Hybrid Assembly-Based Genomic Analysis for the Identification and Characterisation of Flowering Genes in *Musa acuminata* subsp. *sumatrana* (Becc.) Ahmad, Volkaert, Sulist. & Poerba

Dewi Rahmawati[1], *I Made Artika*[1*], *Rahadian Pratama*[1] and *Fajarudin Ahmad*[2]

[1]Biochemistry Department, Faculty of Mathematics and Natural Science, 16680, Indonesia
[2] Research Center for Applied Botany, Organization for Research on Life Sciences and the Environment, BRIN, Indonesia

Abstract. *Musa acuminata* subsp. *sumatrana* exhibits rare productivity advantages (± 20 hands/bunch, 35% > Cavendish). This research aims to characterize the genome of *Musa acuminata* subsp. *sumatrana* through Whole Genome Sequencing (WGS) by comparing long-read PromethION data and short-read Illumina data for in silico identification of flowering genes. The hybrid assembly method integrated PromethION ONT and Illumina (N50 4,097 bp; 13.9 Gb), followed by gene annotation using BRAKER3 and FunAnnotate. The results identified 51,358 genes, including 12 significant genes: MADS-box (*AGL11, SOC1, AGL12*), Myb (*RS1, RS2, NUCT14, PLAT1, PUB13*), AP2/ERF (*RAV2, LOGL3*), and QUIRKY (FTIP7, *CNOT9*). Two new genes (*RAV2* and *LOGL3*) from the AP2/ERF domain were identified to be associated with flowering in bananas for the first time, but their functional roles still need to be validated. High expression of *AGL11* correlates with increased seed formation, while the adaptive divergence of *SOC1* (NJ:0.59) reflects tropical photoperiod selection. GO enrichment confirms the dominance of the "*flower development*" process (FDR < $1.0e^{-7}$), with the regulatory mechanism: environmental stimulus → CO/QUIRKY → phase transition → flower differentiation and fruit initiation by MADS-box and AP2/ERF-Myb. This study concludes that the subspecies has the potential to be a source of superior alleles for high-yield banana breeding.

1 Introduction

Bananas (*Musa* spp.) are the most important tropical fruit commodity globally, with an international trade value of USD 10.2 billion in 2022 [1] and a staple food source for more than 400 million people in developing countries. However, banana production is seriously threatened by a disease called Fusarium Tropical Race 4 (Foc TR4) wilt, which is caused by the fungus *Fusarium oxysporum* f. sp. *cubense*. This soil-borne pathogen has caused annual economic losses of USD 200 million in Southeast Asia, resulting in a 30–50% reduction in production on monoculture plantations [2]. TR4 persists in soil for over 30 years and spreads

* Corresponding author: imart@apps.ipb.ac.id

efficiently through irrigation water, wind, and contaminated agricultural equipment, making it the most destructive biological threat in the history of banana cultivation [3].

As a strategic solution, wild banana germplasm is crucial for developing new, superior varieties. *Musa acuminata* subsp. *sumatrana* (Becc.) Ahmad, Volkaert, Sulist. & Poerba, endemic to Sumatra, Indonesia, stands out with rare phenotypic advantages: the ability to produce bunches with extremely high comb density (±20 combs/bunch), 35% more productive than Cavendish cultivars (<15 combs/bunch). As a member of the AA genome group (diploid, 2n=22), this subspecies has a relatively simple genetic structure for molecular analysis while retaining unique allelic diversity lost in triploid cultivated bananas. Although its fruit is seed-rich and not consumed directly, this trait facilitates conventional crossbreeding to transfer superior traits to seedless commercial cultivars.

Optimal utilisation of this high-density trait requires a deep understanding of flowering regulation and inflorescence development, processes controlled by a complex network of genes with the MADS-box family as the core regulator. The *SOC1*, *AGL11*, and *AGL12* genes are members of the MADS-box family, involved in regulating flowering and fruit development in plants. However, their homologs in *M. acuminata* subsp. *sumatrana* have not been characterised. The lack of genomic and transcriptomic data specific to this subspecies poses a critical challenge in (1) identifying molecular markers for assisted selection, (2) precision introgression of superior traits, and (3) developing gene editing-based genetic engineering strategies.

Given this urgency, this study aims to characterise the genome of *Musa acuminata* subsp. *sumatrana* through whole-genome sequencing (WGS) by comparing long-read PromethION and short-read Illumina data for the in silico identification of flowering-related genes.

2 Method

2.1 Sample Preparation and Sequencing

Leaves of *M. acuminata* subsp. *sumatrana* were obtained from the collection of the Cibinong Botanical Garden—BRIN (National Research and Innovation Agency). Sequencing library preparation was conducted by PT. Genetika Science Indonesia using two complementary platforms: (1) Oxford Nanopore PromethION for long read generation (basecalling via Guppy v6.3.9) and (2) Illumina NovaSeq 6000 for 150 bp paired-end short read generation.

2.2 Quality Control and Hybrid Assembly

Raw sequencing data from both platforms were submitted to Usegalaxy.eu for subsequent processing. The quality control phase comprised (a) assessment of long reads via Nanoplot (v1.42.0) using N50 parameters and (b) filtration of short reads employing Fastp (v0.23.2) with a quality threshold of Q30. Subsequently, hybrid assembly was executed in three consecutive phases: de novo assembly using Flye (v2.9.3; parameters: --nano-raw, minimum overlap = 1000 bp), error correction via Medaka (v1.7.2), and final refinement with Pilon (v1.20.1) employing short reads for indel and substitution rectification. A minimum overlap of 1000 bp was established, informed by the average read length of PromethION (±15 kb), to enhance repeat resolution.

2.3 Genome Annotation and Domain Characterisation

The assembled genome was annotated using two methodologies: structural annotation was conducted using BRAKER3 (v3.0.8) with *Viridiplantae* protein data as homology evidence, and functional annotation was performed using Funannotate (v1.8.15) based on the *Arabidopsis* model. Domain identification was conducted via InterProScan v5.65-97.0 within a Python 3.13.2 environment, incorporating the BioPython and pyHMMER packages. Essential parameters used: E-value ≤1e-10, Z-score ≥25, and a minimum residue overlap of 50 amino acids. Four target domains, MADS-box (**PF00319**), AP2/ERF3 (**PF00847**), QUIRKY (**IPR047259**), and Myb (**PF13837**), were selected due to their biological relevance in regulating flowering. Validation was conducted using the NCBI CDD web interface with default parameters: an expectation threshold of 0.01, compositional correction, and a low-complexity filter.

2.4 Phylogenetic Assessment and Protein Interactions

Homologous genes searches were conducted using NCBI BLASTn v2.14.0 with a dedicated *Musa* spp genomes database. Multiple alignments were performed using MUSCLE v5.1 before constructing a phylogenetic tree with MEGA11 (Neighbour-Joining method, 1000 bootstraps) to elucidate evolutionary relationships. Predictions of protein-protein interactions were conducted using STRING-db with *Musa*-specific parameters (TaxID: 4640) and a confidence level of at least 0.7, followed by Gene Ontology enrichment analysis. Diagrams of flowering regulation were recreated using BioRender, based on KEGG map04712 (*circadian rhythm - plant*) and PPI findings.

3 . Results and Discussion

3.1 Genome Assembly Quality and Characteristics

Long-read sequencing produced a total of 4,387,795 reads with an overall base length reaching 13,897,642,989 base pairs (bp). The N50 read length metric of 4,097.0 bp indicates that half of the total bases are contained in reads with a minimum length of that size. Additionally, 67.9% of the reads have sequencing quality above Q15, which represents a base accuracy of ≥97% (**Table 1**).

Table 1. Overview of long-read sequencing metrics

Parameter	Value
Aggregate count of reads	4,387,795
Total base length (bp)	13,897,642,989
N50 read length (bp)	4,097.0
Percentage > Q15	67.9%

The high N50 read length value (4,097.0 bp) and significant total base length (approaching 14 billion bp) reflect the capability of long-read technology in producing long and comprehensive DNA fragments. This technique is highly advantageous for complex genomics applications, such as de novo assembly or analysis of repetitive regions, because longer reads reduce fragmentation of assembly results and enhance contiguity. However, the number of reads obtained (4.39 million) is considered moderate by large-scale sequencing standards, so the efficiency of genome coverage may need to be evaluated based on the size of the target genome.

The percentage of high-quality reads (>Q15) at 67.9% indicates that one-third of the data has limited accuracy (below Q15), which could potentially affect the reliability of downstream analysis. Although a Q15 value is generally accepted for exploratory studies, the error rate proportion of 32.1% in the remaining reads can hinder the identification of genetic variants or increase the need for bioinformatics corrections. Optimization of sample preparation protocols or instrument calibration may be necessary to improve sequencing quality, particularly when high precision is required, as in clinical diagnostics.

Based on the long-read sequencing results that produced an N50 read length of 4,097 bp and a total base length of ~13.9 billion bp, genome annotation was carried out in two stages: structural annotation with BRAKER3 and functional annotation using FunAnnotate (Galaxy.eu). The high-quality long reads (high N50) and comprehensive base coverage facilitated more accurate gene prediction by BRAKER3, especially for large genes or repetitive regions. The results identified 51,358 genes, consisting of 4,885 annotated genes, 44,729 hypothetical genes, and 1,744 tRNA.

The high proportion of hypothetical genes (44,729 genes; 87.1%) is not solely due to sequencing errors (32.1% of reads ≤ Q15) but also due to the limitations of the functional database for wild banana genomes. Low read quality has the potential to produce frame shifts or premature stop codons, complicating domain prediction, while the lack of reference genomes from close relatives causes tools like FunAnnotate to fail in identifying homology. As a result, the unique specific genes of the *M.acuminata* subsp. *sumatrana* genome that may be important for flowering and development are classified as 'hypothetical' because there are no matches in the database.

3.2 Flowering Gene Identification and Domain Architecture

Domain analysis on the annotated gene fraction (9.5% of the total 51,358 genes) identified 12 genes that regulate flowering, grouped into four domains: MADS-box (***AGL11, SOC1, AGL12***), Myb (***RS1, RS2, NUCT14, PLAT1, PUB13***), AP2/ERF3 (***RAV2, LOGL3***), and QUIRKY ***(FTIP7, CNOT9)***, with the highest distribution in the Myb domain (5 genes) and MADS-box (3 genes) (**Table 2**).

Table 2. Candidate genes potentially influencing plant flowering.

Domain	Genes
AP2/ERF3	*RAV2, LOGL3*
MADS-box	*AGL11, SOC1, AGL12*
QUIRKY	*FTIP7, CNOT9*
Myb	*RS1, RS2, NUCT14, PLAT1, PUB13*

The four primary regulatory domains (MADS-box, Myb, AP2/ERF3, and QUIRKY) were identified by consensus in the literature as the fundamental components of the genetic network regulating flowering in monocotyledonous plants. The selection of these domains was predicated on three principal considerations: first, MADS-box serves as a highly conserved master regulator of floral development in monocots, including bananas [4]; QUIRKY embodies a distinctive mechanism for florigen transport vital in low-light environments; and third, Myb and AP2/ERF3 function as integrators of environmental signals and organ development, a critical strategy for survival in dynamic tropical ecosystems [5].

In the flowering regulation system of bananas, the MADS-box domain (*AGL11, SOC1, AGL12*) plays a central role. Recent studies confirm that *SOC1* is activated by short tropical photoperiods [6], while *AGL11* regulates seed production [7]. The synergy between domains is reinforced by the discovery of *FTIP7* (QUIRKY) as a scaffold protein for transporting florigen [8], a crucial mechanism for adapting to flowering in low-light environments.

The Myb domain exhibits complex functional diversification through the participation of specific genes. RS2 plays a role in suppressing RS1 expression to maintain normal differentiation of lateral organs, while *NUCT14* and *PLAT1* are simultaneously involved in regulating nucleotide and lipid metabolism [9]. In parallel, *PUB13* coordinates the degradation of flowering repressors through ubiquitination mechanisms [10], highlighting the key role of this domain in linking developmental processes to reproductive cycle regulation.

The AP2/ERF domain, known to play a role in growth, fruit ripening, and stress response [11], the discovery of the *RAV2* and *LOGL3* genes in the genome of *Musa acuminata* subsp. *sumatrana* opens new insights. These two genes have not been previously linked to the flowering process. Therefore, this finding suggests a specific function associated with the flower tissues that may be characteristic of subsp. *sumatrana*. The discovery has the potential to reflect an evolutionary innovation in the reproductive regulation mechanisms of the plant.

3.3. Phylogenetic Analysis of MADS-box Genes

The phylogenetic tree of the MADS-box gene in *Musa acuminata* subsp. *sumatrana*, reconstructed using the Neighbor-Joining (NJ) method, reveals three key divergence patterns: (1) *AGL11* forms a monophyletic clade separate from cultivated varieties (AAA group) with an NJ genetic distance of 0.41 and 100% identity; (2) *SOC1* shows the highest divergence level (NJ: 0.59); and (3) *AGL12* is the most conserved (NJ: 0.33). The branching topology is strongly supported by bootstrap values >85%, while the outgroup position (*Helianthus annuus*) at an NJ distance of 1.62 confirms the validity of the tree's root (**Fig. 1**).

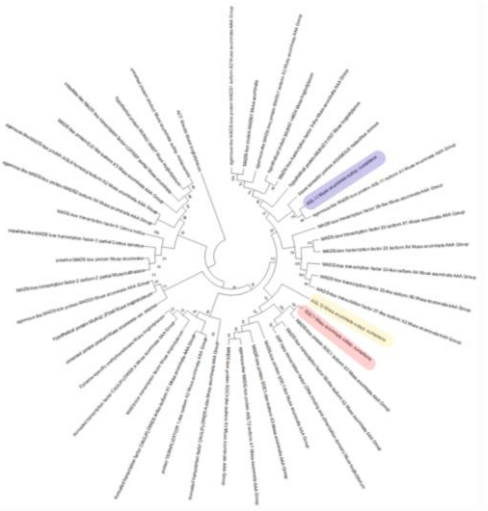

Fig. 1. Neighbor-Joining phylogeny of MADS-box genes in *M. acuminata* subsp. *sumatrana*, showing evolutionary divergence: *AGL11* (NJ:0.41), divergent *SOC1* (NJ:0,59), and conserved *AGL12* (NJ: 0.33). Bootstrap >85%; outgroup: *H. annuus* (NJ: 1.62)

The significant divergence of *SOC1* (NJ: 0.59) indicates the adaptation of subsp. *sumatrana* to tropical environments. As a photoperiod response gene, the alteration of this short light activation mechanism is suspected to support flowering in under-canopy habitats, in line with the report by Teo *et al*. (2019) on the correlation between MADS box gene duplication and tropical environmental adaptation [12]. Meanwhile, the conservation of *AGL12* (NJ: 0.33) indicates a stable function in organ development, such as root formation, which is relatively unaffected by domestication. These different divergence patterns reflect variations in evolutionary rates: *SOC1* undergoes more dynamic changes compared to *AGL12*.

The separation of the *AGL11* clade (100% identity, NJ: 0.41) shows genetic differences between wild and cultivated populations. Considering the presence of seeds in *Musa acuminata* subsp. *sumatrana* and the reduced expression of *AGL11* in seedless cultivars, this pattern supports the hypothesis that *AGL11* continues to function in seed formation in wild varieties. These findings imply that the parthenocarpy mechanism in cultivated bananas has not yet fully inhibited this genetic pathway in their wild relatives.

3.4. Integrated Flowering Regulation Model

Gene Ontology (GO) enrichment analysis for biological processes reveals that differentially regulated genes are significantly involved in "*RNA polymerase II transcription regulation, flower development, and maintenance of inflorescence meristem identity*" (**Fig. 2**). The categories "*Flower development*" and "*Regulation of flower development*" show the most significant enrichment (FDR < $1.0e^{-7}$), marked by bright colors and large circle sizes reflecting a high quantity of genes. Related processes such as reproductive structure development and shoot system development were also identified with moderate significance (FDR ~$1.0e^{-4}$).)

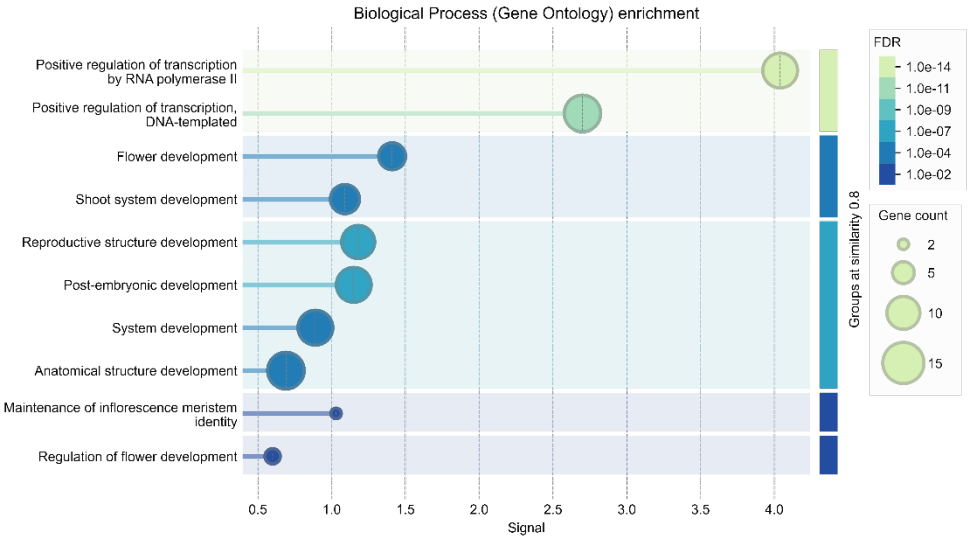

Fig. 2. Visualization of Gene Ontology enrichment analysis results for differentially expressed genes in the regulation of blooming and fruit development in bananas (*Musa* spp.). The circle's size denotes the quantity of genes associated with each biological process category, and the colour reflects the False Discovery Rate (FDR) value; a brighter hue signifies greater statistical significance.

The dominance of categories related to flowering (specifically *flower development, regulation of flower development*, and *maintenance of inflorescence meristem identity*)

directly reflects the activation of key flower formation genes in bananas, including *SOC1*, which functions as an integrator of environmental signals (temperature and photoperiod) to activate LFY (LEAFY) along with SEP (SEPALLATA) factors in the inflorescence meristem [12]. High statistical significance (FDR < $1.0e^{-7}$) and a large number of genes in this category align with the findings of MADS-box genes (*SOC1, AGL11*) and Myb (*PUB13*), which collectively coordinate the differentiation of flower meristems to the maturation of reproductive organs. Fluctuations in expression within the categories of "*Flower development*" and "*Post-embryonic development*" reflect the adaptation of these conserved pathways to environmental selection in tropical habitats. The association with RNA polymerase II transcription regulation confirms that the flowering cascade is regulated through large-scale transcriptional mechanisms. However, the enrichment is weaker in the non-flowering categories (*Shoot system development*, *Post-embryonic development*, *Reproductive structure development*), indicating that flowering genes remain the main drivers, supported by higher significance and a larger number of genes in the flower-related categories.

Bananas (*Musa* spp.) regulate flowering and fruit development through a specific mechanism. External stimuli (temperature and light) trigger the expression of the CONSTANS (CO) gene and the florigen transporter (QUIRKY), which induce the transition from the vegetative phase to flowering through the activation of miR172 and LFY. The MADS-box gene family (*AGL11*, *SOC1*, and *AGL12*) regulates the development of inflorescence meristems and activates AP1. Meanwhile, the AP2/ERF module (*RAV2*, *LOGL3*) and Myb factors (*RS1, RS2, NUCT14, PLAT1, PUB13*) coordinate in the differentiation of inflorescence and the initiation of fruit formation.

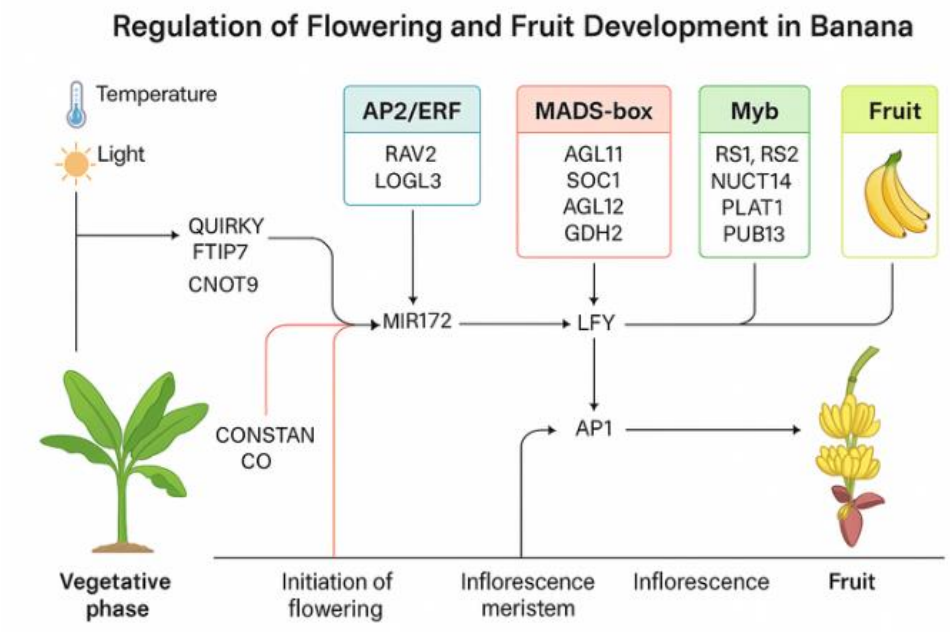

Fig. 3. External stimuli (temperature/light) activate CO and QUIRKY → miR172 and LFY for the vegetative-reproductive transition in *Musa*. MADS-box genes (*AGL11, SOC1, AGL12*) drive the

development of inflorescence meristems and the activation of AP1, coordinating with AP2/ERF (*RAV2, LOGL3*) and Myb transcription factors (*RS1, RS2, NUCT14, PLAT1, PUB13*) for inflorescence differentiation and fruit morphogenesis.

The mechanism in **Fig. 3** reinforces previous findings about the role of *SOC1* as an integrator of environmental signals [12]. The activation of CO-QUIRKY by temperature/light explains the fluctuations in flowering gene expression in the GO enrichment analysis (FDR < $1.0e^{-7}$), while also addressing the adaptation of bananas to the dynamic tropical climate. The vegetative-reproductive transition, mediated by miR172/LFY, is a critical checkpoint where miR172 suppresses the AP2 gene, promoting flowering. The dominance of the "*flower development*" category in GO enrichment aligns with the central function of MADS-box (*SOC1, AGL11*) in the inflorescence meristem, which catalyzes organ differentiation through AP1 activation.

The collaboration of AP2/ERF and Myb in **Fig. 3** (*RAV2, LOGL3, RS2, PLAT1, PUB13*) coordinates flower differentiation-fruit initiation. *PLAT1* is associated with the regulation of lipid-nucleotide metabolism, while PUB13 plays a role in the ubiquitination of flowering repressors [9,10]. Although *NUCT14* and *RS1* were not visualized, both genes are functionally related to *PLAT1* and *RS2* in the reproductive development tissues of bananas.

4 Conclusion

This study successfully characterized the flowering formation in *Musa acuminata* subsp. *sumatrana* through a hybrid whole-genome sequencing (WGS) approach (PromethION ONT and Illumina). The sequencing results yielded 13.9 Gb of data with an N50 of 4,097 bp, enabling the identification of 51,358 genes, including 4,885 annotated genes. Flowering gene analysis identified 12 candidate flowering genes. High expression of *AGL11* correlates with increased seed formation, indicating its role in fertility. Network regulation analysis confirmed the integration of environmental signals such as temperature and light with flower development through the CONSTANS, MADS-box, AP2/ERF, and Myb pathways. Additionally, two new genes (*RAV2* and *LOGL3*) were found to be associated with the flowering cascade in bananas for the first time. This finding highlights the genetic potential of Sumatran bananas as a source of superior alleles for high-yielding and environmentally adaptive banana breeding programs.

Acknowledgment

Sequencing in this study was funded by the Biological and Environmental Research Organization Program, National Research and Innovation Agency, fiscal year 2022, with the title “Exploration of Genetic Diversity of Wild Bananas *Musa acuminata* ssp. *halabenensis*, ssp. *sumatrana*, and ssp. *malaccensis* Using WGS.”

References

1. Food and Agriculture Organization of the United Nations. Banana production and trade data 2022. FAOSTAT (2023)
2. M. Dita, M. Barquero, D. Heck, E.S.G. Mizubuti, C.P. Staver, Fusarium wilt of banana: Current knowledge and future research directions. Front. Plant Sci. 9, (2018). doi :10.3389/fpls.2018.01468
3. T. Munhoz, J. Vargas, L. Teixeira, C. Staver, M. Dita, Fusarium Tropical Race 4 in Latin America and the Caribbean: status and global research advances towards disease management. Front Plant Sci. 16, (2024). doi: 10.3389/fpls.2024.1397617
4. D. Lakhwani, Y. V. Dhar, A. Singh, A. Pandey, P.K. Trivedi. M. H. Asif, Genome wide identification of MADS-box gene family in Musa Balbisiana and their divergence during evolution. Gene. 836, 146666 (2022). doi:10.1016/j.gene.2022.146666
5. W. Li, I. P. Ahn, Y. Ning, C.H. Park, L. Zeng, J. G. A. Whitehill, H. Lu, Q. Zhao, B. Ding, Q. Xie, J. M. Zhou, L. Dai, G. L. Wang, The U-Box/ARM E3 ligase PUB13 regulates cell death, defense, and flowering time in *Arabidopsis*. Plant Physiology. 159(1), 239-250 (2012). doi:10.1104/pp.111.192617
6. L. Dreni, The ABC of Flower Development in Monocots: The Model of Rice Spikelet. Methods Mol Biol. 2686, 59–82 (2023). doi:10.1007/978-1-0716-3299-4_3
7. N. Ocarez, N. Mejía, Suppression of the D-class MADS-box AGL11 gene triggers seedlessness in fleshy fruits. Plant Cell Rep. 35(1), 239–254 (2016). doi:10.1007/s00299-015-1882-x.
8. H. Takagi, A. K. Hempton, T. Imaizumi, Photoperiodic flowering in Arabidopsis: Multilayered regulatory mechanisms of CONSTANS and the florigen FLOWERING LOCUS T. Plant Commun. 4(3), 100552, (2023). doi: 10.1016/j.xplc.2023.100552.
9. M. Kulke, E. Kurtz, D. M. Boren, D. M. Olson, A. M. Koenig, S. Hoffmann-Benning, J. V. Vermaas, PLAT domain protein 1 (PLAT1/PLAFP) binds to the Arabidopsis thaliana plasma membrane and inserts a lipid. Plant Sci. 338, 111900 (2024). doi:10.1016/j.plantsci.2023.111900.
10. J. Yue, X. Zou, Y. Peng, S. Pan, C. Hu, B. Wang, L. Dai, W. Li, The Arabidopsis E3 ubiquitin ligase PUB13 synergistically interacts with BON1 to regulate plant flowering and immunity. Front Plant Sci. 16, 1585221 (2025). doi:10.3389/fpls.2025.1585221.
11. Q. Jiang, Z. Wang, G. Hu, X. Yao, Genome-wide identification and characterization of AP2/ERF gene superfamily during flower development in Actinidia eriantha. BMC Genomics. 23(1), 650 (2022). doi:10.1186/s12864-022-08871-4.
12. Teo ZWN, Zhou W, Shen L. 2019. Dissecting the function of MADS-Box transcription factors in Orchid reproductive development. Front Plant Sci. 10. doi:10.3389/fpls.2019.01474.