genes

Article

Improved High-Quality Genome Assembly and Annotation of Pineapple (*Ananas comosus*) Cultivar MD2 Revealed Extensive Haplotype Diversity and Diversified FRS/FRF Gene Family

Ashley G. Yow [1,2], Hamed Bostan [2], Raúl Castanera [3], Valentino Ruggieri [4], Molla F. Mengist [2], Julien Curaba [2], Roberto Young [5], Nicholas Gillitt [6] and Massimo Iorizzo [1,2,*]

[1] Department of Horticultural Science, North Carolina State University, Raleigh, NC 27695, USA; agyow@ncsu.edu
[2] Plants for Human Health Institute, North Carolina State University, Kannapolis, NC 28081, USA; hbostan@ncsu.edu (H.B.); mmengis@ncsu.edu (M.F.M.); jbcuraba@ncsu.edu (J.C.)
[3] Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, 08193 Barcelona, Spain; raul.castanera@cragenomica.es
[4] Biomeets Consulting, Carrer d'Àlaba 61, 08005 Barcelona, Spain; valentino.ruggieri@biomeets.es
[5] Research Department of Dole, Standard Fruit de Honduras, Zona Mazapan, La Ceiba 31101, Honduras; roberto.young@dole.com
[6] Core Genomics Lab, David H. Murdock Research Institute, Kannapolis, NC 28081, USA; nick@berkleyrd.com
* Correspondence: miorizz@ncsu.edu

Citation: Yow, A.G.; Bostan, H.; Castanera, R.; Ruggieri, V.; Mengist, M.F.; Curaba, J.; Young, R.; Gillitt, N.; Iorizzo, M. Improved High-Quality Genome Assembly and Annotation of Pineapple (*Ananas comosus*) Cultivar MD2 Revealed Extensive Haplotype Diversity and Diversified FRS/FRF Gene Family. *Genes* **2022**, *13*, 52. https://doi.org/10.3390/genes13010052

Academic Editors: Michael Abberton and Rajneesh Paliwal

Received: 22 October 2021
Accepted: 21 December 2021
Published: 24 December 2021

Abstract: Pineapple (*Ananas comosus* (L.) Merr.) is the second most important tropical fruit crop globally, and 'MD2' is the most important cultivated variety. A high-quality genome is important for molecular-based breeding, but available pineapple genomes still have some quality limitations. Here, PacBio and Hi-C data were used to develop a new high-quality MD2 assembly and gene prediction. Compared to the previous MD2 assembly, major improvements included a 26.6-fold increase in contig N50 length, phased chromosomes, and >6000 new genes. The new MD2 assembly also included 161.6 Mb additional sequences and >3000 extra genes compared to the F153 genome. Over 48% of the predicted genes harbored potential deleterious mutations, indicating that the high level of heterozygosity in this species contributes to maintaining functional alleles. The genome was used to characterize the FAR1-RELATED SEQUENCE (FRS) genes that were expanded in pineapple and rice. Transposed and dispersed duplications contributed to expanding the numbers of these genes in the pineapple lineage. Several AcFRS genes were differentially expressed among tissue-types and stages of flower development, suggesting that their expansion contributed to evolving specialized functions in reproductive tissues. The new MD2 assembly will serve as a new reference for genetic and genomic studies in pineapple.

Keywords: pineapple MD2; *Ananas comosus*; improved assembly; haplotype diversity; FHY3/FAR1 genes; FRS/FRF genes

1. Introduction

Pineapple is the second most important tropical fruit globally. The fruit, which contains nutritionally valuable vitamins and bioactive enzymes (e.g., bromelain), is consumed in both fresh and processed forms. Worldwide production was estimated at over 30 million tons in 2019 (http://www.fao.org/, accessed on 23 March 2021). The consumption of pineapple fruit and the use of pineapple-derived supplements has been steadily increasing over the past decade and continues to increase each year [1]. 'MD2' is the most widely-grown fresh fruit market cultivar by major production companies due to its consistently large fruit size and better fruit quality as compared to other cultivars [2].

Developing new cultivars with improved characteristics is key to overcoming production challenges (e.g., disease or abiotic stress) and responding to changes in market demand. The use of genomic resources helps to accelerate the process of cultivar development, allowing breeders to meet consumer and grower needs more quickly than with traditional breeding approaches alone [3]. Incorporation of advanced genetic and genomic resources into a breeding program provides several advantages, including increased efficiency when breeding for specific traits and reduced screening time for determining the presence/absence of those traits in a breeding population [3].

Pineapple is a highly heterozygous diploid species with 25 chromosomes and an estimated haploid genome size of 563 Mb [4]. Multiple genome assemblies for pineapple representing two cultivars (MD2 and F153) and a close relative species (*Ananas bracteatus*, CB5) are currently available. However, the quality of these genomes in terms of sequence contiguity (contig N50) is relatively low (MD2 v1 = 57 kb). The MD2 v1 genome that represents the most commonly grown pineapple variety is not anchored at the chromosome level, and only 56% of the estimated genome size is anchored at the chromosome level for the F153 genome [2,5]. Homologous chromosomes have not been fully phased for any MD2 pineapple genome to date. New long-read sequencing technologies, such as PacBio, allow for the iterative improvement of genome assemblies for economically important crops, such as pineapple [6]. Genomic interaction data, such as Hi-C, facilitate the assembly of high-quality scaffolds and the reconstruction of haplotype phases [7,8]. Also, gene predictions obtained from long-read sequencing technology, such as PacBio Iso-Seq, allow researchers to obtain full-length transcripts, increasing the reliability of the predictions [9].

To continue building on recent advances in pineapple genetics and genomics, here we present a high-quality, phased, chromosome-scale assembly of MD2. Comparative haplotype analysis was performed to study the potential impact of heterozygosity on allele diversity. Comparison with previous pineapple and other crop genomes was performed to highlight assembly and gene prediction improvements, potential chromosome rearrangements, and to characterize regulatory and resistance genes, including the genes that code for the transcription factor (TF) FAR-RED ELONGATED HYPOCOTYL3 (FHY3) and its homolog FAR-RED-IMPAIRED RESPONSE1 (FAR1), as well as FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF). These TFs make up the FRS/FRF family, which has been implicated in multiple developmental functions including chloroplast division and chlorophyll biosynthesis, circadian clock regulation, flowering time regulation, starch biosynthesis, and the biotic and abiotic stress responses [10–13]. The results of this study provide novel resources for genetic and genomic analyses in pineapple that are critical for advancing modern breeding strategies in this crop.

2. Materials and Methods

2.1. Plant Material Collection, DNA and RNA Extraction, and Sequencing

A diploid, commercial fresh-fruit market variety, MD2, was used for whole-genome sequencing with Pacific Biosciences (PacBio, Menlo Park, CA, USA) and Hi-C proximity ligation. Pineapple MD2 plants were shipped to Dr. Iorizzo's lab from Dole in La Ceiba, Honduras and grown in the greenhouse at the North Carolina Research Campus, Kannapolis, NC, USA.

For DNA extraction, young leaves were harvested and then immediately frozen in liquid nitrogen, stored at −80 °C, and freeze-dried for 24 h before use. Dried leaves were ground into powder using a Genogrinder 2000 tissue homogenizer (SPEX SamplePrep, Metuchen, NJ, USA). High molecular weight genomic DNA was extracted from dry powdered leaf tissue using the CTAB method outlined in Charlotte et al. (2016) [14]. The DNA quantity and purity were determined using a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA) and NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), respectively. DNA was also evaluated on a 0.8% agarose gel and Agilent 4200 TapeStation instrument (Agilent Technologies, Santa Clara, CA, USA) to determine the fragment size distribution.

For RNA extraction, fresh tissues from the leaf, meristem, stem, root, flower, and fruit of greenhouse-grown MD2 pineapples were harvested and ground in liquid nitrogen.

Total RNA was extracted from collected tissues using the Sigma Spectrum Total RNA kit (Sigma-Aldrich, St. Louis, MO, USA). RNA integrity was evaluated on a 1% agarose gel and Agilent 2100 Bioanalyzer instrument. RNA was quantified using Qubit and the purity was tested using NanoDrop.

A PacBio genomic library for MD2 was prepared using the SMRTbell Express Template Prep Kit (Cat. #101-357-000) following the protocol for preparing >15 kb libraries (PN 101-397-100) (Pacific Biosciences, Menlo Park, CA, USA). The SMRTbell library was analyzed on the Agilent 4200 TapeStation instrument to determine if size-selection was necessary prior to sequencing with the PacBio Sequel system at the David H. Murdock Research Institute (DHMRI, Kannapolis, NC, USA). PacBio Iso-Seq libraries were prepared with the SMRTbell Express Template Prep Kit 2.0 (Cat. #100-938-900) according to the Iso-Seq Express Template Preparation for Sequel and Sequel II Systems protocol (PN 101-763-800). For Iso-Seq library preparation, RNA from leaf, root, fruit, and flower tissues were maintained separately to prepare one library each, while RNA from meristem and stem tissues were combined in equimolar amounts to prepare one combined library. A total of five libraries were prepared and each library was sequenced on a SMRT cell with the PacBio Sequel system. The Hi-C library for MD2 was prepared using the Phase Genomics Proximo Hi-C Plant Kit (Seattle, WA, USA). Quantity and quality checks were performed by Phase Genomics, including qPCR analysis, electrophoretic assay on the Agilent 2100 Bioanalyzer instrument, and spike-in control. Hi-C libraries were sequenced with the Illumina HiSeq 2500 using 150 bp paired-end (2 × 150 bp) run chemistry.

The PacBio genomic reads were processed for error correction and to generate consensus sequences using the FALCON pipeline [15]. Iso-Seq reads were further processed with IsoSeq3 pipeline (https://github.com/PacificBiosciences/IsoSeq (accessed on 29 January 2019), with the ccs considering -min-rq 0.9, lima considering –isoseq –dump-clips –no-pbi –peek-guess, and refine considering –require-polya and cluster), generating circular consensus sequence (CCS) reads for subsequent data analysis. The Illumina Hi-C sequences were quality-checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed on 23 May 2019) and cleaned using Trimmomatic [16] (parameters: ILLUMINACLIP:2:30:10:2:TRUE SLIDINGWINDOW:10:30 LEADING:30 TRAILING:30 HEADCROP:10 MINLEN:45), which removed adapter sequences. PCR duplicates were flagged for removal using SAMBLASTER [17]. After this step, cleaned sequences were retained for downstream analysis.

2.2. Genome Assembly, Phasing, and Scaffolding

To generate an improved MD2 pineapple genome assembly, newly generated PacBio long reads and Hi-C data were integrated. The new MD2 v2 genome assembly presented here was generated using a de novo assembly approach, and four published pineapple genome assemblies (MD2 v1, F153 v3, F153 v7 and CB5) [2,5,18] were used only for comparative analysis.

A de novo assembly for MD2 was generated using a FALCON assembler [15]. FALCON-UNZIP was later used to reconstruct the un-phased contigs (haplotigs) from the initially assembled contigs [15]. In addition to the sequence error correction implemented in the FALCON pipeline, Pilon v.1.24 [19] was used to further polish the assembled sequences using the Hi-C Illumina short reads. Hi-C data were subsequently incorporated for further phasing of the genome using FALCON-Phase [7], resulting in a fully-phased genome assembly (phases P0 and P1).

Scaffolding of the MD2 v2 genome assembly was performed by Phase Genomics using the Proximo pipeline (https://phasegenomics.com/products/proximo/, accessed on 26 August 2019). The haplotype-phased contigs (haplotigs) obtained from FALCON-Phase were used as the input for scaffolding with Proximo, to create a chromosome-scale genome assembly. A linkage map including 46,860 SSR markers was used to support clustering of contigs into pseudomolecules during the scaffolding analysis [20]. SSR marker sequences developed for pineapple variety F153 were obtained from the Pineapple Genomics Database [20] and aligned against the MD2 v2 FALCON-Phase genome assembly using BWA-aln (default parameters) [5,20]. Note that only the physical coordinates of each SSR marker mapped in the F153 genome were publicly available and not the genetic distance (cM).

The Hi-C interaction heat map in conjunction with the published SSR linkage map were used to identify and correct chimeric regions following two conditions: (1) non-collinearity with the order of the markers in the SSR linkage map, and (2) the presence of gaps in the Hi-C interaction along the heatmap diagonal. The positions of these chimeric regions were corrected to maximize the Hi-C interaction signals along the diagonal. After this correction process, the F153 SSR linkage map aligned with high collinearity to the 25 chromosomes of the MD2 v2 assembly.

In both phases, the longest 25 scaffolds represented the 25 pineapple chromosomes. The P0 and P1 assemblies were used as input files for a final step of error correction and gap filling.

2.3. Assembly Error Correction and Gap-Filling

Assembly errors were manually corrected in Juicebox (v.1.8.8) [21]. Misassemblies and incorrect sequence orientations were detected in the phased, scaffolded MD2 v2 assembly by examining the Hi-C heatmap for zones of depleted genomic interaction (no red signals along diagonal) and by performing a collinearity analysis with the linkage map as described above. The genomic positions of the F153 SSR marker sequences in the MD2 v2 scaffolds were compared to their genomic positions in the F153 v3 reference genome. The Hi-C interaction heatmap provides a visual guide for assessing assembly quality. The red color along the diagonal indicates points of interaction between neighboring genomic regions. Off-diagonal red indicates: (1) highly repetitive genomic regions, such as centromeres or telomeres, or (2) misassembled/chimeric sequences. When performing manual corrections, in those cases where the collinearity analyses with the F153 SSR linkage map and Hi-C interaction heatmap were not in agreement for a region, preference was given to the Hi-C data. The rationale behind this criterion was based on the fact that the Hi-C data represent the MD2 genome while the linkage map represents a different genetic background, and it cannot be excluded that the differences between the linkage map and the MD2 genome represent true structural differences.

The first round of manual corrections was based on the collinearity between the F153 SSR linkage map (using marker locations of the F153 v3 genome) and the scaffolded MD2 v2 genome. Assembly error corrections were only kept if they both improved the collinearity with the SSR map and did not negatively affect the Hi-C interaction heatmap. The second round of manual corrections was based solely on the Hi-C interaction heatmap visualization. These first and second rounds of manual corrections were conducted on each phase separately. Both phases were aligned to each other using the Nucmer module of MUMmer4 [22] to identify regions of non-collinearity between individual haplotype assemblies. Non-collinear regions, including differences in the order and orientation of contigs between phases, were corrected based on the linkage map collinearity and the Hi-C interaction heatmap as described above. After manual correction of the MD2 v2 genome assembly, LR_GapCloser [23] was run with three iterations for gap-filling using the cleaned genomic PacBio reads.

2.4. Genome Quality Analysis

Assembly quality was determined using several metrics that assess correctness, contiguity, and completeness. Assembly correctness was determined using the Hi-C heatmap, F153 SSR linkage map collinearity, and by performing a contamination check against fungal, bacterial, and viral genomes with NCBI BLASTn. Contiguity was determined based on the contig and scaffold N50 lengths, the percentage of contigs anchored to chromosomes, and the ratio of the overall number of contigs assembled to total assembly length. Completeness was determined based on the overall percentage of estimated genome size assembled, long terminal repeat (LTR) assembly index (LAI) score [24], and by using Benchmarking Universal Single-Copy Orthologs (BUSCO) v.3 [25] against the embryophyta_odb10. The completeness of the assembly gene space was assessed using the percentage of mapped transcriptome reads and the BUSCO score. Published transcriptomic data for pineapple were obtained from the NCBI SRA database (project accession nos. PRJNA648819, PRJNA648693, PRJNA494788, PRJNA393610, PRJNA356904, PRJNA331052, PRJNA310033, PRJNA305042, PRJNA237705), and sequences were cleaned using Trimmomatic [16] and quality-checked with FastQC. Transcripts were aligned to the MD2 v2 assembly P0 and P1 haplotypes separately using STAR [26] (–outSAMstrandField intronMotif –outSAMattrIHstart 0 –outFilterMismatchNmax 2 –outSAMtype BAM SortedByCoordinate).

The MD2 v2 assembly was aligned against a previously sequenced pineapple genome (F153 v7) and one representing a wild relative *A. bracteatus* (CB5) [5,18]. Both of these assemblies were assembled at chromosome level. Alignment of the genomes was performed using the Nucmer module of MUMmer4 (default parameters) and visualization of the alignments was performed using the web-based tool, Dot (https://dnanexus.github.io/dot/, accessed on 19 March 2020).

2.5. Genome Annotation

Gene models were predicted by implementing a hybrid strategy using the MAKER pipeline, a tool for annotating a reference genome using empirical and ab initio gene prediction deploying AUGUSTUS v.2.5.5 [27] and SNAP [28], as well as the GeMoMa pipeline [29,30], a homology-based gene prediction program that predicts gene models in target species based on gene models in evolutionary-related reference species. These processes involved several iterations of homology-based and in silico gene prediction steps, using high-throughput short- and long-read sequences as well as gene models obtained from multiple closely related species and model organisms (*A. comosus* [5], *Arabidopsis thaliana* [31], *Carica papaya* [32], *Musa acuminata* [33], *Oryza sativa* [34], *Solanum lycopersicum* [35], *Sorghum bicolor* [36], *Vitis vinifera* [37], and *Zea mays* [38]), downloaded from Phytozome and used as the input for training.

Structures of the predicted gene models were manually verified by aligning PacBio Iso-Seq full-length transcripts using GMAP aligner (–min-identity = 0.99 –min-trimmed-coverage = 0.95 –nosplicing) [39].

The putative function of the predicted genes was annotated using public databases, including NCBI non-redundant protein, KOG, GO, and InterPro. NCBI BLASTx was used to compare the predicted coding sequences (CDS) (e-value $\leq 1 \times 10^{-10}$) with the non-redundant protein database (downloaded in December 2020). Blast2GO v.1.4.11 was used to annotate the GO terms of genes with default parameters. The protein domains were annotated using the Blast2GO InterProScan module [40] based on all available protein databases.

Eukaryotic clustering of orthologous group (KOG) analysis was performed with the eggNOG Mapper v.5.0 Blast2GO extension [41]. The eggNOG Mapper tool assigns orthologs and transfers functions to query genes using phylogenetic inference.

Reciprocal, ungapped BLASTn alignments of MD2 v2 P0 and MD2 v1 CDSs with query coverage and percent identity parameters set to 100 (-ungapped, -qcov_hsp_perc 100, and -perc_identity 100) revealed which genes had structural differences between the two MD2 genomes. Additionally, a reciprocal BLASTn alignment (default parameters) of MD2 v2, MD2 v1, and F153 v3 CDSs indicated which genes predicted in the MD2 v2 genome were not predicted in the previous genomes.

Salmon [42] (default parameters) was used to align Illumina RNA-seq reads from the NCBI SRA database (project accession no. PRJNA305042) with the predicted genes to determine if the new genes identified by BLASTn were expressed (TPM > 0).

Disease resistance genes (R-genes) and regulatory genes were predicted to assess how the new MD2 v2 genome affected the prediction of some known important gene families. Transcription factors (TFs), transcriptional regulators (TRs), and chromatin regulators (CRs) were identified using PlantTFcat [43] on pineapple MD2 v2, MD2 v1, and F153 v3 gene models, as well as for various other species. R-genes were identified using DRAGO2 [44] for the same set of gene models/species. The number of genes in each gene family was compared between MD2 v2 and the published pineapple genomes to determine the number of extra TFs, TRs, CRs and R-genes that were identified in this study.

The EDTA pipeline [45] was used to obtain a non-redundant catalog of TE families using MD2 v2 P0. This step included running LTRharvest [46], LTR_FINDER_parallel [47], and LTR_retriever [48] for detecting LTR-retrotransposons, GRF [49] and TIR-Learner [50] for detecting TIR transposons, and HelitronScanner [51] for detecting Helitrons. Finally, RepeatModeler (http://www.repeatmasker.org/RepeatModeler/, accessed on 20 March 2020) was used to complement the library with non-LTR retrotransposons and other TE families bypassed by the previous tools.

RepeatMasker (http://www.repeatmasker.org, accessed on 3 April 2020) was used to annotate the two MD v2 haplotypes, as well as the MD2 v1 [2], CB5 [18], and F153 v7 [18] genomes using our TE library (Supplementary Dataset S1). In parallel, intact LTR-retrotransposons, TIRs, and Helitrons were identified in the four genomes by running the corresponding *ltr*, *tir*, and *helitron* modules of the EDTA pipeline.

Then, a Perl script (parseRM.pl -p option, available at https://github.com/4ureliek/Parsing-RepeatMasker-Outputs, accessed on 9 April 2020) was used to parse the raw alignment outputs from RepeatMasker and get a detailed summary for each repeat family and class as well as the total amount of repeats per genome. LTR-assembly index (LAI) [24] and LTR-retrotransposon insertion age were calculated using LTR-retriever and compared across the four genomes to provide additional support for the high quality of the MD2 v2 genome assembly.

2.6. Haplotype Comparison

Alignment-based comparison of the phased haplotypes of the MD2 pineapple assembly was conducted using the pipeline recently developed for the phased diploid potato genome [52]. Syntenic regions between the P0 and P1 haplotypes were identified and plotted using MCScanX [53]. Homologous chromosomes were aligned using MUMmer v.4.0 and structural variants ($\geq$100 bp) were detected from differences reported by the *show-diff* function. Presence and absence variation (PAV) genes were identified and defined as genes that lacked a homolog on the complementary haplotype, while its surrounding genes had homologs that were collinear in position between the two haplotypes. SNPs and indels between haplotypes were annotated using SnpEff [54].

2.7. Identification and Characterization of FRS/FRF Transcription Factors

TFs identified with PlantTFcat were compared across multiple monocot and dicot species to determine if any contraction or expansion of specific gene families occurred in pineapple. Based on the prediction of TFs in pineapple and other genomes, and in pineapple and rice, both members of the order Poales within the monocot lineage had a larger number of genes belonging to the FRS/FRF family (see Section 3.6). Therefore, this gene family was characterized in terms of the mode of duplication, shared ancestry, and pattern of expression across different tissue and flower developmental stages.

Ortholog prediction of FRS/FRF genes from pineapple and other species was performed using OrthoMCL v.2.0.9 (https://orthomcl.org, accessed on 7 January 2021) [55]. To explore the evolutionary relationships of FRS/FRF gene family members in pineapple, the FRS/FRF amino acid sequences from pineapple and other species were used. Multiple sequence alignments of all FRS/FRF proteins were performed by using MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/, accessed on 17 August 2021) [56], with default parameters. Subsequently, phylogenetic trees were constructed using MEGA-X v.10.2.6 software (http://www.megasoftware.net, accessed on 17 August 2021) [57] via the maximum-likelihood (ML) method with the following parameters: node robustness was detected using the bootstrap method, and the bootstrap was set to 100 replications.

The tissue-specific expression, level of expression (measured as FPKM), and differential expression between tissues were evaluated with RSEM [58] using RNA-seq data from NCBI Bioprojects 483249 and 656750. An AcFRS gene was considered not to be expressed if it had FPKM <2. AcFRS genes with FPKM ≥2 were considered to be expressed and AcFRS genes with FPKM ≥10 were considered "highly" expressed. NCBI Bioprojects 483249 and 483249 were used for differential expression analysis (up- or downregulated genes). For RNA-seq data analysis and interpretation of the results, each Bioproject was treated as an independent experiment. NCBI Bioproject 483249 is comprised of 11 different pineapple tissues representing the transcriptome during fruit development, and NCBI Bioproject 656750 is comprised of 27 different pineapple floral tissues representing the transcriptome during flowering.

3. Results and Discussion

3.1. Genome Assembly and Quality Assessment

In total, 32.6 Gb of raw PacBio whole-genome sequences, 52 Gb of Hi-C interaction data, and 143.7 Gb of Pacbio Iso-Seq sequences were generated (Supplementary Tables S1 and S2). Error correction and consensus generation of the raw PacBio genomic reads resulted in 2,014,848 (~25 Gb) high-quality error-corrected consensus sequences. A total of 125,054 (~201 Mb) polished, high-quality full-length circular consensus sequence (CCS) Iso-Seq reads were generated for subsequent data analysis (Supplementary Table S1). After trimming the Illumina Hi-C reads and removing PCR duplicates, 49 Gb of cleaned sequences were retained for downstream analysis.

PacBio sequences were used to perform de novo assembly and Hi-C sequences along with a linkage map were used to scaffold and phase the assembly [8], correct chimeric regions, and anchor the assembly to chromosomes (Supplementary Table S3). Gap-filling of the assembly with LR_Gapcloser resulted in ~63 kb additional known sequence added (Supplementary Table S4). The final phased assembly spanned 1.075 Gb, including 543.5 Mb for P0 and 531.6 Mb for P1 (Table 1 and Supplementary Table S3), accounting for ~96.5% (P0) of the estimated genome size (563 Mb per haploid phase) [4]. Each phase of the assembly contained 63 scaffolds and contigs, and 99.7% of the P0 and P1 assembled sequences were anchored to 50 chromosomes (25 chromosomes for each phase) (Figure 1a).

Table 1. Statistics of the MD2 v2 pineapple genome assembly and gene prediction.

	MD2 v2 P0			**MD2 v2 P1**		
	#	**Length [bp]**	**%**	**#**	**Length [bp]**	**%**
Assembly feature						
Sequences	63	543,505,080	96.5 */101.0 **	63	531,615,398	94.4 */98.8 **
Contigs	812	543,433,016	96.5 */101.0 **	805	531,544,088	94.4 */98.8 **
Max. sequence length		43,498,842			42,511,760	
Min. contig length		5678			4092	
Max. contig length		5,969,083			5,971,173	
Contig N50 length		1,524,720			1,521,169	
Scaffold N50 length		21,996,178			23,016,244	
Chromosome-anchored sequence		541,772,120	99.7		529,885,665	99.7
Genome annotation						
Transposable element content		337,179,261	62.0		329,341,567	62.0
Gene models	30,591	35,382,141		29,550	34,906,836	
Genes in pseudomolecules	30,590	35,382,088	100	29,550	34,906,836	100

* Estimated genome size of 563 Mb. ** Estimated genome size of 526 Mb. # indicates number of sequences for a given metric.

The F153 SSR linkage map aligned with high collinearity to the 25 chromosomes, demonstrating correct ordering and orientation of the final MD2 v2 assembly (Supplementary Figure S1). Similarly, the Hi-C heatmap showed a uniform distribution of genomic interactions along the diagonal, demonstrating the proximity of the assembled sequences and the quality of the assembly (Figure 1b and Supplementary Figure S2). The overall N50 in P0 and P1 was >22 Mb and the contig N50 was >1.5 Mb, similar to those of other high-quality genome assemblies, such as *Actinidia chinensis* v3.0 and *S. bicolor* v3.0 [36,59]. The longest chromosome (Chr 1) spanned 43.5 Mb and the shortest (Chr 25) spanned 4.4 Mb (Supplementary Table S5). The length of the longest contig was >5.9 Mb, covering a large part of the chromosome 13 long arm.

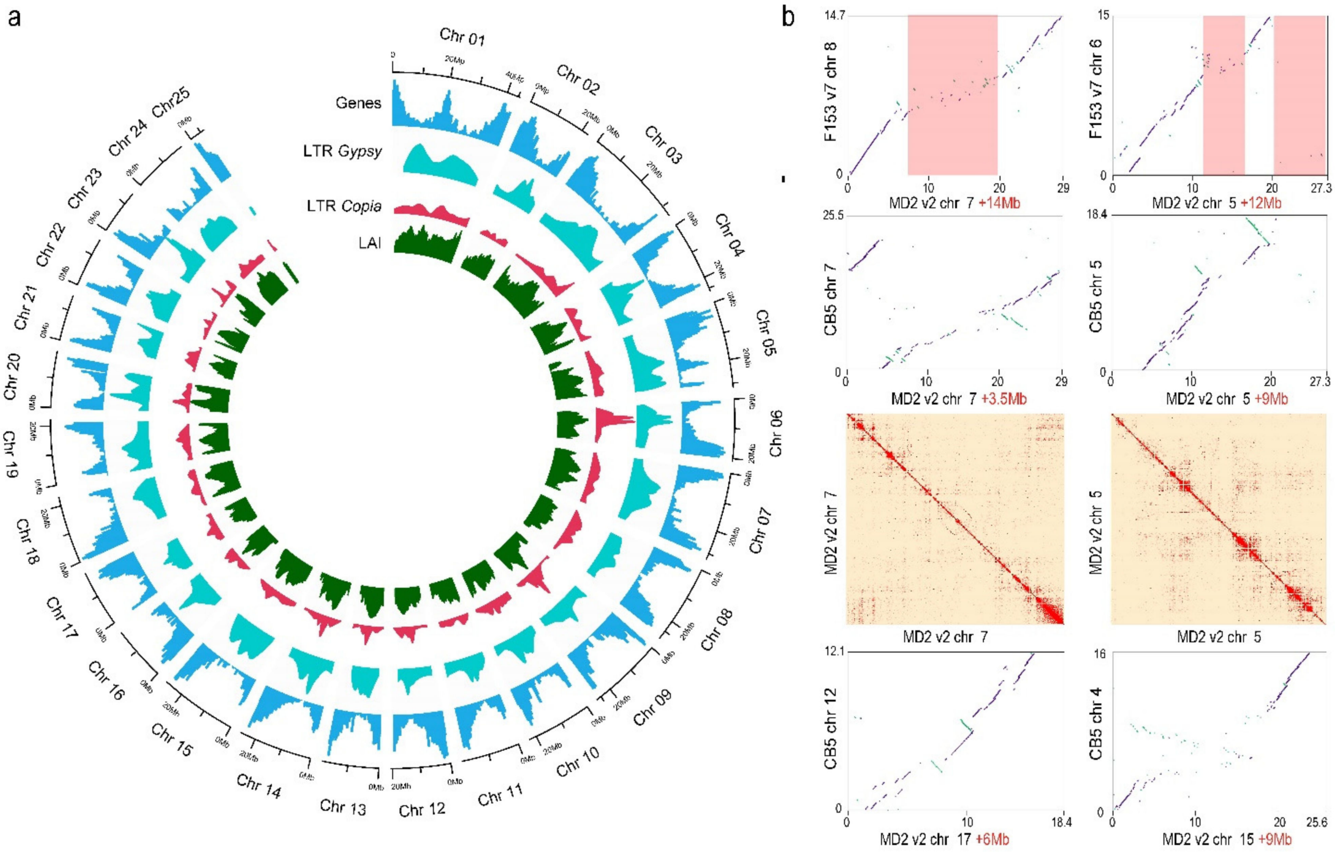

Figure 1. MD2 v2 P0 genomic features' landscape and comparative analysis. (**a**) CIRCOS plot illustrating the location/density of gene models, Ty3/Gypsy LTR, Ty1/Copia LTR, and LTR Assembly Index (LAI) score along the MD2 v2 P0 assembled chromosomes; (**b**) Visual representation of the quality of the MD2 v2 genome compared to other high-quality pineapple genomes (F153 v7 and CB5). Numbers in the *x* and *y* axes represent Mb sequences. Additional sequences assembled in MD2 v2 were primarily located in regions with high amounts of repeats (i.e., centromeres, highlighted with red shaded boxes). Contiguous interaction along the Hi-C plot validates the quality of the newly assembled regions. In each chromosome comparison, the number highlighted in red indicates the additional sequences assembled in the MD2 v2 assembly.

Gene space in the MD v2 assembly was assessed using transcriptome sequences and BUSCO analysis. Mapping 222 sets of transcriptome sequences from the NCBI SRA database indicated that 95.5% aligned with the MD2 v2 genome assembly (Supplementary Table S6). BUSCO analysis indicated that >97% conserved genes had a match in the MD v2 genome, of which >95% were detected as a complete structure. These results demonstrated that this assembly covered the majority of the gene space. Finally, no significant sequence contamination was detected by BLASTn alignment against a custom-made database that included bacterial, viral, and fungal DNA sequences.

3.2. Genome Assembly Comparisons

Comparison of all the currently available *Ananas* sp. genomes highlighted major improvements in the MD2 v2 assembly and some interspecific chromosomal rearrangements.

Compared to the previous MD2 v1 assembly, the MD2 v2 assembly had a 26.6-fold increase in contig N50 length, 19.4 Mb of total additional sequence (including Ns), and 33.5 Mb of additional known sequence at the contig level (Table 2). Most importantly, the MD2 v2 was assembled at the chromosome level and phased, while the MD2 v1 was only assembled at the scaffold level and not phased. The MD2 v2 assembly had a 13.3-fold increase in contig N50 length over the F153 v7 genome, 161.6 Mb of total additional sequence, and 168.3 Mb of additional known sequence at the contig level (Table 2). This represented >30% of the estimated genome size.

Table 2. Assembly statistical comparisons between MD2 v2 and publicly available pineapple genomes.

	MD2 v2 P0 vs. MD2 v1			MD2 v2 P0 vs. F153 v7			MD2 v2 P0 vs. CB5		
	#	**Length [bp]**	**%/Fold Change**	**#**	**Length [bp]**	**%/Fold Change**	**#**	**Length [bp]**	**%/Fold Change**
Assembly feature									
Sequences	−8385	+19,435,418	+3.5% */ +3.6% **	−3070	+161,599,861	+28.7% */ +30.7% **	−40	+30,269,689	+5.4% */ +5.6% **
Contigs	−20,927	+33,521,821	+6.0% */ +6.2% **	−8550	+168,298,756	+29.9% */ 32.0% **	−1158	+30,384,325	+5.4% */ +5.6% **
Max. sequence length		+42,211,785	+33.8 Fold		+26,164,174	+2.5 Fold		+16,366,087	+1.6 Fold
Min. contig length		+5677	+5678 Fold		+5497	+30.4 Fold		−6456	−2.1 Fold
Max. contig length		+4,742,061	+4.9 Fold		+4,957,247	+5.9 Fold		+3,781,837	+2.7 Fold
Contig N50 length		+1,467,419	+26.6 Fold		+1,410,093	+13.3 Fold		+1,098,024	+3.6 Fold
Scaffold N50 length		+21,843,094	+143.7 Fold		+10,290,687	+1.9 Fold		+1,817,735	+1.1 Fold
Chromosome-anchored sequence *		+541,772,120	NA		+225,928,334	+40.1%		+55,160,525	+9.8%
Genome annotation									
Transposable element content		+26,531,143	+2.76%		+136,080,198	+10.03%		+38,062,886	+3.76%
Gene models	+6993	+7,217,109		+3567	+3,729,585		NA	NA	NA
Genes in pseudomolecules	NA	NA		+6698	+6,439,201		NA	NA	NA

* Estimated genome size of 563 Mb. ** Estimated genome size of 526 Mb. NA indicates cases where data for calculating differences were not available for the published genomes. # indicates number of sequences for a given metric.

Compared to the *A. comosus* genomes that are assembled at the chromosome scale [18], MD2 v2 had 40.1% more of its sequence anchored to chromosomes than F153 (Table 2).

At the structural level, the comparison with F153 v7 [18] indicated that the assemblies were highly collinear; however, several novel sequences appeared to be inserted into the MD2 v2 assembly (Supplementary Figure S3). Over 33 Mb (76% of its length) of new known sequences were assembled in chromosome 1, and a total of 10 chromosomes in the MD2 v2 assembly had >10 Mb of additional known sequence compared to the corresponding chromosomes from F153 v7 (Supplementary Table S7). The increase in known sequence content could be partially attributed to the fact that more repetitive regions were sequenced and assembled in the MD2 v2 assembly (Figure 1a,b). Regions of high repeat density in the MD2 v2 chromosomes coincided with newly assembled sequence regions (Figure 1b). MD2 v2 chromosome 25 was the only chromosome in the assembly that was smaller than (and, therefore, had less known sequence) the corresponding chromosome in F153 v7. This difference was due to chimeric sequences identified in the F153 v7 chromosome 25. Indeed, the F153 SSR linkage map and MD2 Hi-C data supported the assembled structure of MD2 v2 chromosome 25 (Supplementary Figure S4). Other potential chimeric regions and very large gaps in the F153 v7 chromosomes also existed (Supplementary Figure S3). For example, the MD2 v2 chromosome 1 was collinear to F153 v7 chromosome 1 and 24, indicating that the F153 v7 chromosome 1 and 24 should be combined into one sequence (Supplementary Figure S3). These results were supported by the contiguous interaction in the Hi-C heatmap for MD2 v2 chromosome 1 (Supplementary Figure S2).

Comparison with the CB5 genome (*A. bracteatus*) [18] revealed moderate collinearity with multiple rearrangements, including a small and large inversion in CB5 chromosomes 12 and 8 (Supplementary Figure S5a,b), respectively, and large translocations in CB5 chromosomes 3, 7, 11, 19, and 20 (Figure 1b and Supplementary Figure S6). To gain some preliminary support for the presence of these chromosome rearrangements, the interaction signals of the MD2 Hi-C data mapped to the CB5 genome were used for verification. Non-collinear regions of CB5 chromosomes 8 and 12 were manually inverted to establish putative collinear regions. After the manual inversions, the sequence spanning the breaking points of the inverted regions had strong Hi-C interaction signals outside of the diagonal, indicating that these regions were more proximal, consistent with the original CB5 assembly (Supplementary Figure S7). These results suggested that the rearrangements represented true interspecific chromosomal differences between the CB5 and MD2 genomes.

3.3. Repetitive Sequence Annotation and Analysis

TEs accounted for about 62% of the MD2 v2 pineapple genome (P0 and P1) (Supplementary Table S8). The two phases slightly differed, with the sequence masked 337.17 Mb for P0 and 329.34 Mb for P1. On average, the MD2 v2 class I elements (LTR, LINE, and SINE) and class II elements (TIR, MITE, and Helitron) occupied 211.85 and 121.4 Mb DNA sequences, accounting for 39.41% and 22.58% of the genome, respectively. Among LTR-TE, Ty3/Gypsy and Ty1/Copia represented approximately 24.94% and 6.72% of the genome, respectively.

Compared to the previous MD2 v1 assembly, the MD2 v2 has about 22.61 Mb additional sequence annotated as TE, with most of this difference being due to an increase (1.76%) in Ty3/Gypsy LTR elements (Figure 1). Wider differences were instead observed in comparisons with the other pineapple genomes considered, showing an increase of about 4% and 10% over CB5 and F153 v7, respectively (Supplementary Table S8).

Up to 12,630 and 12,439 intact TEs were annotated in the MD2 v2 P0 and P1 haplotypes, respectively, in contrast to the 5721 annotated in the previous MD2 v1 genome (Figure 2a). This increment was found in LTR, TIR, MITE, and Helitron, but it was especially evident in the LTR order. These elements occupied up to 61.9% of the MD2 v2 genome and 59.11% in the MD2 v1 genome. The LTR-assembly index (raw LAI) of MD2 v2 ranged from 21.29 (P0) to 21.77 (P1), whereas for MD2 v1 it was 6.63 (Figure 2b), indicating a much higher proportion of intact vs. total LTR-retrotransposons in the MD2 v2 genome, the result of an improved assembly. The LAI score was evenly high across all the chromosomes (Figure 1a). The corrected LAI index of MD2 v2 (19.66, P1) was comparable or superior to other high-quality plant reference genomes (i.e., *Arabidopsis* TAIR10: 14.9, rice MSU7: 21.1, and maize B73 v4: 20.7 [24]). In comparison to the CB5 and F153 v7 genomes, MD2 v2 also showed a significantly higher LAI and proportion of intact elements (Figure 2a,b).

Analysis of the LTR-retrotransposon insertion time in MD2 v2 showed a strong peak of activity at 200 Ky and a very low number of old elements (Figure 2c). A similar profile was found in the CB5 genome, and to a lower extent, in MD2 v1. F153 v7 lacked recent LTR-retrotransposon peaks and had a peak of activity at approximately 1.6 My. The distribution of elements per age suggested that the younger elements, likely forming large regions of repetitive sequences with high similarity, could not be assembled in the F153 v7, but were assembled in the MD2 V2 assembly.

3.4. Gene Prediction and Annotation

In total, 60,141 gene models for the diploid-phased MD2 v2 genome were predicted; 30,591 and 29,550 in P0 and P1, respectively. This represented over 6000 and 3000 extra genes compared with the MD2 v1 and F153 genomes, respectively. Overall, the general statistics (e.g., gene length, intron and exon length) of the predicted MD2 v2 gene models were in line with previously predicted gene sets for pineapple MD2 v1 and F153 v3, as well as gene sets for eight non-pineapple species (Supplementary Table S9). The same set of predicted genes was reported for both F153 v3 [5] and F153 v7 [18]; therefore, one gene prediction was used for comparisons of MD2 v2 and F153.

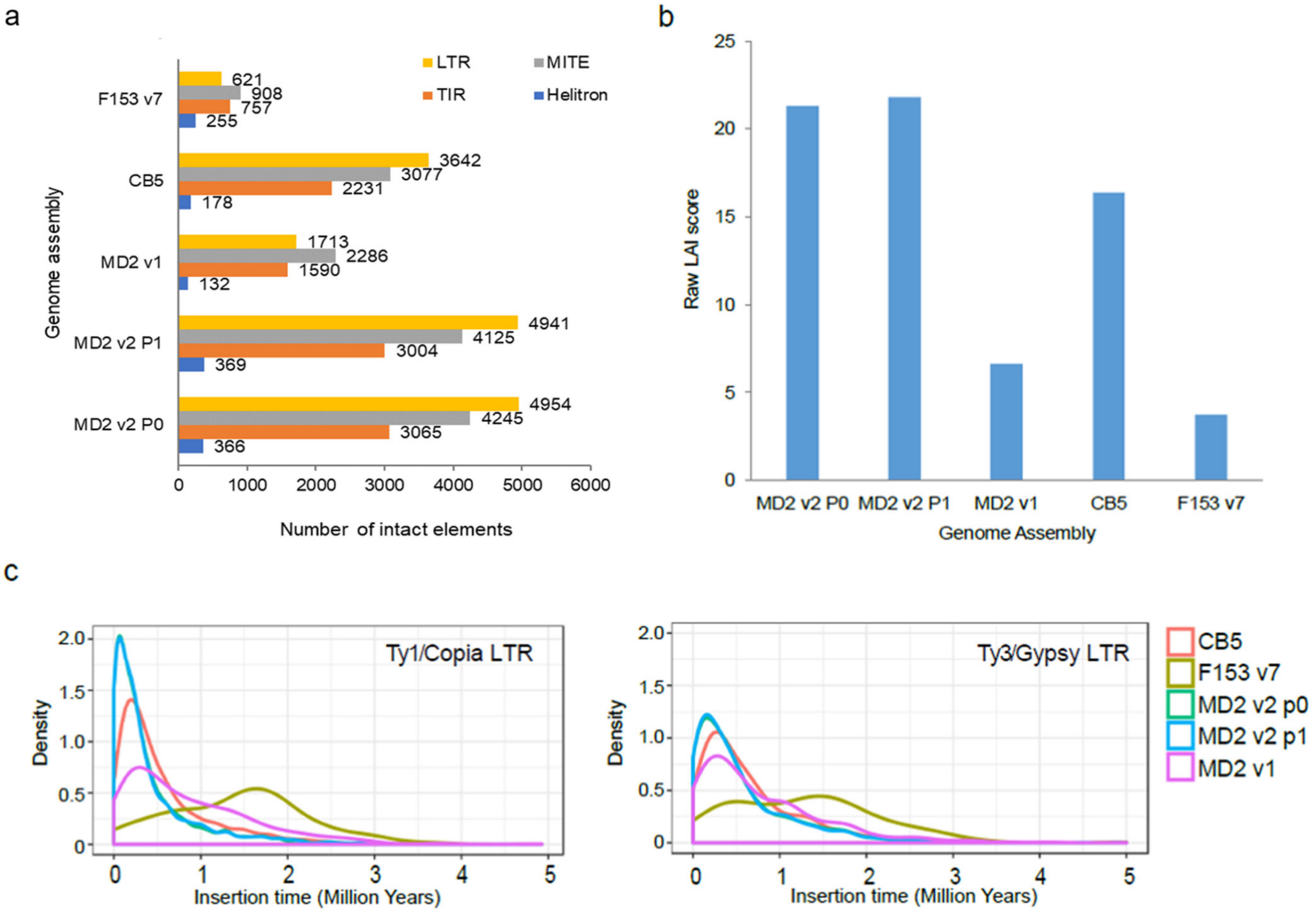

Figure 2. Annotation of repetitive sequences in MD2 v2 and other pineapple genomes (MD2 v1, F153 v7, and CB5). (**a**) Intact transposable elements (TEs) assembled in each pineapple genome; (**b**) Raw LTR Assembly Index (LAI) score for each pineapple genome. The LAI score is calculated as the ratio of intact elements to total elements in the assembly; therefore, a higher LAI score indicates a higher quality assembly with a larger proportion of intact elements assembled; (**c**) Insertion time for Ty3/Gypsy LTR and Ty1/Copia LTR elements in pineapple genomes. Lower insertion time indicates more recent origin. Density was calculated using the function geom_density from ggplot2 using the insertion time of all intact elements in million years.

In the MD2 v2 assembly, the structures of 15,275 and 13,655 genes were supported by PacBio Iso-Seq transcript alignment in the P0 and P1 haplotypes, respectively. BUSCO analysis of MD2 indicated that the v2 predicted gene set had an 11.8% and 4% higher rate of completeness compared to v1 and F153 v3, respectively and the least number of fragmented genes (Supplementary Table S10).

BLASTn alignments of MD2 v2 P0 and MD2 v1 CDSs revealed that only 3612 genes had an identical structure, while the majority of the predicted genes had a different structure. Iso-Seq data supported the quality of the MD2 v2 gene prediction (Supplementary Figure S8a,b). Additionally, reciprocal BLASTn alignment of MD2 v1 and v2 CDS indicated that a total of 5689 genes predicted in the v2 genome were not predicted in the v1 genome. Also, a total of 2060 genes predicted in the MD2 v2 genome were not predicted in the F153 v3 genome and 1483 genes had no BLAST hits to genes in either the MD2 v1 or F153 v3 genomes; therefore, these were designated as novel genes in pineapple.

Alignment of Illumina RNA-seq reads from the NCBI SRA database with the full set of predicted genes showed that 911 of the new genes were expressed, providing additional evidence for the functional relevance of these novel pineapple genes.

In MD2 v2 P0, 19,552 (64%) genes were fully annotated with computationally reliable GO terms, 2709 (9%) genes were assigned a putative GO term, and 6384 (21%) genes had BLAST hits, but no associated GO terms. Overall, only 1946 (6%) genes had no functional annotation assignments. In the MD2 v2 P1 haplotype, 19,832 (67%) genes were annotated with GO terms and 2412 (8%) genes had a putative GO assignment. One hundred and fifty (1%) genes had protein domain hits to the InterProScan database, but did not have any BLAST hits to the non-redundant protein database. Finally, 5729 (19%) genes had BLAST hits but no associated GO terms and 1427 (5%) genes had no functional annotation assignments.

Overall, 26,982 KOGs were obtained from the analysis. A total of 25,368 (83%) genes were assigned to at least one KOG functional category based on orthologous proteins (Supplementary Figure S9), which means that some of the annotated genes were assigned to more than one orthologous group. A large portion (83%) of the predicted genes were successfully assigned a function based on orthologous group clustering, providing additional evidence for the new predictions.

Forty-eight percent of the novel MD2 v2 genes were annotated with at least one of the databases used (Supplementary Figure S10) and the distribution of KOG functional categories assigned to the novel genes (Supplementary Figure S11) was similar to that of the complete set of predicted genes, except the set of novel genes had a higher proportion of sequences associated with DNA replication, recombination, and repair functions. Direct GO counts of the novel genes indicated that they function in a diverse array of biological processes, including DNA integration and repair, telomere maintenance, protease activity, and flowering (Supplementary Figure S12).

The pineapple MD2 v2 P0 assembly encoded 3550 regulatory genes (TF, TR, and CR) and 1467 R-genes (Supplementary Tables S11 and S12). The MD2 v2 predicted gene set encoded 686 and 662 additional regulatory genes (TF, TR, and CR) and 118 and 211 additional R-genes compared to former pineapple gene sets MD2 v1 and F153 v3, respectively.

3.5. Haplotype Comparison

To provide insight into the divergence between the two haplotypes, we estimated polymorphisms between the 25 homologous chromosome pairs. Based on the alignment of the genes on the two haplotypes, 185 syntenic blocks were identified. Between homologous chromosomes, 13,367,404 SNPs, 1,998,259 INDELs, and 12,468 structural variants (SVs, >50 bp) were identified (Figure 3 and Supplementary Figure S13; Supplementary Table S13). Among the SVs identified, 548 SVs spanned >10 kb. Comparison of the two haplotypes exhibited an intragenomic diversity ranging from 1.1% on chromosome 6 to 4.5% on chromosome 18 (Figure 3 and Supplementary Figure S13; Supplementary Table S13). Based on synteny and annotation, out of 60,141 predicted genes in the two haplotypes, 28,811 pairs (56,167 genes, 93% of all predicted genes) were identified as having homologs on the two haplotypes, and therefore, were considered as homologous gene pairs.

Among them, 12,745 gene pairs harbored variants with moderate to high impact on the protein coding regions, which could represent potential alternative alleles. Furthermore, a total of 3974 present and absent (PAV) genes were identified between the two haplotypes.

GO enrichment analysis for the PAV genes identified 26 GO terms including 18 biological processes, five molecular functions and three cellular components that were significantly enriched. Biological processes such as DNA integration, positive regulation of the hydrogen peroxide metabolic process, base-excision repair, and AP site formation via deaminated base removal were the most enriched GO terms (Supplementary Table S14). ADP binding, pre-mRNA 5′-splice site binding, and 5′-deoxyribose-5-phosphate lyase activity were among the molecular functions that were significantly enriched in the GO terms; these biological processes were located in cellular components such as the spindle pole body and eukaryotic translation initiation factor 3 complex. In addition to these genes with GO term annotation, 219 were also identified as R-genes. These results suggested that PAV genes may play an important role in cellular maintenance and defense.

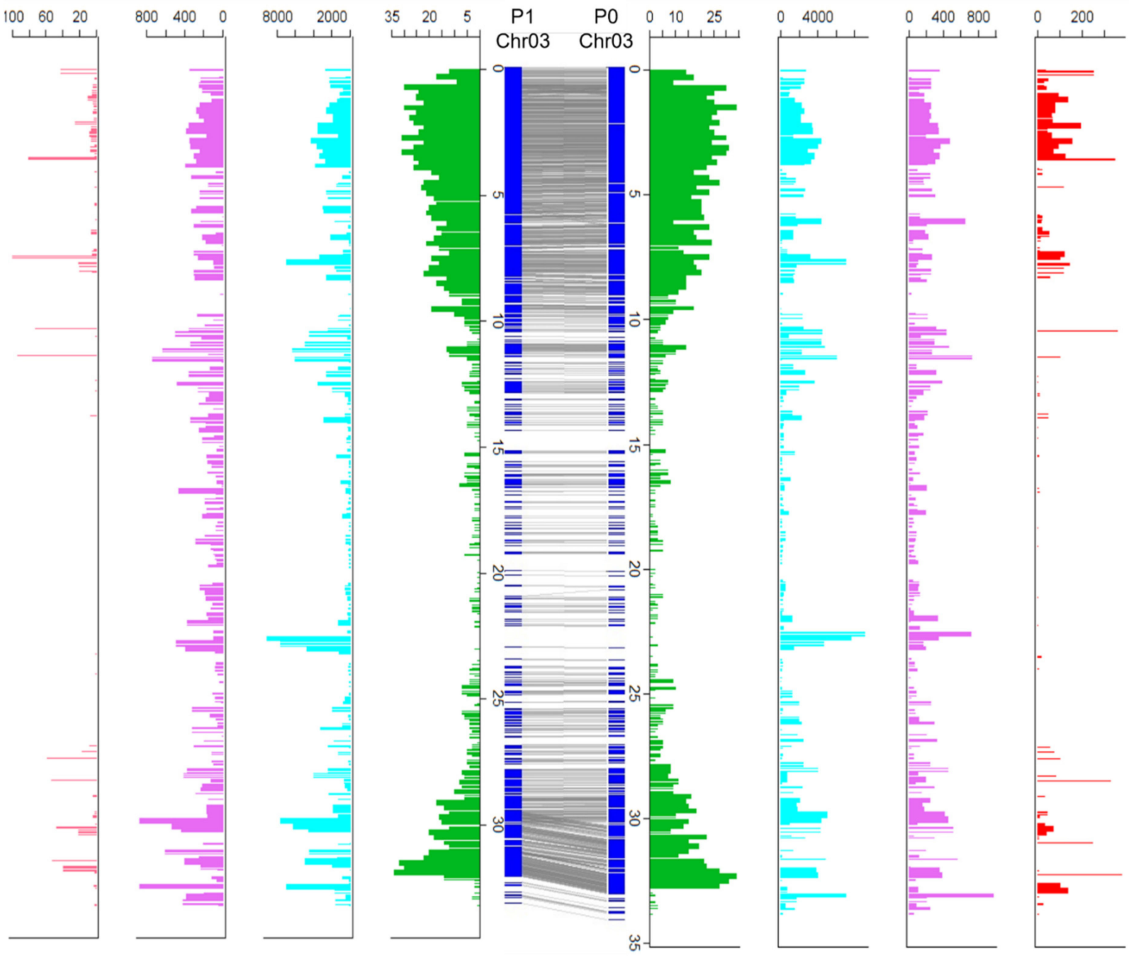

Figure 3. Haplotype diversity in the MD2 chromosome 3. The central blue bars represent the two haplotypes of chromosome 3. The gray lines indicate collinear genes. Outer plots include: distribution of genes (green), SNP density (cyan), INDEL density (pink), and density of potential deleterious effect variants (red). All numbers were determined in 200 kb windows.

3.6. Characterization of FRS/FRF Family Transcription Factors

Comparative analysis of TFs revealed that pineapple and rice harbored a larger number of FRS/FRF genes. A total of 82 pineapple MD2 v2 and 84 rice genes were identified as proteins containing the FAR1 DNA binding domain (IPR004330). Among the other genomes included in this analysis, the number of genes predicted as FRS/FRF family genes was much lower, ranging from six in banana (*M. acuminata*) to 49 in sorghum (*S. bicolor*) (Supplementary Table S11). Interestingly, the number of FRS/FRF genes was also lower in F153. It is unlikely that the larger number of genes detected in MD2 represents an expansion relative to F153 since they represent the same species and germplasm pool. Given the overall lower number of genes predicted in the F153 genome (Table 2), we readily suggest that this difference is due to misprediction in F153. Due to the large number of FRS/FRF genes in pineapple MD2 and rice, we wanted to investigate to what degree the expansion of this gene family was the result of shared duplications versus lineage-specific duplications. Also, we wanted to investigate which specific subgroups expanded their potential function by orthology, and their possible functional specialization. In the model species *Arabidopsis*, the FRS/FRF family has been implicated in multiple developmental functions related to traits of particular importance for pineapple cultivar development [5], including circadian clock regulation and flowering time regulation [10–13]. Recent work in pineapple identified FRS genes as being implicated in flower ovule development [60]. Characterizing this family of genes will lay the foundaition for future genetic studies in pineapple. To this end, the curated annotation, mode of duplication, evolutionary/phylogenetic relationships, and pattern of expression were characterized. Since the FRS/FRF family of TFs is derived from ancient transposases and contains a transposase domain, all AcFRS genes were searched against a plant TE database using BLASTn. None of the predicted AcFRSs had significant similarity with TEs. In addition, predicted AcFRS genes harbored other conserved domains including: Nuclear transcription factor Y subunit A (IPR001289), Zinc finger CCHC-type (IPR001878), WRKY domain (IPR003657), and Zinc finger PMZ-type (IPR006564). These results demonstrated that the predicted AcFRS genes were not TEs. It is also important to note that 19 AcFRS genes (23%) were PAV genes and 10 were part of the novel genes that were not predicted in the previous pineapple genome assemblies. This observation further demonstrated the value of the MD2 v2 assembly and gene prediction improvements.

The AcFRSs were grouped into one of seven groups (A–G) based on their domain content (Supplementary Table S15) and named as AcFRS1-AcFRS82. Specific DNA binding domains within a TF protein determine its potential targets for transcriptional regulation. All identified MD2 v2 FRS/FRF family genes contained at least a FAR1 DNA binding domain (IPR004330), and the majority (62) contained the MULE transposase domain (IPR018289) (Supplementary Figure S14). Group G, containing genes with a FAR1 DNA binding domain (IPR004330) and Zinc finger PMZ-type domain (IPR006564), made up the majority of AcFRS genes (46, 56%). The second largest group (F, 29 genes) contained genes with only the FAR1 DNA binding domain (IPR004330). Groups A–D each contained only a single gene. Group A contained a single gene with the Nuclear transcription factor Y subunit A (IPR001289), FAR1 DNA binding (IPR004330), and Zinc finger PMZ-type (IPR006564) domains. Group B contained a single gene with four domains: Zinc finger CCHC-type (IPR001878), WRKY (IPR003657), FAR1 DNA binding (IPR004330), and Zinc finger PMZ-type (IPR006564). Group C contained a single gene with Zinc finger CCHC-type (IPR001878) and FAR1 DNA binding (IPR004330) domains. Group D contained a single gene with Zinc finger CCHC-type (IPR001878), FAR1 DNA binding (IPR004330), and Zinc finger PMZ-type (IPR006564) domains. Finally, group E contained three genes with WRKY (IPR003657), FAR1 DNA binding (IPR004330), and Zinc finger PMZ-type (IPR006564) domains. We speculate that the differences in domain structure among the AcFRS groups create a complex network of FRS/FRF family TFs with both new and overlapping functions that are capable of binding a wide range of *cis*-elements of target genes. Among the AcFRSs, group G appears to be expanded, and therefore, is more likely to contain genes that serve redundant roles; some may have acquired new functions or new targets as a result of expansion in this group.

Orthologous analysis of FRS TF proteins from all species resulted in 57 total groups (Supplementary Table S16). Twenty-three of those groups included at least one FRS genes from pineapple genomes (MD2 and F153) and nineteen groups included at least one FRS gene from rice. The majority (>65%) of the pineapple FRS genes grouped with orthologous FRS from other species, including some that have been functionally characterized. For instance, orthogroup 533 included nine AcFRSs, as well as *Arabidopsis* orthologs FHY3 and FAR1 that respond to light signals to regulate/modulate multiple processes including flowering time, seed dormancy, starch synthesis, the shade avoidance response, and balance between growth and defense responses under shade conditions [10,61]. Orthogroup 304 included eight AcFRSs and *Arabidopsis* FRS7 and FRS12, which coregulate flowering time and glucosinolate biosynthesis [62,63]. Orthogroup 756 included six AcFRSs and *Arabidopsis* FRS6 and FRS8, which have been suggested to regulate flowering time [64]. Interestingly, in these three orthogroups (304, 533, and 756), the number of AcFRS genes was higher than in the other species, suggesting that pineapple may have evolved FRS orthologs with new or specialized functions. In comparison with rice FRS (OsFRS) genes, we noted that although rice and pineapple share several FRS orthologous groups, the groups that harbor a larger number of FRS genes in each species are different (e.g., group 240: 19 AcFRS and 1 OsFRS). This demonstrated that independent duplication events contributed to accumulating the larger number of FRS genes observed within the Poales lineage.

Phylogenetic analysis clustered FRS/FRF genes into eight clades, four of which were split into two subclades (Figure 4). In this analysis, only phylogenetic clades that contained either pineapple or *Arabidopsis* FRSs were assigned a clade number (clade I–VII). Neither orthologous nor phylogenetic analyses appeared to group the AcFRS proteins based on their conserved domains. Phylogenetic clade VIB contained only pineapple-derived FRS/FRF genes and clade IIB contained primarily pineapple FRS/FRF genes (pineapple and one single papaya gene). One phylogenetic group containing *O. sativa* FRS/FRF genes (clade VB) did not contain any pineapple FRS/FRF genes. Within clades containing both pineapple and rice FRS/FRF genes, there was a clear separation between pineapple FRS/FRF genes and *O. sativa* FRS/FRF genes. These results suggest that the expansion of the FRS/FRF gene family in pineapple and rice was largely the result of independent expansion events. However, some shared signatures of conservation within this lineage were observed. For example, within clade VII, 84 out of 93 (90%) FRS/FRF genes belonged to Poales species, suggesting that the ancestor of this clade was selected within this lineage and duplicated independently after their divergence.

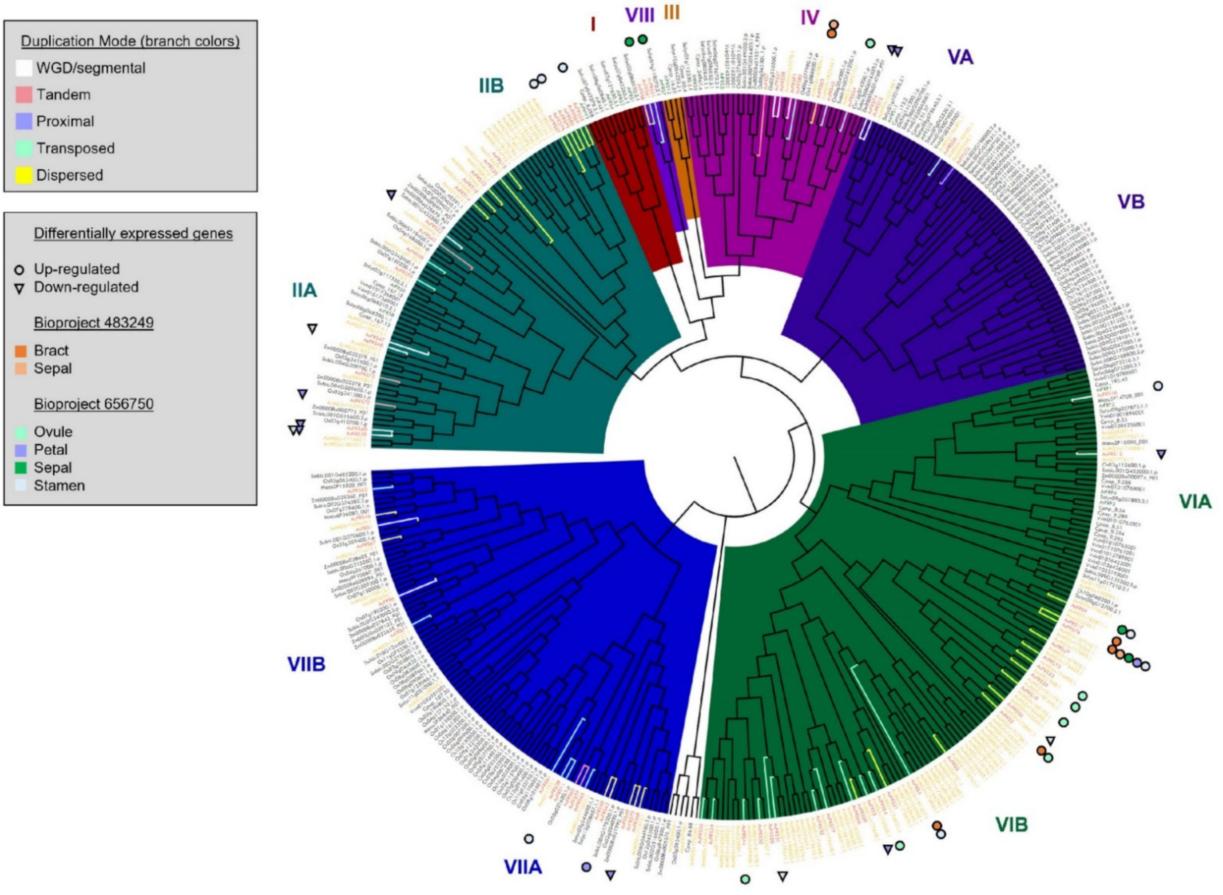

Figure 4. Phylogenetic tree of FRS/FRF genes from MD2 v2, other pineapple genomes (MD2 v1 and F153 v3), and eight non-pineapple species (*A. thaliana*, *C. papaya*, *M. acuminata*, *O. sativa*, *S. lycopersicum*, *S. bicolor*, *V. vinifera*, and *Z. mays*). Only clades containing FRS/FRF genes from MD2 v2 or *Arabidopsis* were numbered (Clades I–VIII). Gene names in orange correspond to pineapple MD2 v2, yellow correspond to other pineapple genomes (MD2 v1 and F153 v3), and green correspond to *A. thaliana*. Duplication mode is indicated by branch color. Differentially expressed (DE) MD2 v2 genes are indicated by colored circles or triangles next to gene names.

To examine the potential mechanism underlying the FRS/FRF family expansion in pineapple, AcFRS duplicated gene pairs were identified with DupGen_finder software [65], which classified each pair into one of five categories of duplicated gene pairs: whole genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DSD) pairs. Among these categories, the DSD category had the most duplicated gene pairs (121), followed by the TRD category (38) (Supplementary Table S17). FRS/FRF genes are derived from TEs [12]; therefore, the results of the duplication mode analysis were consistent with what would be expected for these genes. In other genomes, such as rose, it has been demonstrated that the FRS/FRF gene family likely expanded due to TE-associated DSD [66]. The DSD and TRD modes of duplication contributed to accumulating a larger number of FRS genes in pineapple, which evolved in pineapple-specific clades (IIB and VIB) and may have acquired new specialized function within the pineapple lineage. For instance, 10 AcFRSs grouped in the same subclade (IIA) as *Arabidopsis* AtFRS6 and AtFRS8, but subclade IIB has eleven additional AcFRSs with no proximal *Arabidopsis* FRS/FRF genes. AtFRS6 and AtFRS8 have been shown to negatively regulate flowering time [64]; therefore, AcFRSs in clade II may play a role in regulating the flowering time in pineapple. AcFRSs in subclade IIB may have acquired new functions related to flowering, leading to divergence from the AcFRSs in subclade IIA.

Available RNAseq data were used to evaluate the tissue-specific expression, expression level, and differential expression of AcFRS genes. This analysis revealed that 13 AcFRSs were expressed in all tissues, while 69 were expressed in at least one tissue and 13 were not expressed (Supplementary Table S18). Several AcFRSs were specifically expressed in bract, ovule, petal, sepal, and/or stamen tissue and highly expressed in the ovule, stamen, and gynoecium.

Differential expression analysis across 11 different pineapple tissues representing the transcriptome during fruit development revealed that three AcFRS genes were specifically upregulated in bract tissue and two were upregulated in both the bract and sepal (Supplementary Table S19). Differential expression analysis across 27 different floral tissues at different developmental stages identified 30 differentially expressed AcFRS genes. Of these, 18 were upregulated in one or more tissues, including a subset DE during ovule, sepal and stamen development, and 12 were downregulated, including nine during the final stages of petal development.

Overall, out of 32 unique DE genes, the majority (23, 72%) were derived from pineapple lineage-specific duplication (e.g., clade VIB), through transposed or dispersed duplication mechanisms (Figure 4). Similarly, all eight genes expressed in a tissue-specific manner were duplicated through a dispersed or transposed mechanism. These results demonstrated that expansion of AcFRS genes in the pineapple lineage contributed to evolving specialized expression patterns in reproductive tissues, which may have resulted in specialized functions related to the light-signaling pathway in floral tissues and reproductive organ development. These results complement previous findings, indicating that FRS/FRF genes in pineapple are one of the TF gene families highly expressed in flower tissue [67].

4. Conclusions

The final MD2 v2 assembly presented in this study represents the first MD2 chromosome-scale level assembly and the first pineapple phased genome assembly. Comparative analysis with all available pineapple genomes demonstrated higher completeness in term of sequence contiguity, gene content, and structure. Haplotype analysis, uncovered at the genomic level, the impact that the high level of heterozygosity maintained in this crop due its outcrossing nature has on genes' content and allelic diversity. Comparative analysis with the genome of its close relative *A. bracteatus* highlighted some major genomic differentiation. Overall, the MD2 v2 genome assembly represents the most complete assembly of the pineapple cultivated germplasm. In conclusion, the MD2 v2 genome represents a new valuable resource for genetic and comparative genome analysis in pineapple that is critical for advancing marker-assisted breeding strategies in this crop.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes13010052/s1, Figure S1: Alignment of the SSR marker sequences representing the F153 linkage map against the MD2 v2 chromosomes, Figure S2: MD2 v2 assembly Hi-C heatmap showing a uniform distribution of genomic interactions along the diagonal (indicated by red color on the diagonal), Figure S3: Nucmer alignment of the MD2 v2 P0 and F153 v7 genome assemblies. Large gaps are circled in red to highlight improvements in the MD2 v2 assembly, Figure S4: SSR linkage map (left) and Hi-C data (right) before (a) and after (b) correction of MD2 v2 chromosome 25, Figure S5: Small (a) and large (b) inversions in CB5 chromosome 12 and chromosome 8, respectively, Figure S6: Example of intrachromosomal translocation in CB5 chromosome 19, Figure S7: Manual inversions of non-collinear regions in CB5 chromosome 8, Figure S8: Evidence of Iso-Seq transcript data supporting the quality of the MDv2 gene prediction, Figure S9: KOG (euKaryotic Clustering of Orthologous Groups) analysis of the complete set of genes for MD2 v2 P0, Figure S10: Data distribution for functional annotation results of novel pineapple genes with Blast2GO, Figure S11: KOG (euKaryotic Clustering of Orthologous Groups) analysis of the novel subset of genes for MD2 v2 P0, Figure S12: Direct biological process GO term counts for the novel genes identified in MD2 v2, Figure S13: Haplotype diversity within the pineapple MD2 v2 genome, Figure S14: Representative protein structures for pineapple MD2 v2 FRS/FRF family genes (AcFRSs), Table S1: Summary of the pineapple 'MD2' sequencing data generated in this project using Illumina and PacBio sequencing platforms, Table S2: Statistics of PacBio Iso-Seq data generated by sequencing the pineapple 'MD2' transcriptome, Table S3: Summary statistics of the MD2 v2 Pineapple genome assembly at multiple steps, Table S4: Gap-filling summary results, Table S5: Pineapple MD2 v2 genome sequence assembly organized by chromosomes, Table S6: Assessment of the gene space coverage using publicly available NCBI SRA RNA-seq reads, Table S7: Comparison of total known sequence length between MD2 v2 and F153 v7 assembled chromosomes, Table S8: Transposable element (TE) content and distribution for multiple pineapple genomes, Table S9: General statistics of predicted protein-coding genes of the new MD2 v2 genome and genomes used for gene prediction training and inter-species comparative analyses, Table S10: BUSCO analysis results of the new MD2 v2 gene models and previous pineapple gene models, Table S11: Regulatory genes identified in the pineapple MD2 v2 genome and other plant genomes, Table S12: Disease resistance genes (R-genes) identified in the pineapple MD2 v2 genome and other genomes, Table S13: Results for comparative haplotype analysis of MD2 v2 P0 versus P1, Table S14: Results for gene ontology (GO) term enrichment analysis performed on presence-absence variance (PAV) genes identified between MD2 v2 P0 and P1 haplotypes, Table S15: Summary table for MD2 v2 AcFRS genes, Table S16: Orthologous grouping of FRS/FRF transcription factor genes identified in MD2 v2 and other genomes, Table S17: Duplication mode for FRS/FRF transcription factor genes in MD2 v2 (AcFRS genes), Table S18: FPKM expression results for AcFRS genes using RNA-seq data from two NCBI Bioprojects, Table S19: Summary of differential expression results for AcFRS genes using RNA-seq data from two NCBI Bioprojects, Dataset S1: TE library for pineapple repeat analysis.

Author Contributions: M.I. conceived the study and coordinated the project; M.I., A.G.Y. and H.B. designed the study; A.G.Y. and J.C. prepared the DNA/RNA samples and sequencing libraries; H.B., A.G.Y., V.R., M.F.M. and R.C. performed bioinformatics analysis; A.G.Y., M.I., V.R. and R.C. drafted the manuscript; R.Y. collected and provided plant material; N.G. coordinated and performed PacBio sequencing; all the authors edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Dole Nutrition Institute at Dole Food Company Inc., and M.I. and J.C. were also supported by the United States Department of Agriculture National Institute of Food and Agriculture, Hatch project 1008691.

Data Availability Statement: Data including the genome sequence, annotation, and raw sequencing reads have been deposited in the NCBI under BioProject ID PRJNA719415 and submission ID SUB9421765.

Acknowledgments: We are grateful to the NC State Bioinformatics Research Center for allowing us use of their high-performance computing cluster. Also, we wish to thank Jenny Xiang from the Department of Plant and Microbial Biology, NCSU for her constructive comments.

Conflicts of Interest: Young, who represents the funder (Dole Plc.), was involved in the collection of the plant material but his contribution did not influence the experimental design, data analysis, results or their interpretation, nor the conclusion of the work presented here. All other authors declare no conflict of interest.

References

1. Wali, N. Chapter 3.34—Pineapple (*Ananas comosus*). In *Nonvitamin and Nonmineral Nutritional Supplements*; Nabavi, S.M., Silva, A.S., Eds.; Academic Press: Cambridge, MA, USA, 2019; pp. 367–373. ISBN 978-0-12-812491-8.
2. Redwan, R.M.; Saidin, A.; Kumar, S.V. The draft genome of MD-2 pineapple using hybrid error correction of long reads. *DNA Res.* **2016**, *23*, 427–439. [CrossRef]
3. Ahmar, S.; Gill, R.A.; Ki-Hong, J.; Faheem, A.; Qasim, M.U.; Mubeen, M.; Zhou, W. Conventional and Molecular Techniques from Simple Breeding to Speed Breeding in Crop Plants: Recent Advances and Future Outlook. *Int. J. Mol. Sci.* **2020**, *21*, 2590. [CrossRef]
4. Abdul Rahman, A.; Kumar, V. Estimation of the Pineapple Genome Size by Using Quantitative Real-Time Polymerase Chain Reaction. In Proceedings of the 9th Malaysia Genetics Congress, 9th Malaysia Genetics Congress, Kuching, Malaysia, 28–30 September 2011.
5. Ming, R.; VanBuren, R.; Wai, C.M.; Tang, H.; Schatz, M.C.; Bowers, J.E.; Lyons, E.; Wang, M.-L.; Chen, J.; Biggers, E.; et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **2015**, *47*, 1435–1442. [CrossRef] [PubMed]
6. Li, C.; Lin, F.; An, D.; Wang, W.; Huang, R. Genome Sequencing and Assembly by Long Reads in Plants. *Genes* **2017**, *9*, 6. [CrossRef] [PubMed]
7. Kronenberg, Z.N.; Hall, R.J.; Hiendleder, S.; Smith, T.P.L.; Sullivan, S.T.; Williams, J.L.; Kingan, S.B. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv* **2018**, 327064. [CrossRef]
8. Lajoie, B.R.; Dekker, J.; Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods* **2015**, *72*, 65–75. [CrossRef] [PubMed]
9. Wang, B.; Kumar, V.; Olson, A.; Ware, D. Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. *Front. Genet.* **2019**, *10*, 384. [CrossRef] [PubMed]
10. Ma, L.; Li, G. FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) Family Proteins in Arabidopsis Growth and Development. *Front. Plant. Sci.* **2018**, *9*, 692. [CrossRef] [PubMed]
11. Liu, Y.; Ma, M.; Li, G.; Yuan, L.; Xie, Y.; Wei, H.; Ma, X.; Li, Q.; Devlin, P.F.; Xu, X.; et al. Transcription Factors FHY3 and FAR1 Regulate Light-Induced CIRCADIAN CLOCK ASSOCIATED1 Gene Expression in Arabidopsis. *Plant. Cell* **2020**, *32*, 1464–1478. [CrossRef] [PubMed]
12. Lin, R.; Ding, L.; Casola, C.; Ripoll, D.R.; Feschotte, C.; Wang, H. Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science* **2007**, *318*, 1302–1305. [CrossRef] [PubMed]
13. Tang, W.; Wang, W.; Chen, D.; Ji, Q.; Jing, Y.; Wang, H.; Lin, R. Transposase-derived proteins FHY3/FAR1 interact with PHYTOCHROME-INTERACTING FACTOR1 to regulate chlorophyll biosynthesis by modulating HEMB1 during deetiolation in Arabidopsis. *Plant. Cell* **2012**, *24*, 1984–2000. [CrossRef]
14. Charlotte, A.O.A.; Gbènato, A.-D.E.; Clément, A. Optimizing Genomic DNA Isolation in Pineapple (*Ananas comosus* L.). *J. Plant. Breed. Genet.* **2016**, *4*, 11–18.
15. Chin, C.-S.; Peluso, P.; Sedlazeck, F.J.; Nattestad, M.; Concepcion, G.T.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A.; et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **2016**, *13*, 1050. [CrossRef] [PubMed]
16. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
17. Faust, G.G.; Hall, I.M. SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **2014**, *30*, 2503–2505. [CrossRef] [PubMed]
18. Chen, L.-Y.; VanBuren, R.; Paris, M.; Zhou, H.; Zhang, X.; Wai, C.M.; Yan, H.; Chen, S.; Alonge, M.; Ramakrishnan, S.; et al. The bracteatus pineapple genome and domestication of clonally propagated crops. *Nat. Genet.* **2019**, *51*, 1549–1558. [CrossRef]
19. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **2014**, *9*, e112963. [CrossRef] [PubMed]
20. 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]
21. Durand, N.C.; Robinson, J.T.; Shamim, M.S.; Machol, I.; Mesirov, J.P.; Lander, E.S.; Aiden, E.L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **2016**, *3*, 99–101. [CrossRef] [PubMed]

22. Marçais, G.; Delcher, A.L.; Phillippy, A.M.; Coston, R.; Salzberg, S.L.; Zimin, A. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **2018**, *14*, e1005944. [CrossRef] [PubMed]
23. Xu, G.-C.; Xu, T.-J.; Zhu, R.; Zhang, Y.; Li, S.-Q.; Wang, H.-W.; Li, J.-T. LR_Gapcloser: A tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **2019**, *8*, giy157. [CrossRef] [PubMed]
24. Ou, S.; Chen, J.; Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **2018**, *46*, e126. [CrossRef] [PubMed]
25. Waterhouse, R.M.; Seppey, M.; Simão, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **2017**, *35*, 543–548. [CrossRef]
26. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
27. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **2008**, *24*, 637–644. [CrossRef] [PubMed]
28. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **2004**, *5*, 59. [CrossRef]
29. Keilwagen, J.; Hartung, F.; Paulini, M.; Twardziok, S.O.; Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform.* **2018**, *19*, 189. [CrossRef]
30. Keilwagen, J.; Wenk, M.; Erickson, J.L.; Schattat, M.H.; Grau, J.; Hartung, F. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **2016**, *44*, e89. [CrossRef]
31. 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.* **2017**, *89*, 789–804. [CrossRef] [PubMed]
32. Ming, R.; Hou, S.; Feng, Y.; Yu, Q.; Dionne-Laporte, A.; Saw, J.H.; Senin, P.; Wang, W.; Ly, B.V.; Lewis, K.L.T.; et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **2008**, *452*, 991–996. [CrossRef] [PubMed]
33. Droc, G.; Larivière, D.; Guignon, V.; Yahiaoui, N.; This, D.; Garsmeur, O.; Dereeper, A.; Hamelin, C.; Argout, X.; Dufayard, J.-F.; et al. The banana genome hub. *Database* **2013**, *2013*, bat035. [CrossRef]
34. Jain, R.; Jenkins, J.; Shu, S.; Chern, M.; Martin, J.A.; Copetti, D.; Duong, P.Q.; Pham, N.T.; Kudrna, D.A.; Talag, J.; et al. Genome sequence of the model rice variety KitaakeX. *BMC Genom.* **2019**, *20*, 905. [CrossRef] [PubMed]
35. Ni, X.; Yang, J.; Sun, S.; Yang, W. Identification and Analysis of Resistance-like Genes in the Tomato Genome. *J. Phytopathol.* **2014**, *162*, 137–146. [CrossRef]
36. McCormick, R.F.; Truong, S.K.; Sreedasyam, A.; Jenkins, J.; Shu, S.; Sims, D.; Kennedy, M.; Amirebrahimi, M.; Weers, B.D.; McKinley, B.; et al. The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant. J.* **2018**, *93*, 338–354. [CrossRef] [PubMed]
37. Jaillon, O.; Aury, J.-M.; Noel, B.; Policriti, A.; Clepet, C.; Casagrande, A.; Choisne, N.; Aubourg, S.; Vitulo, N.; Jubin, C.; et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **2007**, *449*, 463–467. [CrossRef] [PubMed]
38. Hirsch, C.N.; Hirsch, C.D.; Brohammer, A.B.; Bowman, M.J.; Soifer, I.; Barad, O.; Shem-Tov, D.; Baruch, K.; Lu, F.; Hernandez, A.G.; et al. Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *Plant. Cell* **2016**, *28*, 2700–2714. [CrossRef] [PubMed]
39. Wu, T.D.; Watanabe, C.K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **2005**, *21*, 1859–1875. [CrossRef] [PubMed]
40. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]
41. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314. [CrossRef]
42. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef]
43. Dai, X.; Sinharoy, S.; Udvardi, M.; Zhao, P.X. PlantTFcat: An online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinform.* **2013**, *14*, 321. [CrossRef] [PubMed]
44. Osuna-Cruz, C.M.; Paytuvi-Gallart, A.; Di Donato, A.; Sundesha, V.; Andolfo, G.; Aiese Cigliano, R.; Sanseverino, W.; Ercolano, M.R. PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* **2018**, *46*, D1197–D1201. [CrossRef]
45. Ou, S.; Su, W.; Liao, Y.; Chougule, K.; Agda, J.R.A.; Hellinga, A.J.; Lugo, C.S.B.; Elliott, T.A.; Ware, D.; Peterson, T.; et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **2019**, *20*, 275. [CrossRef]
46. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **2008**, *9*, 18. [CrossRef] [PubMed]
47. Ou, S.; Jiang, N. LTR_FINDER_parallel: Parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **2019**, *10*, 48. [CrossRef]
48. Ou, S.; Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant. Physiol.* **2018**, *176*, 1410–1422. [CrossRef]

49. Shi, J.; Liang, C. Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection. *Plant. Physiol.* **2019**, *180*, 1803–1815. [CrossRef]
50. Su, W.; Gu, X.; Peterson, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol. Plant.* **2019**, *12*, 447–460. [CrossRef]
51. Xiong, W.; He, L.; Lai, J.; Dooner, H.K.; Du, C. HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10263–10268. [CrossRef]
52. Zhou, Q.; Tang, D.; Huang, W.; Yang, Z.; Zhang, Y.; Hamilton, J.P.; Visser, R.G.F.; Bachem, C.W.B.; Robin Buell, C.; Zhang, Z.; et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **2020**, *52*, 1018–1023. [CrossRef]
53. Wang, Y.; Tang, H.; Debarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*, e49. [CrossRef]
54. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef]
55. Fischer, S.; Brunk, B.P.; Chen, F.; Gao, X.; Harb, O.S.; Iodice, J.B.; Shanmugam, D.; Roos, D.S.; Stoeckert, C., Jr. J. Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. *Curr. Protoc. Bioinform.* **2011**, *35*, 6–12. [CrossRef]
56. Edgar, R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **2004**, *5*, 113. [CrossRef] [PubMed]
57. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
58. 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]
59. Wu, H.; Ma, T.; Kang, M.; Ai, F.; Zhang, J.; Dong, G.; Liu, J. A high-quality Actinidia chinensis (kiwifruit) genome. *Hortic. Res.* **2019**, *6*, 117. [CrossRef]
60. Wang, L.; Li, Y.; Jin, X.; Liu, L.; Dai, X.; Liu, Y.; Zhao, L.; Zheng, P.; Wang, X.; Liu, Y.; et al. Floral transcriptomes reveal gene networks in pineapple floral growth and fruit development. *Commun. Biol.* **2020**, *3*, 500. [CrossRef]
61. Liu, S.; Yang, L.; Li, J.; Tang, W.; Li, J.; Lin, R. FHY3 interacts with phytochrome B and regulates seed dormancy and germination. *Plant. Physiol.* **2021**, *187*, 289–302. [CrossRef]
62. Ritter, A.; Iñigo, S.; Fernández-Calvo, P.; Heyndrickx, K.S.; Dhondt, S.; Shi, H.; De Milde, L.; Vanden Bossche, R.; De Clercq, R.; Eeckhout, D.; et al. The transcriptional repressor complex FRS7-FRS12 regulates flowering time and growth in Arabidopsis. *Nat. Commun.* **2017**, *8*, 15235. [CrossRef]
63. Fernández-Calvo, P.; Iñigo, S.; Glauser, G.; Vanden Bossche, R.; Tang, M.; Li, B.; De Clercq, R.; Nagels Durand, A.; Eeckhout, D.; Gevaert, K.; et al. FRS7 and FRS12 recruit NINJA to regulate expression of glucosinolate biosynthesis genes. *New Phytol.* **2020**, *227*, 1124–1137. [CrossRef] [PubMed]
64. Lin, R.; Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of Arabidopsis development. *Plant. Physiol.* **2004**, *136*, 4010–4022. [CrossRef]
65. Qiao, X.; Li, Q.; Yin, H.; Qi, K.; Li, L.; Wang, R.; Zhang, S.; Paterson, A.H. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **2019**, *20*, 38. [CrossRef]
66. Zhong, M.; Jiang, X.-D.; Weihua, C.; Hu, J.-Y. Expansion and expression diversity of FAR1/FRS-like genes provides insights into flowering time regulation in roses. *Plant. Divers.* **2020**, *43*, 173–179. [CrossRef]
67. Sharma, A.; Wai, C.M.; Ming, R.; Yu, Q. Diurnal Cycling Transcription Factors of Pineapple Revealed by Genome-Wide Annotation and Global Transcriptomic Analysis. *Genome Biol. Evol.* **2017**, *9*, 2170–2190. [CrossRef]