horticulturae

MDPI

Article

Lily Database: A Comprehensive Genomic Resource for the Liliaceae Family

Manosh Kumar Biswas [1,2,*], Sathishkumar Natarajan [2,3], Dhiman Biswas [4], Jewel Howlader [2], Jong-In Park [2] and Ill-Sup Nou [2]

[1] Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK
[2] Department of Horticulture, Sunchon National University, 255, Jungang-ro, Suncheon 57922, Republic of Korea; sathisbioinfo@gmail.com (S.N.); jewel.howlader81@gmail.com (J.H.); jipark@scnu.ac.kr (J.-I.P.); nis@sunchon.ac.kr (I.-S.N.)
[3] 3BIGS Co., Ltd., B-831, Geumgang Penterium IX Tower, Hwaseong 18469, Republic of Korea
[4] Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata 700064, West Bengal, India; bdcse12@gmail.com
* Correspondence: manosh24@yahoo.com

Abstract: The Lily database is an online genomic resource which is composed of a Korean Lily germplasm collection, transcriptome sequences, molecular markers, transcription factors (TFs) and DEGs (Differentially Expressed Genes) data. A total of ~0.23 gb of RNA-sequencing data were used for gene identification, marker development and gene expression analysis. As a result, 103,929 genomic, 47,863 EST-SSR, 20,929 SNP and 1213 COS-marker were developed. A total of 1327 TF genes were identified and characterized. This is the first unique, user-friendly, genomic resource database for Lilium species. It is a relational database based on a 'three-tier architecture' that catalogs all the information in a MySQL table and a user-friendly query interface and data visualization page developed using JavaScript, PHP and HTML code. The search parameters are highly flexible; users can retrieve data by using either single or multiple search parameters. Data present in this database can be used for germplasm characterization, gene discovery, population structure analysis, QTL mapping, and accelerating lily variety improvements.

Keywords: genetic diversity; germplasm; molecular markers; transcription factors; DEGs; genes

Citation: Biswas, M.K.; Natarajan, S.; Biswas, D.; Howlader, J.; Park, J.-I.; Nou, I.-S. Lily Database: A Comprehensive Genomic Resource for the Liliaceae Family. *Horticulturae* **2024**, *10*, 23. https://doi.org/10.3390/horticulturae10010023

Academic Editor: Jiafu Jiang

Received: 7 November 2023
Revised: 12 December 2023
Accepted: 14 December 2023
Published: 25 December 2023

1. Introduction

The Liliaceae family is one of the most economically and culturally important plants worldwide [1]. According to Chase et al. [2], there are approximately 70 genera and over 3000 species in this family. Among them, many are grown as ornamental plants, while others have medicinal, culinary, and industrial uses. Lily (*Lilium* spp.) is the most popular species of this family, and they are commercially cultivated in The Netherlands, France, Chile, USA, Japan, New Zealand, and China for cut flowers [1,3,4]. Germplasm collection, characterization and maintenance are the key steps prior to the cultivar development of any plant species. Lily germplasm is collected and maintained by many organizations around the globe. Notable among them are KEW, USDA, The Kunming Institute of Botany in China, The National Institute of Floricultural Science in Korea, and The Instituto de Botánica Darwinion in Argentina [5]. Genomic and genetic resources play a crucial role in crop development. For the Lilium species, complete chloroplast genome sequences [6–9], RNA sequences [10,11] and molecular markers [1,3,12] are available as genetics and genomic resources.

Freely accessible genomic resource databases have been a driving force in advancing plant research over the past decade. There are many publicly available genomic resource databases for various plant species. For instance, notable genomic resource databases include TAIR (https://www.arabidopsis.org/); Banana Genome Hub (https://banana-genome-hub.southgreen.fr/) [13]; Sol Genomics Network (https://solgenomics.

net/) [14]; Ensete knowledge base (http://www.genomicsres.org/enknbase/); BRAD (http://brassicadb.cn/#/) [15]; cottonFGD (https://cottonfgd.net/); GDR (https://www.rosaceae.org/) [16]; and citrus genome database (https://www.citrusgenomedb.org/) [17]. These databases offer comprehensive and diverse datasets, including whole genome sequences, transcriptomes, and metabolomes which can accessed and utilized by researchers to address biological questions. However, such databases are currently lacking for the Lilium species.

The development of genomic resource databases provides a centralized and comprehensive collection of genetic information for plant species that typically contain DNA sequences, RNA sequences, gene structures, functional annotations, molecular markers, phenotypic data, germplasm information, and more. These data are valuable for the research community and accelerate research activity in several ways. Firstly, the data are freely accessible and ready to use, which reduces the cost, time, and labor required for data production. Secondly, the databases offer a diverse set and large amounts of data, which can be used to test new methods for analyzing and interpreting genomic information. This can ultimately lead to new insights into the genetic basis of complex traits and diseases. Freely accessible RNA-seq databases are particularly powerful for comparative genomics, gene identification, gene expression study, alternate splicing event discovery, molecular marker development, and phylogenomic and GWA studies.

Molecular markers have played a potential role in lily breeding compared to traditional methods. Various DNA markers, such as RAPD (Random Amplified Polymorphic DNA), ISSR (Inter-Simple Sequence Repeats), AFLP (Amplified Fragment Length Polymorphism), DArT (Diversity Arrays Technology), SSR (Simple Sequence Repeats), and SNP (Single-Nucleotide Polymorphisms), have been developed for genetic diversity analysis, germplasm characterization, hybrid identification, mutant detection, and genetic mapping in lilies [18–21]. However, the applicability of RAPD and ISSR markers is limited due to their low reproducibility and dominance issues. RAPD markers were employed as linked markers for *Fusarium* resistance in 150 Asiatic hybrid individuals. Nevertheless, only three RAPD markers were polymorphic, and they accounted for 24% of the resistance, highlighting constraints in lily breeding. Varshney et al. [22] failed to detect variation in micro-propagated Lilium species using RAPD markers. In contrast, ISSR markers were suggested by Xi et al. [23] as potentially useful for identifying *L. longiflorum* mutants instead of RAPD. However, Yin et al. [24] were unable to identify mutants in the oriental hybrid 'Siberia' using ISSR markers and found a very low-frequency polymorphism using AFLP markers among regenerated 'Siberia.' Therefore, to enhance Molecular Marker-Assisted Breeding (MMAB), there is a need for user-friendly, cost-effective, and transferable molecular markers for lily species.

This study was conducted to develop genomic resources for the *Lilium* species. which is a comprehensive genomic resource for the Liliaceae family, providing access to a variety of molecular markers, transcription factor genes, gene expression data, and phenotypic data. These resources are expected to play a crucial role in advancing our understanding of the genomics of *Lilium* species and the improvement of this species.

2. Materials and Methods

2.1. Data Sources and Data Processing

In this study, one genome assembly datum (*Lilium candidum*; GCA_031763035.1) and seven sets of RNA sequence data were used. Among them, 4 sets of RNA sequences and the genome assembly data were acquired from the NCBI. The remaining three sets were generated from RNA-sequencing libraries prepared from three genotypes of Lilium species exposed to various treatments (as listed in Table S1). Leaf samples from greenhouse-grown plants were collected from each treatment and immediately transferred to liquid nitrogen, where they were stored at −80 °C until RNA extractions. RNA was extracted using the RNeasy mini kit (Qiagen, Hilden, Germany), following the manufacturer's guidelines. The RNA library was prepared according to Illumina's guidelines (San Diego, CA, USA), and

the library was sequenced using the Illumina Hiseq 2000 platform. The quality of the raw reads was assessed using the FastQC tool [25], and de novo assembly was performed using Trinity [26] and RSEM tools [27]. Afterward, all transcriptomes were pooled based on the treatment and species name (sample) and assembled into non-redundant lily unigenes using CAP3 [28]. Furthermore, to access publicly available nucleotide sequences of Lilium species, we also downloaded data from the NCBI database (dated 1 November 2023). All the sequences were then categorized into four groups based on the respective Lilium species: Longiflorum (L), Asiatic (A), Oriental (O), and hybrid (H) species sequences. The list of species summary is presented in Table 1. Phenotypic data, including morphological characters, taxonomic information, and flower images, were collected from 141 lily cultivars at six different research institutes in South Korea.

Table 1. Lily species sequence data summary.

Species	No of Sequences	Total No of Bases	GC Count	GC Content %	Data Source	Lily Group
L. formosanum	1339	397,156	181,219	45.63%	NCBI	Asiatic
L. longiflorum	1336	920,102	430,617	46.80%	NCBI	Longiflorum
L. longiflorum Easter	179,988	113,297,779	49,127,334	43.36%	SNU	Longiflorum
L. longiflorum White	85,647	58,051,294	26,248,028	45.22%	SNU	Longiflorum
L. regale	1171	581,552	271,740	46.73%	NCBI	Oriental
Lily Hybrid	953	681,293	327,827	48.12%	NCBI	Hybrid
L. formolongi Sinnapal	90,115	60,473,109	27,533,192	45.53%	SNU	Hybrid
L. candidum	458,622	345,516,278	109,031,745	31.56%	NCBI	Oriental

Note: NCBI (National Center for Biotechnology Information) and SNU (Suncheon National University) represent different data origins.

2.2. Marker Development

In this study, we conducted microsatellite (SSR) mining and primer design using the LSAT pipeline. We identified microsatellite repeat motifs in the Lily transcriptome sequences and genome assembly, with a minimum of 10 repeat units for mononucleotides, 5 for dinucleotides, and 4 for other repeat types. After identifying the SSRs, we extracted flanking regions around them using custom Perl scripts. Primer design was accomplished using Primer3 executables with default parameters: a melting temperature range of 55–56 °C, GC content between 40 and 60%, primer size ranging from 18 to 25 nucleotides, and a desired product length of 150–280 base pairs. All SSR markers were then in silico-characterized and experimentally validated. Further details can be found in our previous study published by Biswas et al. [29].

SNP discovery and characterization were conducted using seven lily datasets, as detailed in Table 1. Initially, all data underwent clustering with the cd-hit tool employing a 95% identity threshold and a 5% length difference parameter. Subsequently, clusters containing a minimum of four representative sequences were selected, and the longest representative sequences from these chosen clusters were extracted using a combination of Bash and Perl scripts. These extracted representative sequences served as the reference sequences for SNP mining. To facilitate SNP discovery, the reference sequences were indexed using Bowtie2 with default parameters. Subsequently, all seven datasets were aligned to these reference sequences using Bowtie2 with its default settings. SNP calling was carried out using samtools, following established protocols. The identified SNPs were subjected to comprehensive characterization, considering various attributes such as the base mutation type, allele types, and their distribution across lily species. Additionally, the heterozygosity and homozygosity of these SNPs were estimated as part of the characterization process.

To identify COS loci, the transcriptome data sets of *Lilium formolongi* cv. "Sinnapal", *Lilium longiflorum* cv. "Napal", and *Lilium longiflorum* "Easter lily" were BLASTed against

each other in a "round-robin" fashion, and reciprocal best BLAST hits were retained. The following blast parameters were used: e-10, Hit-3. Potential COS loci were identified where all comparisons successfully produced a reciprocal best BLAST hit with identity at 90% and query coverage at 85%. The COS sequences were then used to develop COS markers. Primer pairs for COS sequences were designed using primer3 software with default parameters.

2.3. TF Gene Analysis

All the data sets (sequences) were searched against the Plant transcription factor Database version 5.0 [30] using an E-value cut-off of e-10. A cut-off value of 50% for query coverage and 80% for identity was set to filter TF gene-encoded transcripts. After that, putative TF-encoded lily RNA sequences were analyzed using the online tool iTAK [31] (Plant transcription factor & Protein Kinase Identifier and Classifier, http://itak.feilab.net/cgi-bin/itak/online_itak.cgi).

2.4. Expression Analysis

To ensure quality control, raw reads underwent trimming to remove adapters and eliminate low-quality reads with a Q value of less than 20, utilizing the command-line tools 'trimmomatic' [32] and fastQC [25]. Subsequently, the high-quality reads (transcriptome) were assembled using the de novo approach with the Trinity program v3 [26], and their transcript abundances were quantified using RSEM [27]. Since lily lacks a reference genome, RSEM v06 was employed for quantification purposes. The transcript expression levels were quantified in terms of FPKM (fragments per kilobase of transcript per million mapped reads) values, ranging from greater than 0 to over 104, along with the estimation of fold change (FC). The differential expressed genes (DEGs) were analyzed using the DESeq2 tool [33]. Transcripts meeting the criteria of FC $\geq$ 2 and FC $\leq$ –2 were considered significantly upregulated and downregulated DEGs (Differentially Expressed Genes), respectively, in at least one of the total comparisons. This rigorous analysis allowed for the identification of key transcripts that showed substantial changes in expression, shedding light on the crucial regulatory mechanisms involved in the studied lily species.

2.5. Database Architecture and Web Interface Design

The Three-Level Schema Architecture has been employed to develop the lily database. This architecture is designed to divide the database into three distinct tiers: the front-end, middle, and back-end. The front-end tier is responsible for the client-side interface of the application and is mostly developed using web technologies such as HTML and JavaScript. Its main purpose is to provide users with a visually attractive and user-friendly interface for interacting with the database. The middle tier, on the other hand, is developed using the PHP language. Its main function is to act as a bridge between the front-end tier and the back-end tier. This tier receives requests from the client side, processes them, and then communicates with the database. The back-end tier is where the database itself resides. It is responsible for storing, managing, and retrieving data from the database. We use the database management system MySQL to store all the lily-db data.

3. Result and Discussion

3.1. Content of the Lily-Db

To better organize all the data in Lily-db, we have developed user-friendly interfaces and easy-to-use tools. We have categorized all the data into four main groups based on their nature and utility. These categories can be accessed from the main menu: Genotype, Expression, Marker, and TF (transcription factor). The marker menu also contains sub-menus based on marker types, such as SSR, SNP, and COS-marker, which lead users to specific search pages.

Additionally, we have created Home, Download, and About Us pages. On the Home page, we provide a brief introduction to this website and the database. The Download page

allows users to download bulk data, and the About Us page contains information about our research group and credits for the individuals involved in this project.

3.1.1. Germplasm Information

To develop the Lilium species germplasm information database, morphological characteristics of 141 cultivars maintained at various research institutes in South Korea were collected from published manuals and books. The morphological and taxonomical information of these cultivars was then compiled and stored in this user-friendly database. To easily navigate the phenotypic data (Figure S1), two search options were incorporated: cultivar name and lily taxonomic group. Users can search for specific cultivars by name or browse through different taxonomic groups to access information from this database.

3.1.2. Molecular Markers Developments and Database Features

In the lily database, three distinct types of molecular markers, namely SSR, SNP, and COS, have been introduced. SSR markers were mined from eight datasets, and their summarized results are presented in Table 2. The result reveals that approximately 12% of the analyzed RNA sequences and 23% of genomic sequences contain SSR motifs, with an average of one SSR found in 2793 bp of sequences in Lily species. As expected, Class II SSR types (SSR loci less than 20 nucleotides) dominate over Class I SSR types (SSR loci greater than 20 nucleotides). Additionally, AT-rich motifs are more abundant than GC-rich motifs, and tri-nucleotide repeats predominate among all other SSR repeats in this study. Similar results have also been reported in many other plant species, including citrus, banana, lily, and others [34,35]. A total of 47,863 primers from the RNA sequence and 103,929 primer pairs from the genome assembly were designed and integrated into the SSR marker database. This database offers a sophisticated search interface with three search criteria that can be used individually or in different combinations. The search results are visualized on a HTML page, providing users with quick access to detailed information about each marker (Figure 1a). Each SSR markers contains 41 attributes that are unique compared to other SSR marker databases. Each set of SSR markers provides comprehensive information, including SSR types, SSR motifs, SSR size, SSR position coordinates in the sequences, flanking SSR sequences, SSR locus, SSR class, and three sets of SSR primer sequences with their annealing temperature, lengths, and predicted PCR product size. All these attributes are visualized on the detailed search result page. Most of the published SSR marker databases only present the forward and reverse primer sets [36–40], while in some cases, genomic coordinates and flanking sequences are presented. In the Lily SSR maker datasets, all 41 attributes were obtained from the in silico characterization of the markers. This information will help in selecting the best primer sets based on the objectives of the research.

Table 2. SSR mining and marker development summary.

Species (Dataset)	***L. formosanum***	***L. longiflorum***	***L. longiflorum* Easter**	***L. longiflorum* White**	***L. regale***	**Lily Hybrid**	***L. formolongi* Sinnapal**	**Over All for RNA Data**	***L. candidum* (Genomic Data)**
No of SSR containing sequences	438	434	22,723	9792	444	242	11,066	45,139	106,749
% of SSR sequences	32.71	32.49	12.62	11.43	37.92	25.39	12.28	12.52	23.28
No of Sequences have more than one SSR	62	67	3778	1638	77	44	1872	7538	17,249

Table 2. *Cont.*

Species (Dataset)	*L. formosanum*	*L. longiflorum*	*L. longiflorum* Easter	*L. longiflorum* White	*L. regale*	Lily Hybrid	*L. formolongi* Sinnapal	Over All for RNA Data	*L. candidum* (Genomic Data)
Total No of SSR identify	515	512	27,332	11,760	533	295	13,374	54,321	106,749
SSR density (per bp)	771.18	1797.07	4138.66	4929.05	1091.09	2309.47	4514.95	2793.07	3236.71
No compound SSR	1	2	235	158	1	2	169	568	2821
Class II SSR	137	393	25,131	10,514	168	235	11,967	48,545	83,250
Class I SSR	377	117	1966	1088	364	58	1238	5208	20,678
AT-rich	475	365	15,177	4761	451	147	5573	26,949	83,369
GC rich	34	124	6458	4148	55	104	4605	15,528	3282
Balance	5	21	5462	2693	26	42	3027	11,276	17,277
Mono	477	307	10,541	2688	432	114	3363	17,922	14,260
Di	5	30	7651	3474	31	47	3899	15,137	66,903
Tri	33	173	8471	5289	63	131	5792	19,952	21,821
Tetra	0	2	357	117	2	2	127	607	2783
Penta	0	0	133	46	2	1	46	228	608
Hexa	0	0	179	146	3	0	147	475	374
No of SSR primer modeling	456	506	26,943	11,511	521	290	7636	47,863	103,929

Note: Class II SSRs are those with a repeat motif length of less than 20 bp. Class I SSRs are 20 nt or longer. Balanced motifs represent an equal proportion of AT and GC content. The abbreviations are as follows: Mono for mononucleotide repeats, Di for dinucleotide repeats, Tri for trinucleotide repeats, Tetra for tetranucleotide repeats, Penta for pentanucleotide repeats, and Hexa for hexanucleotide repeats.

A comprehensive summary of the SNP mining results is presented in Figure 2. Our findings revealed that the majority of these SNPs can be classified into two primary categories: Transitions and Transversions, with Transitions being slightly more abundant (13,006 vs. 7466 SNPs). Within the Transition category, we observed distinct allele combinations, including A/G, C/A, C/G, C/T, and T/A. Similarly, Transversions exhibited a diverse array of allele combinations, such as A/C, A/T, G/A, G/C, T/C, and T/G. Notably, a significant proportion of these SNPs were found to be biallelic, accounting for 99.36% of the total. A similar result has also been reported in potatoes [41]. Furthermore, we conducted an analysis of allele distribution among various lily species. Our observations indicated a prevalence of heterozygous conditions, with heterozygotes being more abundant than homozygotes. These results provide valuable insights into the genetic diversity and allelic distribution within the lily species, offering essential information for further genetic research and breeding programs.

A total of 20,929 SNPs were successfully identified and stored in a dedicated database. The SNP-database search interface enable users to search SNP markers with three search criteria options which include SNP type, SNP class, and SNP position, which can be easily used to filter and retrieve relevant data. The interface (Figure 1b) was designed to provide an intuitive and user-friendly experience. The search criteria can be selected from the dropdown list to submit the queries. The search results are displayed in a user-friendly format, allowing users to easily navigate and analyze the data. A total of 27 attributes are presented for each SNP, encompassing details such as the position of the SNP locus, reference base, SNP type, SNP class, Allele Type, SNP quality, SNP depth coverage at the SNP position, and SNP locus variation among different lily species. These attributes offer valuable insights that facilitate the identification of optimal markers for subsequent analysis.

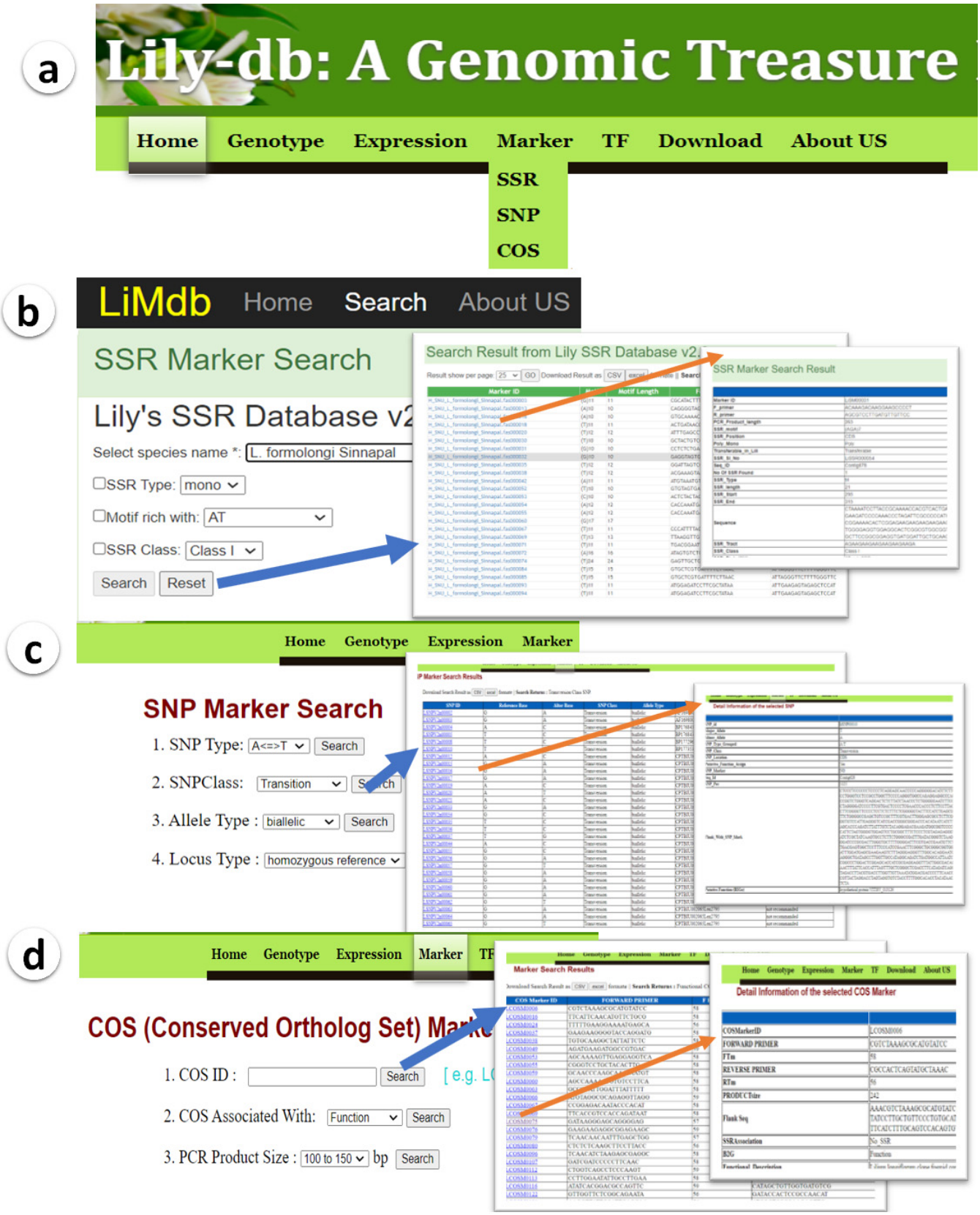

Figure 1. Lily database—(**a**) The frame of the lily database website; (**b**) SSR marker search page; (**c**) SNP markers and (**d**) COS markers search pages. By inputting specific search criteria, users can easily locate relevant markers. The search results are presented in a user-friendly tabular format on a dedicated page. Each entry in the table is associated with a primer ID link. When users click on a

primer ID, they gain access to more detailed information about the corresponding marker. This valuable feature empowers researchers to access comprehensive data linked to their selected microsatellite markers, aiding them in their studies effectively.

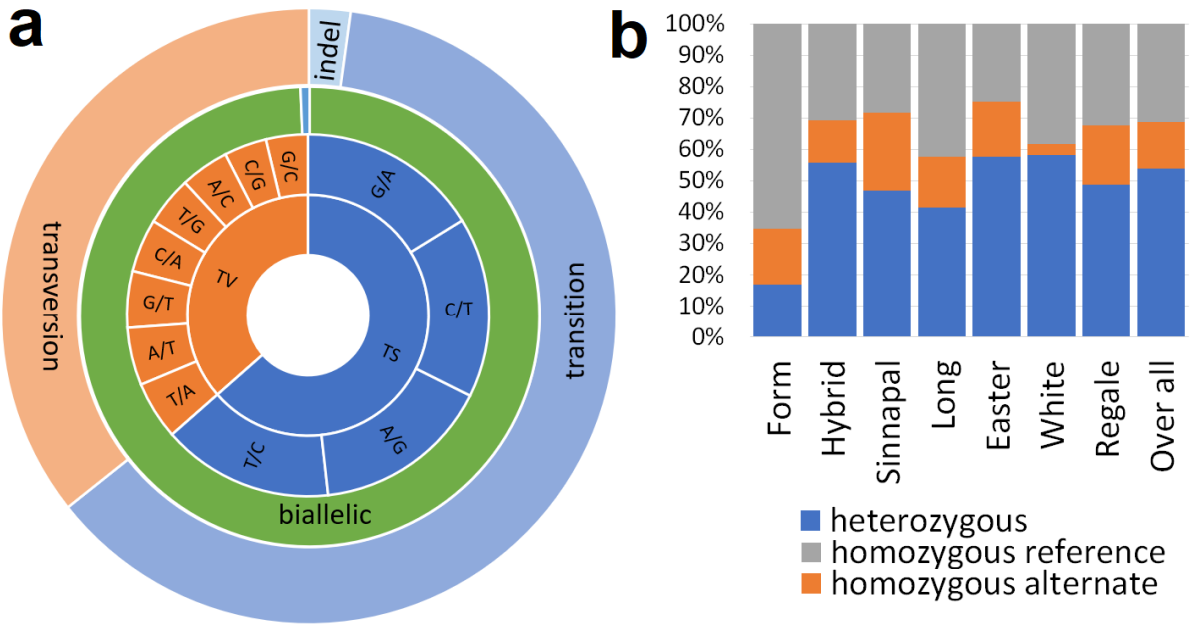

Figure 2. Lily SNP mining and characterization summary. (**a**) Distribution of lily SNP in different classes and types. (**b**) SNP allele distribution in different lily species (here, Form = *L. formosanum; Long* = *L. longiflorum;* Easter = *L. longiflorum* Easter; White = *L. longiflorum* White; Regale = *L. regale;* Sinnapal =*L. formolongi* Sinnapal).

In this study, a total of 179,988; 85,647, and 90,115 transcriptome sequences from *L. longiflorum* "Easter", *L. longiflorum* "White", and *L. formolongi* "Sinnapal" were reciprocally compared, resulting in the identification of 12,695 common sequences that were conserved among the three lily species (Figure S2). Subsequently, we developed 1213 COS markers, representing 9.65% of the conserved sequences (Table S2), for potential marker development. These markers have been stored in a searchable database for future use. The search interface for COS (Conserved Ortholog Sequences) markers offers users an efficient and effective way to search for COS marker sets with 10 attributes. The user-friendly interface permits users to search the entire COS dataset based on three individual search criteria (Figure 1c). Once the user has selected their desired search criteria, they can submit their query and the results will appear in a table format.

3.1.3. Transcription Factor Genes

A comprehensive search for putative transcription factor (TF) genes was conducted on transcriptome sequences. Initially, all sequences were subjected to a BLAST search against the plant TF-database v5, resulting in the identification of 2151 TF genes. Subsequently, these genes underwent further analysis using iTAKdb tools, leading to the annotation of 1327 TF genes. These annotations accounted for approximately 1% of all genes within the transcriptomes (Table S3). These TF genes were subsequently categorized into 46 to 49 families based on their distinctive DNA binding domains, as detailed in Tables S4 and S5 and Figure 3. Notably, the MYB superfamily emerged as the largest family, followed by the bHLH, ERF, NAC, C2H2, WRKY, C3H, bZIP, and G2-like families. Interestingly, the copy numbers of each family varied, with the identification of between 1 and 65 copies per family. To enhance the utility of these TF genes, a user-friendly search page was developed, organized by TF family. All identified families are presented in a tabulated

format, accompanied by links to the TF database. Users can easily access comprehensive details for each TF family by clicking on the respective links. This search page (Figure S3) facilitates the efficient exploration of plant regulatory networks, allowing users to identify specific factors involved in crucial responses in a streamlined manner.

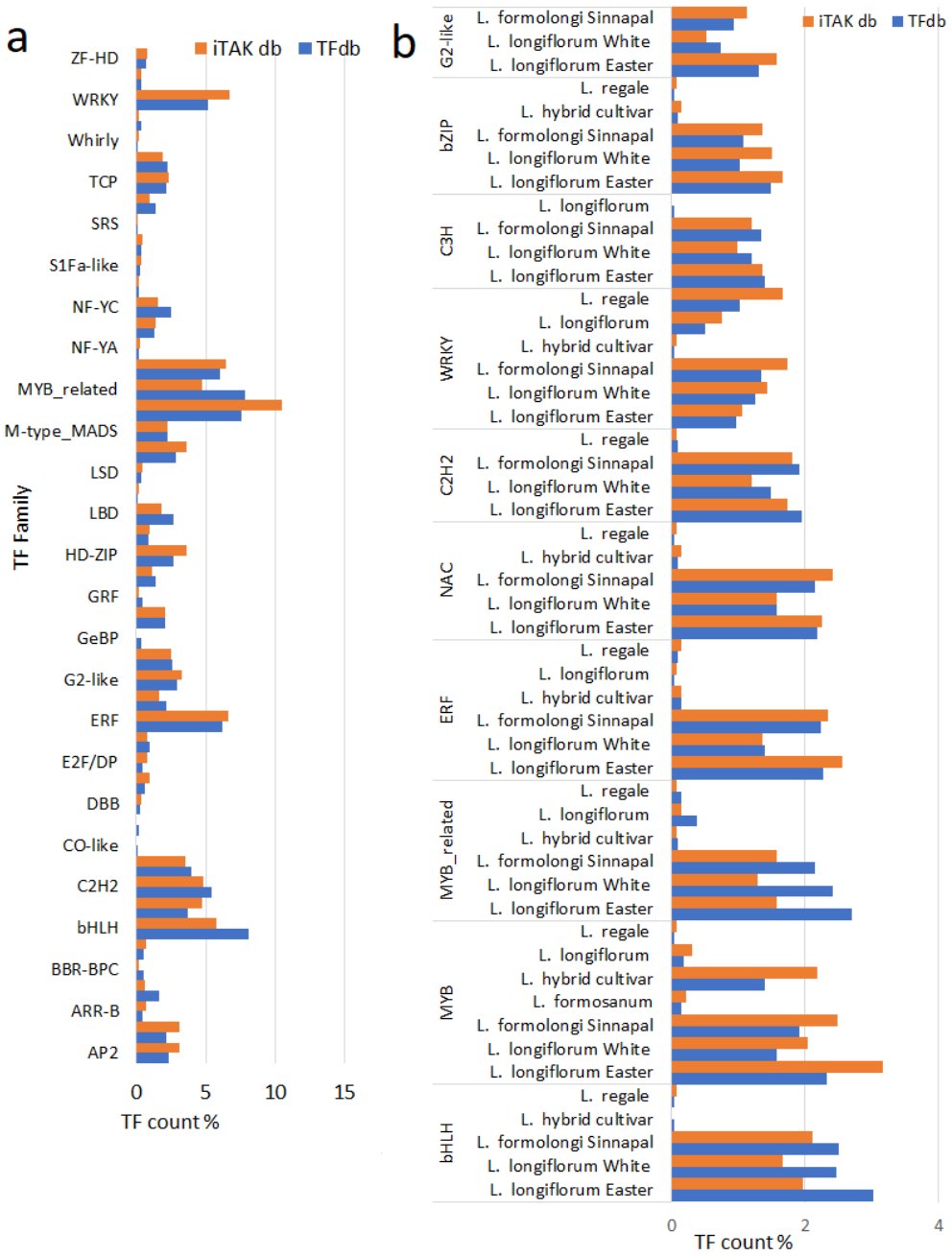

Figure 3. Transcription factor (TF) gene distribution in the lily transcriptome. (**a**) Distribution of transcription factor (TF) genes in the lily transcriptome. (**b**) Distribution of the top 10 TF Families Among the Seven Lily Datasets. The results are presented as percentages of the total TF genes identified in the seven datasets.

3.1.4. Gene Expression Data

The gene expression database is a comprehensive repository that highlights the details of gene regulation [42–44]. Transcriptome data from 22 cDNA libraries were analyzed to calculate the FPKM value for each identified gene, and the data were grouped based on the experiments, including treatments and cultivars. These experiments focused on three

specific lily species: Sinnapal Lily (data from *L. formolongi* cv. Sinnapal Lily transcriptome of 6 cDNA libraries), White tower Lily (data from *L. longiflorum* White tower Lily transcriptome of 10 cDNA libraries), and Easter Lily (data from *L. longiflorum* Easter Lily transcriptome). This valuable resource provides crucial insights into the gene expression patterns of these lily species.

To facilitate easy accessibility and data exploration, the database has been precisely organized into groups according to the specific experiments conducted (Figures S4–S6). Users can navigate through the vast dataset using various search criteria, such as Unigene ID, gene status, gene name, and Gene Ontology (GO) ID, enabling them to pinpoint and analyze gene expression profiles relevant to their research interests. Search results are presented in tabular form, allowing users to copy or download the data in XLS and CSV formats for further use.

This database serves as an indispensable tool for the scientific community, fostering a deeper understanding of gene expression dynamics and paving the way for innovative discoveries in the realm of molecular biology and genetics. It holds the potential to accelerate research efforts and contribute to significant advancements in our knowledge of gene regulation and its implications in various biological processes.

3.2. Applications, Limitations and Future Directions

The current version of Lily-db contains a wide range of data that can be applied in several ways to improve Lilium species' breeding programs. One notable example is the morphological data of the cultivars, which can be utilized by researchers to select and characterize germplasm. Additionally, molecular markers found within the database can be employed to distinguish between different genotypes. This is particularly useful in situations where the morphology of the cultivars is similar or when the cultivars have mixed identities. Biswas et al. [1] have used 46 SSR makers from this database for genetic diversity, population structure, and phylogenetic studies of Korean Lily germplasm. Expression and transcriptome data presented in this database help to understand the gene expression in Lilium species during biotic stress.

In the first version of Lily-db, it only contained the data generated by our research group; recently, many other transcriptome data are now available in the public domain. We will update the Lily database periodically, incorporating public data. More features will be included in the next version of this database.

4. Conclusions

Lily-db is a compressive database that contains various types of data such as morphological, molecular markers, gene expression data, and transcription factor genes. This comprehensive resource offers researchers a broad range of information that can be used in numerous research applications, making it an essential tool for those studying Lilium species in the fields of genetics, genomics, and breeding programs.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/horticulturae10010023/s1. Figure S1. Lily germplasm search user interface: This is designed to facilitate easy access to the diverse germplasm resources available in this database. The user interface offers a user-friendly and efficient platform to explore and retrieve information related to Lily germplasm. Search returns a list of the germplasm with morphological information and images of the cultivars/variety of the germplasm. Figure S2. Vine diagram for conserved ortholog sequences identification from the RNA-seq data of *Lilium* species. Figure S3. TF data search user interface: This figure illustrates the user interface of the TF data search platform. The user interface offers an intuitive and user-friendly experience, enabling researchers to efficiently access valuable information related to TFs. Figure S4. Expression data search user interface for Sinnapal Lily cultivar. Figure S5. Expression data search user interface for White tower Lily cultivar. Figure S6. Expression data search user interface for Easter Lily cultivar. Table S1. List of the RNA-sequencing libraries and read information. Table S2. Lily conserved orthologous sequences (COS) marker development and characterization. Table S3. Summary of the TF analysis.

Table S4. Distribution of Lily TF genes among the TF family. Table S5. Distribution of Lily TF family genes in different family and species level.

Author Contributions: The work presented here was carried out in collaboration among all authors. M.K.B. was involved in transcriptome assembly and annotation; microsatellite mining, database development, and drafted the manuscript; S.N. and D.B. wrote and managed the server to host the database; J.H. and J.-I.P. grew the plant materials and maintained the greenhouse, extracted RNA and performed quality control; M.K.B. and I.-S.N. conceived and designed the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: The Golden Seed Project (Center for Horticultural Seed Development, No. 213007-05-2-CG100) of the Ministry of Agriculture, Food and Rural Affairs (MAFRA), South Korea.

Data Availability Statement: Supporting data can be found is the Supplementary Materials.

Acknowledgments: We are grateful to the National Institute of Horticultural & Herbal Science, Agricultural Research & Extension Services of Gangwondo, Jeollabuk-do, Chungcheongbuk-do, and Gyeongsangnam-do in South Korea for their invaluable support to access essential lily germplasm information and literature.

Conflicts of Interest: Author Sathishkumar Natarajan was employed by the company 3BIGS Co., Ltd., B-831, Geumgang Penterium IX Tower, Hwaseong 18469, Korea. The remaining authors wish to disclose that, at the time of conducting this research and submitting the manuscript, they were not subject to any commercial or financial relationships that could be perceived as a potential conflict of interest.

References

1. Biswas, M.K.; Bagchi, M.; Nath, U.K.; Biswas, D.; Natarajan, S.; Jesse, D.M.I.; Park, J.; Nou, I. Transcriptome wide SSR discovery cross-taxa transferability and development of marker database for studying genetic diversity population structure of Lilium species. *Sci. Rep.* **2020**, *10*, 18621. [CrossRef] [PubMed]
2. Angiosperm Phylogeny Group; Chase, M.W.; Christenhusz, M.J.; Fay, M.F.; Byng, J.W.; Judd, W.S.; Soltis, D.E.; Mabberley, D.J.; Sennikov, A.N.; Soltis, P.S. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **2016**, *181*, 1–20. [CrossRef]
3. Biswas, M.K.; Nath, U.K.; Howlader, J.; Bagchi, M.; Natarajan, S.; Kayum, M.A.; Kim, H.; Park, J.; Kang, J.; Nou, I. Exploration and exploitation of novel SSR markers for candidate transcription factor genes in *Lilium* species. *Genes* **2018**, *9*, 97. [CrossRef] [PubMed]
4. Buschman, J. Globalisation-flower-flower bulbs-bulb flowers. In *IX International Symposium on Flower Bulbs 673*; ISHS: Leuven, Belgium, 2004; pp. 27–33.
5. Wilford, R.; Gardens, K.R.B. *The Kew Gardener s Guide to Growing Bulbs: The Art and Science to Grow Your Own Bulbs*; White Lion Publishing: London, UK, 2019.
6. Li, Y.; Zhang, L.; Wang, T.; Zhang, C.; Wang, R.; Zhang, D.; Xie, Y.; Zhou, N.; Wang, W.; Zhang, H. The complete chloroplast genome sequences of three lilies: Genome structure, comparative genomic and phylogenetic analyses. *J. Plant Res.* **2022**, *135*, 723–737. [CrossRef] [PubMed]
7. Du, Y.; Bi, Y.; Yang, F.; Zhang, M.; Chen, X.; Xue, J.; Zhang, X. Complete chloroplast genome sequences of *Lilium*: Insights into evolutionary dynamics and phylogenetic analyses. *Sci. Rep.* **2017**, *7*, 5751. [CrossRef] [PubMed]
8. Liu, H.; Yu, Y.; Deng, Y.; Li, J.; Huang, Z.; Zhou, S. The chloroplast genome of *Lilium henrici*: Genome structure and comparative analysis. *Molecules* **2018**, *23*, 1276. [CrossRef] [PubMed]
9. Kim, J.H.; Lee, S.I.; Kim, B.R.; Choi, I.Y.; Ryser, P.; Kim, N.S. Chloroplast genomes of *Lilium lancifolium*, *L. amabile*, *L. callosum*, and *L. philadelphicum*: Molecular characterization and their use in phylogenetic analysis in the genus Lilium and other allied genera in the order Liliales. *PLoS ONE* **2017**, *12*, e0186788.
10. Howlader, J.; Robin, A.H.K.; Natarajan, S.; Biswas, M.K.; Sumi, K.R.; Song, C.Y.; Park, J.; Nou, I. Transcriptome analysis by rna–seq reveals genes related to plant height in two sets of parent-hybrid combinations in easter lily (*Lilium longiflorum*). *Sci. Rep.* **2020**, *10*, 9082. [CrossRef]
11. Du, F.; Wu, Y.; Zhang, L.; Li, X.; Zhao, X.; Wang, W.; Gao, Z.; Xia, Y. *De novo* assembled transcriptome analysis and SSR marker development of a mixture of six tissues from *Lilium* Oriental hybrid 'Sorbonne'. *Plant Mol. Biol. Rep.* **2015**, *33*, 281–293. [CrossRef]
12. Sun, M.; Zhao, Y.; Shao, X.; Ge, J.; Tang, X.; Zhu, P.; Wang, J.; Zhao, T. EST–SSR Marker Development and Full-Length Transcriptome Sequence Analysis of Tiger Lily (*Lilium lancifolium* Thunb). *Appl. Bionics Biomech.* **2022**, *2022*, 7641048. [CrossRef]
13. Droc, G.; Lariviere, D.; Guignon, V.; Yahiaoui, N.; This, D.; Garsmeur, O.; Dereeper, A.; Hamelin, C.; Argout, X.; Dufayard, J. The banana genome hub. *Database* **2013**, *2013*, bat035. [CrossRef]

14. Fernandez-Pozo, N.; Menda, N.; Edwards, J.D.; Saha, S.; Tecle, I.Y.; Strickler, S.R.; Bombarely, A.; Fisher-York, T.; Pujar, A.; Foerster, H. The Sol Genomics Network (SGN)—From genotype to phenotype to breeding. *Nucleic Acids Res.* **2015**, *43*, D1036–D1041. [CrossRef] [PubMed]
15. Chen, H.; Wang, T.; He, X.; Cai, X.; Lin, R.; Liang, J.; Wu, J.; King, G.; Wang, X. BRAD V3. 0: An upgraded Brassicaceae database. *Nucleic Acids Res.* **2022**, *50*, D1432–D1441. [CrossRef] [PubMed]
16. Jung, S.; Lee, T.; Cheng, C.; Zheng, P.; Bubble, K.; Crabb, J.; Gasic, K.; Yu, J.; Humann, J.; Hough, H. Resources for peach genomics, genetics and breeding research in GDR, the Genome Database for Rosaceae. In *X International Peach Symposium 1352*; ISHS: Leuven, Belgium, 2022; pp. 149–156.
17. Liu, H.; Wang, X.; Liu, S.; Huang, Y.; Guo, Y.; Xie, W.; Liu, H.; ul Qamar, M.T.; Xu, Q.; Chen, L. Citrus Pan-Genome to Breeding Database (CPBD): A comprehensive genome database for citrus breeding. *Mol. Plant* **2022**, *15*, 1503–1505. [CrossRef] [PubMed]
18. Lee, S.; Nguyen, X.T.; Kim, J.; Kim, N. Genetic diversity and structure analyses on the natural populations of diploids and triploids of tiger lily, *Lilium lancifolium* Thunb., from Korea, China, and Japan. *Genes Genom.* **2016**, *38*, 467–477. [CrossRef]
19. Wen, C.S.; Hsiao, J.Y. Altitudinal genetic differentiation and diversity of Taiwan lily (*Lilium longiflorum* var. formosanum; Liliaceae) using RAPD markers and morphological characters. *Int. J. Plant Sci.* **2001**, *162*, 287–295. [CrossRef]
20. Shahin, A.; Smulders, M.J.; van Tuyl, J.M.; Arens, P.; Bakker, F.T. Using multi-locus allelic sequence data to estimate genetic divergence among four *Lilium* (Liliaceae) cultivars. *Front. Plant Sci.* **2014**, *5*, 567. [CrossRef]
21. Yuan, S.; Ge, L.; Liu, C.; Ming, J. The development of EST-SSR markers in Lilium regale and their cross-amplification in related species. *Euphytica* **2013**, *189*, 393–419. [CrossRef]
22. Varshney, A.; Sharma, M.P.; Adholeya, A.; Dhawan, V.; Srivastava, P.S. Enhanced growth of micropropagated bulblets of *Lilium* sp. inoculated with arbuscular mycorrhizal fungi at different P fertility levels in an alfisol. *J. Hortic. Sci. Biotechnol.* **2002**, *77*, 258–263. [CrossRef]
23. Xi, M.; Sun, L.; Qiu, S.; Liu, J.; Xu, J.; Shi, J. In vitro mutagenesis and identification of mutants via ISSR in lily (*Lilium longiflorum*). *Plant Cell Rep.* **2012**, *31*, 1043–1051. [CrossRef]
24. Yin, Z.; Zhao, B.; Bi, W.; Chen, L.; Wang, Q. Direct shoot regeneration from basal leaf segments of *Lilium* and assessment of genetic stability in regenerants by ISSR and AFLP markers. *Vitr. Cell. Dev. Biol.-Plant* **2013**, *49*, 333–342. [CrossRef]
25. Brown, J.; Pirrung, M.; McCue, L.A. FQC Dashboard: Integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **2017**, *33*, 3137–3139. [CrossRef] [PubMed]
26. Hancock, B. Trinity v3, a DDoS tool, hits the streets. *Comput. Secur.* **2000**, *19*, 574. [CrossRef]
27. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
28. Huang, H.; Wang, L.; Tak, B.C.; Wang, L.; Tang, C. Cap3: A cloud auto-provisioning framework for parallel processing using on-demand and spot instances. In Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, CA, USA, 28 June–3 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 228–235.
29. Biswas, M.K.; Natarajan, S.; Biswas, D.; Nath, U.K.; Park, J.; Nou, I. LSAT: Liliaceae Simple Sequences Analysis Tool, a web server. *Bioinformation* **2018**, *14*, 181. [CrossRef]
30. Zheng, Y.; Jiao, C.; Sun, H.; Rosli, H.G.; Pombo, M.A.; Zhang, P.; Banf, M.; Dai, X.; Martin, G.B.; Giovannoni, J.J. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **2016**, *9*, 1667–1670. [CrossRef]
31. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
32. Jin, J.; Tian, F.; Yang, D.; Meng, Y.; Kong, L.; Luo, J.; Gao, G. PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **2017**, *45*, D1040–D1045. [CrossRef]
33. Love, M.; Anders, S.; Huber, W. Differential analysis of count data–the DESeq2 package. *Genome Biol.* **2014**, *15*, 550.
34. Biswas, M.K.; Xu, Q.; Mayer, C.; Deng, X. Genome wide characterization of short tandem repeat markers in sweet orange (*Citrus sinensis*). *PLoS ONE* **2014**, *9*, e104182. [CrossRef]
35. Biswas, M.K.; Chai, L.; Mayer, C.; Xu, Q.; Guo, W.; Deng, X. Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in Citrus. *Mol. Biol. Rep.* **2012**, *39*, 5373–5386. [CrossRef] [PubMed]
36. Arora, V.; Kapoor, N.; Fatma, S.; Jaiswal, S.; Iquebal, M.A.; Rai, A.; Kumar, D. BanSatDB, a whole-genome-based database of putative and experimentally validated microsatellite markers of three Musa species. *Crop J.* **2018**, *6*, 642–650. [CrossRef]
37. Xu, H.; Yu, Q.; Shi, Y.; Hua, X.; Tang, H.; Yang, L.; Ming, R.; Zhang, J. PGD: Pineapple genomics database. *Hortic. Res.* **2018**, *5*, 66. [CrossRef]
38. Moyle, R.L.; Crowe, M.L.; Ripi-Koia, J.; Fairbairn, D.J.; Botella, J.R. PineappleDB: An Online Pineapple Bioinformatics Resource. *BMC Plant Biol.* **2005**, *5*, 21. [CrossRef] [PubMed]
39. Mokhtar, M.M.; Atia, M.A.M. SSRome: An integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Res.* **2019**, *47*, D244–D252. [CrossRef] [PubMed]
40. Yu, J.; Dossa, K.; Wang, L.; Zhang, Y.; Wei, X.; Liao, B.; Zhang, X. PMDBase: A database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Res.* **2017**, *45*, D1046–D1053. [CrossRef] [PubMed]
41. Pham, G.M.; Newton, L.; Wiegert-Rininger, K.; Vaillancourt, B.; Douches, D.S.; Buell, C.R. Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J.* **2017**, *92*, 624–637. [CrossRef]

42. Yu, Y.; Zhang, H.; Long, Y.; Shu, Y.; Zhai, J. Plant public RNA-seq database: A comprehensive online database for expression analysis of ~45 000 plant public RNA-seq libraries. *Plant Biotechnol. J.* **2022**, *20*, 806. [CrossRef]
43. Ma, X.; Yan, H.; Yang, J.; Liu, Y.; Li, Z.; Sheng, M.; Cao, Y.; Yu, X.; Yi, X.; Xu, W. PlantGSAD: A comprehensive gene set annotation database for plant species. *Nucleic Acids Res.* **2022**, *50*, D1456–D1467. [CrossRef]
44. Zhou, Z.; Tan, C.; Chau, M.H.K.; Jiang, X.; Ke, Z.; Chen, X.; Cao, Y.; Kwok, Y.K.; Bellgard, M.; Leung, T.Y. TEDD: A database of temporal gene expression patterns during multiple developmental periods in human and model organisms. *Nucleic Acids Res.* **2023**, *51*, D1168–D1178. [CrossRef]