

RESSALVA

Atendendo solicitação do(a)
autor(a), o texto completo desta tese
será disponibilizado somente a partir
de 04/08/2027.

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**UNRAVELING GENOMIC AND PHENOMICS INSIGHTS INTO
FLOWERING-RELATED TRAITS IN SUGARCANE THROUGH
MACHINE LEARNING APPROACHES**

Paulo Henrique da Silva Santos

Master of Science in Genetics and Plant Breeding

2025

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**UNRAVELING GENOMIC AND PHENOMICS INSIGHTS INTO
FLOWERING-RELATED TRAITS IN SUGARCANE THROUGH
MACHINE LEARNING APPROACHES**

Paulo Henrique da Silva Santos

Advisor: Profa. Dra. Luciana Rossini Pinto Machado da Silva

Co-advisor: Prof. Dr. Elisson Antonio da Costa Romanel

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Agronomia (Genética e Melhoramento de Plantas).

2025

S237u Santos, Paulo Henrique da Silva
Unraveling genomic and phenomics insights into flowering-related traits in sugarcane through machine learning approaches / Paulo Henrique da Silva Santos. -- Jaboticabal, 2025
121 p. : il., tabs., fotos

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientadora: Luciana Rossini Pinto Machado da Silva
Coorientador: Elisson Antonio da Costa Romanel

1. Genética quantitativa. 2. Cana-de-açúcar. 3. Drone. 4. Bioinformática. 5. Redes neurais (Computação). I. Título.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: UNRAVELING GENOMIC AND PHENOMICS INSIGHTS INTO FLOWERING-RELATED TRAITS IN SUGARCANE THROUGH MACHINE LEARNING APPROACHES

AUTOR: PAULO HENRIQUE DA SILVA SANTOS

ORIENTADORA: LUCIANA ROSSINI PINTO

COORIENTADOR: ELISSON ANTÔNIO DA COSTA ROMANEL

Aprovado como parte das exigências para obtenção do Título de Doutor em Agronomia (Genética e Melhoramento de Plantas), pela Comissão Examinadora:



Documento assinado digitalmente

LUCIANA ROSSINI PINTO MACHADO DA SILVA

Data: 05/08/2025 09:36:56-0300

Verifique em <https://validar.iti.gov.br>

Pesquisadora Dra. LUCIANA ROSSINI PINTO (Participação Presencial)
Centro Avançado de Pesquisa Tecnológica do Agronegócio de Cana / IAC Ribeirão Preto/SP



Documento assinado digitalmente

RICARDO JOSE GONZAGA PIMENTA

Data: 05/08/2025 11:07:47-0300

Verifique em <https://validar.iti.gov.br>

Pesquisador Dr. RICARDO JOSÉ GONZAGA PIMENTA (Participação Virtual)
Department of Research Crops / Agriculture and Food Development Authority (Teagasc) - Oak Park/County Carlow/Ireland



Documento assinado digitalmente

GUSTAVO VITTI MORO

Data: 05/08/2025 19:03:42-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. GUSTAVO VITTI MÔRO (Participação Presencial)
Departamento de Produção Vegetal / FCAV UNESP Jaboticabal



Documento assinado digitalmente

JOAO RICARDO BACHEGA FEIJÓ ROSA

Data: 06/08/2025 08:22:56-0300

Verifique em <https://validar.iti.gov.br>

Dr. JOÃO RICARDO BACHEGA FEIJÓ ROSA (Participação Presencial)
RB Genetics & Statics Consulting / Jaú/SP



UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal



Documento assinado digitalmente

DAVID LUCIANO ROSALEN

Data: 06/08/2025 14:36:13-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. DAVID LUCIANO ROSALEN (Participação Presencial)
Departamento de Engenharia / FCAV / UNESP - Jaboticabal

Jaboticabal, 04 de agosto de 2025

I would like to dedicate this thesis to my biological grandfather and raising father, Manuel Julio da Silva (*in memoriam*), who always supported me in pursuing my studies. Despite having limited formal education, he consistently encouraged me and showed a deep interest in my academic journey.

AUTHOR'S INFORMATION

PAULO HENRIQUE DA SILVA SANTOS - Born on February 13, 1992, in Barreira (SP), Brazil, he graduated with a Bachelor's Degree in Biological Sciences from the Centro Universitário Barão de Mauá, where he also obtained his Teacher's Degree in Biology. During his undergraduate studies, he began an academic journey in 2016 with a scientific initiation scholarship at the Biotechnology Laboratory of the Instituto Agrônomo de Campinas (IAC) in Ribeirão Preto, SP, Brazil. Since then, he has worked as an external professional at the institution and in 2020, earned his Master's Degree in Agronomy (Genetics and Plant Breeding) at the São Paulo State University "Júlio de Mesquita Filho" (UNESP), Faculty of Agricultural and Veterinary Sciences, Jaboticabal Campus. His Master's research focused on gene expression in the sugarcane photoperiodic pathway. In 2021, he started his doctoral research in the same institution and postgraduate program. During his Ph.D., he completed a sandwich Ph.D. internship at the Queensland Alliance for Agriculture and Food Innovation (QAAFI), University of Queensland, Australia. His current main research interests include bioinformatics, high-throughput phenotyping, and data science, focusing on applying these approaches to plant breeding.

ACKNOWLEDGEMENTS

First of all, I would like to thank my family, especially my grandparents, also known as my parents—Tereza and Manuel (*in memoriam*), my mother Silvana, my father José, my uncle Luiz, my brothers Wilson and Ivanildo, my wife Barbara, and my daughter Sophia, who brought light into our lives upon her arrival.

I thank all of them for their unwavering support since the beginning of my academic journey, even during the toughest moments. I am truly grateful to have you all in my life—thank you!

All the progress made in experimental design and phenotypic data collection for this work would not have been possible without the collaboration of many people and institutions. I would like to thank Dr. Marcos Landell for welcoming me to the “Instituto Agronômico de Campinas” – IAC (Centro de Cana), and Dr. Mauro Xavier for making the operational logistics possible to implement the experimental design at the sugarcane breeding station in Uruçuca-BA.

Regarding the field trials, my sincere thanks go to Dr. Carlos Kantack and Casimiro, and their team, nicknamed as: Kel, Quinho, Amarelo, Cambota, Macumba, —thank you!

To my advisor, Dr. Luciana Rossini, and co-advisor, Dr. Elisson Romanel—thank you for all your efforts in supporting me throughout the PhD. Your guidance and discussions greatly contributed to the success of the experiments, and beyond my development as a researcher, your mentorship also helped me grow personally.

I am grateful to Dra. Anete Pereira for receiving me at CBMEG/UNICAMP for GBS library preparation, and to Dr. Aline Moraes and Dr. Ricardo Pimenta for their assistance during this step. A special thanks to Dr. Ricardo Pimenta as well, for helping me navigate the bioinformatics analyses.

I would like to thank all my colleagues and collaborators at the Biotechnology Laboratory at IAC, including Dr Alexandre Boer, Dr João Ricardo, Dr Marcel Fernando, and the lab technicians Thais, Maicon and Kátia, for their partnership throughout the project. Whether in the lab, the field, or life in general, it was always a pleasure to have you around.

I would also like to thank Dr. Craig Hardner and Dra. Elizabeth Ross for accepting me at the University of Queensland, and my colleagues Daniel Edge-Garza, Shashi, Stephan, Phu Khang Ha, and Norman for making my stay much easier with their warm welcome and continuous support.

I am grateful to the Graduate Program in Agronomy (Genetics and Plant Breeding) – FCAV/UNESP for their support during my studies. To conclude, I would like to express my sincere gratitude to the examination committee — Dr. Ricardo José Gonzaga Pimenta, Prof.Dr. Gustavo Vitti Môro, Dr. João Ricardo Bachega Feijó Rosa, and Prof. Dr. David Luciano Rosalen — for their invaluable contributions. If for any reason you contributed to this work and were not mentioned, please accept my sincere apologies, and thank you!

This research was funded by FAPESP (grant number 2022/16415-4); the National Council for Scientific and Technological Development (CNPq/FNDCT/MCTI, grant number 408043/2022-9); and was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

SUMÁRIO

RESUMO	iv
ABSTRACT	vi
1 General Considerations	1
1.1 Introduction	1
1.2 Literature Review	2
1.2.1 A Brief History of Sugarcane	2
1.2.2 Genomics in Sugarcane Breeding	3
1.2.3 General Aspects of Flowering in Sugarcane	5
1.2.4 High-throughput phenotyping and machine learning in sugarcane	7
References	10
2 High-Throughput Phenotyping for the Prediction and Quantification of Flower-Related Traits in Sugarcane¹	21
Abstract	22
2.1 Introduction	22
Introduction	22
2.2 MATERIALS AND METHODS	24
2.2.1 Experimental design and study area	24
2.2.2 Unmanned aerial vehicle data collection and extraction	24
2.2.3 Phenotyping and data analysis	26
2.2.4 General Machine Learning modelling and Data Analysis	27
2.2.5 Machine Learning Classification Model	28
2.2.6 Machine Learning Regression Models	28
2.2.7 Comparison of days to flowering observed in the field and the orthomosaic database	29
2.2.8 Deep learning model for inflorescence counting	29
2.3 Results	30
2.3.1 Mixed Linear Model and Descriptive analysis	30
2.3.2 Heritability and Repeatability	31
2.3.3 Selection of vegetation index and metrics	34
2.3.4 Machine Learning Classification Models	34
2.3.5 Machine Learning Regression Predictive Models	37
2.3.6 Comparisons of days to flowering on field and orthomosaic observations	38

2.3.7	A comparative analysis of observed flower counts versus artificial counts utilizing computer vision techniques	39
2.4	Discussion	41
2.5	Conclusion	44
	Acknowledgments	44
	References	45
3	Genome-Wide Association Studies of Flowering Time in Sugarcane Germplasm	49
	Abstract	49
3.1	Introduction	50
	Introduction	50
3.2	Materials and Methods	51
	Materials and Methods	51
3.2.1	Experiment design and Phenotypic data	51
3.2.2	Phenotyping Model and Statistical Analysis	51
3.2.3	High Throughput Phenotypic Data Acquisition	52
3.2.4	Genotyping by sequencing library preparation	52
3.2.5	Variant calling	53
3.2.6	Trap Markers for Flowering Locus T Genes	53
3.2.7	Structure and Diversity Analysis	54
3.2.8	Linkage Disequilibrium	54
3.2.9	Genome wide association studies	54
3.2.10	Mapping and Annotation	55
3.2.11	Feature Selection and Machine Learning Analysis	55
3.3	Results	56
3.3.1	Phenotypic analysis	56
3.3.2	Variant calling and Population Structure Analysis	56
3.3.3	Association analysis for ground measures	58
3.3.4	Association analysis on phenomics data	59
3.3.5	TRAP marker association on Fixed and random model Circulating Probability Unification	60
3.3.6	Mapping and Annotation analysis	60
3.3.7	Variant annotation and functional analysis	63
3.3.8	Machine learning and feature selection	64
3.4	Discussion	66
	References	68
4	Final considerations	75

A APPENDIX A	76
APPENDIX A-Supplementary material chapter 2	76
A APPENDIX B	83
APPENDIX B-Supplementary material chapter 3	83

DESVENDANDO PERSPECTIVAS GENÔMICAS E FENÔMICAS SOBRE CARACTERÍSTICAS RELACIONADAS À FLORAÇÃO EM CANA-DE-AÇÚCAR USANDO MÉTODOS DE APRENDIZADO DE MÁQUINA

RESUMO -

Cana-de-açúcar (*Saccharum spp.*) é uma gramínea de alta capacidade fotossintética e acúmulo de sacarose, sendo de suma importância para o setor sucroenergético e o agronegócio brasileiro. O florescimento é um dos processos fisiológicos que afeta negativamente, tanto a qualidade da matéria prima como a produtividade, porém é importante para o melhoramento genético na obtenção de novas cultivares, sendo portanto relevante uma busca de padrões fenotípicos e genotípicos associados à característica. Neste sentido, objetivou-se a implementação de fenotipagem de alto rendimento para avaliação dos caracteres envolvidos na floração e um estudo de associação genômica ampla (GWAS) em um painel de acessos utilizando caracteres fenotípicos coletados tradicionalmente e via fenômica. Um painel de 154 genótipos foi instalado na estação de hibridação do Programa de Melhoramento Genético de Cana-de-Açúcar do Instituto agrônômico de Campinas (IAC), localizado em Urucu-BÁ. Foram avaliados fenotipicamente em cana-planta (2023) e primeira soca (2024) com voos de drone realizados durante as fases de florescimento com posterior processamento de ortomosaicos e métricas para análises com abordagens de aprendizado de máquina. A genotipagem foi realizada via sequenciamento para em conjunto com dados fenotípicos realizar-se o GWAS. O primeiro capítulo abrange uma pequena revisão introdutória sobre os principais temas da tese como um histórico da cana-de-açúcar, além dos principais aspectos relacionados a floração, mapeamento associativo, fenômica e aprendizado de máquina. O segundo capítulo aborda a fenômica, onde um modelo de aprendizado profundo de redes neurais convolucionais (CNN) para detecção e contagem de inflorescências obteve uma acurácia de 84% e sua correlação com dados de campo variaram de $R^2 = 0.29 - 0.72$. Já o modelo de aprendizado de máquina supervisionado para classificação de resposta ao florescimento por redes neurais perceptron multicamadas obtiveram acurácia de até 87%. Por fim, dentre os modelos de regressão testados, o algoritmo xgboost teve a melhor performance com $R^2 = 0.52$, indicando que fenômica é uma abordagem promissora para complementação de dados fenotípicos a nível de campo. O capítulo três aborda o mapeamento associativo, por intermédio da prospecção de 19.139 single nucleotide polymorphism (SNPs), na análise de associação foram encontrados 54 marcadores associados significativamente com dados fenotípicos tradicionais e 69 marcadores associados a dados de fenotipagem de alto rendimento. Além disso, um modelo de classificação por redes neurais após a etapa de seleção de atributos, com um conjunto de dados de apenas 136 SNPs selecionados, atingiu 100% de acurácia na classificação em resposta ao tempo para o florescimento a modelagem para mesma tarefa com dados de fenômica atingiu 87%. Por fim o capítulo 4 traz as considerações finais destacando os principais resultados que ressaltam o poder preditivo de dados de fenotipagem de alto rendimento e GWAS, como ferramentas conjuntas que podem

auxiliar os melhoristas e produtores a lidarem com a característica do florescimento em cana-de-açúcar.

Palavras-chave: Fenômica, GWAS, Índices vegetativos, Mapeamento associativo, Seleção genômica

UNRAVELING GENOMIC AND PHENOMICS INSIGHTS INTO FLOWERING-RELATED TRAITS IN SUGARCANE THROUGH MACHINE LEARNING APPROACHES

ABSTRACT -

Sugarcane (*Saccharum spp.*) is a grass distinguished by its remarkable photosynthetic efficiency and notable sucrose accumulation, making it essential to the sugar-energy sector and a cornerstone of Brazilian agribusiness. The physiological process of flowering adversely affects both the quality of the raw material and the overall yield; nevertheless, it holds significant importance for genetic breeding endeavours aimed at developing novel varieties. Therefore, it is crucial to delineate phenotypic and genotypic patterns associated with this trait. In this regard, the primary objective of the present research was to implement high-throughput phenotyping to evaluate traits pertinent to flowering, in conjunction with conducting a genome-wide association study (GWAS) utilizing a diverse panel of germplasm accessions. Both traditional and phenomics methodologies were used to assess the phenotypic traits. A total of 154 genotypes were planted at the sugarcane crossing station of the Instituto Agronômico de Campinas (IAC) located in Urucuá, Bahia. These cultivars underwent phenotypic evaluation during the plant-cane stage in 2023 and the first ratoon stage in 2024, employing Unmanned Aerial Vehicles (UAV) flights throughout the flowering season. This process was complemented by the generation of orthomosaics and the computation of vegetative indices and metrics for subsequent analysis utilizing machine learning techniques. Genotyping was conducted via genotyping by sequencing (GBS) to support the GWAS. The first chapter provides an introductory overview of the main thesis topics, including sugarcane history, as well as aspects of flowering, association mapping, phenomics, and machine learning. The second chapter of this thesis specifically addresses high-throughput phenotyping. A sophisticated deep learning framework employing a convolutional neural network (CNN) for the detection and quantification of inflorescences achieved an accuracy of 84%, with correlation to the ground truth ranging from $R^2 = 0.29$ to 0.72 . Moreover, neural network classification models, particularly multi-layer perceptrons (MLP), attained classification accuracies of up to 87%. Among the regression models assessed, the XGBoost algorithm demonstrated superior efficacy, achieving an $R^2 = 0.52$, indicating its potential as an ancillary approach to phenotypic data acquisition in field environments. The third chapter is dedicated to association mapping. Through the examination of 19,139 single-nucleotide polymorphisms (SNPs) via GBS, 54 markers were identified in the GWAS as significantly correlated with field data, while an additional 69 markers displayed associations with high-throughput phenotyping data. Notably, post-feature selection, a neural network classification model achieved a classification accuracy of 100% concerning flowering time responses with a restricted dataset of 136 selected. Lastly, the fourth chapter brings the final considerations, highlighting the key findings that underscore the predictive capacity of integrating high-throughput phenotyping data with GWAS methodologies, which can significantly assist breeders and producers in the management of

flowering traits in sugarcane.

Keywords: Phenomics, GWAS, Vegetation indices, Association mapping, genomic selection

CHAPTER 1 – GENERAL CONSIDERATIONS

1.1 Introduction

Climate change, primarily driven by greenhouse gas emissions, is linked to an increase in the frequency of extreme catastrophic events (Zhao et al. 2015). Within this framework, bioenergy crops are posited to play a pivotal role; specifically, bioethanol is recognized as a sustainable energy source capable of reducing greenhouse gas emissions by approximately 85% when compared to fossil fuels like gasoline (Goldemberg and Guardabassi 2010). However, climate change poses potential risks to the production yield of sugarcane (Linnenluecke, Nucifora, and Thompson 2018). In 2024, the Brazilian sugarcane harvest yield was recorded at 79.95 tons per hectare, culminating in a total production yield of 689.8 million tons. The sugar market generated revenues totalling 5.6 billion United States dollars (USD), indicating a 24% increase. Conversely, ethanol exports saw a decline of 17.2%, totalling 440.1 million liters, which has significantly influenced Brazil's Gross Domestic Product (GDP).

Flowering represents a significant challenge in sugarcane cultivation, adversely impacting yield by diminishing the accumulation of sugar within the stalk (Berding and Hurney 2005). This phenomenon may lead to stalk desiccation, commonly termed pithiness (Evans 1966; Caputo et al. 2007). In contrast, flowering is a critical trait for breeders to monitor during the development of new varieties; in particular, although many sugarcane varieties successfully flower in equatorial regions, some genotypes may experience flowering recalcitrance and lack of synchrony, making it difficult to select the best crossbreeding options. This restricts the choice to genotypes that flower simultaneously, presenting a significant limitation for breeding programs (Glassop, Anne L Rae, and Bonnett 2014). Currently, sugarcane is classified as an Intermediate-short Day Plant (IDP), indicating that its flowering occurs when the photoperiod is shorter than its critical flowering photoperiod (Paul H. Moore and Berding 2013). Moreover, the sugarcane flowering response exhibits substantial variability and is significantly influenced by genotype, even under optimal natural conditions (Melloni et al. 2015). Given the significance of flowering in this critical agricultural crop, it is imperative to develop a comprehensive understanding of this trait to aid breeders and producers in effective management.

High-throughput phenotyping (HTP) employing unmanned aerial vehicles (UAVs) represents a promising, cost-effective, and scalable approach for the acquisition of extensive phenotypic data (Luo, Liu, and Lakshmanan 2023). This methodology is particularly advantageous for traits demonstrating temporal variations, such as flowering. HTP operates in concert with high-throughput genotyping techniques such as Genotyping by Sequencing (GBS) (Elshire et al. 2011), which efficiently addresses the analysis of a large number of single nucleotide polymorphism (SNP) markers. The feasibility of GBS is further enhanced by the reduced costs associated with next-generation sequencing (NGS), thus facilitating marker-trait associations (MTA). This contemporary approach seeks to synergize HTP-generated phenotypic data with NGS-derived genotypic data, thereby improving genetic gains in sugarcane breeding effectively, par-

ticularly by streamlining the preliminary selection phase. Currently, after significant yield improvements in recent decades primarily through traditional breeding methods, sugarcane is experiencing a plateau phase characterized by minimal substantial progress in critical yield traits (Yadav et al. 2020). Therefore, both genomic selection (GS) and genome-wide association studies (GWAS) are important strategies for future advancements. (Hong, W. Huang, et al. 2024). There are also more specific approaches, such as using breeding values to guide crossbreeding through genomic prediction (Gonçalves et al. 2025).

Association mapping (AM) strategies offer notable advantages compared to traditional linkage maps (LM), which generally require detailed pedigree information and often display low resolution. In contrast, AM not only delivers improved genomic resolution but also enables the conduct of genetic population studies and provides access to germplasm information independent of previous genealogical data. Furthermore, AM is capable of identifying several alleles at each locus and recording historical events that have occurred throughout evolutionary history. It exhibits significant power in terms of linkage markers; however, this effectiveness relies on the utilization of a high number of SNP markers, accompanied by careful curation and extensive genome coverage, to realize its full potential (Banerjee et al. 2020).

In contemporary research, alongside omics, a notable emerging methodology for data analysis is machine learning (ML), which significantly enhances analytical capabilities for managing large datasets, such as those derived from NGS and HTP. This advancement enables the application of both supervised and unsupervised models to uncover latent patterns within extensive data collections. Consequently, the primary objectives of this thesis are: i) to examine traits associated with flowering and their correlations with vegetative indices and metrics obtained through high-throughput phenotyping, aiming to develop machine learning and deep learning models; and ii) to conduct a GWAS to identify markers linked to flowering traits, utilizing an integrated approach that combines GWAS and HTP methodologies.

2.5 Conclusion

High-throughput phenotyping (HTP) serves as an effective and scalable methodology for conducting phenotypic analysis in sugarcane, which has conventionally emphasized traits pertaining to productivity and phytopathology. This study aims to bridge the gap in research regarding flowering-related traits, elucidating significant findings in this area. Vegetation index and metrics, including canopy cover (CC), excess greenness (EXG), plant height (PH), and plant volume (PV), can yield valuable insights into flowering traits. These vegetation index and metrics demonstrated superior performance in classification models compared to regression models. For labor-intensive traits, such as flowering count and flowering intensity in cane (PC), as well as days to flowering (DTF) and days to flag leaf emergence (DFTL) in the first ratoon (FR), regression models employing lasso and ridge regularization outperformed ensemble models, including random forest (RF) and XGBoost (XGB). Notably, DTF can be accurately assessed utilizing orthomosaic datasets from extensive field trials, which exhibit strong correlations with ground truth data. Vision-based computational models for the detection and enumeration of inflorescences in sugarcane remain in the developmental stage. These models show strong correlations with ground truth in sparse flowering plots at the beginning of the flowering season. However, their accuracy decreases significantly as canopy flowering coverage increases, making detection and counting difficult. Future enhancements could tackle these limitations by including multi-year data and training models across a broader spectrum of sugarcane varieties and flowering stages. The future combination of high-throughput phenotypic data with genotypic information will facilitate genomic selection related to flowering traits. This integration promises to expedite breeding cycles and enhance the efficiency of sugarcane improvement programs.

Acknowledgments

This research was funded by FAPESP, grant count 2022/16415-4 and by National Council for Scientific and Technological Development (CNPq/FNDCT/MCTI) - grant count 408043/2022-9 and financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

References

- Alexander, David H, John Novembre, and Kenneth Lange (2009a). Fast model-based estimation of ancestry in unrelated individuals. **Genome research** 19.9, pp. 1655–1664.
- (Sept. 2009b). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. **Genome Research** 19.9, pp. 1655–1664. ISSN: 1088-9051. DOI: 10.1101/gr.094052.109.
- Amadeu, Rodrigo R. et al. (Nov. 2016). AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. **The Plant Genome** 9.3. ISSN: 1940-3372. DOI: 10.3835/plantgenome2016.01.0009.
- Andrés, Fernando and George Coupland (Aug. 2012). The Genetic Basis of Flowering Responses to Seasonal Cues. **Nature Reviews Genetics** 2012 13:9 13.9, pp. 627–639. ISSN: 1471-0064. DOI: 10.1038/nrg3291. PMID: 22898651.
- Andrews, S (2010). FastQC: a quality control tool for high throughput sequence data. abraham ioinformatics, abraham Institute. **Cambridge, United Kingdom**.
- Aono, Alexandre Hild, Estela Araujo Costa, et al. (Nov. 2020). Machine Learning Approaches Reveal Genomic Regions Associated with Sugarcane Brown Rust Resistance. **Scientific Reports** 10.1, p. 20057. ISSN: 2045-2322. DOI: 10.1038/s41598-020-77063-5.
- Aono, Alexandre Hild, Ricardo José Gonzaga Pimenta, et al. (2024). Multiomic insights into sucrose accumulation in sugarcane. **bioRxiv**, pp. 2024–06.
- Araldi, Rosilaine et al. (Mar. 2010). Florescimento Em Cana-de-Açúcar. **Ciência Rural** 40.3, pp. 694–702. ISSN: 1678-4596. DOI: 10.1590/S0103-84782010005000033.
- Bhat, Shripad R and SS Gill (1985). The implications of 2n egg gametes in nobilization and breeding of sugarcane. **Euphytica** 34.2, pp. 377–384.

- Breiman, Leo (Aug. 1996). Bagging Predictors. **Machine Learning** 24.2, pp. 123–140. ISSN: 0885-6125. DOI: 10.1007/BF00058655.
- (2001). Random Forests. **Machine Learning** 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324.
- Camacho, Christiam et al. (2009). BLAST+: architecture and applications. **BMC bioinformatics** 10, pp. 1–9.
- Catchen, Julian et al. (June 2013). Stacks: An Analysis Tool Set for Population Genomics. **Molecular Ecology** 22.11, pp. 3124–3140. ISSN: 09621083. DOI: 10.1111/mec.12354.
- Cenci, Alberto and Mathieu Rouard (Mar. 2017). Evolutionary Analyses of GRAS Transcription Factors in Angiosperms. **Frontiers in Plant Science** 8, p. 229588. ISSN: 1664462X. DOI: 10.3389/FPLS.2017.00273/BIBTEX.
- Chen, Tianqi and Carlos Guestrin (Aug. 2016). XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.
- Coelho, P et al. (2014). Putative Sugarcane FT / TFL1 Genes Delay Flowering Time and Alter Reproductive Architecture in Arabidopsis. 5 (May), pp. 1–12. DOI: 10.3389/fpls.2014.00221.
- Cover, T. and P. Hart (Jan. 1967). Nearest Neighbor Pattern Classification. **IEEE Transactions on Information Theory** 13.1, pp. 21–27. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1053964.
- Cristianini, Nello and John Shawe-Taylor (2004). Support Vector Machines and other kernel-based learning methods. Cambridge.
- Danecek, Petr et al. (Aug. 2011). The Variant Call Format and VCFtools. **Bioinformatics (Oxford, England)** 27.15, pp. 2156–2158. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr330.
- Deren, CW (1992). Stability and heritability of pith in sugarcane and its influence on yield. **Plant breeding** 109.3, pp. 242–247.

- Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular ecology** 14.8, pp. 2611–2620.
- Fan, Jianqing and Jinchi Lv (Jan. 2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. **Statistica Sinica** 20.1, p. 101. ISSN: 10170405. PMID: 21572976.
- Federer, Walter T et al. (1956). Augmented (or hoonuiaku) designs.
- Freund, Yoav and Robert E Schapire (Aug. 1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences** 55.1, pp. 119–139. ISSN: 00220000. DOI: 10.1006/jcss.1997.1504.
- Frichot, Eric and Olivier François (Aug. 2015). LEA: An R Package for Landscape and Ecological Association Studies. **Methods in Ecology and Evolution** 6.8, pp. 925–929. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12382.
- Friedman, Nir, Dan Geiger, and Moises Goldszmidt (1997). A. **Machine Learning** 29.2/3, pp. 131–163. ISSN: 08856125. DOI: 10.1023/A:1007465528199.
- Fujiwara, Sumire et al. (Dec. 2008). Circadian Clock Proteins LHY and CCA1 Regulate SVP Protein Accumulation to Control Flowering in Arabidopsis. **The Plant Cell** 20.11, pp. 2960–2971. ISSN: 1040-4651. DOI: 10.1105/TPC.108.061531. PMID: 19011118.
- Garcia, Antonio AF et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Scientific reports** 3.1, p. 3399.
- Geurts, Pierre, Damien Ernst, and Louis Wehenkel (Apr. 2006). Extremely Randomized Trees. **Machine Learning** 63.1, pp. 3–42. ISSN: 0885-6125. DOI: 10.1007/s10994-006-6226-1.
- Glaubitz, Jeffrey C. et al. (Feb. 2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. **PLoS ONE** 9.2. Ed. by Nicholas A. Tinker, e90346. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0090346.
- Grativol, Clícia et al. (July 2014). Sugarcane Genome Sequencing by Methylation Filtration Provides Tools for Genomic Research in the Genus *Saccharum*.

charum. **The Plant Journal** 79.1, pp. 162–172. ISSN: 0960-7412. DOI: 10.1111/tpj.12539.

Hastie, Trevor (2009). The elements of statistical learning: data mining, inference, and prediction.

Healey, A. L. et al. (Apr. 2024). The Complex Polyploid Genome Architecture of Sugarcane. **Nature** 628.8009, pp. 804–810. ISSN: 0028-0836. DOI: 10.1038/s41586-024-07231-4.

Hu, Jinguo et al. (2003). Assessing genetic variability among sixteen perennial Helianthus species using PCR-based TRAP markers. In: *Proc of 25th Sunflower Research Workshop, Fargo, North Dakota, USA*, pp. 1–5.

Jombart, Thibaut (June 2008). *Adegenet* : A R Package for the Multivariate Analysis of Genetic Markers. **Bioinformatics (Oxford, England)** 24.11, pp. 1403–1405. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn129.

Karlgren, A. et al. (2011). Evolution of the PEBP Gene Family in Plants: Functional Diversification in Seed Plant Evolution. **Plant Physiology** 156.4, pp. 1967–1977. ISSN: 0032-0889. DOI: 10.1104/pp.111.176206. PMID: 21642442.

Kodama, Yutaka, Noriyuki Suetsugu, and Masamitsu Wada (2011). Novel protein-protein interaction family proteins involved in chloroplast movement response. **Plant Signaling & Behavior** 6.4, pp. 483–490.

LaValle, Steven M, Michael S Branicky, and Stephen R Lindemann (2004). On the relationship between classical grid search and probabilistic roadmaps. **The International Journal of Robotics Research** 23.7-8, pp. 673–692.

Li, H. and R. Durbin (July 2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. **Bioinformatics (Oxford, England)** 25.14, pp. 1754–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324.

Li, Yu-Long and Jin-Xian Liu (2018). StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. **Molecular ecology resources** 18.1, pp. 176–177.

Linhares-Neto, Manoel Viana et al. (Sept. 2023). Molecular Screening Reveals a Photoperiod Responsive Floral Regulator in Sugarcane. **Theoretical and Experimen-**

tal Plant Physiology 35.3, pp. 199–214. ISSN: 21970025. DOI: 10.1007/s40626-023-00276-2/FIGURES/6.

Liu, Xiaolei et al. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS genetics** 12.2, e1005767.

Luo, Ting, Xiaoyan Liu, and Prakash Lakshmanan (Jan. 2023). A Combined Genomics and Phenomics Approach Is Needed to Boost Breeding in Sugarcane. **Plant phenomics (Washington, D.C.)** 5. ISSN: 2643-6515. DOI: 10.34133/plantphenomics.0074.

Manechini, João Ricardo Vieira et al. (June 2021). Transcriptomic Analysis of Changes in Gene Expression During Flowering Induction in Sugarcane Under Controlled Photoperiodic Conditions. **Frontiers in Plant Science** 12. ISSN: 1664-462X. DOI: 10.3389/fpls.2021.635784.

Mendiburu, Felipe de (2021). agricolae tutorial (Version 1.3-5). **Universidad Nacional Agraria: La Molina, Peru.**

Miller, JD and NI James (1974). The influence of stalk density on cane yield.

Ming, Ray et al. (2002). Comparative Analysis of QTLs Affecting Plant Height and Flowering among Closely-Related Diploid and Polyploid Genomes. **Genome** 45.5, pp. 794–803.

Moore, Paul H. and Frederik C. Botha, eds. (Dec. 2013). Sugarcane: Physiology, Biochemistry, and Functional Biology. Chichester, UK: John Wiley & Sons Ltd. ISBN: 978-1-118-77128-0. DOI: 10.1002/9781118771280.

OECD and FAO (2023). OECD-FAO Agricultural Outlook 2023-2032. Accessed: 2025-06-17. Paris: OECD Publishing. DOI: 10.1787/08801ab7-en. URL: <https://doi.org/10.1787/08801ab7-en>.

Pereira, Guilherme S., Antonio Augusto F. Garcia, and Gabriel R. A. Margarido (Dec. 2018). A Fully Automated Pipeline for Quantitative Genotype Calling from next Generation Sequencing Data in Autopolyploids. **BMC Bioinformatics** 19.1, p. 398. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2433-6.

Pimenta, Ricardo José Gonzaga et al. (Dec. 2021). Genome-Wide Approaches for the Identification of Markers and Genes Associated with Sugarcane Yellow Leaf Virus

- Resistance. **Scientific Reports** 11.1, p. 15730. ISSN: 2045-2322. DOI: 10.1038/s41598-021-95116-1.
- Popescu, Marius-Constantin et al. (2009). Multilayer perceptron and neural networks. **WSEAS Transactions on Circuits and Systems** 8.7, pp. 579–588.
- Quinlan, J. R. (Mar. 1986). Induction of Decision Trees. **Machine Learning** 1.1, pp. 81–106. ISSN: 0885-6125. DOI: 10.1007/BF00116251.
- R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Raboin, Louis-Marie et al. (2008). Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. **Theoretical and Applied Genetics** 116.5, pp. 701–714.
- Raj, Anil, Matthew Stephens, and Jonathan K Pritchard (June 2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. **Genetics** 197.2, pp. 573–589. ISSN: 1943-2631. DOI: 10.1534/genetics.114.164350.
- Rasmussen, Carl Edward (2004). Gaussian Processes in Machine Learning. In: pp. 63–71. DOI: 10.1007/978-3-540-28650-9_4.
- Rasool, Saiema and Rozi Mohamed (Sept. 2015). Plant Cytochrome P450s: Nomenclature and Involvement in Natural Product Biosynthesis. **Protoplasma** 2015 253:5 253.5, pp. 1197–1209. ISSN: 1615-6102. DOI: 10.1007/s00709-015-0884-4. PMID: 26364028.
- Rosyara, Umesh R. et al. (July 2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. **The Plant Genome** 9.2. ISSN: 1940-3372. DOI: 10.3835/plantgenome2015.08.0073.
- Scott, R. A. and G. A. Milliken (July 1993). A SAS Program for Analyzing Augmented Randomized Complete-Block Designs. **Crop Science** 33.4, pp. 865–867. ISSN: 0011-183X. DOI: 10.2135/cropsci1993.0011183X003300040046x.
- Slater, Anthony T. et al. (Nov. 2016). Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. **The Plant Genome** 9.3. ISSN: 1940-3372. DOI: 10.3835/plantgenome2016.02.0021.

- Venail, J. et al. (2022). Erratum: Analysis of the PEBP Gene Family and Identification of a Novel FT Orthologue in Sugarcane (Journal of Experimental Botany (Erab539) DOI: 10.1093/Jxb/Erab539). **Journal of Experimental Botany** 73.12. ISSN: 14602431. DOI: 10.1093/jxb/erac106.
- Voorrips, R. E. (Jan. 2002). MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. **Journal of Heredity** 93.1, pp. 77–78. ISSN: 14718505. DOI: 10.1093/jhered/93.1.77.
- Wei, Xianming et al. (2010). Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane. **Genome** 53.11, pp. 973–981.
- Weir, Bruce S and C Clark Cockerham (1984). Estimating F-statistics for the analysis of population structure. **evolution**, pp. 1358–1370.
- Wellmer, Frank and José L. Riechmann (Dec. 2010). Gene Networks Controlling the Initiation of Flower Development. **Trends in Genetics** 26.12, pp. 519–527. ISSN: 01689525. DOI: 10.1016/j.tig.2010.09.001. PMID: 20947199.
- Wright, Sewall (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution**, pp. 395–420.
- Zhang, Jisen et al. (2025). The highly allo-autopolyploid modern sugarcane genome and very recent allopolyploidization in *Saccharum*. **Nature Genetics** 57.1, pp. 242–253.
- Zhao, Hui et al. (Oct. 2015). Holocene Climate Changes in Westerly-Dominated Areas of Central Asia: Evidence from Optical Dating of Two Loess Sections in Tianshan Mountain, China. **Quaternary Geochronology** 30, pp. 188–193. ISSN: 18711014. DOI: 10.1016/j.quageo.2015.04.002.