forests

MDPI

Article

Single-Molecule Long-Read Sequencing of *Zanthoxylum bungeanum* Maxim. Transcriptome: Identification of Aroma-Related Genes

Jieyun Tian [1,2], Shijing Feng [1,2], Yulin Liu [1,2], Lili Zhao [1,2], Lu Tian [1,2], Yang Hu [1,2], Tuxi Yang [1,2] and Anzhi Wei [1,2,*]

1 College of Forestry, Northwest A&F University, Yangling, Xianyang 712100, China; tjytianjieyun@163.com (J.T.); shijing0554010112@163.com (S.F.); lyl12504001@126.com (Y.L.); zllblue@126.com (L.Z.); t1anlu@nwafu.edu.cn (L.T.); huyang19960312@163.com (Y.H.); yangtuxi@nwafu.edu.cn (T.Y.)

2 Research Centre for Engineering and Technology of *Zanthoxylum* State Forestry Administration, Yangling, Xianyang 712100, China

* Correspondence: weianzhi@nwafu.edu.cn; Tel: +86-29-8708-2211

Received: 12 November 2018; Accepted: 9 December 2018; Published: 12 December 2018

Abstract: *Zanthoxylum bungeanum* Maxim. is an economically important tree species that is resistant to drought and infertility, and has potential medicinal and edible value. However, comprehensive genomic data are not yet available for this species, limiting its potential utility for medicinal use, breeding programs, and cultivation. Transcriptome sequencing provides an effective approach to remedying this shortcoming. Herein, single-molecule long-read sequencing and next-generation sequencing approaches were used in parallel to obtain transcript isoform structure and gene functional information in *Z. bungeanum*. In total, 282,101 reads of inserts (ROIs) were identified, including 134,074 full-length non-chimeric reads, among which 65,711 open reading frames (ORFs), 50,135 simple sequence repeats (SSRs), and 1492 long non-coding RNAs (lncRNAs) were detected. Functional annotation revealed metabolic pathways related to aroma components and color characteristics in *Z. bungeanum*. Unexpectedly, 30 transcripts were annotated as genes involved in regulating the pathogenesis of breast and colorectal cancers. This work provides a comprehensive transcriptome resource for *Z. bungeanum*, and lays a foundation for the further investigation and utilization of *Zanthoxylum* resources.

Keywords: aroma metabolic pathway; long non-coding RNA; single-molecule real-time sequencing; *Zanthoxylum bungeanum* Maxim.

1. Introduction

Zanthoxylum bungeanum Maxim. is a member of the evergreen and deciduous trees and shrubs of the citrus (Rutaceae) family that is native to China, where it is mainly used as a condiment and oilseed crop [1]. Although the popularity of *Z. bungeanum* as a condiment is relatively low outside of China, its medicinal value merits attention. This species has been used as an ancient herbal medicine in China, and its pericarp shows anti-inflammatory and anti-bacterial activities [2], relieves arthritis [3], protects against gastric mucosal damage [4], lowers blood lipid levels [5], regulates antithrombotic effects [6], and enhances immunity [7]. A recent study found that collagen extracted from *Z. schinifolium* exerts positive anti-tuberculosis effects [8]. More interestingly, genes associated with breast and colorectal cancers were identified during the functional annotation of *Z. bungeanum* in this work. The discovery of artemisinin is a powerful reminder of the need to devote more attention to seemingly niche plants that possess unique medicinal qualities in order to tap their potential value.

Despite its long history of use, *Z. bungeanum* is a plant that remains poorly understood. This species grows on sloping land and benefits from strong resistance to stress. However, while physiological studies of this phenomenon have been performed, in-depth molecular knowledge is lacking. In addition, *Z. bungeanum* is an apomictic plant whose male flowers are not found in cultivars [9,10], yet research on this aspect is also scarce. Currently, available *Z. bungeanum* resources are mostly cultivars. It remains difficult to find wild species, which impedes the origin and classification analysis. Moreover, *Z. bungeanum* contains numerous chromosomes and its ploidy is still unclear. Due to its complex genetic information, the genomics research of *Z. bungeanum* has progressed slowly. The lack of genomic and high-quality transcriptome information acts as a barrier to such studies and hampers a full exploration of the medicinal value of this plant.

The advent of single-molecule real-time (SMRT) sequencing technology has injected new vitality into genome and transcriptome research, which provides useful gene structure and function information for many organisms [11,12]. High-quality SMRT transcriptome sequencing is particularly powerful for investigating alternative splicing, alternative polyadenylation, novel genes, and non-coding RNAs [13]. Most non-model organisms lack genomic data, and full-length transcripts derived using the PacBio platform can greatly facilitate basic and applied research on gene function, gene expression regulation, and evolutionary relationships [14–16]. Next-generation sequencing (NGS) technology also has advantages, including high-throughput, improved accuracy, and almost universal application [17–19]. NGS and SMRT sequencing approaches complement each other, and their combination facilitates more complete transcriptome resources [20].

The present study reports the first full-length transcriptome of *Z. bungeanum* from five tissues using SMRT sequencing technology. Short reads obtained by NGS were used to correct the transcript isoforms that were obtained using SMRT sequencing. Annotation of the *Z. bungeanum* transcript isoforms sheds light on the complexity of the molecular mechanisms underpinning the medicinal and edible functions. This study provides genetic resources and transcriptome information to support further research on *Z. bungeanum*.

2. Materials and Methods

2.1. Plant Materials and Sample Preparation

The *Z. bungeanum* cultivar "Fuguhuajiao" was grown in an experimental field at the Research Centre for Engineering and Technology of *Zanthoxylum*, State Forestry Administration, Northwest A&F University, Fengxian, Shaanxi Province, China. Shoots (sprouting stage), leaves (leaf expansion stage), stems (branching stage), flowers (flowering stage), and fruits (fruiting stage) were collected in approximately equal weights from the east, west, south, and north directions from a 6-year-old plant (Figure 1). Collected samples were quickly frozen in liquid nitrogen and stored at −80 °C until used.

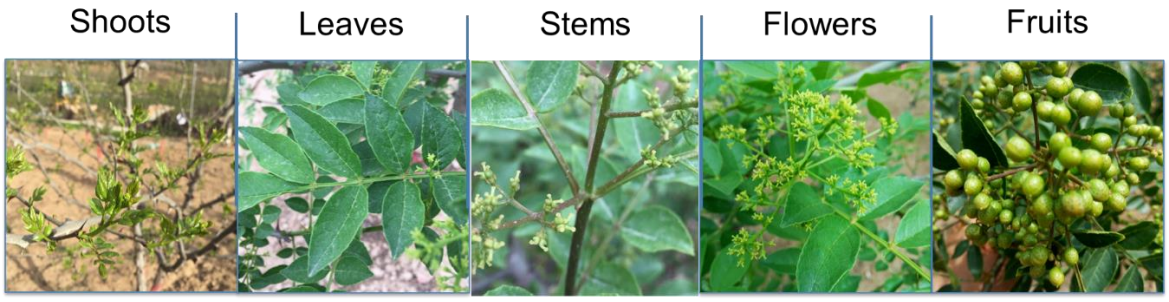

Figure 1. *Zanthoxylum bungeanum* Maxim. samples collected from the shoots, leaves, stems, flowers, and fruits of a 6-year-old individual in this study.

2.2. RNA Preparation

The total RNA was extracted from each tissue using an RNeasy Plus Mini Kit (Qiagen, Valencia, CA, USA). The RNA quality, integrity, and quantities were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and an Agilent Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, USA). For all samples, the RNA integrity number was >7.5. Qualified RNA samples were subsequently used for constructing complementary DNA (cDNA) libraries.

2.3. PacBio cDNA Library Construction and Sequencing

The cDNA was synthesized using mixed RNA from all five tissues with a SMARTer PCR cDNA Synthesis Kit (Clontech, Palo Alto, CA, USA; Cat. No. 634926). Full-length cDNA fragments were screened using a BluePippin Size Selection System (Sage Science, Beverly, MA, USA), and three cDNA libraries (1−2 kb, 2−3 kb, and >3 kb) were constructed. Selected full-length cDNA was re-amplified by PCR and the SMRT bell hairpin loop was ligated to the cDNA. A Qubit 2.0 fluorimeter (Life Technologies, Carlsbad, CA, USA) was then used for accurate quantification of the libraries after a secondary screening with the BluePippin system. Library size was measured using an Agilent Bioanalyzer 2100 system, and sequencing was performed on a PacBio RSII instrument (Pacific Bioscience, Menlo Park, CA, USA).

2.4. Illumina cDNA Library Construction and Sequencing

The mRNA was purified and enriched from mixed total RNA using oligo (dT) magnetic beads. A fragmentation buffer was added to break the mRNA into random fragments using divalent cations under elevated temperatures. The resultant mRNA fragments were used as templates to synthesize the first-strand cDNA using random hexamers. The second cDNA strand was synthesized using buffer, deoxyribonucleotide triphosphates (dNTPs), Ribonuclease H (RNase H), and DNA polymerase I, after which the cDNA was purified using AMPure XP (Beckman, Brea, CA, USA) beads. Double-stranded cDNAs were further end-repaired, poly(A)-tailed, and ligated to sequencing adapters. The purified and repaired double-stranded cDNA fragments were size-selected again with AMPure XP beads and enriched by PCR. Qualifying libraries were sequenced using an Illumina HiSeq instrument (Illumina Inc., San Diego, CA, USA).

2.5. Quality Filtering and Error Correction of PacBio Long Reads

Amplified sequences obtained by SMRT sequencing which were smaller than 50 bp and had a quality score of less than 0.75 were removed by filtering. The application of full passes that were $\geq$0 resulted in the extraction of reads of inserts (ROIs) from filtered subreads. ROIs were separated into full-length and non-full-length reads by detecting whether 5′ primers, 3′ primers, and polyA tails were included in ROIs. Iterative Clustering for Error Correction (ICE) [21] was then used to obtain consensus isoforms, and high-quality consensus isoforms from ICE were polished using Quiver. Clean reads obtained after the removal of adapters and low-quality reads from raw Illumina NGS data were used to correct low-quality full-length transcript isoforms using Proovread software [22]. High-quality and corrected low-quality transcript isoforms were confirmed as de-redundant using CD-HIT software [23] to obtain non-redundant isoforms.

2.6. Analysis of Detecting Coding Sequence (CDS), Simple Sequence Repeat (SSR), and Long Non-Coding RNA (lncRNA) Features

Each open reading frame (ORF) detected by the TransDecoder v.3.0.0 (https://github.com/TransDecoder/TransDecoder/releases) was defined as a putative detecting coding sequence (CDS). Predicted CDS types were classified as complete (both start and stop codons were predicted), 5′ end

(5′-partial, only the start codon was predicted), 3′ end (3′-partial, only the stop codon was predicted), or internal segment (neither the start codons nor the stop codons were predicted).

Transcripts >500 bp were screened from the identified transcripts, and simple sequence repeat (SSR) analysis was performed on the selected transcripts using the MIcroSAtellite identification tool (MISA; http://pgrc.ipk-gatersleben.de/misa/) [24] from which seven types of SSR (mono-nucleotide, di-nucleotide, tri-nucleotide, tetra-nucleotide, penta-nucleotide, hexa-nucleotide, and compound SSR) were detected.

Transcripts with lengths >200 nt and more than two exons were selected as long non-coding RNA (lncRNA) candidates. Four computational approaches—the Coding Potential Calculator (CPC), Coding-Non-Coding Index (CNCI), Coding Potential Assessment Tool (CPAT), and Protein family database (Pfam)—were used to distinguish protein-coding and non-coding genes [25–27].

2.7. Functional Annotation

The obtained non-redundant transcript sequences were aligned using eight NCBI protein and nucleotide databases, namely non-redundant protein sequences (NR), Evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG), Protein family (Pfam), Swiss-Prot, Gene Ontology (GO), euKaryotic Ortholog Groups (KOG), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups (COG). Gene functional annotation was performed using BLAST v.2.2.26 [28].

3. Results

3.1. General Properties of Single-Molecule Long-Reads

The total RNA from five tissues (shoots, leaves, stems, flowers, and fruits) was combined in equal amounts to acquire a wide coverage of full-length transcripts in *Z. bungeanum* for single-molecule long-read sequencing. Three cDNA libraries of different sizes (1−2 kb, 2−3 kb, and >3 kb) were constructed and sequenced using a PacBio RSII sequencing system. Five SMRT cells generated 113,885, 102,613, and 65,603 ROIs from the three libraries. A total of 282,101 ROIs were generated, and >47% (134,400) were full-length ROIs, based on the presence of five primers, three primers, and poly(A) tails. The average length of the ROIs was 2819 bp (Table 1 and Figure 2).

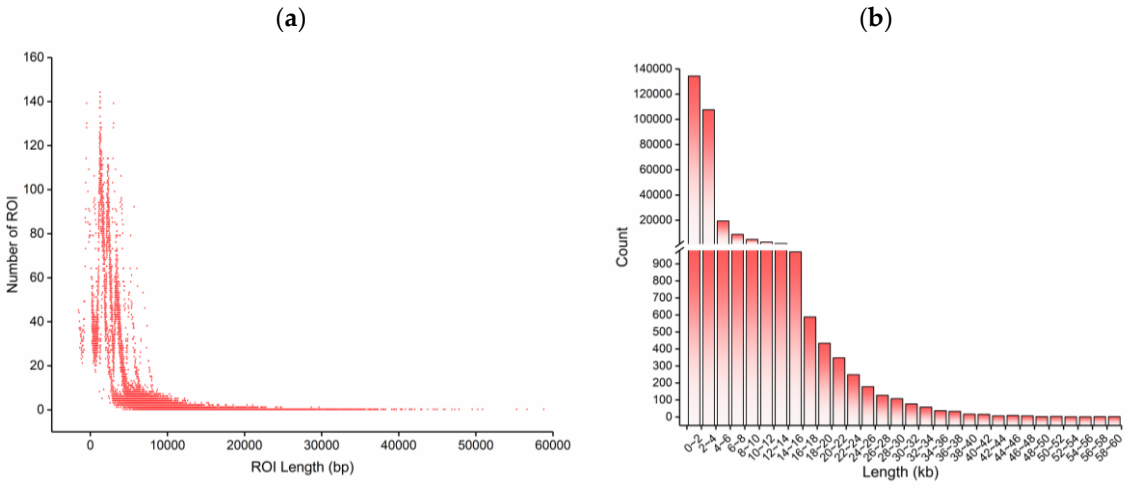

Figure 2. Length distribution of 282,101 reads of inserts (ROIs) from *Z. bungeanum*; (**a**) number and length distribution of ROIs; (**b**) histogram of ROI read length frequency distribution.

Table 1. ROI statistics for *Z. bungeanum.*

cDNA Size	1–2 kb	2–3 kb	>3 kb	Total
No. of ROIs	113,885	102,613	65,603	282,101
No. of five primer reads	62,689	56,205	40,479	159,373
No. of three primer reads	69,924	60,719	42,195	441,474
No. of poly(A) reads	68,221	60,120	42,172	170,513
No. of filtered short reads	15,681	8132	1886	25,699
No. of full-length reads	52,960	46,744	34,696	134,400
No. of full-length non-chimeric reads	52,785	46,629	34,660	134,074
Full-length percentage (%)	46.50	45.55	52.89	47.64
Mean length of ROIs	2000	2891	3567	2819

3.2. Acquisition of High-Quality Sequences and Error Correction of Long-Reads Using Illumina Data

In order to generate high-quality transcripts, an ICE-based algorithm was used to iteratively cluster sequences. After clustering similar sequences, consensus isoforms were obtained for each cluster. In combination with non-full-length sequences, we used the quiver program to correct consensus sequences for each cluster, and finally obtained high-quality transcripts with an accuracy rate of >99%. We eventually obtained 75,973 consensus isoform sequences with 58,445 high-quality transcript isoforms (Table 2).

Table 2. Results of iterative clustering for error correction in *Z. bungeanum.*

Size	No. of Consensus Isoforms	Average Consensus Isoforms Read Length	No. of Polished High-Quality Isoforms	No. of Polished Low-Quality Isoforms	Percentage of Polished High-Quality Isoforms (%)
0–1 kb	1821	903	1608	213	88.30
1–2 kb	26,917	1442	23,043	3874	85.61
2–3 kb	24,913	2387	19,181	5732	76.99
3–6 kb	21,823	3582	14,603	7220	66.92
>6 kb	499	8757	10	489	2.00
Total	75,973	3414	58,445	17,528	-

In parallel, 23,869,770 clean reads were sequenced using an Illumina platform and subsequently used for correcting low-quality consensus isoforms from the PacBio data. A total of 68,710 transcriptional sequences were obtained after the removal of redundant transcripts, which were suitable for further structural and functional analysis.

3.3. Predictions of ORFs, SSRs, and lncRNAs

A total of 65,711 ORFs were predicted, 52,331 of which were complete CDSs. The number and length distribution of proteins encoded by CDS regions are shown in Figure 3.

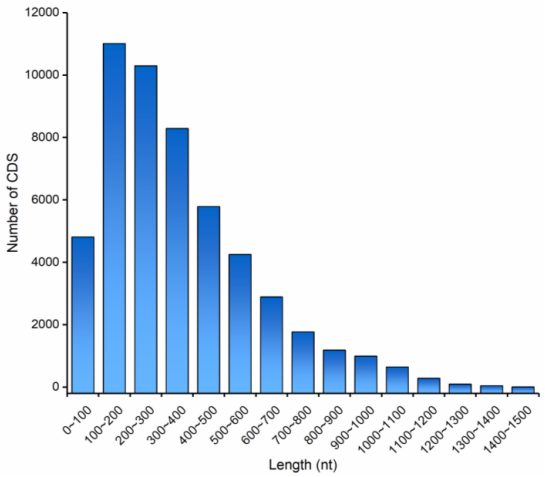

Figure 3. Distribution of 52,331 complete coding sequences (CDSs) of complete open reading frames (ORFs) from *Z. bungeanum*.

SSR markers can serve as useful tools for the marker-assisted breeding of numerous organisms [29]. Herein, 50,135 SSRs and 30,745 SSR-containing sequences were detected across 68,698 transcripts (>500 bp) from *Z. bungeanum*. Of these, 11,707 transcripts contained more than one SSR, and 6273 were compound SSRs. The number of SSRs per transcript ranged from 1 to 14. Mononucleotide repeat transcripts were the most frequent type (25,123, 57.28%) with 10−549 repeats, followed by 6858 (15.64%) tri-type transcripts with 5−33 repeats, 5611 (12.79%) di-type transcripts with 6−66 repeats, and 514 (1.17%) tetra-type transcripts with 5−20 repeats. Fewer hexa-type transcripts (334, 0.76%) and penta-type transcripts (194, 0.44%) were observed, with 5−12 and 5−8 repeats, respectively (Figure 4).

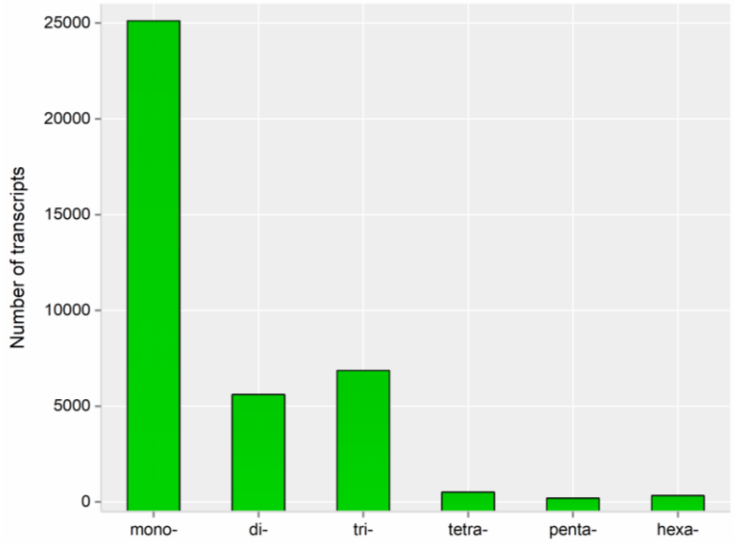

Figure 4. Distribution statistics for six types of simple sequence repeats (SSRs) from *Z. bungeanum*. The six types are mononucleotide (mono-), dinucleotide (di-), trinucleotide (tri-), tetranucleotide (tetra-), pentanucleotide (penta-), and hexanucleotide (hexa-).

Four computational approaches (CPC, CNCI, CPAT, and Pfam) were combined to distinguish non-protein-coding RNA candidates from putative protein-coding RNAs among the unknown transcripts. From the four analysis approaches, 3261, 3054, 10,520, and 9877 transcripts with lengths >200 nt and more than two exons were respectively selected as lncRNA candidates. Together, 1492 lncRNA transcripts were predicted for subsequent lncRNA analysis (Figure 5).

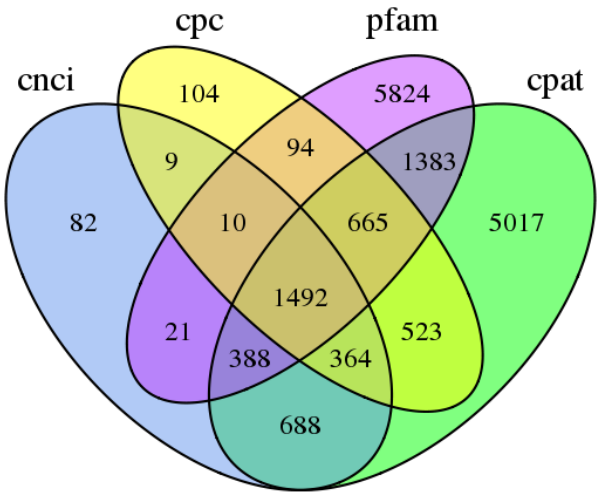

Figure 5. The number of long non-coding RNA (lncRNA) transcripts predicted in *Z. bungeanum* by the Coding Potential Calculator (CPC), Coding-Non-Coding Index (CNCI), Coding Potential Assessment Tool (CPAT), and Protein family database (Pfam) computational analyses.

3.4. Functional Annotation of Transcripts

Transcript function was annotated using eight databases (NR, EggNOG, Pfam, Swiss-Prot, GO, KOG, KEGG, and COG). Of 68,710 transcripts, 66,823 (97.25%) matched at least one of the above databases, and 66,696 (97.07%), 66,325 (96.53%), 55,555 (80.85%), 51,340 (74.72%), 45,982 (66.92%), 43,438 (63.22%), 30,714 (44.70%), and 29,469 (42.89%) transcripts were annotated in each respective database, while 1887 transcripts did not return any matches and could potentially be novel transcripts in the *Z. bungeanum* transcriptome.

The comparison of the *Z. bungeanum* transcripts in the NR database with other species revealed that *Z. bungeanum* shares homology with *Citrus sinensis* Osb., *Citrus clementine* Hort. ex. Tanaka, *Theobroma cacao* L., *Vitis vinifera* L., *Jatropha curcas* L., *Ricinus communis* L., *Populus trichocarpa* Torr. & Gray, *Populus euphratica* Oli., *Medicago truncatula* Gaertner, and *Gossypium arboretum* L.. The highest homology is shared with *C. sinensis* Osb. (49.51%) and *C. clementina* Hort. ex. Tanaka (36.08%) that belong to the same genus, while homology with the other species is relatively low (0.41%−1.41%; Figure 6).

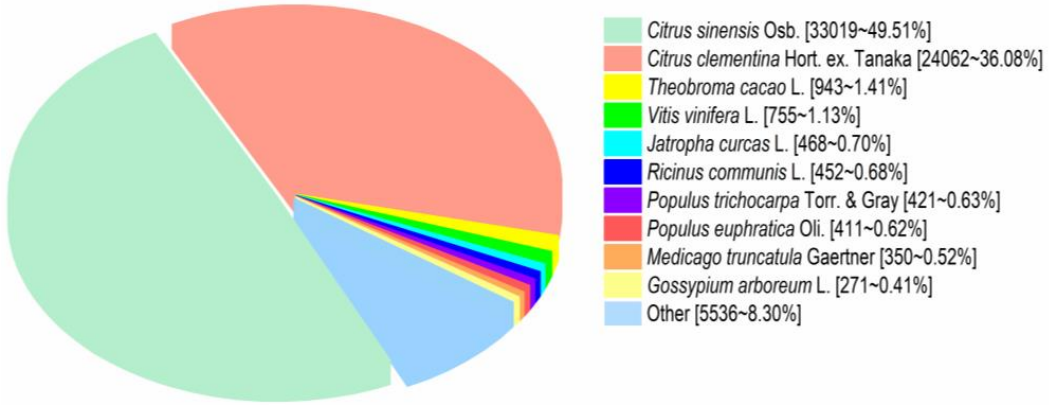

Figure 6. Species most closely related to *Z. bungeanum* in the NCBI database of non-redundant protein sequences database.

3.5. Gene Ontology (GO) Annotation

Using all ORF-containing transcripts for functional annotation, 4796 gene ontology (GO) terms were assigned to 45,982 PacBio transcript isoforms. These GO terms were classified into three main biological categories (cellular component, molecular function, and biological process). Biological process (2403, 50.10%) and molecular function (1932, 40.28%) represented the largest number of transcript isoforms, followed by cellular component (461, 9.61%).

Among the cellular component category, the highest proportion of transcript isoforms were associated with cell part (50.88%), cell (50.60%), organelle (37.60%), and membrane (28.32%) categories. Among molecular function terms, the largest number of transcript isoforms was assigned to catalytic activity (56.50%), binding (54.08%), transporter activity (6.82%), nucleic acid binding transcription factor activity (2.54%), structural molecule activity (2.34%), and molecular transducer activity (2.14%). In the biological process category, metabolic process was dominant (69.54%), followed by cellular process (62.07%), single-organism process (53.11%), response to stimulus (23.86%), and biological regulation (22.22%; Figure 7).

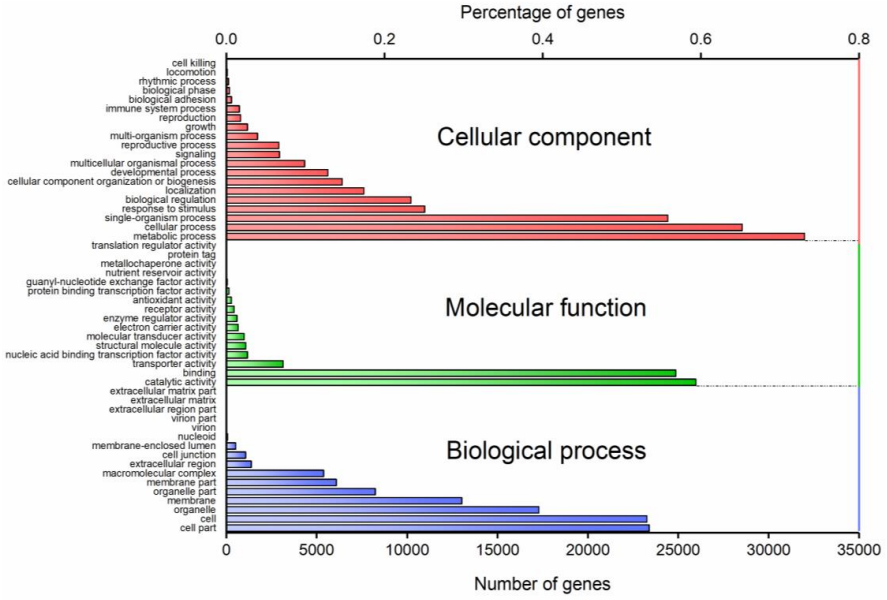

Figure 7. Gene Ontology (GO) classification of annotated transcripts from *Z. bungeanum.* GO terms were classified into three main categories: cellular component, molecular function, and biological process. The bottom x-axis indicates the number of transcripts, and the top x-axis indicates the percentage of transcripts.

3.6. EggNOG Annotation

EggNOG is a database of orthologous proteins and functional annotations that extends the non-supervised homology group based on the COG database [30]. After searching transcript isoforms against the EggNOG database, 68,999 with functional information were annotated and divided into 25 categories. Among these, the "general function prediction only" group accounted for the largest number of transcripts (10,883, 15.77%), followed by signal transduction mechanisms (6797, 9.85%), posttranslational modification, protein turnover, and chaperones (6081, 8.81%), transcription (5257, 7.62%), carbohydrate transport and metabolism (3456, 5.01%), and intracellular trafficking, secretion, and vesicular transport (2606, 3.78%). Only two transcripts were annotated in the cell motility category (Figure 8).

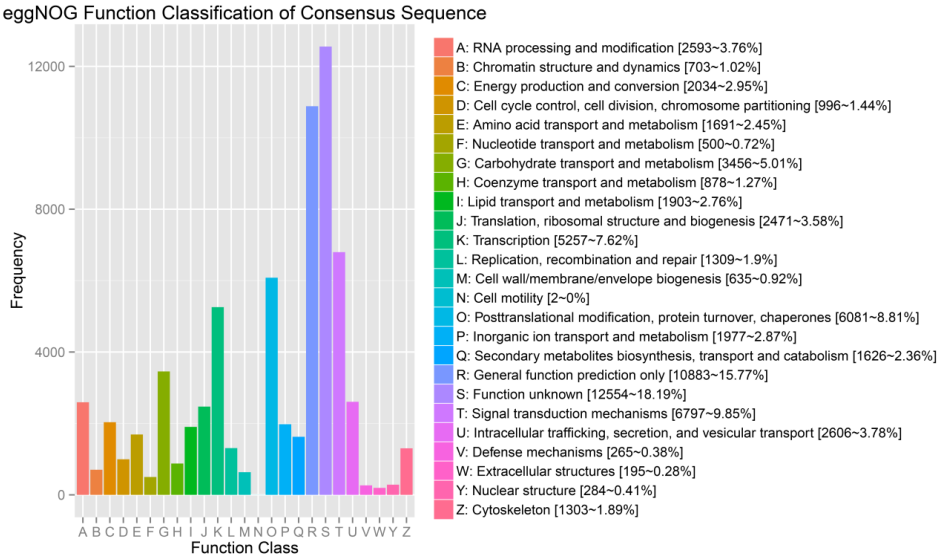

Figure 8. Evolutionary genealogy of genes based on Non-supervised Orthologous Groups (EggNOG) annotation of *Z. bungeanum* transcripts. The x-axis indicates EggNOG categories, and the y-axis indicates the number of transcripts.

3.7. Analysis of KEGG Pathways and Gene Annotation Information

A total of 30,947 transcripts from *Z. bungeanum* were annotated to 126 reference metabolic pathways in the KEGG database. Pathways with more annotated transcripts were mainly related to the normal growth of plants, namely carbon metabolism (1471), biosynthesis of amino acids (1212), protein processing in the endoplasmic reticulum (1015), plant hormone signal transduction (953), RNA transport (900), spliceosome (896), starch and sucrose metabolism (823), mRNA surveillance pathway (773), and oxidative phosphorylation (737; Supplementary Figure S1).

Functional annotation revealed 30 cancer-related transcripts, with 18 transcripts annotated in the Pfam database, 11 transcripts in the SwissProt database, 10 transcripts in the KEGG database, and one in the NR database. These transcripts were colon cancer-associated protein Mic1-like and breast cancer susceptibility 1 and 2 (BRCA1 and BRCA2) homologs (Supplementary Table S1).

4. Discussion

4.1. Evaluation of SMRT Sequencing Quality

In this study, we performed high-quality, full-length transcript isoform sequencing using a combination of SMRT sequencing and NGS technology to investigate gene information in *Z. bungeanum.* In total, 282,101 ROIs including 134,074 full-length non-chimeric reads were produced, and 75,973 consensus isoforms with 58,445 high-quality transcript isoforms were identified using ICE analysis of full-length non-chimeric reads. After error correction using reads sequenced with the Illumina platform, and the removal of redundant transcripts by way of CD-HIT software, 68,710 non-redundant transcript isoforms were obtained.

The high quality of the SMRT data was confirmed based on the comprehensive sequencing results. First, the average transcript lengths in the 1−2 kb, 2−3 kb, and 3−6 kb libraries were 2000, 2891, and 3567 bp, respectively, and the long total average length was 2819 bp, which indicates that ROIs were long enough to represent full-length transcripts [31]. Second, the Illumina data were used to correct low-quality SMRT reads, which can reduce the error rate of SMRT sequencing [32]. Thus, the resulting comprehensive transcriptome database for *Z. bungeanum* provided insight into the structures and functions of genes.

4.2. Application of lncRNA and NR Annotation

In this work, 1,492 lncRNAs were predicted in *Z. bungeanum* using all four computational approaches. Unexpectedly, the predicted lncRNAs were exceptionally long, with an average length of 2.5 kb (range = 0.3−10.0 kb). LncRNAs play an important role in the physiology and development of plants, especially in some key biological processes. However, only a fraction of the functions of lncRNAs have been discovered [33,34]. This almost untouched gene pool could include genes associated with agronomically relevant traits related to unresolved issues.

For example, the *PN-LNC-N13* lncRNA discovered in *Paspalum notatum* Flugge is only expressed in apomictic plants, and there may be structural differences between apomixis and sexual types [34]. Studies have confirmed that lncRNAs participate in abiotic stress responses and act as regulatory factors [35,36]. Cui et al. [37] found that an lncRNA participates in responses to *Phytophthora infestans* (Montagne) de Bary in tomato, further suggesting that lncRNAs may also play a role in enhancing biological resistance. The above research suggests that lncRNA data can provide useful gene resource information of potential medicinal value, as well as apomixis and stress tolerance in *Z. bungeanum*. In addition, lncRNAs are species- and tissue-specific, and can provide inspiration for studying the characteristic traits of *Z. bungeanum,* such as numb taste components [38,39].

NR analysis revealed that *Z. bungeanum* is highly homologous to *C. sinensis* and *C. clementina*, and all belong to the family Rutaceae. Since the study of *Z. bungeanum* is still in its infancy, research on many problems must draw on knowledge of other plants. Fortunately, research on citrus plants is more in-depth, and sources for apomictic reproduction and stress resistance have been developed [40,41].

4.3. Excavation of KEGG Annotation Pathways Gene Annotation Information in Z. bungeanum

A large number of transcripts from *Z. bungeanum* were associated with metabolic pathways, indicating that the growth and development of *Zanthoxylum* requires varied metabolic support. This also shows that there are multiple functional metabolites in *Z. bungeanum*, many of which may be of medicinal value. Although some pathways were associated with fewer transcripts, they may still be worth noting.

Aroma is an important quality attribute of *Zanthoxylum* products. Volatile oils are mainly responsible for aroma in *Zanthoxylum* and the main components are terpenoids [42]. In the KEGG annotation results, we identified the limonene and pinene degradation pathway (ko00903; Supplementary Figure S2A), the arachidonic acid metabolism pathway (ko00590; Supplementary Figure S2B), the degradation of aromatic compounds (ko01220; Supplementary Figure S2C), the sesquiterpenoid and triterpenoid biosynthesis pathway (ko00909; Supplementary Figure S2D), the monoterpenoid biosynthesis (ko00902; Supplementary Figure S2E), and the diterpenoid biosynthesis (ko00904; Supplementary Figure S2F). These seven pathways are most likely related to the metabolism of aroma components in the pericarp of *Zanthoxylum*, which provides an effective resource for future studies in this area.

Another decisive factor affecting the quality of *Zanthoxylum* is the color of the fruit. The color of ripe *Zanthoxylulm* fruit varies in different varieties. Interestingly, five of the annotated transcripts were associated with the anthocyanin biosynthesis pathway (ko00942; Supplementary Figure S2G). Thus, studies on the anthocyanin metabolic pathway could improve the color of *Zanthoxylum* fruit, benefitting production.

Human disease genes have been reported in plants, where they may perform different functions. For example, the BRCA1 homolog in *Arabidopsis* has a most conserved region PHD (plant homeodomain), which is absent in mammals [43]. Numerous genes that participate in genome stability and cancer predisposition in animals are well conserved in plants [44]. Herein, functional annotation revealed the presence of 30 cancer-related transcripts in *Z. bungeanum*. A study has shown that the BRCA2-RAD51 (recombination protein) complex plays a transcriptional regulatory role in plant immune response [45]. This provides another idea for us to study their roles in genome stability and transcriptional regulation of plant growth in *Zanthoxylum* resources.

5. Conclusions

In the present study, high-quality Full-length non-chimeric transcripts, ORFs, SSRs, and lncRNAs were detected in *Z. bungeanum* by SMRT sequencing. Metabolic pathways related to the aroma and color of *Z. bungeanum* pericarp were discovered through gene functional annotation. Unexpectedly, tumor suppressors involved in the pathogenesis of human breast and colorectal cancers were also annotated in *Z. bungeanum*. Our work provides comprehensive genetic resources for further molecular research on the cultivation, genetics, and quality characteristics of *Z. bungeanum*. Meanwhile, the discovery of anti-oncogene information corroborates the potential medicinal value of *Z. bungeanum*.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4907/9/12/765/s1, Figure S1: KEGG database pathway assignment of 30,714 transcripts from *Zanthoxylum bungeanum*, Figure S2: KEGG pathways related to the main characteristics of *Z. bungeanum*, Table S1: Functional annotation of 30 cancer-related transcripts.

Author Contributions: A.W. and J.T. conceived and designed the experiments. S.F., Y.L., and T.Y. aided in study design. J.T. performed the experiments, analyzed the results, and wrote the manuscript. L.Z., L.T., and Y.H. performed some of the experiments. All authors reviewed and approved the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2016YFC0501706).

Acknowledgments: We thank Chenglin Ju and Shengqi Wang for their technical assistance in the samples collection.

Conflicts of Interest: The authors declare no conflicts of interest. Data Availability: Transcriptome datasets supporting the conclusions of this article are available in the NCBI SRA repository under the accession number SRP149639.

References

1. Xiang, L.; Liu, Y.; Xie, C.; Li, X.; Yu, Y.; Ye, M.; Chen, X. The chemical and genetic characteristics of Szechuan pepper (*Zanthoxylum bungeanum* and *Z. armatum*) cultivars and their suitable habitat. *Front. Plant Sci.* **2016**, *7*, 467. [CrossRef] [PubMed]
2. Guo, T.; Deng, Y.; Xie, H.; Yao, C.; Cai, C.; Pan, S.; Wang, Y. Antinociceptive and anti-inflammatory activities of ethyl acetate fraction from *Zanthoxylum armatum* in mice. *Fitoterapia* **2011**, *82*, 347–351. [CrossRef]
3. Lei, X.; Cheng, S.; Peng, H.; He, Q.; Zhu, H.; Xu, M.; Wang, Q.; Liu, L.; Zhang, C.; Zhou, Q.; et al. Anti-inflammatory effect of *Zanthoxylum bungeanum*-cake-separated moxibustion on rheumatoid arthritis rats. *Afr. J. Tradit. Complement. Altern. Med.* **2016**, *13*, 45–52. [CrossRef]
4. Barkatullah; Ibrar, M.; Muhammad, N.; Khan, A.; Khan, S.A.; Zafar, S.; Jan, S.; Riaz, N.; Ullah, Z.; Farooq, U.; et al. Pharmacognostic and phytochemical studies of *Zanthoxylum armatum* DC. *Pak. J. Pharm. Sci.* **2017**, *30*, 429–438. [PubMed]
5. Alam, F.; Saqib, Q.N.U.; Ashraf, M. *Zanthoxylum armatum* DC extracts from fruit, bark and leaf induce hypolipidemic and hypoglycemic effects in mice- in vivo and in vitro study. *BMC Complement. Altern. Med.* **2018**, *18*, 68. [CrossRef] [PubMed]
6. Yang, Q.; Cao, W.; Zhou, X.; Cao, W.; Xie, Y.; Wang, S. Anti-thrombotic effects of alpha-linolenic acid isolated from *Zanthoxylum bungeanum* Maxim seeds. *BMC Complement. Altern. Med.* **2014**, *14*, 348. [CrossRef] [PubMed]
7. Park, D.; Park, Y.; Li, Y.; Xu, H.; Lee, J.; Kim, Y.; Jang, S.; Park, S.; Lee, Y.; Ahn, J.; et al. Biological safety and B cells activation effects of *Zanthoxylum schinifolium*. *Mol. Cell. Toxicol.* **2011**, *7*, 157–162. [CrossRef]
8. Kim, S.; Seo, H.; Al Mahmud, H.; Islam, M.I.; Lee, B.E.; Cho, M.L.; Song, H.Y. In vitro activity of collinin isolated from the leaves of *Zanthoxylum schinifolium* against multidrug- and extensively drug-resistant mycobacterium tuberculosis. *Phytomedicine* **2018**, *46*, 104–110. [CrossRef]
9. Cai, X.; Mu, X.; Zhang, Z.; Hua, Z.; Huang, Y.; Sun, Q. Polyembryony and multiple seedlings in the apomictic plants. *Acta Bot. Sin.* **1997**, *39*, 590–595.
10. Liu, Y. Apomixis in *Zanthoxylum bungeanum* and *Z. simulans*. *J. Genet. Genomics* **1987**, *14*, 107–113.
11. Li, Y.; Dai, C.; Hu, C.; Liu, Z.; Kang, C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *Plant J.* **2017**, *90*, 164–176. [CrossRef] [PubMed]
12. Wang, M.; Wang, P.; Liang, F.; Ye, Z.; Li, J.; Shen, C.; Pei, L.; Wang, F.; Hu, J.; Tu, L.; et al. A global survey of alternative splicing in allopolyploid cotton: Landscape, complexity and regulation. *New Phytol.* **2018**, *217*, 163–178. [CrossRef] [PubMed]
13. Wang, T.; Wang, H.; Cai, D.; Gao, Y.; Zhang, H.; Wang, Y.; Lin, C.; Ma, L.; Gu, L. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). *Plant J.* **2017**, *91*, 684–699. [CrossRef] [PubMed]
14. Hoang, N.V.; Furtado, A.; Mason, P.J.; Marquardt, A.; Kasirajan, L.; Thirugnanasambandam, P.P.; Botha, F.C.; Henry, R.J. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* **2017**, *18*, 395. [CrossRef] [PubMed]
15. Liu, X.; Mei, W.; Soltis, P.S.; Soltis, D.E.; Barbazuk, W.B. Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Mol. Ecol. Resour.* **2017**, *17*, 1243–1256. [CrossRef] [PubMed]
16. Ning, G.; Cheng, X.; Luo, P.; Liang, F.; Wang, Z.; Yu, G.; Li, X.; Wang, D.; Bao, M. Hybrid sequencing and map finding (HySeMaFi): Optional strategies for extensively deciphering gene splicing and expression in organisms without reference genome. *Sci. Rep.* **2017**, *7*, 43793. [CrossRef] [PubMed]
17. Pang, T.; Ye, C.Y.; Xia, X.; Yin, W. *De novo* sequencing and transcriptome analysis of the desert shrub, *Ammopiptanthus mongolicus*, during cold acclimation using Illumina/Solexa. *BMC Genomics* **2013**, *14*, 488. [CrossRef] [PubMed]
18. Sinha, S.; Raxwal, V.K.; Joshi, B.; Jagannath, A.; Katiyar-Agarwal, S.; Goel, S.; Kumar, A.; Agarwal, M. *De novo* transcriptome profiling of cold-stressed siliques during pod filling stages in Indian mustard (*Brassica juncea* L.). *Front. Plant Sci.* **2015**, *6*, 932. [CrossRef]

19. Chen, E.; Wei, D.; Shen, G.; Yuan, G.; Bai, P.; Wang, J. *De novo* characterization of the *Dialeurodes citri* transcriptome: Mining genes involved in stress resistance and simple sequence repeats (SSRs) discovery. *Insect Mol. Biol.* **2014**, *23*, 52–66. [CrossRef]
20. Xu, Z.; Peters, R.J.; Weirather, J.; Luo, H.; Liao, B.; Zhang, X.; Zhu, Y.; Ji, A.; Zhang, B.; Hu, S.; et al. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia* miltiorrhiza and tanshinone biosynthesis. *Plant J.* **2015**, *82*, 951–961. [CrossRef]
21. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **2009**, *323*, 133–138. [CrossRef] [PubMed]
22. Hackl, T.; Hedrich, R.; Schultz, J.; Foester, F. *Proovread*: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **2014**, *30*, 3004–3011. [CrossRef]
23. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
24. Beier, S.; Thiel, T.; Muench, T.; Scholz, U.; Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef]
25. Kong, L.; Zhang, Y.; Ye, Z.; Liu, X.; Zhao, S.; Wei, L.; Gao, G. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **2007**, *35*, W345–W349. [CrossRef]
26. Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **2013**, *41*, e166–e166. [CrossRef] [PubMed]
27. Wang, L.; Park, H.J.; Dasari, S.; Wang, S.; Kocher, J.; Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **2013**, *41*, e74. [CrossRef] [PubMed]
28. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
29. Biswas, M.; Nath, U.; Howlader, J.; Bagchi, M.; Natarajan, S.; Kayum, M.A.; Kim, H.-T.; Park, J.I.; Kang, J.-G.; Nou, S., III. Exploration and exploitation of novel SSR markers for candidate transcription factor genes in *Lilium* species. *Genes* **2018**, *9*, 97. [CrossRef]
30. Powell, S.; Szklarczyk, D.; Trachana, K.; Roth, A.; Kuhn, M.; Muller, J.; Arnold, R.; Rattei, T.; Letunic, I.; Doerks, T.; et al. eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **2011**, *40*, D284–D289. [CrossRef]
31. Minoche, A.E.; Dohm, J.C.; Schneider, J.; Holtgräwe, D.; Viehöver, P.; Montfort, M.; Sörensen, T.R.; Weisshaar, B.; Himmelbauer, H. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* **2015**, *16*, 184. [CrossRef] [PubMed]
32. Chen, S.; Deng, F.; Jia, X.; Li, C.; Lai, S. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* **2017**. [CrossRef] [PubMed]
33. Liu, J.; Jung, C.; Xu, J.; Wang, H.; Deng, S.; Bernad, L.; Arenas-Huertero, C.; Chua, N.-H. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **2012**, *24*, 4333–4345. [CrossRef]
34. Ochogavía, A.; Galla, G.; Seijo, J.G.; González, A.M.; Bellucci, M.; Pupilli, F.; Barcaccia, G.; Albertini, E.; Pessino, S. Structure, target-specificity and expression of *PN_LNC_N13*, a long non-coding RNA differentially expressed in apomictic and sexual *Paspalum notatum*. *Plant Mol. Biol.* **2018**, *96*, 53–67. [CrossRef] [PubMed]
35. Deng, F.; Zhang, X.; Wang, W.; Yuan, R.; Shen, F. Identification of *Gossypium hirsutum* long non-coding RNAs (lncRNAs) under salt stress. *BMC Plant Biol.* **2018**, *18*, 23. [CrossRef] [PubMed]
36. Laporte, P.; Merchan, F.; Amor, B.B.; Wirth, S.; Crespi, M. Riboregulators in plant development. *Biochem. Soc. Trans.* **2007**, *35*, 1638–1642. [CrossRef] [PubMed]
37. Cui, J.; Luan, Y.; Jiang, N.; Bao, H.; Meng, J. Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lncRNA16397 conferring resistance to *Phytophthora infestans* by co-expressing glutaredoxin. *Plant J.* **2017**, *89*, 577–589. [CrossRef] [PubMed]
38. Dinger, M.E.; Amaral, P.P.; Mercer, T.R.; Pang, K.C.; Bruce, S.J.; Gardiner, B.B.; Askarian-Amiri, M.E.; Ru, K.; Solda, G.; Sunkin, S.M.; et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **2008**, *18*, 1433–1445. [CrossRef]

39. Wang, B.; Tseng, E.; Regulski, M.; Clark, T.A.; Hon, T.; Jiao, Y.; Lu, Z.; Olson, A.; Stein, J.C.; Ware, D. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **2016**, *7*, 11708. [CrossRef]
40. Xu, Q.; Chen, L.; Ruan, X.; Chen, D.; Zhu, A.; Chen, C.; Bertrand, D.; Jiao, W.; Hao, B.; Lyon, M.P.; et al. The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **2013**, *45*, 59–66. [CrossRef]
41. Deng, Z.; Huang, S.; Ling, P.; Chen, C.; Yu, C.; Weber, C.A.; Moore, G.A.; Gmitter, F.G., Jr. Cloning and characterization of NBS-LRR class resistance-gene candidate sequences in citrus. *Theor. Appl. Genet.* **2000**, *101*, 814–822. [CrossRef]
42. Diao, W.; Hu, Q.; Feng, S.; Li, W.; Xu, J. Chemical composition and antibacterial activity of the essential oil from Green Huajiao (*Zanthoxylum schinifolium*) against selected foodborne pathogens. *J. Agr. Food Chem.* **2013**, *61*, 6044–6049. [CrossRef] [PubMed]
43. Trapp, O.; Seeliger, K.; Puchta, H. Homologs of Breast Cancer Genes in Plants. *Front. Plant Sci.* **2011**, *2*, 19. [CrossRef] [PubMed]
44. Puchta, H.; Kobbe, D.; Wanieck, K.; Knoll, A.; Suer, S.; Focke, M.; Hartung, F. Role of human disease genes for the maintenance of genome stability in plants. In *Induced Plant Mutations in the Genomics Era*; Food and Agriculture Organization of the United Nations (FAO): Rome, Italy, 2009; pp. 129–132.
45. Wang, S.; Durrant, W.E.; Song, J.; Spivey, N.W.; Dong, X. *Arabidopsis* BRCA2 and RAD51 proteins are specifically involved in defense gene transcription during plant immune responses. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 22716–22721. [CrossRef] [PubMed]