Dkhar, J.; Kumaria, S.; Rao, S.R.; Tandon, P. Molecular phylogenetics and taxonomic reassessment of four Indian representatives of the genus Nymphaea. *Aquat. Bot.* **2010**, *93*, 135–139. [CrossRef]
28. Poczai, P.; Mátyás, K.K.; Szabó, I.; Varga, I.; Hyvönen, J.; Cernák, I.; Gorji, A.M.; Decsi, K.; Taller, J. Genetic variability of *thermal Nymphaea* (Nymphaeaceae) populations based on ISSR markers: Implications on relationships, hybridization, and conservation. *Plant Mol. Biol. Rep.* **2011**, *29*, 906–918. [CrossRef]
29. Sun, L.D.; Yang, W.R.; Zhang, Q.X.; Cheng, T.R.; Pan, H.T.; Xu, Z.D.; Zhang, J.; Chen, C.G. Genome-wide characterization and linkage mapping of simple sequence repeats in Mei (*Prunus mume* Sieb. et Zucc.). *PLoS ONE* **2013**, *8*, e59562. [CrossRef] [PubMed]
30. Patil, P.G.; Singh, N.V.; Sharma, J.; Bohra, A.; Raghavendra, K.P.; Mane, R.; Mundewadikar, D.M.; Babu, K.D.; Sharma, J. Comprehensive characterization and validation of chromosome-specific highly polymorphic SSR markers from pomegranate (*Punica granatum* L.) cv. Tunisia Genome. Front. *Plant Sci.* **2021**, *12*, 337. [CrossRef]
31. Zhang, Q.; Ma, B.; Hui, L.; Chang, Y.S.; Han, Y.Y.; Li, J.; Wei, G.C.; Zhao, S.; Khan, M.A.; Zhou, Y.; et al. Identification, characterization, and utilization of genome-wide simple sequence repeats to identify a QTL for acidity in apple. *BMC Genom.* **2012**, *1*, 537. [CrossRef]
32. Morgante, M.; Hanafey, M.; Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **2002**, *30*, 194–200. [CrossRef]
33. Yang, M.; Han, Y.; VanBuren, R.; Ming, R.; Xu, L.; Han, Y.; Liu, Y. Genetic linkage maps for Asian and American lotus constructed using novel SSR markers derived from the genome of sequenced cultivar. *BMC Genom.* **2012**, *13*, 653. [CrossRef]
34. Cavagnaro, P.F.; Senalik, D.A.; Yang, L.M.; Simon, P.W.; Harkins, T.T.; Kodira, C.D.; Huang, S.W.; Weng, Y.Q. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genom.* **2010**, *11*, 569. [CrossRef] [PubMed]
35. Lu, Q.; Hong, Y.B.; Li, S.X.; Liu, H.; Li, H.F.; Zhang, J.N.; Lan, H.F.; Liu, H.Y.; Liu, X.Y.; Wen, S.J.; et al. Genome-wide identification of microsatellite markers from cultivated peanut (*Arachis hypogaea* L.). *BMC Genom.* **2019**, *20*, 799. [CrossRef] [PubMed]
36. Liu, C.Y.; Li, J.Y.; Qin, G.H. Genome-wide distribution of simple sequence repeats in pomegranate and their application to the analysis of genetic diversity. *Tree Genet. Genomes* **2020**, *16*, 36. [CrossRef]
37. Li, S.F.; Zhang, G.J.; Li, X.; Wang, L.J.; Yuan, J.H.; Deng, C.L.; Gao, W.J. Genome-wide identification and validation of simple sequence repeats (SSRs) from *Asparagus officinalis*. *Mol. Cell. Probes* **2016**, *30*, 153–160. [CrossRef] [PubMed]
38. Hancock, J.M. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* **1995**, *41*, 1038–1047. [CrossRef]
39. Mokhtar, M.M.; Adawy, S.S.; El-Assal, S.D.; Hussein, E.H. Genic and intergenic SSR database generation, SNPs determination and pathway annotations, in date palm (*Phoenix dactylifera* L.). *PLoS ONE* **2016**, *11*, e0159268. [CrossRef]
40. Wang, X.T.; Zhang, Y.J.; Qiao, L.; Chen, B. Comparative analyses of simple sequence repeats (SSRs) in 23 mosquito species genomes: Identification, characterization and distribution (Diptera: Culicidae). *Insect Sci.* **2019**, *26*, 607–619. [CrossRef]
41. Trivedi, S.; Wills, C.; Metzgar, D. Analysis of simple sequence repeats in mammalian cell cycle genes. *Recent. Adv. DNA Gene Seq.* **2014**, *8*, 20–29. [CrossRef]
42. Liu, W.C.; Xu, Y.T.; Li, Z.K.; Fan, J.; Yang, Y. Genome-wide mining of microsatellites in king cobra (*Ophiophagus hannah*) and cross-species development of tetranucleotide SSR markers in Chinese cobra (*Naja atra*). *Mol. Biol. Rep.* **2019**, *46*, 6087–6098. [CrossRef]
43. Stone, J. Plant protein kinase families and signal transduction. *Plant Physiol.* **1995**, *108*, 451–457. [CrossRef]
44. Navari-Izzo, F.; Quartacci, M.F.; Sgherri, C. Lipoic acid: A unique antioxidant in the detoxification of activated oxygen species. *Plant Physiol. Biochem.* **2002**, *40*, 463–470. [CrossRef]
45. Zheng, X.F.; You, Y.N.; Diao, Y.; Zheng, X.W.; Xie, K.Q.; Zhou, M.Q.; Hu, Z.L.; Wang, Y.W. Development and characterization of genic-SSR markers from different Asia lotus (*Nelumbo nucifera*) types by RNA-seq. *Genet. Mol. Res.* **2015**, *14*, 11171–11184. [CrossRef]
46. Fu, Y.R.; Liu, F.L.; Li, S.; Tian, D.K.; Dong, L.; Chen, Y.C.; Su, Y. Genetic diversity of wild Asian lotus (*Nelumbo nucifera*) from northern China. *Hortic. Plant J.* **2021**, *7*, 13. [CrossRef]
47. Nurainee, S.; Seiji, T.; Nakao, K.; Samak, K. Molecular phylogeny and postharvest morphology of petals in two major *Nelumbo nucifera* cultivars in Thailand. *Agric. Nat. Resour.* **2018**, *52*, 45–52. [CrossRef]
48. Rai, M.K.; Phulwaria, M.; Shekhawat, N.S. Transferability of simple sequence repeat (SSR) markers developed in guava (*Psidium guajava* L.) to four Myrtaceae species. *Mol. Biol. Rep.* **2013**, *40*, 5067–5071. [CrossRef] [PubMed]
49. Endo, C.; Yamamoto, N.; Kobayashi, M.; Nakamura, Y.; Yokoyama, K.; Kurusu, T.; Yano, K.; Tada, Y. Development of simple sequence repeat markers in the halophytic turf grass *Sporobolus virginicus* and transferable genotyping across multiple grass genera/species/genotypes. *Euphytica* **2017**, *213*, 56. [CrossRef]
50. Tuler, A.C.; Carrijo, T.T.; Nóia, L.R.; Ferreira, A.; Peixoto, A.L.; da Silva Ferreira, M.F. SSR markers: A tool for species identification in *Psidium* (Myrtaceae). *Mol. Biol. Rep.* **2015**, *42*, 1501–1513. [CrossRef]