data

MDPI

Data Descriptor

Gene Expression Datasets for Two Versions of the *Saccharum spontaneum* AP85-441 Genome

Nicolás López-Rozo [1,2,†], Mauricio Ramirez-Castrillon [2,†], Miguel Romero [1,2], Jorge Finke [1,2] and Camilo Rocha [1,2,*]

[1] Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali 760031, Colombia
[2] OMICAS Program, Pontificia Universidad Javeriana, Cali 760031, Colombia
* Correspondence: camilo.rocha@javerianacali.edu.co
† These authors contributed equally to this work.

Abstract: Sugarcane is a species of tall grass with high biomass and sucrose production, and the world's largest crop by production quantity. Its evolutionary environment adaptation and anthropogenic breeding response have resulted in a complex autopolyploid genome. Few efforts have been reported in the literature to document this organism's gene co-expression and annotation, and, when available, use different gene identifiers that cannot be easily associated across studies. This data descriptor paper presents a dataset that consolidates expression matrices of two *Saccharum spontaneum* AP85-441 genome versions and an algorithm implemented in Python to mechanically obtain this dataset. The data are processed from the allele-level information of the two sources, with BLASTn used bidirectionally to suggest feasible mappings between the two sets of alleles, and a graph-matching optimization algorithm to maximize global identity and uniqueness of genes. Association tables are used to consolidate the expression values from alleles to genes. The contributed expression matrices comprise 96 experiments and 109,050 and 35,516 from the two genome versions. They can represent significant computational cost reduction for further research on, e.g., sugarcane co-expression network generation, functional annotation prediction, and stress-specific gene identification.

Dataset: https://github.com/mauriciogeteg/sugarcane-gene-expression.

Keywords: sugarcane; expression matrix; allele expression; graph flow

Citation: López-Rozo, N.; Ramirez-Castrillon, M.; Romero, M.; Finke, J.; Rocha, C. Gene Expression Datasets for Two Versions of the *Saccharum spontaneum* AP85-441 Genome. *Data* **2023**, *8*, 1. https://doi.org/10.3390/data8010001

Academic Editor: Pu-Feng Du

Received: 5 September 2022
Revised: 17 November 2022
Accepted: 30 November 2022
Published: 20 December 2022

1. Summary

Modern sugarcane is the world's largest crop by production quantity. It is a high biomass producer hybrid with verified photosynthetic efficiency [1], obtained mainly from *Saccharum spontaneum* and *Saccharum oficinarum* [2]. Nonetheless, few efforts have been reported in the literature to document this organism's gene sequence, expression, and annotation, thus making the availability of public genomic data scarce.

The work of Zhang et al. [3] is among the few and recent efforts to bridge the gap caused by the missing sugarcane genomic information. They report on two genome sequences of the *S. spontaneum* cultivar AP85-441, identified here as v2018 [4] and v2019 [5]. Each sequencing effort uses different sets of alleles and obtains distinct total number of alleles. The v2018 transcriptome data was further used, e.g., to characterize allelic expression dominance [3], specific genes related to abiotic stress and response to hormones [6], synthesis of starch [7], and plant-specific transcription factors involved in growth and development [8–12].

Because of the identification mismatch between v2018 and v2019, such experiments cannot be directly interpreted in the context of the complete dataset. For example, it would

be useful to understand the mechanisms of heterosis in modern hybrid crops (usually cultured to produce sugar and ethanol), such as cultivars CC-01-1940 in Colombia [13] or SP80-3280 in Brazil [14], with the help of these datasets, given that expression of most genes exhibit a direct relationship with the frequency of alleles in the genome [15]. Conveniently, the computational cost and time effort required to integrate both datasets for such explorations, if the information on events of polyploidization and chromosome reduction in monoploid/diploid cultivars of the species were to be used, would need to be paid once and amortized in subsequent experiments.

This paper presents a dataset and an algorithm to construct it from the genome sequences in v2018 and v2019. The contributed dataset comprises several consolidated matrices of allele and gene expression for the cultivar APS85-441. The contributed algorithm is implemented in the Python 3 programming language; it standardizes the gene and allele identifiers with unique nomenclature, and consolidates the expression values from the allele to the gene level.

The overall process includes global alignment with BLASTn [16], applied bidirectionally to find a suitable matching between alleles of v2018 and v2019. The mapping information is fed into a graph-matching optimization algorithm designed to maximize the number of alleles in v2018 assigned to the alleles in v2019. The optimization goal is to maximize the global identity between pairs of selected alleles on the two genome versions. The algorithm ensures that an allele in v2018 is assigned to at most one allele in v2019, thus avoiding duplication of the expression values in the final dataset.

Allele expression values from each genome version are consolidated separately to the gene level using correspondence tables available from the original sources [4,5]. The resulting expression matrices comprise 96 experiments, and 109,050 and 35,516 genes for the v2018 and v2019 genomes, respectively. The dataset and the implementation presented in this paper can be seen as a new effort to bridge the gap of missing genomic information for sugarcane, thus adding new potential for further genomic experimentation and exploration on this organism.

2. Data Description

Figure 1 presents the workflow, input files, processing steps, and output (i.e., databases, intermediate files, and matrices) adopted in the effort of combining the datasets v2018 [4] and v2019 [5] into a single consolidated dataset. A detailed explanation of each step and file is given in Section 3. The input and output files, and the code are available from the repository at https://github.com/mauriciogeteg/sugarcane-gene-expression (accessed on 3 December 2022).

Each one of the datasets v2018 and v2019 contains FASTA sequences of coding DNA sequences (CDS) and proteins, and a table with the allele-gene association information. The dataset v2018 also includes several expression matrices. However, the current reference genome for this cultivar is v2019. The expression matrices were concatenated and adjusted for alleles and genes, as described in Table 1.

In total, all matrices comprise 96 experiments (columns) with expression data obtained from the original files. In the allele expression matrix of v2018, there are 112,788 transcripts reported (rows), corresponding to all expression profiles of four alleles per gene. The allele expression matrix of v2019 consists of 83,821 transcripts. In particular, expression values for 28,967 alleles are consolidated by the algorithm. Regarding the expression matrices at gene level, the dataset v2019 was consolidated into 35,516 genes, with only a discrepancy of 6 genes when compared to the previously reported data in [3]. Overall, v2018 consists of 109,050 genes, suggesting that few alleles are consolidated.

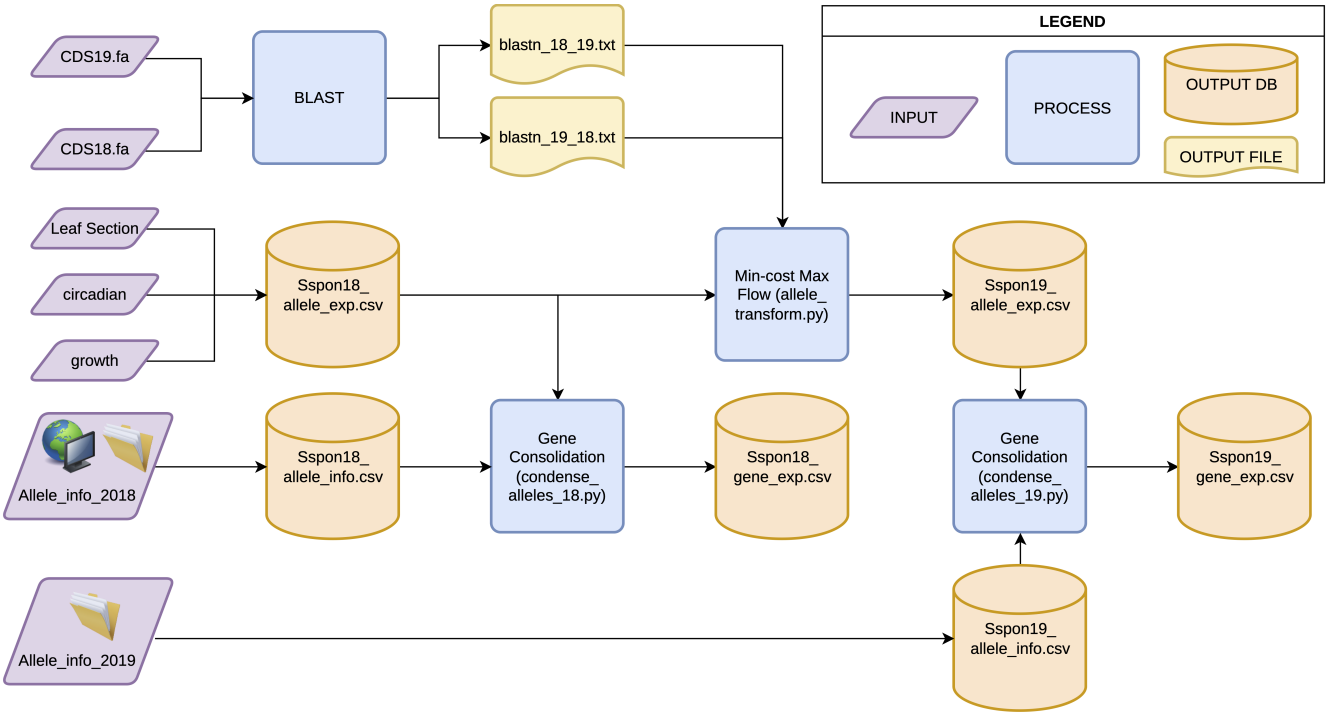

Figure 1. Representation of the workflow, input files, processing steps, and output (i.e., databases, intermediate files, and matrices) adopted in the effort of combining the datasets v2018 and v2019 into a single consolidated dataset.

Table 1. Composition of expression matrices for the two *Saccharum spontaneum* AP85-441 genomes in v2018 and v2019. The matrices comprise 96 columns of experimental data.

Matrix Name	Number of Alleles/Genes	Reference
Alleles v2018	112,788	[3]
Genes v2018	109,050	This data descriptor
Alleles v2019	83,821	This data descriptor
Genes v2019	35,516	This data descriptor

Figure 2 presents a histogram of absolute values (FPKM data) of allele (left) and gene expression (right). Note that the shape of the histogram in both datasets is similar, suggesting that no significant changes in the final consolidated matrices are added. The consolidation from allele expression to gene expression in the v2019 genome reduces considerably the frequency of genes (from 83,821 to 35,516), shrinking the histogram vertically, but expanding it horizontally because some expression values increase in magnitude.

Table 2 presents details of the expression matrices. The maximum expression value remains constant in the v2018 dataset. In contrast, consolidating expression data for the v2019 dataset increases its maximum expression value. Data aggregation is, in general, higher in the v2019 genome due to its annotation system [3]. The number of zeros is approximately 45% for the v2019 allele expression matrix. In comparison, the other matrices have fewer zeros (approximately 33%).

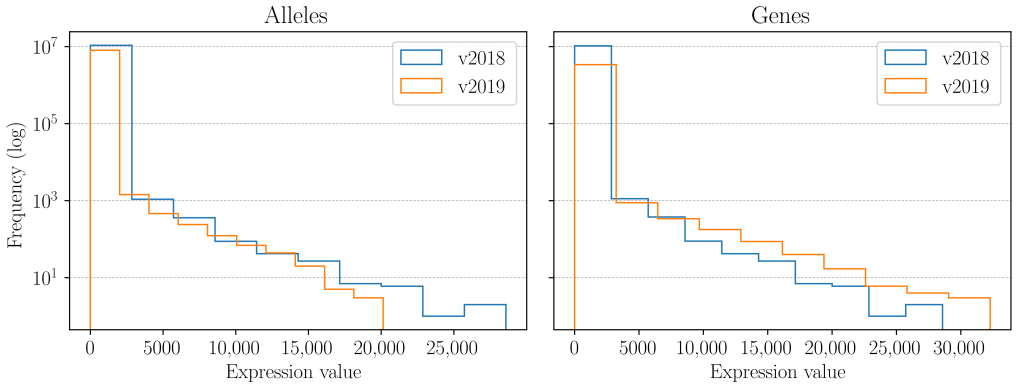

Figure 2. Histogram contour of the expression data (FPKM) of v2018 and v2019: (**left**) allele expression and (**right**) gene expression.

Figure 3 presents the distribution of the expression values for both alleles and genes. Note that allele information for v2019 is highly skewed to small values due to an abundance of zeros, with median 0.13. However, if the values less than or equal to 0.001 are removed, the median 2.12 of the allele expression values approaches the median 3.96 of the gene information.

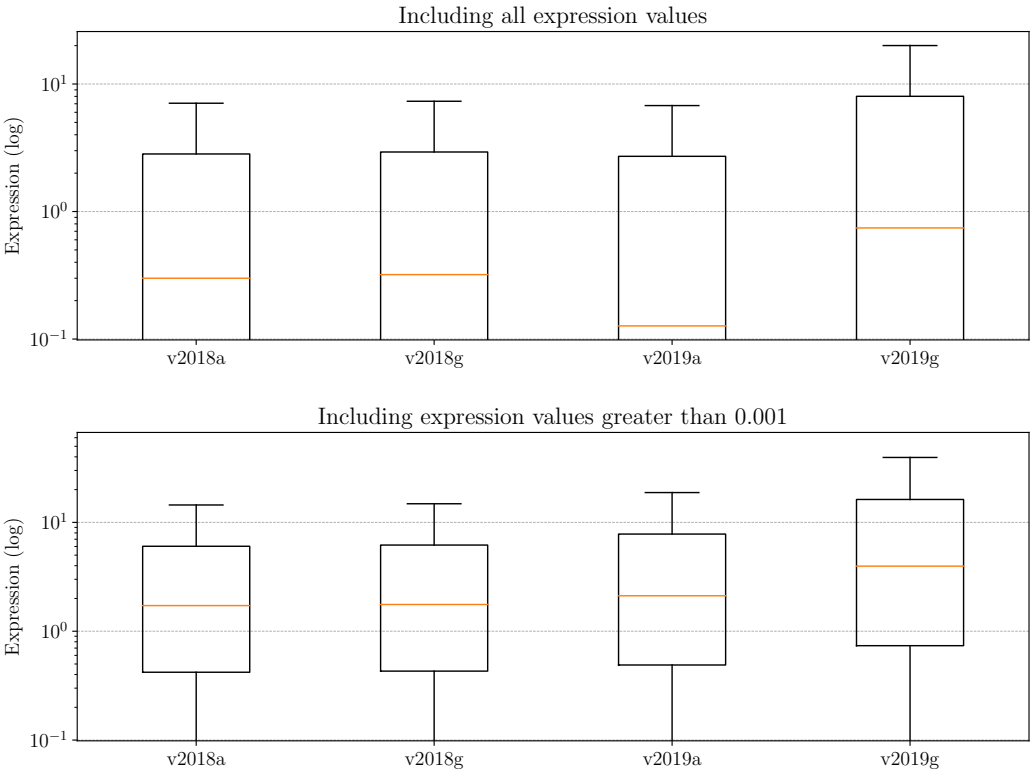

Figure 3. Boxplot of the distribution of expression values for alleles (v2018a, v2019a) and genes (v2018g, v2019g) data. Vertical magnitudes are displayed in logarithmic axis: (**top**) including all expression values and (**bottom**) including only expression values greater than 0.001.

Table 2. Description of expression values for the two *Saccharum spontaneum* AP85-441 genomes.

Matrix Name	Max Value	All Values		Values > 0.001	
		Count	Median	Count	Median
Alleles v2018	28,579	10,827,650	0.300	6,809,283	1.720
Genes v2018	28,579	10,468,800	0.320	6,621,587	1.761
Alleles v2019	20,135	8,046,816	0.127	4,460,747	2.117
Genes v2019	32,282	3,409,536	0.744	2,277,266	3.963

Tools to detect homologous genes are reported in the literature. Zhu et al. [17], for instance, used a similar approach to the one presented here to detect alleles in wheat (*Triticum aestivum*) from two different genome versions, including the bidirectional use of BLASTp. In contrast, in the proposed approach, BLASTn is used from predicted CDS. However, the parameters of BLASTn used for the dataset contributed in this paper are stricter than the mentioned work on wheat. They also used the MCScanX package [18] to identify syntenic blocks among subgenomes, and dynamic programming and a scoring schema to find the highest-scoring paths between similar gene pairs. In contrast, the approach presented here relies on graph theory and uses a min-cost max-flow algorithm to select the maximal number of connections possible.

Kallisto [19] is another package used to quantify RNA-seq reads belonging to the same gene. In this sense, the alignment against a reference genome is not strictly specific (i.e., it is a pseudo-alignment), avoiding the formation of variants of the same gene (alleles). In Kallisto, however, the generation of reads pseudoalignments belonging to a probable "core" transcript requires raw FastQC data and an index (estimated from a reference genome). In the approach presented here, two FASTA files from CDS for each genome version are used. In this sense, the target of the contributed Python code is different to other tools published previously. Also, the approach presented in this paper consolidates genes from multiple annotated genomes, combining expression data without the need for an alignment with a reference genome.

3. Methods

3.1. Data Sources

Recall that the datasets correspond to two genome sequences of the *S. spontaneum* cultivar AP85-441, identified here as v2018 [4] and v2019 [5].

The website [4] contains three genomes: AP85-441 (v2018), Np-X, and L-Purple. The FASTA file in v2018 contains the CDS and three files with gene expression level data. The information about alleles from each gene is retrieved directly from this data source. The resulting file can be found in the supplementary material as `Sspon18_allele_info.csv`.

The website at [5] contains a second release of the AP85-441 genome (v2019). No available data about expression could be found for this dataset. The FASTA file is used, including the CDS and a CSV file registering the relationship between representative genes and their alleles (`Saccharum_spont_alleleTable-Jan_2019.csv`).

3.2. Pre-Processing

The pre-processing of the expression data in v2018 has a single file as goal: it consolidates the expression experiments leaf section (`sspon.leafsection.zip`), circadian rhythm (`day_night.zip`), and growth and development (`growth_period.zip`). The resulting file, `Sspon18_allele_exp.csv`, is part of the supplementary material.

The allele information for the genome in v2019 is organized and filtered to only store relevant data, i.e., the representative gene model identifier, the Sorghum homolog identifier (if available), and four columns containing the information of the alleles. Each allele can contain more than one copy in the genome. These duplicated alleles are separated in the same column by a vertical bar ("|"). The resulting file can be found in the supplementary material as `Sspon19_allele_info.csv`.

3.3. Gene Expression Consolidation

3.3.1. Blast

Two databases are created from the FASTA files containing the CDS of each genome version (files `CDS18.fa` and `CDS19.fa`). Then, two BLAST nucleotide alignments (also known as BLASTn) are executed using each dataset against the other. The resulting files can be found in the folder `blast_results`. The following parameters are used for the command-line program: e-value at most 0.000001, minimal percent identity of 90, minimal percent coverage of 90, and only the best 10 alignments by query are permitted (using the best hit in each alignment).

3.3.2. Optimized Matching

Based on the output of BLAST, the associations among the alleles in v2018 and v2019 are found to have repetitions. In the case of the mappings between v2018 to v2019, a CDS in the source could be associated with several CDS in the target. To generate a reasonable coverage, both mappings are combined by modeling the problem as a graph flow optimization problem [20] with multiple sources (v2018 alleles) and multiple targets (v2019 alleles).

A min-cost max-flow problem requires to compute a graph-matching (i.e., match a the level of nodes/vertices) with maximal cardinality (i.e., maximal number of connections), thus ensuring a maximal covering of the source-target associations. If more than one maximal matching is possible, then the cost of producing that maximal flow is to be minimized. In this case, identity scores can be considered to identify the matching with the greatest sum of identity scores, while still ensuring that a v2018 allele expression is used at most once. Since the algorithm implemented in networkx minimizes cost, the artificial cost fed to the min-cost max-flow algorithm is pident (i.e., percent identity) on each possible association between the two versions of the alleles.

An overview of the algorithm is presented next:

- Each CDS in v2018 is represented as a node u in the group S of sources.
- Each CDS in v2019 is represented as a node v in the group T of targets.
- If node $u \in S$ and node $v \in T$ appear as a match in either of the mappings, then they are connected by an edge (u, v) with capacity 1 and cost $-pident$, corresponding to the highest BLAST identity value between them.
- An additional source node u_S is added and edges (u_S, s_i), for each $s_i \in S$, are created with capacity 1 and cost 0, ensuring that each CDS in v2018 can be used at most once.
- An additional target node v_T is added and edges (t_i, v_T), for each $t_i \in T$, are created with infinite capacity and cost 0, ensuring that each CDS in v2019 can be used several times.
- Finally, a min-cost max-flow algorithm is executed taking u_S as the source node and v_T as the sink node. This will have the effect of most nodes in S being used and each node in T having at least one possible incoming connection.

The final matching of pairs between CDS in v2018 and v2019 is used to transfer the expression levels from the v2018 genome to the v2019 genome. Expression value consolidation for each CDS in v2019 is carried out by experiment-wise addition. The resulting file can be found in the supplementary material as `Sspon19_allele_exp.csv`; it consists of 83821 rows and 96 experimental expression values per row.

3.3.3. Gene Expression Consolidation

The allele information for the genome in v2018 has no representative gene identifier. In this case, the first gene in lexicographical order is selected as the representative. The list of representative genes is used for the allele to gene expression consolidation by adding the expression values of each experiment. The resulting file can be found in the supplementary material as `Sspon18_gene_exp.csv` and the Python code for generating this file is `condense_alleles_18.py`.

The expression information in v2019 is processed using the prefix of each gene. The prefix structure is explained in the file `README-SsponAnnotation.pdf` from [5]. If a CDS has the form `Sspon.01G0000010-1A`, then its representative gene is `Sspon.01G0000010`. The expression values of different alleles of the same gene is consolidated by adding the expression values in each experiment. A total of 35,516 genes were identified. The resulting file is `Sspon19_gene_exp.csv` and the code generating this file is in `condense_alleles_19.py`.

3.4. Metadata

Table 3 presents the associated metadata with the dataset published in the repository under commit `60f9344`. A total of 16 files are reported: seven input files, four of them being intermediate files related to BLAST results and pre-processing steps, and the remaining files are the final expression matrices.

Table 3. Metadata associated with the repository.

Specifications	Description
Subject area	Biological science, computer science
More specific subject area	Bioinformatics, Genomics, Sugarcane, Expression analysis
Type of data	Data spreadsheets, plain text, Python code
How data was acquired	Compiled from open access databases and websites
Data source location	Global
Data accesibility	The data presented in this article is freely and publicly available for any academic, educational, and research purpose. The public repository is located at https://github.com/mauriciogeteg/sugarcane-gene-expression (accessed on 3 December 2022)
Folders included in the dataset	4 (`blast_results`; `Codes`; `figures`; `inputs`)
Files included as inputs	7 (`Allele_info_2018.csv`; `Allele_info_2019.csv`; `circadian.7z`; `growth.xlsx`; `Leaf_Section.7z`; `Sspon.v20180123.cds.fasta.7z`; `Sspon.v20190103.cds.fasta.7z`)
Files included as blast_results	2 (`blastn_2018_2019.txt`; `blastn_2019_2018.txt`)
Files included as figures	3 (`Figure1.pdf`; `Figure2.pdf`; `Figure3.pdf`)
Codes included in the dataset	3 (`allele_transform.py`; `condense_alleles_18.py`; `condense_alleles_19.py`)
Expression matrices	4 (`Sspon18_allele_.7z`; `Sspon18_gene_.7z`; 2 compressed matrices in: `Sspon19_.7z`)

Author Contributions: Conceptualization, N.L.-R. and M.R.-C.; methodology, N.L.-R., M.R. and M.R.-C.; software, N.L.-R. and M.R.; investigation, N.L.-R., M.R. and M.R.-C.; data curation, N.L.-R. and M.R.; writing—original draft preparation, N.L.-R. and M.R.-C.; writing—review and editing, C.R. and J.F.; project administration, C.R. and J.F.; funding acquisition, C.R. and J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the "OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y validación en Arroz y Caña de Azúcar)" Scientific Ecosystem belonging to the Colombia Científica Program, sponsored by The World Bank, The Ministry of Science, Technology and Innovation (MINCIENCIAS), ICETEX, the Colombian Ministry of Education and the Colombian Ministry of Commerce, Industry and Turism, under GRANT ID: FP44842-217-2018, OMICAS Award ID: 792-61187.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets, intermediate files, and codes used for this data descriptor are available at https://github.com/mauriciogeteg/sugarcane-gene-expression (accessed on 3 December 2022) with the commit "3fbcf98". The original files were downloaded from the next websites: http://sugarcane.zhangjisenlab.cn/sgd/html/download.html (accessed on 16 November 2022) and https://www.life.illinois.edu/ming/downloads/Spontaneum_genome/ (accessed on 16 November 2022).

Acknowledgments: The authors would like to thanks Camila Riccio and Chrystian Sosa for help in performing statistical validations (data not shown), and valuable and constructive suggestions after reading the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

CDS	Coding DNA Sequence
ASE	Allele-Specific Expression
CSV	Comma-Separated Value
FPKM	Fragments Per Kilobase of exon per million Mapped
OMICAS	Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles

References

1. Henry, R.J.; Kole, C. Basic information on the sugarcane plant. In *Genetics, Genomics and Breeding of Sugarcane*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2010; Volume 9, pp. 1–8. [CrossRef]
2. Kim, C.; Wang, X.; Lee, T.H.; Jakob, K.; Lee, G.J.; Paterson, A.H. Comparative analysis of Miscanthus and Saccharum reveals a shared whole-genome duplication but different evolutionary fates. *Plant Cell* **2014**, *26*, 2420–2429. [CrossRef] [PubMed]
3. Zhang, J.; Zhang, X.; Tang, H.; Zhang, Q.; Hua, X.; Ma, X.; Zhu, F.; Jones, T.; Zhu, X.; Bowers, J.; et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* **2018**, *50*, 1565–1573. [CrossRef] [PubMed]
4. Saccharum Genome Database. 2018. Available online: http://sugarcane.zhangjisenlab.cn/sgd/html/download.html (accessed on 7 August 2022).
5. The Ming Laboratory, Saccharum Spontaneum AP85-441 Genome. 2019. Available online: https://www.life.illinois.edu/ming/downloads/Spontaneum_genome/ (accessed on 7 August 2022).
6. Cai, M.; Lin, J.; Li, Z.; Lin, Z.; Ma, Y.; Wang, Y.; Ming, R. Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. *PLoS ONE* **2020**, *15*, 1–24. [CrossRef] [PubMed]
7. Ma, P.; Yuan, Y.; Shen, Q.; Jiang, Q.; Hua, X.; Zhang, Q.; Zhang, M.; Ming, R.; Zhang, J. Evolution and Expression Analysis of Starch Synthase Gene Families in Saccharum spontaneum. *Trop. Plant Biol.* **2019**, *12*, 158–173. [CrossRef]
8. Lin, J.; Zhu, M.; Cai, M.; Zhang, W.; Fatima, M.; Jia, H.; Li, F.; Ming, R. Identification and Expression Analysis of TCP Genes in Saccharum spontaneum L. *Trop. Plant Biol.* **2019**, *12*, 206–218. [CrossRef]
9. Li, Z.; Hua, X.; Zhong, W.; Yuan, Y.; Wang, Y.; Wang, Z.; Ming, R.; Zhang, J. Genome-Wide Identification and Expression Profile Analysis of WRKY Family Genes in the Autopolyploid Saccharum spontaneum. *Plant Cell Physiol.* **2019**, *61*, 616–630. [CrossRef] [PubMed]
10. Li, P.; Chai, Z.; Lin, P.; Huang, C.; Huang, G.; Xu, L.; Deng, Z.; Zhang, M.; Zhang, Y.; Zhao, X. Genome-wide identification and expression analysis of AP2/ERF transcription factors in sugarcane (*Saccharum spontaneum* L.). *BMC Genom.* **2020**, *21*, 685. [CrossRef] [PubMed]
11. Feng, X.; Wang, Y.; Zhang, N.; Zhang, X.; Wu, J.; Huang, Y.; Ruan, M.; Zhang, J.; Qi, Y. Systematic Identification, Evolution and Expression Analysis of the SPL Gene Family in Sugarcane (*Saccharum spontaneum*). *Trop. Plant Biol.* **2021**, *14*, 313–328. [CrossRef]
12. Ali, A.; Javed, T.; Zaheer, U.; Zhou, J.R.; Huang, M.T.; Fu, H.Y.; Gao, S.J. Genome-Wide Identification and Expression Profiling of the bHLH Transcription Factor Gene Family in Saccharum spontaneum Under Bacterial Pathogen Stimuli. *Trop. Plant Biol.* **2021**, *14*, 283–294. [CrossRef]
13. Trujillo-Montenegro, J.H.; Cubillos, M.J.R.; Loaiza, C.D.; Quintero, M.; Espitia-Navarro, H.F.; Villareal, F.A.S.; Valens, C.A.V.; Barrios, A.F.G.; Vega, J.D.; Duitama, J.; et al. Unraveling the genome of a high yielding colombian sugarcane hybrid. *Front. Plant Sci.* **2021**, *12*, 694859. [CrossRef] [PubMed]
14. Souza, G.M.; Sluys, M.A.V.; Lembke, C.G.; Lee, H.; Margarido, G.R.A.; Hotta, C.T.; Gaiarsa, J.W.; Diniz, A.L.; Oliveira, M.D.M.; Ferreira, S.D.S.; et al. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience* **2019**, *8*, giz129. [CrossRef] [PubMed]
15. Margarido, G.R.A.; Correr, F.H.; Furtado, A.; Botha, F.C.; Henry, R.J. Limited allele-specific gene expression in highly polyploid sugarcane. *Genome Res.* **2022**, *32*, 297–308. [CrossRef] [PubMed]

16. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef] [PubMed]
17. Zhu, T.; Wang, L.; Rimbert, H.; Rodriguez, J.C.; Deal, K.R.; Oliveira, R.D.; Choulet, F.; Keeble-Gagnère, G.; Tibbits, J.; Rogers, J.; et al. Optical maps refine the bread wheat Triticum aestivum cv. Chinese Spring genome assembly. *Plant J.* **2021**, *107*, 303–314. [CrossRef] [PubMed]
18. Wang, Y.; Tang, H.; Debarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.H.; 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] [PubMed]
19. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef] [PubMed]
20. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, USA, 2022.