plants

MDPI

Article

Updated Gene Prediction of the Cucumber (9930) Genome through Manual Annotation

Weixuan Du [1], Lei Xia [1], Rui Li [1], Xiaokun Zhao [1], Danna Jin [1], Xiaoning Wang [1], Yun Pei [1,2], Rong Zhou [3], Jinfeng Chen [1] and Xiaqing Yu [1,*]

[1] State Key Laboratory of Crop Genetics & Germplasm Enhancement and Utilization, Nanjing Agricultural University, No. 1 Weigang, Nanjing 210095, China; jfchen@njau.edu.cn (J.C.)
[2] College of Agriculture, Guizhou University, Guiyang 550025, China
[3] Department of Food Science, Plant, Food & Climate, Aarhus University, Agro Food Park 48, DK-8200 Aarhus, Denmark; rong.zhou@food.au.dk
* Correspondence: xqyu@njau.edu.cn; Tel.: +86-025-8439-6279

Abstract: Thorough and precise gene structure annotations are essential for maximizing the benefits of genomic data and unveiling valuable genetic insights. The cucumber genome was first released in 2009 and updated in 2019. To increase the accuracy of the predicted gene models, 64 published RNA-seq data and 9 new strand-specific RNA-seq data from multiple tissues were used for manual comparison with the gene models. The updated annotation file (V3.1) contains an increased number (24,145) of predicted genes compared to the previous version (24,317 genes), with a higher BUSCO value of 96.9%. A total of 6231 and 1490 transcripts were adjusted and newly added, respectively, accounting for 31.99% of the overall gene tally. These newly added and adjusted genes were renamed (CsaV3.1_XGXXXXX), while genes remaining unaltered preserved their original designations. A random selection of 21 modified/added genes were validated using RT-PCR analyses. Additionally, tissue-specific patterns of gene expression were examined using the newly obtained transcriptome data with the revised gene prediction model. This improved annotation of the cucumber genome will provide essential and accurate resources for studies in cucumber.

Keywords: cucumber; manual reannotation; transcriptome; gene model; tissue-specific expression

Citation: Du, W.; Xia, L.; Li, R.; Zhao, X.; Jin, D.; Wang, X.; Pei, Y.; Zhou, R.; Chen, J.; Yu, X. Updated Gene Prediction of the Cucumber (9930) Genome through Manual Annotation. *Plants* **2024**, *13*, 1604. https://doi.org/10.3390/plants13121604

Academic Editor: Ming Chen

Received: 12 April 2024
Revised: 2 June 2024
Accepted: 3 June 2024
Published: 9 June 2024

1. Introduction

The utility of genomic data relies on precise and comprehensive genome annotation. This process involves multifaceted gene prediction, drawing from diverse data sources [1]. Common strategies for genome annotation encompass de novo prediction, homology prediction via cross-species protein sequences, and automated prediction utilizing RNA data from varied tissues [2,3]. Genomes of some important species have been subjected to repeated reannotation efforts to enrich genetic information exploration. Notably, recent years have witnessed multiple versions of genome annotations for Arabidopsis, citrus, rice, peach, and maize [4–8]. Manual refinement of gene models through RNA data visualization to repair shortages of algorithmic software has proven efficacious in enhancing the accuracy of annotation files. WebApollo software (version 2.0.2) has been employed to manually adjust gene models for the kiwifruit genome [9], and the GSAman software (v.0.8.2) has been utilized to refine gene structure models for the peach and sweet potato genomes [8,10].

Cucumber (*Cucumis sativus* L.) is a widely cultivated annual herbaceous plant recognized as one of the top ten globally grown vegetables due to its distinctive taste and nutritionally dense profile. Cucumber was the first vegetable crop to be successfully sequenced in 2009, marking the onset of the genomic era in vegetable research and facilitating molecular breeding advancements [11]. Through whole-genome shotgun sequencing techniques, including Sanger and next-generation sequencing, a comprehensive genome sequence of 243.5 Mb was assembled for the cucumber 9930 ('Chinese Long' line). This

effort validated the evolutionary transformation of cucumber's seven chromosomes from twelve ancestral chromosomes and laid the foundation for subsequent comparative genomics studies on the identification of trait-related genes that are favored during cucumber domestication [11,12]. Key genes linked to the generation of bitterness in cucumbers were pinpointed, effectively resolving the issue of bitter cucumbers in the southern region of China and producing notable societal advantages [13].

The release of the 9930_V3 version of the cucumber reference genome, generated using PacBio third-generation sequencing technology, represents a significant improvement in the research in cucumber-related fields. However, the current gene annotation for the 9930_V3 reference genome is based on algorithmic predictions, highlighting the need for further enhancements in both the accuracy and comprehensiveness of gene annotations [2]. Ensuring high genomic stability and coherence is paramount in functional genomics studies, while precise annotated files are indispensable for extracting the vast array of information present within genomic information. The currently utilized 9930 V3 version stands as the most complete reference genome for cucumber, attaining a BUSCO score of 95.4% based on angiosperm embryophyte benchmarks. Yet its gene structure annotations exhibit inadequacies in their comprehensiveness. In this study, we manually reannotated the cucumber gene models by incorporating 64 previously published RNA datasets and 9 newly generated strand-specific RNA datasets, covering various developmental stages (e.g., stem, flower, and fruit tissues) and experimental conditions. The quality of the updated annotation file, which includes 24,145 protein-coding genes, was evaluated using the BUSCO tool, resulting in a completion rate of 96.9%. The nomenclature for the newly added or revised genes followed the format of CsaV3.1_XGXXXXX. Validation of the annotations was performed through the identification of 21 genes using PCR assays and Sanger sequencing. Moreover, the gene expression across diverse tissues was investigated, revealing varying expression patterns, especially those specific to particular tissues. In conclusion, the newly established gene annotation file and its related expression data will serve as a solid and reliable basis for functional studies on cucumber.

2. Results

2.1. Reannotation of the Cucumber Genome through Manual Operation

To enhance the utility of the V3 reference genome, a thorough manual reannotation was conducted through the integration of publicly accessible RNA-seq data alongside newly acquired strand-specific RNA-seq information. By leveraging 64 RNA-seq libraries derived for the Chinese Long cucumber variety, encompassing diverse tissue types and experimental conditions, a substantial body of data comprising 1.6 billion paired-end reads was generated. Nine strand-specific sequencing data were employed to analyze the gene orientation in root, stem, and leaf tissues, with three replicates for each. Through meticulous examination of the RNA data, sequences showing transcriptional activity were manually curated. This effort led to the creation of an updated annotation file containing 24,145 predicted protein-coding genes (Figure 1).

In contrast to the prior annotation, the updated annotation file reflects alterations or supplements to 7721 genes, constituting 31.99% of the overall gene tally. Genes remaining unaltered preserve their original designations in the revised iteration, while adjusted or newly integrated genes are denoted by the CsaV3.1_XGXXXXX nomenclature. Essential statistical comparisons between pre- and post-modifications have been tabulated in Table 1. The mean exon count per gene witnessed an increment from 5.22 to 5.27, and the average coding sequence (CDS) length elongated from 1189 to 1247 nucleotides. The BUSCO tool is recognized for its reliability in evaluating the completeness of genome annotations, with a higher score indicating a more comprehensive annotation. The updated genome annotation attained a completion rate of 96.9% according to BUSCO, exceeding the previous score of 95.4%. This objectively confirms the enhanced completeness of the new annotation version in comparison to the previous version.

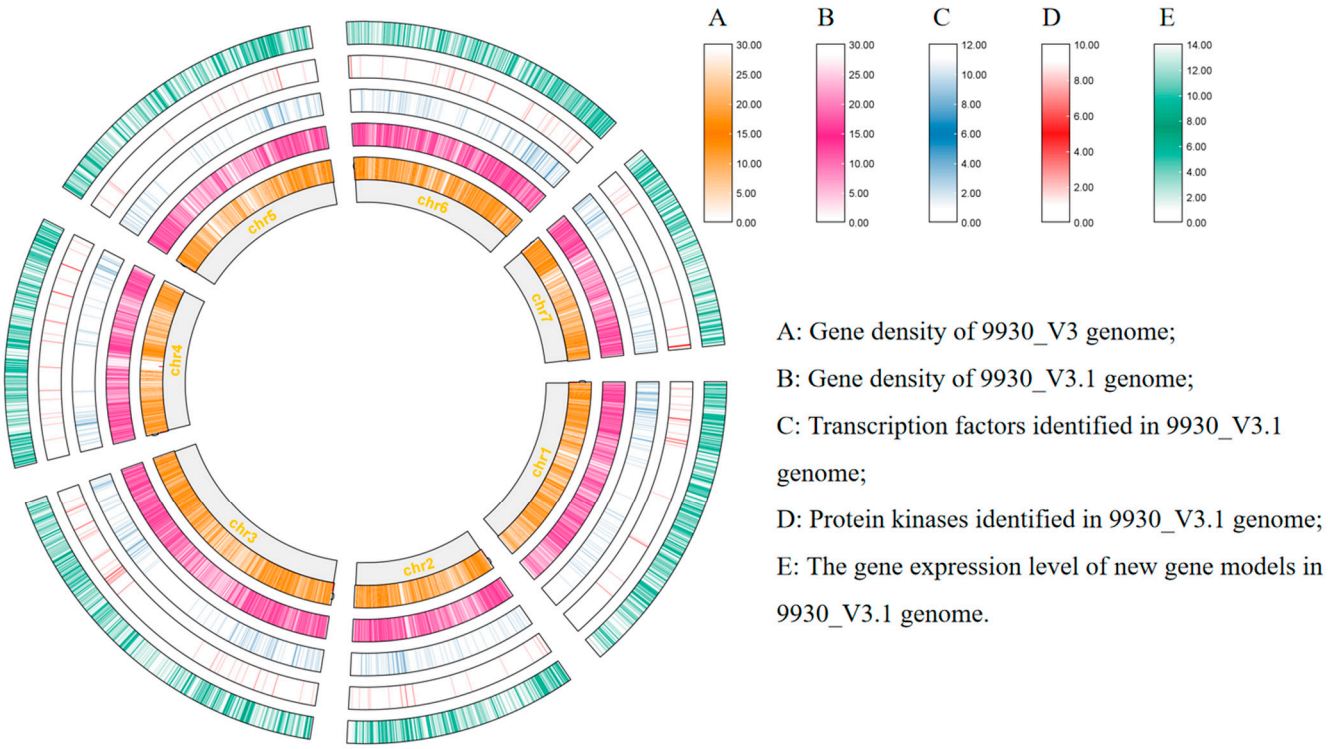

Figure 1. Characterization of the updated gene annotation of the 9930 cucumber genome. (**A**) Gene density in the 9930_V3 reference genome; (**B**) Gene density in the updated 9930_V3.1 reference genome; (**C**) Transcription factors identified in the 9930_V3.1 reference genome; (**D**) Protein kinases discovered in the 9930_V3.1 reference genome; (**E**) Gene expression levels of the newly annotated gene models in the 9930_V3.1 reference genome. The image was generated by TBtools [14] with default parameters.

Table 1. Summary of the new cucumber genome annotation.

	V3	**V3.1**
Number of genes	24,317	24,145
Mean mRNA length (bp)	1716	1819
Mean exon number	5.22	5.27
Mean CDS length (bp)	1128	1131
Genes with GO terms	11,086	11,142
Transcripts with KEGG terms	20,167	20,317
Complete BUSCOs	95.4%	96.9%
Complete and single-copy BUSCOs	91.3%	92.6%
Complete and duplicated BUSCOs	4.1%	4.3%
Fragmented BUSCOs	1.5%	0.7%
Missing BUSCOs	3.1%	2.2%

2.2. Functional Annotation of Genes in the New Annotation File

The updated annotation file was scrutinized for alterations in the coding sequences. Functional annotation of all the protein sequences was conducted utilizing the GO, KEGG, and iTAK databases. Gene categorization into specific GO terms was accomplished via eggNOG-mapper, yielding an assignment of 46.15% (11,142 transcripts) of the total 24,145 transcripts to distinct GO terms, indicating enhancement from the previous assignment of 45.58% (11,086 transcripts). Furthermore, annotation of protein sequences using the Kobas online tool led to the assignment of 84.15% (20,317 transcripts) to specific KEGG pathways, surpassing the previous assignment of 82.93% (20,167 transcripts).

The iTAK software tool (http://itak.feilab.net/cgi-bin/itak/index.cgi, accessed on 12 April 2024) was employed for the identification of transcription factors (TFs), transcriptional regulators (TRs), and protein kinases (PKs) based on protein or nucleotide sequences, with subsequent classification of individual TFs. The updated annotation revealed a total of 1792 TFs/regulators and 801 protein kinases. Modifications to the gene models of 772 TFs/regulators and 349 protein kinases were made in the revised annotation. Noteworthy alterations were observed in the abundance of specific TF families; for instance, the AP2/ERF family decreased from 144 to 139 genes, the C2H2 zinc finger family decreased from 105 to 101 genes, and the C3H family decreased from 47 to 46 genes, while the C2C2 family increased from 81 to 82 genes. Additionally, structural revisions were implemented for 43 RLK-Pelle_DLSV genes, 20 RLK-Pelle_LRR-III genes, and 16 RLK-Pelle_LRR-XI-1 genes in the protein kinase analysis. In summary, the new annotation file underwent substantial modifications. All the predicted gene functions, as determined using the plaBi database, are provided in Table S7.

2.3. Identification of Novel Gene Features in the New Annotation File

Moreover, apart from reannotating established genes, a comparative assessment was carried out utilizing the gffcompare tool (v0.12.6) to analyze modifications in the annotation datasets. The findings indicated that 16,424 genes retained their original annotations, while 1490 newly identified genes were detected and 6231 original genes were refined. The dispersion of these novel genes on the cucumber chromosomes displayed an uneven distribution, with 853 genes located on Chromosome 2 and 1334 genes on Chromosome 3.

Out of the 7721 newly identified gene structures, 5400 (69.93%) were associated with KEGG terms, and 2716 (35.18%) were linked to GO terms. Using the RNA-seq data and the updated annotation file, the expression profiles of these novel genes were evaluated in roots, stems, and leaves. Among the total genes, 5746 presented an average expression level above 2 transcripts per million (TPM) in the analyzed tissues. Subsequently, a visual representation was constructed to illustrate the top 50 genes with the highest expression levels in each tissue (Figure 2A). The majority of these genes exhibited expression values ranging from 2 to 200 TPM, reflecting distinct tissue-specific expression patterns. Notably, of the top 50 highly expressed genes, 13 were predominantly expressed in the leaves, 21 in the stems, and 16 in the roots. According to the GO enrichment analysis, they are related to pathways such as environmental response and cytoplasmic activities (Figure 2B).

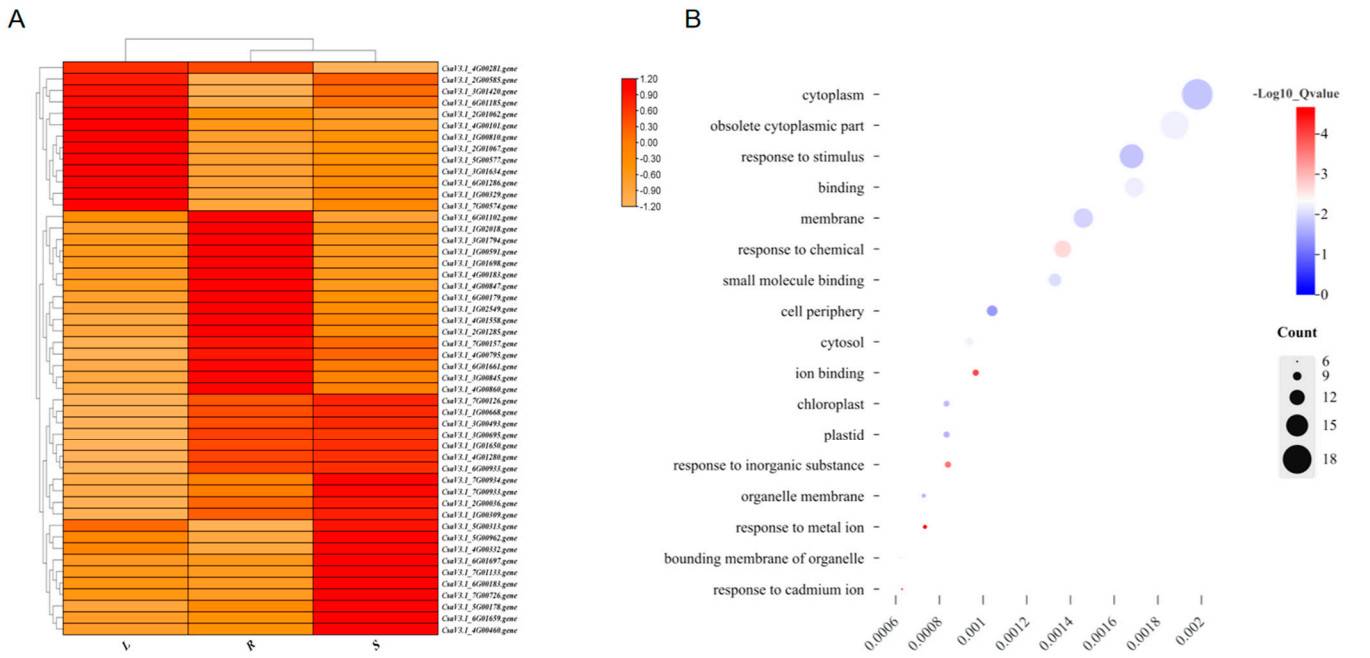

Figure 2. Heatmap (**A**) illustrating the top 50 novel genes with the highest relative expression levels among the 7721 novel discovered genes and their GO enrichment (**B**). (**A**) was generated using TBtools [14] with default parameters, and (**B**) was made with ggplot2 package in R (v4.3.3).

The results reveal new gene characteristics in the cucumber genome, offering insights into their distribution and tissue-specific expression patterns. Additional research is necessary to clarify the specific roles and importance of these recently identified gene features in cucumber biology.

2.4. Validation of Selected Novel Transcripts

In order to validate the accuracy and dependability of the revised annotation file, a total of 21 newly annotated genes at the genomic level were randomly selected for verification. The coding sequences (CDS) of these predicted genes were amplified using the primers detailed in Supplementary Table S1, followed by confirmation through Sanger sequencing.

The preceding annotation file demonstrated multiple instances of gene annotation inaccuracies, errors, and omissions, as depicted in Figure 3. Initially, there were instances of excessively extended untranslated regions (UTRs), notably exemplified in the initial annotation file for the gene *CasV3_1G000090*. Secondly, numerous genes were annotated with an incorrect exon count, as exemplified by the gene *CsaV3_4G028190*, which exhibited three erroneously annotated exons in the original file. Thirdly, misconceptions in gene annotation were identified, such as the merging of gene entities, as illustrated by *CasV3.1_1G02457*, initially annotated as separate genes, *CasV3_1G00097* and *CasV3_1G00098*. In addition, the absence of partial UTR structures in previous annotations was detected, notably in the gene *CasV3_3G03778*. Furthermore, discrepancies in gene expression profiling led to the exclusion of highly expressed genes, like *CasV3.1_6G01238*. Moreover, instances of missing exons were noticed, such as in the case of *CasV3.1_7G00744*, which lacked a terminal exon. Lastly, discrepancies in gene length were observed compared to the RNA data, particularly evident in the shortened length of the gene *CsaV3_7G00815* at one end according to the RNA data.

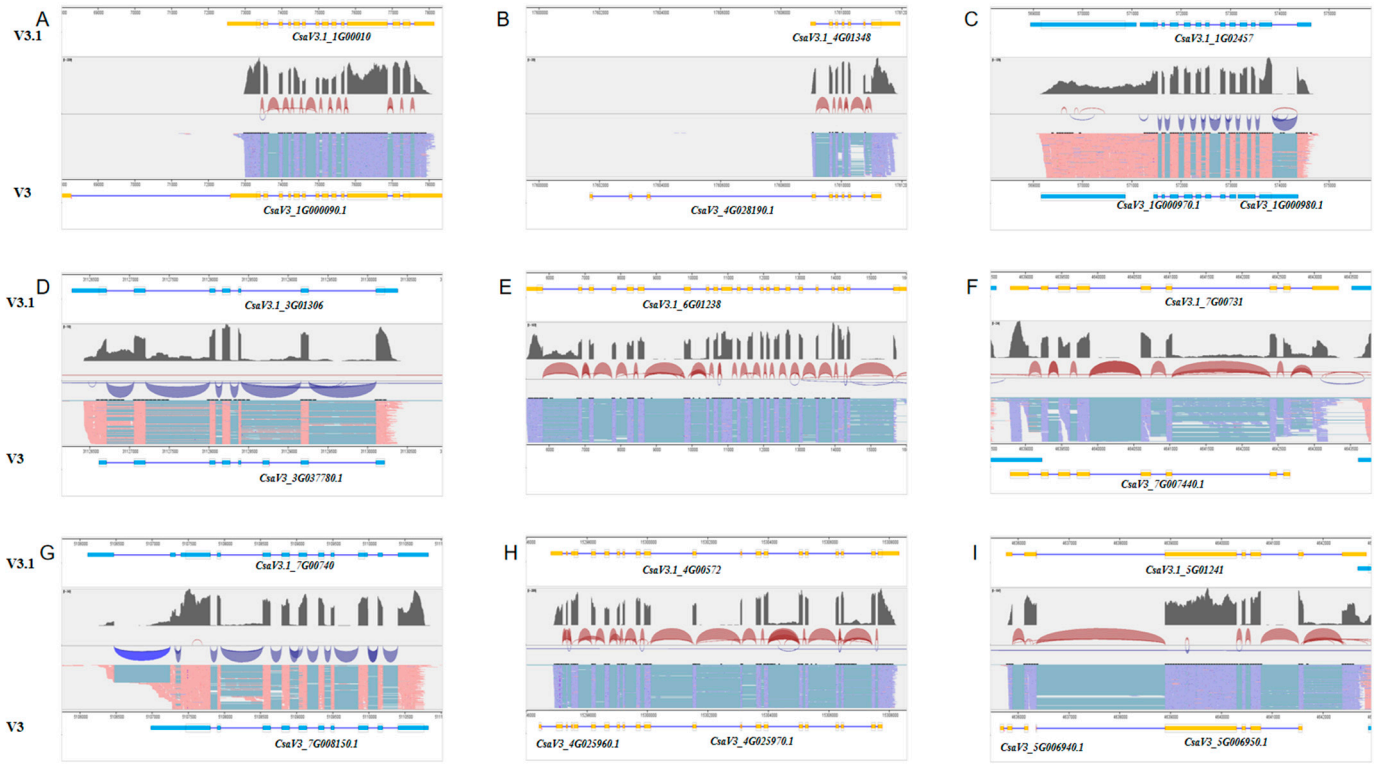

Figure 3. Gene adjustments and new additions to the updated annotation based on RNA-seq mapping results. (**A**) Premature termination of the *CasV3_1G000090* gene prediction. (**B**) Reduction of three exons in the *CsaV3_4G028190* gene. (**C**) Fusion of *CsaV3_1G00097* and *CsaV3_1G00098* into a single gene. (**D**). Delayed termination of the *CsaV3_3G03778* gene prediction. (**E**). Addition of the *CsaV3.1_6G01238* gene. (**F**). Delayed termination and the inclusion of an additional intron in the predicted *CsaV3_7G00744* gene. (**G**). Delayed termination and the inclusion of two additional introns in the predicted *CsaV3_7G00815* gene. (**H**). Fusion of *CsaV3_4G02596* and *CsaV3_4G02597* into a single gene with a delayed termination. (**I**) Mutual impact of gene expression results among *CsaV3_5G00694*, *CsaV3_5G00695*, and *CsaV3_5G00696*. The images were output from GSAman (v.0.8.2).

Through the implementation of validation procedures, our objective is to verify the accuracy of the updated annotation document and rectify any flaws, discrepancies, or oversights detected in the prior annotation. These results enhance the knowledge of the cucumber genome, particularly its genetic architecture, in a more thorough and reliable manner.

2.5. Expression Profiles of Genes in the Novel Annotation File Based on RNA-Seq Data

Highlighting expression profiling is an effective tool for gene recognition. The study calibrated the expression matrix encompassing roots, stems, and leaves derived from the "9930" cucumber inbred line using TPM. Among these plant parts, this research observed 22,190 genes exhibiting significant expression, with a mean TPM value exceeding 1 in 17,575 genes. Each sampled portion contained an approximation of 19,466 to 20,326 expressed genes, with 19,842 genes consistently expressed across all three types of tissues (Figure S3).

The study categorized the expressed genes (22,190 in total) into six distinct clusters considering their unique expression tendencies across various tissues (Figure 4). The groups comprised 3553, 5444, 4240, 2292, 3969, and 2978 genes, respectively. These results reinforce the inherent tissue specificity of gene expression. Interestingly, groups 1 and 5 manifested marked expression in the stem tissues, while groups 3 and 4 showcased escalated expression in the leaf tissues. Predominantly, roots were the major sites of expression for groups 2 and 6.

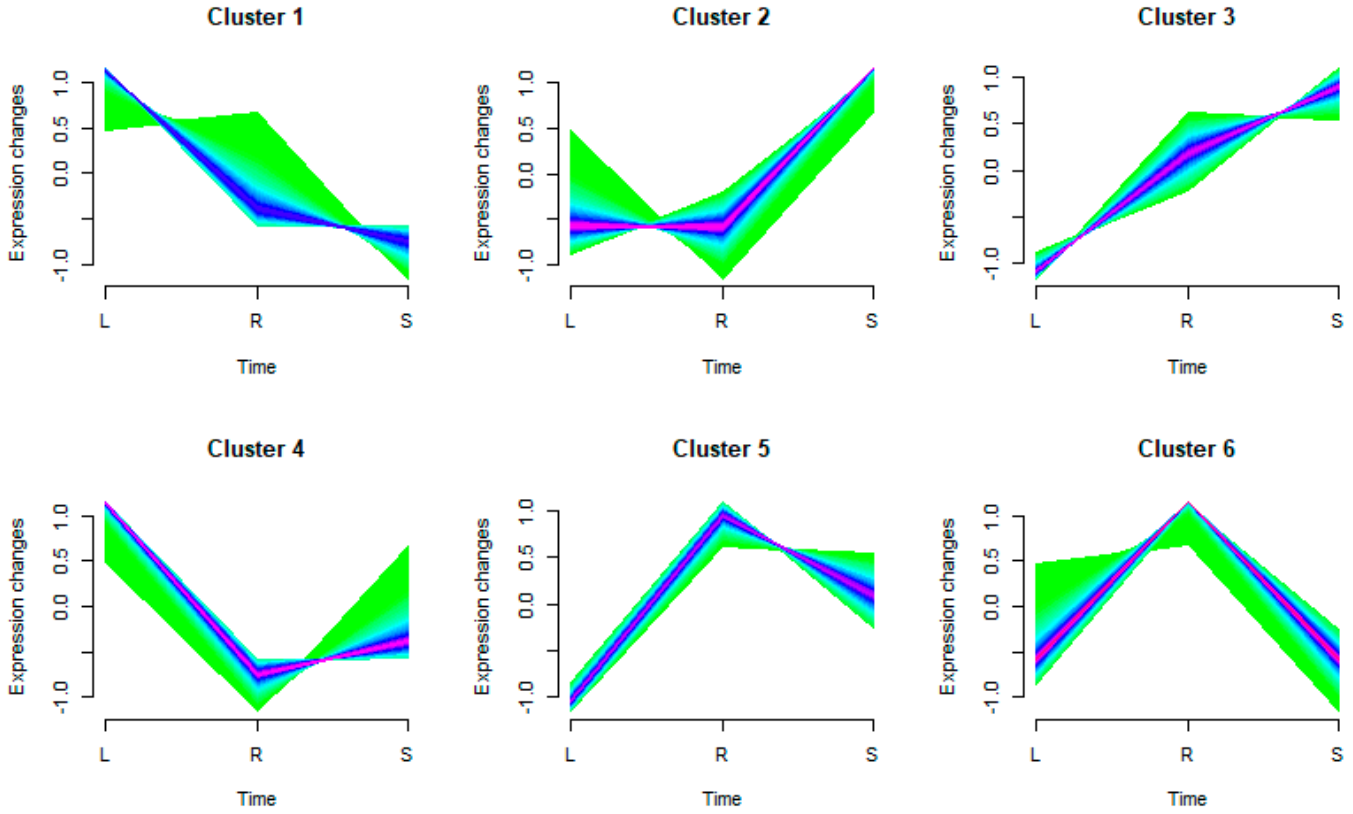

Figure 4. Genes were clustered according to their expression patterns in the transcriptome data from three tissue samples. The shades of color represent the density of genes, the darker the color, the more genes are concentrated. The image was generated using TBtools [14] with default parameters.

For an in-depth understanding of the functional role of the grouped genes, exhaustive annotations were carried out (Figure S4). Group 3 and group 4 genes—highly expressed in the leaves—showed a rich presence in metabolic pathways linked to plant hormone signaling, photosynthesis, and metabolism (Figure S4C,D). Meanwhile, the genes from group 6, mainly found expressed in the roots, indicated an enrichment in energy-related metabolic pathways, including polysaccharide and lipid metabolism (Figure S4F).

A tissue specificity index analysis was applied to delve into tissue-specific genes in the roots, stems, and leaves using a Perl script. Genes with TAU values exceeding 0.5 were deemed to show tissue-specific expression. In total, 11,868 genes exhibited unique tissue specificity. Such tissue-specific genes were consistently distributed over all seven chromosomes (Figure 5A), and their expression levels were illustrated using a heatmap (Figure 5B). In light of their high expression, functional annotations of genes using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses elucidated the tissue-specific functions. For instance, of the 4540 genes identified as being specifically expressed in the stems, functional annotation highlighted their role in plant hormone signaling and genetic signal transduction.

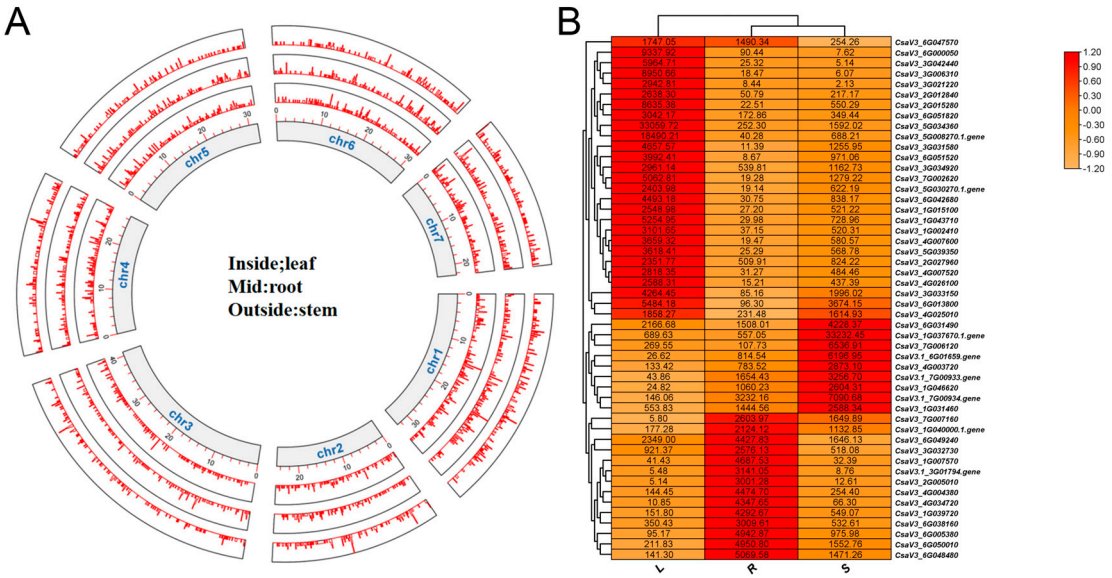

Figure 5. Tissue-specific genes located across the seven chromosomes (**A**), and heatmap showing the expression levels of genes expressed in specific tissues: leaf, root, and stem (**B**). The images were generated using TBtools [14] with default parameters.

2.6. Transcripts with Diverse Expression Levels across Different Tissues

The analysis of a total of 17,575 genes revealed an impressive subset of 14,185 genes associated with KEGG information, having consistently shown an average TPM value greater than 1 across all three tissue types. These annotated genes were primarily engaged in crucial metabolic pathways, which include polysaccharide metabolism, lipid metabolism, plant hormone signaling, photosynthesis, and broader cellular metabolism. Interestingly, within these annotated genes, a noteworthy group of 237 were identified as transcription factors, and 148 were categorized as protein kinases. Some unique transcription factors and protein kinases demonstrated remarkably similar expression patterns, indicating near-identical expression levels in the roots, stems, and leaves (Figure 6).

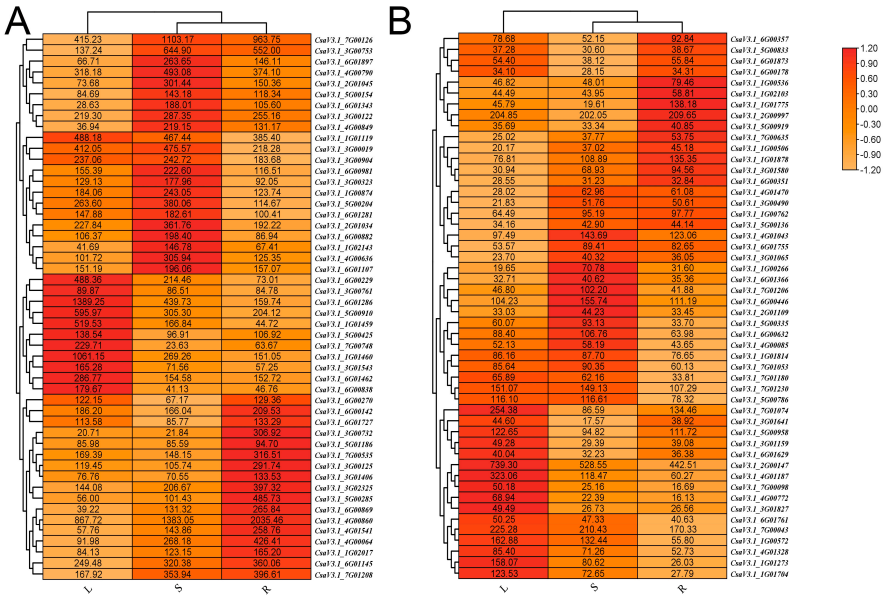

Figure 6. Heatmaps depicting the expression levels of transcription factor (**A**) and protein kinase (**B**) transcripts across three tissues. The images were generated using TBtools [14] with default parameters.

3. Discussion

3.1. Enhancing Functional Genomics Research into Cucumber through Accurate Gene Annotation

Cucumber, a major global vegetable, exhibits unique genetic characteristics, such as a rather compact genome size (approximately 260 Mb) and few chromosomes ($2n$ = 14). Prior research has emphasized the potency of top-tier genomic resources in pinpointing crucial Quantitative Trait Loci (QTLs) and potential genes related to agronomic traits [15–19]. Notably, fluctuations in the expression of five particular genes (*CsaV3_1G028310*, *CsaV3_2G006960*, *CsaV3_3G009560*, *CsaV3_5G031320*, and *CsaV3_6G031260*) are associated with stem thickness in cucumber [20]. Also, a specific 1449 bp sequence insert within CsMLO8's coding section at the pm-s5.1 location notably bolsters powdery mildew resistance in cucumber stems [21]. Past phenotypic research stretching across the previous century has elucidated that the fruit neck length in cucumber is predominantly governed by additive genetic components rather than environmental influences. This revelation, in conjunction with the identification of the *HECATE1* gene in cucumber (*CsHEC1*) and its pronounced correlation with fruit neck length, has greatly augmented our insight into the regulatory controls over the variation in fruit neck length in cucumber [22,23]. In essence, high-grade assembled genomic datasets function as invaluable tools for examining crop characteristics and gene operations, thereby galvanizing progress in both scientific investigations and industrial advancements within the vegetable sector.

The criticality of accurate gene model annotations for efficiently utilizing genomic information, enhancing the analysis precision of extensive biological datasets, and forwarding molecular biology experimentations cannot be overstated [24]. Previously, gene structures were predominantly annotated according to computational predictions, de novo prediction, homology-based prediction, and RNA-seq methodology included. However, it is important to note that these techniques often generate annotation files replete with numerous miscalculations, resulting from the inherent restrictions of the utilized algorithms. Historically, extensive algorithmic resources were essential for alignment and assembly tasks. Insufficient algorithmic precision in meeting intricate annotation demands inevitably results in alignment inaccuracies and reduced assembly precision. Particularly in species with intricate genomes, the error rates tend to be elevated, leading to frequent challenges in region annotation. To offset these anomalies, researchers have found employing an iterative reannotation process using varied methodologies and datasets beneficial for obtaining more expansive genomic and transcriptomic data. Encouragingly, these tactics have shown promising results in model plants and an array of economically valuable crops, such as Arabidopsis, orange, rice, maize, peach, and strawberry [4–8,25].

As well as automated annotation grounded in algorithms, manual adjustment via the visualization of RNA data stands as a notably pragmatic and efficacious mode for attaining precise annotation files. A case in point is the research on the Red5 kiwifruit genome, in which gene structures were meticulously modified employing WebApollo software [9]. Moreover, GSAman software has triumphed in annotating the peach genome and decoding the genetics of sweet potatoes [8,10]. This strategic method for annotation significantly boosts the accuracy of gene models. On the downside, the task of manipulating gene structures at the genomic level presents a strenuous endeavor, with the use of manual annotation software adding to the complexity in an appreciable way.

Nevertheless, it is widely recognized that the manual refinement of gene models enables the most precise predictions about gene behavior. Our study has led to the emergence of updated genome annotation, meticulously tailored and drawing upon 64 open-source RNA-seq datasets and 9 stranded-specific RNA-seq sets collected from root, stem, and leaf samples. Meticulous alignment of these collections with the cucumber 9930_V3 reference genome led to the calibration of 7721 earlier non-annotated gene structures. For validation, Sanger sequencing was conducted on 21 distinct genes, affirming their perfect congruence with the anticipated gene results. The outcomes of this study equip other scientists in similar domains with a potent and enhanced gene structure annotation file, driving forward innovation in the bioinformatics and molecular biological investigations of cucumber.

3.2. Gene Expression Profiles Provide Support for Gene Identification

Transcriptomic expression analysis serves as a fundamental tool for pinpointing gene candidates associated with various factors. A plethora of studies on fruit quality have underscored its significance by utilizing this method to discern primary gene candidates contributing to cucumber's resistance to powdery mildew [26]. Additionally, through the employment of expression analysis, an array of genes contributing to the buildup of compounds such as anthocyanins, carotenoids, terpenes, acids, and flavonoids across different fruit organs have been accurately established [27–33]. Offering a wealth of information on the gene expression levels across varied tissues and conditions, transcriptomic expression analysis vitally assists targeted screening for specific candidate genes.

This study annotated an impressive number of 24,145 genes, and 7721 of these genes experienced modifications or additions. This work carried out an RNA-seq expression analysis of roots, stems, and leaves and found that over 70%—precisely 17,575 transcripts—showed expression levels above 1 TPM. Interestingly, around 16% of genes, accounting for 3957, sustained their gene structure models without changes, despite having minimal expression. It is worth mentioning that the Version 1.0 genome's gene structures were computationally anticipated via de novo and homology-based approaches. This indicates these genes may stand unexpressed or their expression may occur under conditions not investigated in our study—highlighting the likely impact of including RNA-seq data from various tissues and experimental setups [25]. Analyzing RNA expression to reveal tissue-specific expression trends creates fundamental groundwork for handpicking appropriate genes contributing to particular traits. For instance, by using such tissue-specific expression layouts, several genes linked to wilt disease resistance have been detected in watermelon [34]. Similarly, a study into the tissue-specific spatiotemporal expression patterns in lotus disclosed genes associated with phenolic compounds [35]. Further, analysis of tissue-specific expression facilitated the pinpointing of genes associated with head formation in Chinese cabbage [36]. The findings from these studies highlight the potential role of genes expressed uniquely in different tissues in various metabolic pathways within those specific tissues.

4. Materials and Methods

4.1. Construction and Sequencing of RNA Libraries

To update the annotation of the cucumber 9930_V3 reference genome, 64 RNA datasets from diverse tissues and experimental conditions were employed, sourced from the National Center for Biotechnology Information (NCBI) website (Table S2). Moreover, new strand-specific RNA sequencing was conducted on root, stem, and leaf tissues, with three replicates per tissue, to document the transcript orientation.

4.2. Reannotation of the 9930_V3 Reference Genome

In order to enhance gene structure annotation and overcome algorithmic limitations, manual annotation based on RNA-seq data was employed. Initially, a genome index was created using hisat2-Build, and clean reads were aligned with the cucumber 9930_V3 reference genome using Hisat2. The resulting alignments were then converted into BAM files using SAMtools (v1.13) [37]. Subsequently, duplicate reads were eliminated using SAMtools' markdup function, and the BAM files were sorted and indexed. The sorted BAM files were then utilized for gene structure adjustment via GSAman (v.0.8.2, available online: https://tbtools.cowtransfer.com/S/a11146181df14f (accessed on 8 October 2023)). The structures of all the expressed genes underwent visual confirmation through inspection of the BAM files.

4.3. Validation of Novel Transcripts

Total RNA was isolated from the foliage of the Chinese North variety of the cultivated cucumber inbred line 9930 using the accuracy® Universal Plant RNA Extraction Kit (ACCURATE BIOTECHNOLOGY(HUMAN) Co., Ltd., Changsha, China). Subsequently, the isolated RNA was transcribed in the opposite direction to generate complementary

DNA (cDNA) through the use of the Evo M-MLV Plus 1st Strand cDNA Synthesis Kit. The coding regions (CDS) of 21 genes selected at random were amplified using the specific primers detailed in Table S1 and the accuracy High-Fidelity DNA polymerase. The resulting amplified segments underwent Sanger sequencing, followed by alignment with the novel annotation sequences for verification.

4.4. Functional Annotation of Genes in the Novel Annotation File

Functional annotation of genes in the new annotation file was conducted through the utilization of the eggNOG-mapper (v2.1.12), Kobas (v.3.0), iTAK Online (v.1.6), and Mercator (v.3.6) platforms [38–41]. Protein sequences underwent analysis on the eggNOG-mapper and Kobas websites using the default parameters. Alignment of the CDS sequences with databases such as TAIR10, SwissProt/UniProt plant proteins, Augustus Models (JGI Chlamy Release 4), TIGR 5 rice proteins, and the NCBI conserved domain database was performed using Mercator (v.3.6).

4.5. Gene Expression Profiles in the Novel Annotation File

RNA-seq data obtained from three tissue specimens underwent alignment with the 9930_V3 reference genome and normalization employing transcripts per million (TPM) values, which were informed by the updated gene structure annotation. Utilization of the Mfuzz package in the R programming language facilitated the categorization of the expressed genes into six distinct clusters. Subsequently, Perl scripts were employed to ascertain the tissue specificity coefficient for individual genes [42], with genes exhibiting a coefficient surpassing the threshold of 0.5 being selected for further analysis.

5. Conclusions

The unveiling of the high-quality 9930_V3 cucumber genome in 2019 denoted a significant landmark. To elevate the precision of the gene prediction model, comprehensive refinements of the genome structure annotation file were undertaken. This included detailed reannotation of the cucumber 9930_V3 reference genome through a robust dataset composed of 64 publicly accessed RNA-seq datasets and 9 strand-specific RNA-seq datasets. After stringent validation via GO, KEGG, BUSCO, and RT-PCR analyses, the enhanced superiority of the newly annotated file to its predecessor was unequivocally established. It is worth mentioning that this research facilitated the amendment and incorporation of 7721 gene structures into the improved annotation file. By taking advantage of this augmented annotation, the research performed a deeper investigation into the expression profiles of genes in the tissues of the roots, stems, and leaves. This enhanced version of the cucumber 9930_V3 genome annotation gives researchers in bioinformatics and molecular biology an extraordinarily accurate gene structure prediction model, paving the way for more profound insights into cucumber biology.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/plants13121604/s1. Figure S1: Gene models verified using RT-PCR; Figure S2: Gene models adjusted according to RNA-seq mapping result; Figure S3: Expressed genes in RNA-seq data derived from three tissues; Figure S4: KEGG annotation of genes in Cluster 1 to Cluster 6; Figure S5: Tissue-specific expressed genes in different tissues; Table S1: Primers used in this study; Table S2: Published data used in this study; Table S3: GO terms of genes in the 9930_V3.1 cucumber genome; Table S4: KEGG annotation of genes in the 9930_V3.1 cucumber genome; Table S5: Transcription factors in the 9930_V3.1 cucumber genome; Table S6: Protein kinases identified in the 9930_V3.1 genome; Table S7: Gene functional annotation of all genes in the 9930_V3.1 genome; Table S8: Comparison of the gene models between 9930_V3 and 9930_V3.1 genomes; Table S9: Adjusted or newly added genes in the 9930_V3.1 annotation; Table S10: The expression profiles of all genes in the three tissues; Table S11: Tissue-specific expressed genes identified in the three tissues.

Author Contributions: Conceptualization, W.D. and X.Y.; methodology, W.D.; software, W.D.; validation, W.D. and R.L.; formal analysis, W.D., L.X. and R.L.; investigation, W.D., X.W. and D.J.; resources, W.D.; data curation, W.D.; writing—original draft preparation, W.D.; writing—review and editing, L.X., X.Z., Y.P., R.Z., X.Y. and J.C.; visualization, W.D.; supervision, X.Y.; project administration, X.Y.; funding acquisition, X.Y., R.Z. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China (2023YFF1002000), the Natural Science Foundation of Jiangsu Province (BK2022148), and the National Natural Science Foundation of China (32372697). This project was also funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Data Availability Statement: The raw reads of nine strand-specific RNA-seq data were submitted to the National Genomics Data Center under the accession number PRJCA025018. The updated version of the annotation in gff3 file format can be found in the uploaded Supplementary Files.

Acknowledgments: We thank Chengjie Chen for his help with the use of GSAman software. This project is supported by the Bioinformatics Center of Nanjing Agricultural University.

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

References

1. Haas, B.J.; Wortman, J.R.; Ronning, C.M.; Hannick, L.I.; Smith, R.K.; Maiti, R.; Chan, A.P.; Yu, C.; Farzad, M.; Wu, D.; et al. Complete Reannotation of the Arabidopsis Genome: Methods, Tools, Protocols and the Final Release. *BMC Biol.* **2005**, *3*, 7. [CrossRef]
2. Li, Q.; Li, H.; Huang, W.; Xu, Y.; Zhou, Q.; Wang, S.; Ruan, J.; Huang, S.; Zhang, Z. A Chromosome-Scale Genome Assembly of Cucumber (*Cucumis sativus* L.). *GigaScience* **2019**, *8*, giz072. [CrossRef] [PubMed]
3. Li, Z.; Zhang, Z.; Yan, P.; Huang, S.; Fei, Z.; Lin, K. RNA-Seq Improves Annotation of Protein-Coding Genes in the Cucumber Genome. *BMC Genom.* **2011**, *12*, 540. [CrossRef]
4. Andorf, C.M.; Cannon, E.K.; Portwood, J.L.; Gardiner, J.M.; Harper, L.C.; Schaeffer, M.L.; Braun, B.L.; Campbell, D.A.; Vinnakota, A.G.; Sribalusu, V.V.; et al. MaizeGDB Update: New Tools, Data and Interface for the Maize Model Organism Database. *Nucleic Acids Res.* **2016**, *44*, D1195–D1201. [CrossRef] [PubMed]
5. Cheng, C.-Y.; Krishnakumar, V.; Chan, A.P.; Thibaud-Nissen, F.; Schobel, S.; Town, C.D. Araport11: A Complete Reannotation of the Arabidopsis Thaliana Reference Genome. *Plant J. Cell Mol. Biol.* **2017**, *89*, 789–804. [CrossRef]
6. Liu, H.; Wang, X.; Liu, S.; Huang, Y.; Guo, Y.-X.; Xie, W.-Z.; Liu, H.; Tahir Ul Qamar, M.; Xu, Q.; Chen, L.-L. Citrus Pan-Genome to Breeding Database (CPBD): A Comprehensive Genome Database for Citrus Breeding. *Mol. Plant* **2022**, *15*, 1503–1505. [CrossRef]
7. Sang, J.; Zou, D.; Wang, Z.; Wang, F.; Zhang, Y.; Xia, L.; Li, Z.; Ma, L.; Li, M.; Xu, B.; et al. IC4R-2.0: Rice Genome Reannotation Using Massive RNA-Seq Data. *Genom. Proteom. Bioinform.* **2020**, *18*, 161–172. [CrossRef]
8. Zhang, H.; Feng, B.; Wang, C.; Lian, X.; Wang, X.; Zheng, X.; Cheng, J.; Wang, W.; Zhang, L.; Ye, X.; et al. Manually Annotated Gene Prediction of the CN14 Peach Genome. *Sci. Hortic.* **2023**, *321*, 112242. [CrossRef]
9. Pilkington, S.M.; Crowhurst, R.; Hilario, E.; Nardozza, S.; Fraser, L.; Peng, Y.; Gunaseelan, K.; Simpson, R.; Tahir, J.; Deroles, S.C.; et al. A Manually Annotated *Actinidia chinensis* Var. *chinensis* (Kiwifruit) Genome Highlights the Challenges Associated with Draft Genomes and Gene Prediction in Plants. *BMC Genom.* **2018**, *19*, 257. [CrossRef]
10. Liang, B.; Zhou, Y.; Liu, T.; Wang, M.; Liu, Y.; Liu, Y.; Li, Y.; Zhu, G. Genome Reannotation of the Sweetpotato (*Ipomoea batatas* (L.) Lam.) Using Extensive Nanopore and Illumina-Based RNA-Seq Datasets. *Trop. Plants* **2024**, *3*, e008. [CrossRef]
11. Huang, S.; Li, R.; Zhang, Z.; Li, L.; Gu, X.; Fan, W.; Lucas, W.J.; Wang, X.; Xie, B.; Ni, P.; et al. The Genome of the Cucumber, *Cucumis sativus* L. *Nat. Genet.* **2009**, *41*, 1275–1281. [CrossRef] [PubMed]
12. Qi, J.; Liu, X.; Shen, D.; Miao, H.; Xie, B.; Li, X.; Zeng, P.; Wang, S.; Shang, Y.; Gu, X.; et al. A Genomic Variation Map Provides Insights into the Genetic Basis of Cucumber Domestication and Diversity. *Nat. Genet.* **2013**, *45*, 1510–1515. [CrossRef] [PubMed]
13. Shang, Y.; Ma, Y.; Zhou, Y.; Zhang, H.; Duan, L.; Chen, H.; Zeng, J.; Zhou, Q.; Wang, S.; Gu, W.; et al. Plant Science. Biosynthesis, Regulation, and Domestication of Bitterness in Cucumber. *Science* **2014**, *346*, 1084–1088. [CrossRef] [PubMed]
14. Chen, C.; Wu, Y.; Li, J.; Wang, X.; Zeng, Z.; Xu, J.; Liu, Y.; Feng, J.; Chen, H.; He, Y.; et al. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant.* **2023**, *16*, 1733–1742. [CrossRef] [PubMed]
15. Palmer, D.; Fabris, F.; Doherty, A.; Freitas, A.A.; de Magalhães, J.P. Ageing Transcriptome Meta-Analysis Reveals Similarities and Differences between Key Mammalian Tissues. *Aging* **2021**, *13*, 3313–3341. [CrossRef] [PubMed]
16. Huang, H.; Du, Y.; Long, Z.; Li, Y.; Kong, W.; Wang, H.; Wei, A.; Du, S.; Yang, R.; Li, J.; et al. Fine Mapping of a Novel QTL CsFSG1 for Fruit Skin Gloss in Cucumber (*Cucumis sativus* L.). *Mol. Breed. New Strateg. Plant Improv.* **2022**, *42*, 25. [CrossRef] [PubMed]
17. Li, X.; Lin, S.; Xiang, C.; Liu, W.; Zhang, X.; Wang, C.; Lu, X.; Liu, M.; Wang, T.; Liu, Z.; et al. CUCUME: An RNA Methylation Database Integrating Systemic mRNAs Signals, GWAS and QTL Genetic Regulation and Epigenetics in Different Tissues of Cucurbitaceae. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 837–846. [CrossRef] [PubMed]

18. Lin, Y.-C.; Mansfeld, B.N.; Tang, X.; Colle, M.; Chen, F.; Weng, Y.; Fei, Z.; Grumet, R. Identification of QTL Associated with Resistance to Phytophthora Fruit Rot in Cucumber (*Cucumis sativus* L.). *Front. Plant Sci.* **2023**, *14*, 1281755. [CrossRef] [PubMed]
19. Sun, J.; Nie, J.; Xiao, T.; Guo, C.; Lv, D.; Zhang, K.; He, H.-L.; Pan, J.; Cai, R.; Wang, G. CsPM5.2, a Phosphate Transporter Protein-like Gene, Promotes Powdery Mildew Resistance in Cucumber. *Plant J. Cell Mol. Biol.* **2024**, *117*, 1487–1502. [CrossRef]
20. Zhang, R.-J.; Liu, B.; Song, S.-S.; Salah, R.; Song, C.-J.; Xia, S.-W.; Hao, Q.; Liu, Y.-J.; Li, Y.; Lai, Y.-S. Lipid-Related Domestication Accounts for the Extreme Cold Sensitivity of Semiwild and Tropic Xishuangbanna Cucumber (*Cucumis sativus* L. Var. Xishuangbannanesis). *Int. J. Mol. Sci.* **2024**, *25*, 79. [CrossRef]
21. Yang, Y.; Dong, S.; Miao, H.; Liu, X.; Dai, Z.; Li, X.; Gu, X.; Zhang, S. Genome-Wide Association Studies Reveal Candidate Genes Related to Stem Diameter in Cucumber (*Cucumis sativus* L.). *Genes* **2022**, *13*, 1095. [CrossRef] [PubMed]
22. Dong, S.; Liu, X.; Han, J.; Miao, H.; Beckles, D.M.; Bai, Y.; Liu, X.; Guan, J.; Yang, R.; Gu, X.; et al. CsMLO8/11 Are Required for Full Susceptibility of Cucumber Stem to Powdery Mildew and Interact with CsCRK2 and CsRbohD. *Hortic. Res.* **2024**, *11*, uhad295. [CrossRef] [PubMed]
23. Wang, Z.; Zhou, Z.; Wang, L.; Yan, S.; Cheng, Z.; Liu, X.; Han, L.; Chen, G.; Wang, S.; Song, W.; et al. The CsHEC1-CsOVATE Module Contributes to Fruit Neck Length Variation via Modulating Auxin Biosynthesis in Cucumber. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2209717119. [CrossRef] [PubMed]
24. Xu, X.; Wei, C.; Liu, Q.; Qu, W.; Qi, X.; Xu, Q.; Chen, X. The Major-Effect Quantitative Trait Locus Fnl7.1 Encodes a Late Embryogenesis Abundant Protein Associated with Fruit Neck Length in Cucumber. *Plant Biotechnol. J.* **2020**, *18*, 1598–1609. [CrossRef]
25. Campbell, M.S.; Yandell, M. An Introduction to Genome Annotation. *Curr. Protoc. Bioinforma.* **2015**, *52*, 4.1.1–4.1.17. [CrossRef] [PubMed]
26. Li, Y.; Pi, M.; Gao, Q.; Liu, Z.; Kang, C. Updated Annotation of the Wild Strawberry Fragaria Vesca V4 Genome. *Hortic. Res.* **2019**, *6*, 61. [CrossRef] [PubMed]
27. Berg, J.A.; Hermans, F.W.K.; Beenders, F.; Lou, L.; Vriezen, W.H.; Visser, R.G.F.; Bai, Y.; Schouten, H.J. Analysis of QTL DM4.1 for Downy Mildew Resistance in Cucumber Reveals Multiple subQTL: A Novel RLK as Candidate Gene for the Most Important subQTL. *Front. Plant Sci.* **2020**, *11*, 569876. [CrossRef] [PubMed]
28. Chen, J.; Yuan, Z.; Zhang, H.; Li, W.; Shi, M.; Peng, Z.; Li, M.; Tian, J.; Deng, X.; Cheng, Y.; et al. Cit1,2RhaT and Two Novel CitdGlcTs Participate in Flavor-Related Flavonoid Metabolism during Citrus Fruit Development. *J. Exp. Bot.* **2019**, *70*, 2759–2771. [CrossRef] [PubMed]
29. Hu, D.-G.; Sun, C.-H.; Ma, Q.-J.; You, C.-X.; Cheng, L.; Hao, Y.-J. MdMYB1 Regulates Anthocyanin and Malate Accumulation by Directly Facilitating Their Transport into Vacuoles in Apples. *Plant Physiol.* **2016**, *170*, 1315–1330. [CrossRef]
30. Ma, B.; Liao, L.; Fang, T.; Peng, Q.; Ogutu, C.; Zhou, H.; Ma, F.; Han, Y. A Ma10 Gene Encoding P-Type ATPase Is Involved in Fruit Organic Acid Accumulation in Apple. *Plant Biotechnol. J.* **2019**, *17*, 674–686. [CrossRef]
31. Tian, Y.; Thrimawithana, A.; Ding, T.; Guo, J.; Gleave, A.; Chagné, D.; Ampomah-Dwamena, C.; Ireland, H.S.; Schaffer, R.J.; Luo, Z.; et al. Transposon Insertions Regulate Genome-Wide Allele-Specific Expression and Underpin Flower Colour Variations in Apple (*Malus* spp.). *Plant Biotechnol. J.* **2022**, *20*, 1285–1297. [CrossRef] [PubMed]
32. Zhang, H.; Chen, J.; Peng, Z.; Shi, M.; Liu, X.; Wen, H.; Jiang, Y.; Cheng, Y.; Xu, J.; Zhang, H. Integrated Transcriptomic and Metabolomic Analysis Reveals a Transcriptional Regulation Network for the Biosynthesis of Carotenoids and Flavonoids in "Cara Cara" Navel Orange. *BMC Plant Biol.* **2021**, *21*, 29. [CrossRef] [PubMed]
33. Zhang, H.; Chen, M.; Wen, H.; Wang, Z.; Chen, J.; Fang, L.; Zhang, H.; Xie, Z.; Jiang, D.; Cheng, Y.; et al. Transcriptomic and Metabolomic Analyses Provide Insight into the Volatile Compounds of Citrus Leaves and Flowers. *BMC Plant Biol.* **2020**, *20*, 7. [CrossRef] [PubMed]
34. Xuan, C.; Feng, M.; Li, X.; Hou, Y.; Wei, C.; Zhang, X. Genome-Wide Identification and Expression Analysis of Chitinase Genes in Watermelon under Abiotic Stimuli and Fusarium Oxysporum Infection. *Int. J. Mol. Sci.* **2024**, *25*, 638. [CrossRef] [PubMed]
35. Liu, G.; Fu, J.; Wang, L.; Fang, M.; Zhang, W.; Yang, M.; Yang, X.; Xu, Y.; Shi, L.; Ma, X.; et al. Diverse O-Methyltransferases Catalyze the Biosynthesis of Floral Benzenoids That Repel Aphids from the Flowers of Waterlily Nymphaea Prolifera. *Hortic. Res.* **2023**, *10*, uhad237. [CrossRef] [PubMed]
36. Yue, X.; Su, T.; Xin, X.; Li, P.; Wang, W.; Yu, Y.; Zhang, D.; Zhao, X.; Wang, J.; Sun, L.; et al. The Adaxial/Abaxial Patterning of Auxin and Auxin Gene in Leaf Veins Functions in Leafy Head Formation of Chinese Cabbage. *Front. Plant Sci.* **2022**, *13*, 918112. [CrossRef] [PubMed]
37. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-Level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. [CrossRef] [PubMed]
38. Bu, D.; Luo, H.; Huo, P.; Wang, Z.; Zhang, S.; He, Z.; Wu, Y.; Zhao, L.; Liu, J.; Guo, J.; et al. KOBAS-i: Intelligent Prioritization and Exploratory Visualization of Biological Functions for Gene Enrichment Analysis. *Nucleic Acids Res.* **2021**, *49*, W317–W325. [CrossRef] [PubMed]
39. Cantalapiedra, C.P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **2021**, *38*, 5825–5829. [CrossRef]
40. Lohse, M.; Nagel, A.; Herter, T.; May, P.; Schroda, M.; Zrenner, R.; Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: A Fast and Simple Web Server for Genome Scale Functional Annotation of Plant Sequence Data. *Plant Cell Environ.* **2014**, *37*, 1250–1258. [CrossRef]

41. Zheng, Y.; Jiao, C.; Sun, H.; Rosli, H.G.; Pombo, M.A.; Zhang, P.; Banf, M.; Dai, X.; Martin, G.B.; Giovannoni, J.J.; et al. 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] [PubMed]
42. Zhou, H.; Liao, L.; Xu, S.; Ren, F.; Zhao, J.; Ogutu, C.; Wang, L.; Jiang, Q.; Han, Y. Two Amino Acid Changes in the R3 Repeat Cause Functional Divergence of Two Clustered MYB10 Genes in Peach. *Plant Mol. Biol.* **2018**, *98*, 169–183. [CrossRef] [PubMed]