Multi-omics data integration and computational approaches to enhance gene annotations and decipher function

by

Shatabdi Sen

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Carson M. Andorf, Co-major Professor
Justin Walley, Co-major Professor
Sarah N Andersen
Lynna Chu
Qi Li

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2023

DEDICATION

I would like to dedicate this dissertation to my mother, Bandana Sen, and father, Jia Ram Sen, who have always stood by me, sacrificed throughout their lives to make sure I get the best of everything in life. Further, I would like to dedicate this dissertation to my elder brother, Shantanu Sen, without whose active moral support and encouragement, this dissertation would not have been complete. I would not be where I am today without their constant love, support, and guidance.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

NOMENCLATURE

SQL	Structured Query Language
VCF	Variant Calling File
TSS	Transcription Start Site
FPKM	Fragments per kilobase of transcript per million of mapped reads
MFS	Maize Feature Store
ML	Machine Learning
WGD	Whole Genome Duplicate
GFF	Genome Feature Format
GUI	Graphical User Interface

ACKNOWLEDGMENTS

I would like to start off by acknowledging my major advisor and mentor during my doctoral training, Dr Carson M. Andorf. He has been a wonderful mentor, and I would like to thank him for his constant guidance, support and encouragement throughout my graduate life. I am highly indebted to Dr Andorf for always believing in me in the toughest of situations and helping me in shaping my dissertation. I am fortunate to have been mentored by such a humble and knowledgeable guide. Dr Andorf's constant support has helped me grow and develop as an independent researcher. I would like to extend my gratitude to my co-major professor, Dr Justin Walley, for his consistent support and constructive feedback. Dr Walley's unwavering encouragement, prompt assistance, and engaging discussions have been instrumental in shaping my research. Special thanks to Dr Sarah N Andersen for her mentorship in plant biology, guiding me with insightful questions and enhancing my skills in the field. A heartfelt acknowledgment goes to Dr Lynna Chu, whose continuous support and guidance have played a crucial role in my understanding of machine learning. Dr Chu's teaching in the DS 303 course laid the foundation for my robust knowledge in machine learning. I am also thankful to Dr Qi Li, whose input during my prelims significantly influenced the development of my final chapter.

I extend my appreciation to Margaret R Woodhouse, my co-author and mentee, for engaging discussions and lab meetings that fueled my projects and contributed to my first paper. Special thanks to John L Portwood from USDA ARS for his consistent assistance with IT-related queries, and in making my research accessible to a wider audience.

I would like to acknowledge Dr Kelley Dior for her invaluable feedback and guidance during lab presentations, which contributed to the improvement of my presentation and speaking

skills. Gratitude to the past and current program coordinators, Trish Stauble and Carla Harris, for their assistance with various paperwork.

A big thank you to all the past and current members of the Carson and Walley lab, including Dr Nancy Manchanda, Dr Sagnik Banerjee, Dr Rita Hayford, Dr Olivia Haley, and Dr Shikha Malik, for their support and assistance at different stages of my dissertation.

ABSTRACT

The big-data analysis of multi-omics data associated with maize genomes is increasingly utilized to accelerate genetic research and improve agronomic traits. As a result, efforts have increased to integrate diverse datasets and extract meaning from these measurements. For my Ph.D. dissertation, I have evaluated the current pitfalls of multi-omics data integration and analysis and built platforms that automatically analyze these omics' datasets. One such platform is qTeller, now designed to handle pan-genome level transcriptomics and proteomics datasets and extract meaningful interpretation from them by providing an interactive user interface. Although genomics and transcriptomics have been more extensively used, other omics technologies, such as epigenomics, variomics, and proteomics, are now often incorporated into standard research methodologies.

Therefore, I designed a fully automated platform, called Maize Feature Store (MFS), that allows the integration of complex omics to construct models that can be used to predict complex gene traits or annotations. To demonstrate the utility of the MFS, I critically discussed the application of MFS in pan-genome analysis using only a single maize genome (B73v5) as a multi-omics utility case study. I also aim to utilize these large-scale omics data to solve several other complex biological problems associated with the maize genome and phenome. I aim to continue improving the tools and assisting users in implementing them.

CHAPTER 1. GENERAL INTRODUCTION

In recent years, the field of plant biology has been revolutionized by the advent of omics data and computational approaches (Dai and Shen 2022; Mahmood et al. 2022; Scossa, Alseekh, and Fernie 2021; Van Emon 2016). The Human Genome Project and the subsequent advent of next-generation sequencing (NGS) technologies have been the catalysts for an explosion of genomics-related information that is having a profound effect on plant research (Koboldt et al. 2013; Ray and Satya 2014; Satam et al. 2023).

Multiple "omics" approaches have emerged as successful technologies for plant systems over the last few decades. Omics data refers to large-scale datasets that capture information about various biological molecules and processes, such as genomics, proteomics, metabolomics, and transcriptomics (Perez-Riverol et al. 2019). Multi-omics approaches with high throughput techniques provide a comprehensive view of the molecular components and interactions within a plant, enabling researchers to gain insights into the complex biological processes underlying growth, senescence, yield, and the responses to biotic and abiotic stress in numerous crops. These omics approaches have been implemented in some important crops including wheat (Triticum aestivum L.), soybean (Glycine max), tomato (Solanum lycopersicum), barley (Hordeum vulgare L.), maize (Zea mays L.), millet (Setaria italica L.), cotton (Gossypium hirsutum L.), Medicago truncatula, and rice (Oryza sativa L.) (Yang et al. 2021; Yang et al. 2023). The integration of functional genomics with other omics highlights the relationships between crop genomes and phenotypes under specific physiological and environmental conditions.

The need for omics data in current plant breeding and agriculture is paramount. With the growing global population and the increasing demand for food, there is a pressing need to

develop crops that are more resilient, productive, and nutritious. Omics data can provide valuable information about the genetic makeup of plants, allowing breeders to identify desirable traits and develop improved varieties through targeted breeding programs (AbuQamar, Moustafa, and Tran 2016). Additionally, omics data can help in understanding the molecular mechanisms underlying plant responses to biotic and abiotic stresses, such as diseases, pests, drought, and temperature extremes. This knowledge can inform the development of strategies to enhance plant resilience and productivity in the face of changing environmental conditions (Zogli et al. 2020).

The integration and application of omics data have become essential for solving complex biological problems in plants. The combination of these approaches has allowed significant steps forward in all phases of the breeding process, from the discovery of novel genetic variation to more extensive and detailed phenotyping, until the elucidation (and introgression) of a myriad of growth-related, life-history, stress resistance and metabolic traits (Sreeman et al. 2018). By combining multiple omics datasets, researchers can gain a more comprehensive understanding of the molecular networks and regulatory mechanisms that govern plant biology, identify key genes, proteins, and metabolic pathways that are involved in specific biological processes or traits of interest (Qi et al. 2023). Computational approaches, such as network analysis and machine learning algorithms, can be applied to omics data to uncover hidden patterns and predict gene functions (Picard et al. 2021). These computational tools facilitate the annotation of genes and the deciphering of their functions, which is crucial for advancing our understanding of plant biology and for guiding crop improvement efforts.

1.1 Different facets of plant omics

Among the various types of omics data, genomics, transcriptomics, proteomics and epigenomics have emerged as major focuses of research in plant biology (Chao et al. 2023).

Therefore, in this dissertation the primary focus has been on these four major types of omics datasets.

1.1.1 Plant genomics:

Plant genomes exhibit remarkable features such as genome size variation, gene content conservation, and the presence of repetitive sequences (Pellicer et al. 2018). The variation in genome size is primarily attributed to the accumulation of repetitive DNA sequences and polyploidy events. By unraveling the complexities of plant genomes, researchers can gain insights into the evolution and biology of different plant species (Li, Jain, et al. 2017; You 2023). The advent of Next-Generation Sequencing (NGS) techniques has revolutionized the study of plant genomes (Nguyen et al. 2019). These techniques have made it possible to sequence, assemble, and analyze the genomes of numerous plant species (Dmitriev, Pushkova, and Melnikova 2022).

The release of the first plant genome sequence, belonging to Arabidopsis, in 2000, marked a significant milestone in our understanding of plant genomics (Initiative, 2000). This breakthrough provided new insights and perspectives into the field (Initiative, 2000). Since then, rapid progress has been made in plant genomics, with the sequencing of not only model organisms but also a wide variety of species of ecological, agricultural, or economic importance (Song et al. 2023). This has resulted in the generation of a vast amount of genomic data. To make these data accessible to the scientific community, various web portals have been established, such as the Ensembl Plants portal and the NCBI genome portal (Cunningham et al. 2022). These portals provide researchers with easy access to the publicly available plant genomic data.

The availability of plant genomic data has opened new avenues for research in plant biology. For example, the sequencing of the tomato genome has provided insights into the

evolution of fleshy fruits. By comparing the genome sequences of domesticated tomato and its closest wild relative, Solanum pimpinellifolium, researchers have gained a better understanding of the genetic basis of fruit development (Tomato Genome 2012). Similarly, the sequencing of the Brachypodium genome has made it a valuable model system for studying plant biology. The genome sequence of Brachypodium distachyon, the flagship species of the genus, has led to significant advancements in our understanding of plant chromosomes and biology (Scholthof et al. 2018).

Plant genomics has become a vital component of understanding gene functions, and developing techniques like gene-targeted mutational forward genetics, sequence-based markers, and microarray platforms for gene expression studies (You 2023). These tools are instrumental in molecular breeding and the identification of economically important genes, addressing the challenges of providing food, fiber, and fuel for the growing global population.

1.1.2 Plant transcriptomics:

Plant transcriptomics is a field of study that focuses on understanding the dynamic changes in the transcriptome of plants, which play a crucial role in their inherent adaptive potential (Tyagi et al. 2022).As sessile organisms, plants have evolved mechanisms to respond and adapt to various developmental and environmental signals. These responses are reflected in the dynamic nature of the transcriptome, which represents the collective expression of genes in a given cell or tissue (Cha, Yang, and Lee 2022). The extensive gene duplication events in plant genomes have contributed to the formation of large gene families, allowing for sub-functionalization and the creation of specialized networks (Panchy, Lehti-Shiu, and Shiu 2016). The differential regulation of paralogs and their interactions across the genome contribute to the diverse transcriptome configurations that define different adaptive responses in plants.

Understanding the dynamics of the plant transcriptome is crucial for unraveling the molecular mechanisms underlying plant adaptation (Lian et al. 2020).

Over the past three decades, the field of plant transcriptomics has undergone significant advancements in technology and approaches used for profiling the transcriptome. These advancements have allowed researchers to gain a deeper understanding of the dynamic changes in gene expression in plants. Initially, semi-global clone-by-clone sequencing of expressed sequence tags (ESTs) provided insights into specific genes, followed by global hybridization-based profiling using microarrays and first-generation global sequencing-based platforms like massively parallel signature sequencing (MPSS) (Wang and Chekanova 2019). However, the most recent and transformative innovation in plant transcriptomics has been the application of next-generation sequencing (NGS) technology. This paradigm shift has enabled an even more comprehensive scope of profiling the spatio-temporal transcriptome fluxes by directly sampling and deep-sequencing transcripts, a feat not possible with earlier technologies (Wang, Gerstein, and Snyder 2009).

This technology, known as RNA-seq, has revolutionized the field by allowing researchers to obtain a more comprehensive and detailed view of gene expression patterns in plants. It has provided researchers with a universal tool for profiling gene expression in different tissues, developmental stages, and under various environmental conditions. By analyzing the transcriptome, researchers can identify key genes involved in specific biological pathways, unravel regulatory networks, and gain insights into plant responses to different stimuli.

1.1.3 Plant proteomics:

Proteomics offers direct insight into cellular functions by analyzing the proteome—the entire complement of proteins expressed by a cell, tissue, or organism—thereby revealing the network of molecular interactions that make up biological systems. Unlike genetic codes or

messenger molecules, proteomics allows researchers to investigate the actual proteins present in cells and understand their roles in various biological processes. In the context of plant biology, knowledge of plant proteins and their dynamics in response to environmental and biological stressors can have a significant impact on improving crop yield and nutritional properties. Functional genomic tools, including proteomics, have become indispensable in plant research (Eldakak et al. 2013). Proteomics is routinely employed to comprehensively profile complex protein extracts from plant organisms, providing valuable qualitative and quantitative information on protein dynamics. By studying the proteome, researchers can gain insights into the abundance, modifications, interactions, and functions of proteins in plants. The application of proteomics in plant research has led to significant advancements in understanding plant biology and addressing agricultural challenges (Liu, Lu, et al. 2019). Proteomic studies have shed light on the identification and characterization of key proteins involved in various biological processes, such as photosynthesis, metabolism, stress responses, and signal transduction pathways (Zhou et al. 2022). This knowledge has the potential to enhance our understanding of plant physiology and improve crop traits.

1.1.4 Plant epigenomics:

Plant epigenomics is a rapidly advancing field that focuses on the study of epigenetic modifications and their impact on gene expression and plant development. Epigenetic modifications, for example the tri-methylation of lysine 27 on histone H3 protein (H3K27me3), play a crucial role in regulating tissue-specific expression patterns and determining cell fate in plants (Zhao et al. 2020). The development of advanced technologies has enabled high-resolution mapping of plant epigenomes, providing valuable insights into the distribution and dynamics of epigenetic modifications. Understanding plant epigenomics, the study of epigenetic modifications and their impact on gene expression and plant biology, has important implications

for various aspects of plant research. Epigenetic modifications can contribute to phenotypic variation and adaptation in plants, influencing traits such as stress tolerance, growth, and development. Epigenomic variation can also influence the heritability of adaptive traits, although the extent of this influence is still not fully understood (Dar et al. 2022).

Epigenomic studies have revealed the role of epigenetic modifications in plant responses to environmental stimuli, including stressors such as biotic and abiotic factors. These modifications can regulate the expression of genes involved in stress responses and adaptation, providing insights into the molecular mechanisms underlying plant resilience (Rajpal et al. 2022). Furthermore, epigenomic variation has been observed in natural populations and non-model plant species, highlighting the importance of epigenetic modifications in adaptation and evolution. Epigenetic changes can contribute to phenotypic variation and influence fitness, potentially subjecting them to natural selection (Ashe, Colot, and Oldroyd 2021).

1.2 Challenges in omics data integration

In the era of Big Data, massive waves of ‘omics’ data have revolutionized the way we do science. These multi-dimensional large data sets, termed as ‘Big Data’- constitute a collection of huge structured and unstructured data sets. Complete extraction of information from such huge raw data stimulates scientific inventions in the field of precision agriculture and crop breeding (Popescu, Noutsos, and Popescu 2016). Plant science researchers are no longer analyzing one data set at a time but are moving towards multi-disciplinary integrative biology. It has been demonstrated that integration of different ‘omics’ data types (such as on genomes, transcriptomes, proteomes, epigenomes, etc..), boosts biological discoveries and improves predictions of the underlying interactions and regulation among molecular entities. Integrating different ‘omics’ datasets is a challenging task that relies heavily on data mining and machine learning algorithms (Flores et al. 2023). One must account for the specificities of each data type,

solve problems associated with processing data across different platforms, and consider the variable reliability levels of heterogeneous data.

Heaviness of such data sets makes it very complex to connect and correlate relationships among them, maintaining the hierarchies and multiple data linkage. It is very complex to manage large volumes of multidimensional raw data smoothly. Since ‘Big Data’ includes structured and unstructured data sets, it is very difficult to store, transfer, process, search the raw data and it cannot be managed using conventional database systems and software tools. The heterogeneous ‘Big data’ set may contain petabytes or exabytes of raw data consisting of billions to trillions of archives (Xia, Wang, and Niu 2020).

Integrating heterogeneous multi-omics data presents a cascade of challenges involving the unique data scaling, normalization, and transformation requirements of each individual dataset. Any effective integration strategy will also have to account for the regulatory relationships between datasets from different omics layers in order to accurately and holistically reflect the nature of this multidimensional data.

Furthermore, there is the issue of integrating omics and non-omics (OnO) data, like numerical or imaging data, for example, in order to enhance analytical productivity and to access richer insights into biological events and processes (Lopez de Maturana et al. 2019). Currently, the large-scale integration of non-omics data with high-throughput omics data is extremely limited due to a range of factors, including heterogeneity and the presence of sub phenotypes, for instance. The crux of the matter is that without effective and efficient data integration, multi-omics analysis will only tend to become more complex and resource-intensive without any proportional or even significant augmentation in productivity, performance, or insight generation (Subramanian et al. 2020).

1.3 Maize as model organism for plant omics research

Maize (*Zea mays*) has emerged as a valuable model organism for plant multi-omics research due to its genetic architecture, complex responses to abiotic stress, and its significance as a staple crop (Farooqi et al. 2022). As a model organism, maize provides researchers with a unique opportunity to study fundamental aspects of plant biology, including genetic inheritance, genomic properties, domestication, epigenetics, evolution, and chromosome structure (Hake and Ross-Ibarra 2015). Additionally, maize serves as a model for investigating complex traits such as hybrid vigor and quantitative trait loci (Wallace, Larsson, and Buckler 2014). An integrated developmental atlas of the transcriptome, proteome, and phosphoproteome of maize has been generated, highlighting the use of maize as a model organism for multiomics research. This atlas has allowed for the construction of transcriptome- and proteome-based networks, providing a comprehensive understanding of maize development (Walley et al. 2016).

The use of multi-omics approaches in maize research has greatly enhanced our understanding of various aspects of its biology. These approaches, which integrate genomics, transcriptomics, proteomics, metabolomics, and phenomics, have provided insights into maize crop growth, senescence, yield, and responses to biotic and abiotic stresses (Zenda et al. 2021). By employing multi-omics technologies, researchers can generate multi-layered information that allows for a comprehensive understanding of the interactions between maize and its environment.

Furthermore, the application of multi-omics approaches in maize research has facilitated the development of breeding strategies for abiotic stress tolerance. By studying the metabolic responses of maize to stressors such as drought, heat, and nutrient deficiencies, researchers can identify key genes and pathways involved in stress tolerance and develop improved maize varieties (Roychowdhury et al. 2023). Maize's status as a model organism also extends to its role

in advancing plant systems biology. The integration of multi-omics data in maize research has led to the development of computational tools and methodologies for data analysis and mining (Pazhamala et al. 2021). These tools enable researchers to extract meaningful insights from large-scale multi-omics datasets and uncover hidden patterns and relationships within the data.

1.4 Computational approaches for multi-omics data integration and analysis

In recent years, there has been a significant advancement in high-throughput omics (HTO) technologies, such as genomics, transcriptomics, proteomics, metabolomics, and phenomics which have revolutionized the field of crop science (Shaw et al. 2021). These omics techniques have provided researchers with a wealth of data, but the integration, visualization, and analysis of multi-omics data pose significant challenges due to their heterogeneity and non-structured nature. To address these challenges, various computational approaches and tools have been developed, including bioinformatics resources, software packages, and databases (Krassowski et al. 2020). These resources have become indispensable for data production, mining, integration, and extraction of valuable information. The integration of omics resources in crop breeding holds great promise for the development of better-designed crops. However, to fully harness the potential of multi-omics data, innovative analytical approaches are required.

Machine learning has emerged as an effective solution for large-scale data analysis in plant biology (Liakos et al. 2018). Machine learning algorithms can handle the complexity and volume of multi-omics data by creating data compatible with parallel computing infrastructures. This enables researchers to extract meaningful patterns and insights from the data. By leveraging machine learning algorithms, researchers can effectively integrate and interpret multi-omics platforms, facilitating the study of plant molecular omics interactions (Cembrowska-Lech et al. 2023). Machine learning offers promising computational and analytical solutions for the

integrative analysis of large, heterogeneous, and unstructured datasets on the Big-Data scale and is gradually gaining popularity in plant biology.

Despite the progress made so far, there are still several challenges that need to be addressed. For instance, the integration of the entire omics by employing the 'phenotype to genotype' and 'genotype to phenotype' concept is yet to be fully realized (Hardiman 2020). Additionally, there is a lack of minimum information standards for multi-omics experiments. Currently, such standards are only available for single omics datasets, which hinders the comparability and reproducibility of multi-omics studies. Therefore, it is essential to develop novel hosting options that can accommodate the multi-platform and multi-layered nature of omics data (Conesa and Beck 2019). These hosting options should provide a unified framework for integrating and analyzing multiple omics data types, thereby facilitating better accessibility and collaboration among the genomics community. The purpose of this dissertation is to address these challenges and propose innovative approaches for preprocessing, quality control, hosting, and access of multiomics datasets. By embracing the nature of multi-platform and multi-layered omics data, this research aims to create frameworks that can effectively integrate and analyze various omics data types. These frameworks will not only enhance the accessibility of multiomics datasets but also enable researchers to gain deeper insights into the complex relationships between genotype and phenotype.

1.5 Dissertation Organization

Chapter 1 serves as a general introduction to multi-omics data and computational techniques related to the work in this dissertation. It provides an overview of plant omics and big data in the fields of plant and crop biology, discusses each omics layer individually, including genomics, transcriptomics, proteomics and epigenomics, highlights the challenges associated with handling and integration of omics dataset, covers the importance of maize as model species

to understand the plant multi-omics biology, introduces biological resources, datasets, online bioinformatics tools and machine learning approaches that are in the public domain.

Chapter 2 consists of a published manuscript titled “qTeller: a tool for comparative multi-genomic gene expression analysis” published in Bioinformatics (Woodhouse, Sen, et al. 2021b). In this report, we performed gene-level comparative analysis of gene expressions across many different conditions, tissues using RNA-Seq datasets from the 26 maize genomes. The detailed workflow of qTeller is outlined in (Fig 1.1). We developed an easy-to-use web application for in-depth analysis of gene expression data. With this MaizeGDB instance of qTeller we encompassed expression dataset generated from 200 different RNA-seq mapping pipelines, a modernized interface and back-end database and an optimized framework for adoption by other organisms’ databases. The focus of this work is to make data generated from individual genomic analysis accessible and reusable at a gene-level scale and allow for comparative analysis between genes, across different genomes and meta-analyses.

Chapter 3 consists of a published manuscript titled “Maize Feature Store (MFS): A centralized resource to manage and analyze curated maize multi-omics features for machine learning applications” published in Oxford Database (Sen et al. 2023). In this work, we introduce a framework that hosts gene-based machine learning features built on multi-omics data to facilitate the exploration and modeling of classification problems. We populated an instance of this framework at MaizeGDB, called the Maize Feature Store (MFS), with over 14,000 gene-based features based on published genomic, transcriptomic, epigenomic, variomic, and proteomics data sets. In addition, we also integrated supervised and unsupervised machine-learning algorithms that can significantly simplify the analysis and prediction of complex

genome annotations. Last, we demonstrated the tool's utility by achieving high classification accuracy for distinguishing core and non-core genes in the maize pan-genome (Fig 1.2).

Chapter 4 consists of a manuscript named “Predicting genes associated with biotic or abiotic stress across different maize lines and related species”. In this report, we created a computational model that predicted stress-responsive genes associated with abiotic and biotic stresses in Maize (Zea Mays) as well as across other maize genome assemblies (e.g., maize inbred lines W22 and Zea mays ssp. mexicana L (TIL-18, TIL-25)) by performing meta-analyses of a comprehensive set of multi-omics datasets. The workflow provided a framework that yielded insight into the possible characteristics of specific genes and the role they play in response to different environmental stimuli (Fig 1.3).

Chapter 5 consists of the conclusions of the thesis and discusses the future directions of the work.

1.6 Main Figures

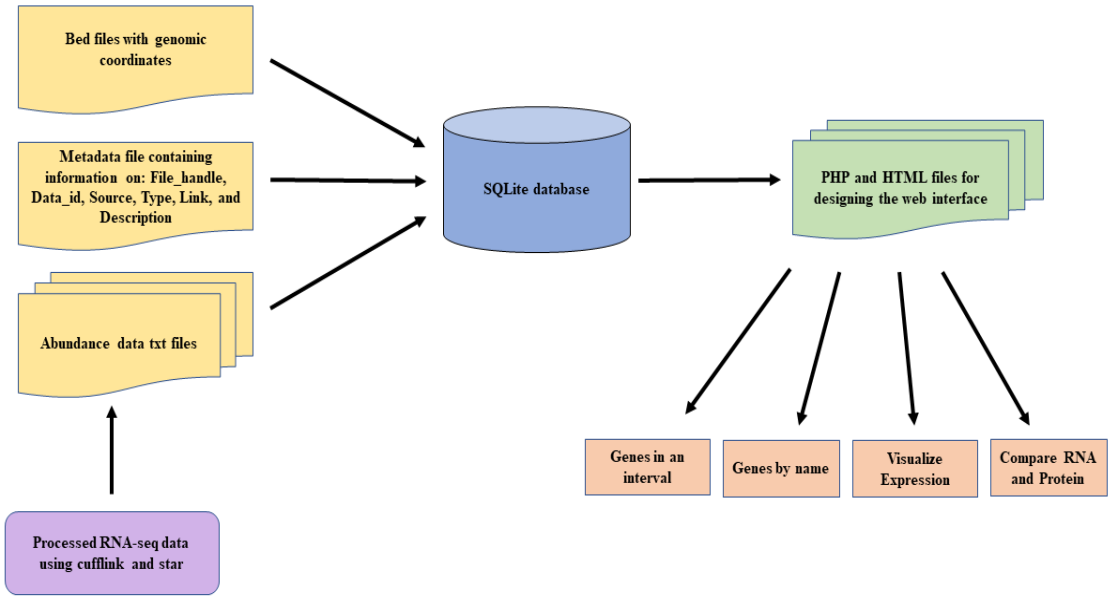

Figure 1.1 Graphical workflow of the interactive qTeller comparative RNA-seq expression platform.

Overall graphical workflow of the interactive qTeller comparative RNA-seq expression platform. qTeller assembles preprocessed RNA-seq expression datasets from varied sources, collected from multiple tissues and conditions. The assembling requires three inputs: a gff or bed file of the gene models of interest; a directory containing files with RNA-Seq and/or protein abundances by experiment; and a metadata file structured. The assembled data is later stored in a SQLite database and is presented interactively via multiple HTML and PHP files.

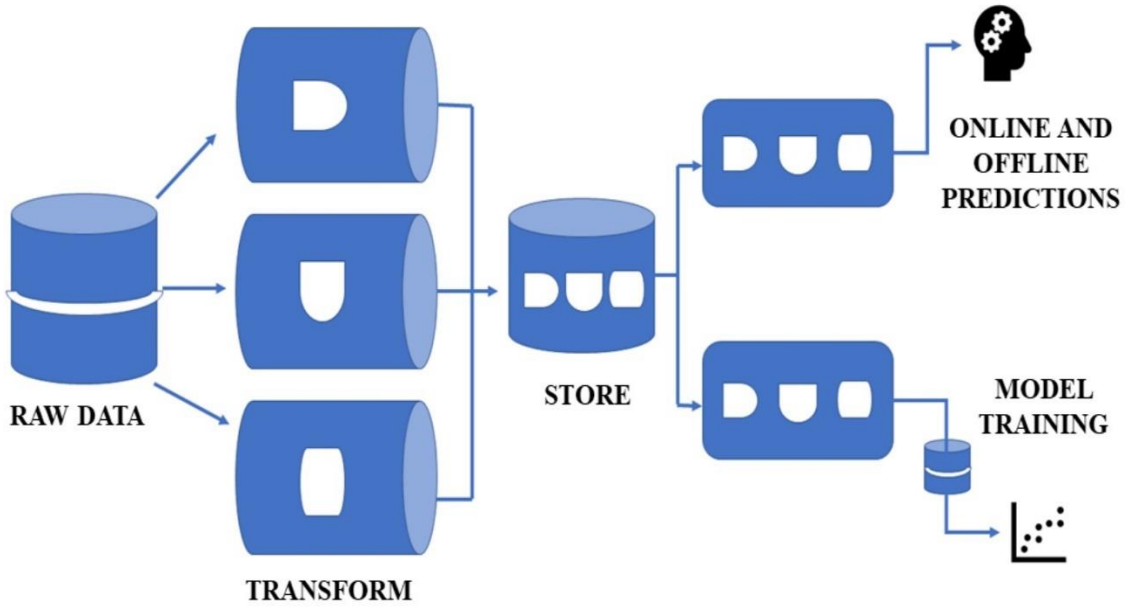

Figure 1.2 Maize Feature Store Data Flow: a central place to transform, store, and serve raw data for both online and offline predictions, model training, and exploratory analyses.

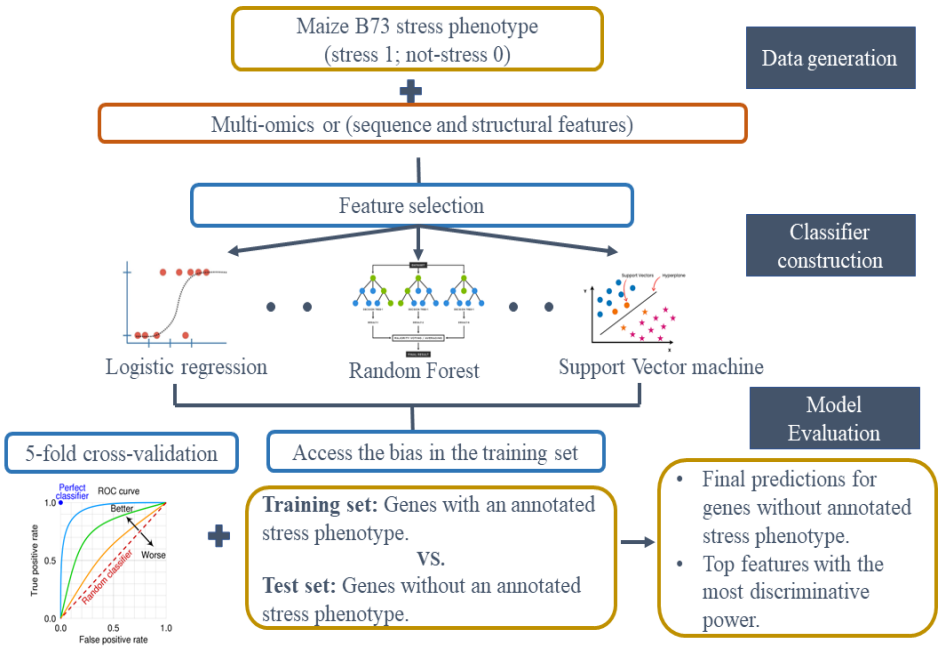

Figure 1.3 Computational model to predict stress-responsive genes associated with abiotic and biotic stresses in maize as well as across other maize lines.

1.7 References

AbuQamar, S. F., K. Moustafa, and L. S. Tran. 2016. "Omics' and Plant Responses to Botrytis cinerea', *Front Plant Sci*, 7: 1658.

Ashe, A., V. Colot, and B. P. Oldroyd. 2021. 'How does epigenetics influence the course of evolution?', *Philos Trans R Soc Lond B Biol Sci*, 376: 20200111.

Cembrowska-Lech, D., A. Krzeminska, T. Miller, A. Nowakowska, C. Adamski, M. Radaczynska, G. Mikiciuk, and M. Mikiciuk. 2023. 'An Integrated Multi-Omics and Artificial Intelligence Framework for Advance Plant Phenotyping in Horticulture', *Biology (Basel)*, 12.

Cha, O. K., S. Yang, and H. Lee. 2022. 'Transcriptomics Using the Enriched Arabidopsis Shoot Apex Reveals Developmental Priming Genes Involved in Plastic Plant Growth under Salt Stress Conditions', *Plants (Basel)*, 11.

Chao, H., S. Zhang, Y. Hu, Q. Ni, S. Xin, L. Zhao, V. A. Ivanisenko, Y. L. Orlov, and M. Chen. 2023. 'Integrating omics databases for enhanced crop breeding', *J Integr Bioinform*.

Conesa, A., and S. Beck. 2019. 'Making multi-omics data accessible to researchers', *Sci Data*, 6: 251.

Cunningham, F., J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugan, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Pilizota, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Perez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M. M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gomez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, I. Isley GR, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, and P. Flicek. 2022. 'Ensembl 2022', *Nucleic Acids Res*, 50: D988-D95.

Dai, X., and L. Shen. 2022. 'Advances and Trends in Omics Technology Development', *Front Med (Lausanne)*, 9: 911861.

Dar, F. A., N. U. Mushtaq, S. Saleem, R. U. Rehman, T. U. H. Dar, and K. R. Hakeem. 2022. 'Role of Epigenetics in Modulating Phenotypic Plasticity against Abiotic Stresses in Plants', *Int J Genomics*, 2022: 1092894.

Dmitriev, A. A., E. N. Pushkova, and N. V. Melnikova. 2022. '[Plant Genome Sequencing: Modern Technologies and Novel Opportunities for Breeding]', *Mol Biol (Mosk)*, 56: 531-45.

Eldakak, M., S. I. Milad, A. I. Nawar, and J. S. Rohila. 2013. 'Proteomics: a biotechnology tool for crop improvement', *Front Plant Sci*, 4: 35.

Farooqi, M. Q. U., G. Nawaz, S. H. Wani, J. R. Choudhary, M. Rana, R. P. Sah, M. Afzal, Z. Zahra, S. A. Ganie, A. Razzaq, V. P. Reyes, E. A. Mahmoud, H. O. Elansary, T. K. Z. El-Abedin, and K. H. M. Siddique. 2022. 'Recent developments in multi-omics and breeding strategies for abiotic stress tolerance in maize (Zea mays L.)', *Front Plant Sci*, 13: 965878.

Flores, J. E., D. M. Claborne, Z. D. Weller, B. M. Webb-Robertson, K. M. Waters, and L. M. Bramer. 2023. 'Missing data in multi-omics integration: Recent advances through artificial intelligence', *Front Artif Intell*, 6: 1098308.

Hake, S., and J. Ross-Ibarra. 2015. 'Genetic, evolutionary and plant breeding insights from the domestication of maize', *Elife*, 4.

Hardiman, G. 2020. 'An Introduction to Systems Analytics and Integration of Big Omics Data', *Genes (Basel)*, 11.

Koboldt, D. C., K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis. 2013. 'The next-generation sequencing revolution and its impact on genomics', *Cell*, 155: 27-38.

Krassowski, M., V. Das, S. K. Sahu, and B. B. Misra. 2020. 'State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing', *Front Genet*, 11: 610798.

Li, G., R. Jain, M. Chern, N. T. Pham, J. A. Martin, T. Wei, W. S. Schackwitz, A. M. Lipzen, P. Q. Duong, K. C. Jones, L. Jiang, D. Ruan, D. Bauer, Y. Peng, K. W. Barry, J. Schmutz, and P. C. Ronald. 2017. 'The Sequences of 1504 Mutants in the Model Rice Variety Kitaake Facilitate Rapid Functional Genomic Studies', *Plant Cell*, 29: 1218-31.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. 'Machine Learning in Agriculture: A Review', *Sensors (Basel)*, 18.

Lian, S., Y. Zhou, Z. Liu, A. Gong, and L. Cheng. 2020. 'The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences', *BMC Plant Biol*, 20: 277.

Liu, Y., S. Lu, K. Liu, S. Wang, L. Huang, and L. Guo. 2019. 'Proteomics: a powerful tool to study plant responses to biotic stress', *Plant Methods*, 15: 135.

Lopez de Maturana, E., L. Alonso, P. Alarcon, I. A. Martin-Antoniano, S. Pineda, L. Piorno, M. L. Calle, and N. Malats. 2019. 'Challenges in the Integration of Omics and Non-Omics Data', *Genes (Basel)*, 10.

Mahmood, U., X. Li, Y. Fan, W. Chang, Y. Niu, J. Li, C. Qu, and K. Lu. 2022. 'Multi-omics revolution to promote plant breeding efficiency', *Front Plant Sci*, 13: 1062952.

Nguyen, K. L., A. Grondin, B. Courtois, and P. Gantet. 2019. 'Next-Generation Sequencing Accelerates Crop Gene Discovery', *Trends Plant Sci*, 24: 263-74.

Panchy, N., M. Lehti-Shiu, and S. H. Shiu. 2016. 'Evolution of Gene Duplication in Plants', *Plant Physiol*, 171: 2294-316.

Pazhamala, L. T., H. Kudapa, W. Weckwerth, A. H. Millar, and R. K. Varshney. 2021. 'Systems biology for crop improvement', *Plant Genome*, 14: e20098.

Pellicer, J., O. Hidalgo, S. Dodsworth, and I. J. Leitch. 2018. 'Genome Size Diversity and Its Impact on the Evolution of Land Plants', *Genes (Basel)*, 9.

Perez-Riverol, Y., A. Zorin, G. Dass, M. T. Vu, P. Xu, M. Glont, J. A. Vizcaino, A. F. Jarnuczak, R. Petryszak, P. Ping, and H. Hermjakob. 2019. 'Quantifying the impact of public omics data', *Nat Commun*, 10: 3512.

Picard, M., M. P. Scott-Boyer, A. Bodein, O. Perin, and A. Droit. 2021. 'Integration strategies of multi-omics data for machine learning analysis', *Comput Struct Biotechnol J*, 19: 3735-46.

Popescu, G. V., C. Noutsos, and S. C. Popescu. 2016. 'Big Data in Plant Science: Resources and Data Mining Tools for Plant Genomics and Proteomics', *Methods Mol Biol*, 1415: 533-47.

Qi, S., J. Wang, Y. Zhang, M. Naz, M. R. Afzal, D. Du, and Z. Dai. 2023. 'Omics Approaches in Invasion Biology: Understanding Mechanisms and Impacts on Ecological Health', *Plants (Basel)*, 12.

Rajpal, V. R., P. Rathore, S. Mehta, N. Wadhwa, P. Yadav, E. Berry, S. Goel, V. Bhat, and S. N. Raina. 2022. 'Epigenetic variation: A major player in facilitating plant fitness under changing environmental conditions', *Front Cell Dev Biol*, 10: 1020958.

Ray, S., and P. Satya. 2014. 'Next generation sequencing technologies for next generation plant breeding', *Front Plant Sci*, 5: 367.

Roychowdhury, R., S. P. Das, A. Gupta, P. Parihar, K. Chandrasekhar, U. Sarker, A. Kumar, D. P. Ramrao, and C. Sudhakar. 2023. 'Multi-Omics Pipeline and Omics-Integration Approach to Decipher Plant's Abiotic Stress Tolerance Responses', *Genes (Basel)*, 14.

Satam, H., K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das, and S. K. Malonia. 2023. 'Next-Generation Sequencing Technology: Current Trends and Advancements', *Biology (Basel)*, 12.

Scholthof, K. B. G., S. Irigoyen, P. Catalan, and K. K. Mandadi. 2018. 'Brachypodium: A Monocot Grass Model Genus for Plant Biology', *Plant Cell*, 30: 1673-94.

Scossa, F., S. Alseekh, and A. R. Fernie. 2021. 'Integrating multi-omics data for crop improvement', *J Plant Physiol*, 257: 153352.

Sen, S., M. R. Woodhouse, J. L. Portwood, 2nd, and C. M. Andorf. 2023. 'Maize Feature Store: A centralized resource to manage and analyze curated maize multi-omics features for machine learning applications', *Database (Oxford)*, 2023.

Shaw, R. K., Y. Shen, J. Wang, X. Sheng, Z. Zhao, H. Yu, and H. Gu. 2021. 'Advances in Multi-Omics Approaches for Molecular Breeding of Black Rot Resistance in Brassica oleracea L', *Front Plant Sci*, 12: 742553.

Song, B., W. Ning, D. Wei, M. Jiang, K. Zhu, X. Wang, D. Edwards, D. A. Odeny, and S. Cheng. 2023. 'Plant genome resequencing and population genomics: Current status and future prospects', *Mol Plant*, 16: 1252-68.

Sreeman, S. M., P. Vijayaraghavareddy, R. Sreevathsa, S. Rajendrareddy, S. Arakesh, P. Bharti, P. Dharmappa, and R. Soolanayakanahally. 2018. 'Introgression of Physiological Traits for a Comprehensive Improvement of Drought Adaptation in Crop Plants', *Front Chem*, 6: 92.

Subramanian, I., S. Verma, S. Kumar, A. Jere, and K. Anamika. 2020. 'Multi-omics Data Integration, Interpretation, and Its Application', *Bioinform Biol Insights*, 14: 1177932219899051.

Tomato Genome, Consortium. 2012. 'The tomato genome sequence provides insights into fleshy fruit evolution', *Nature*, 485: 635-41.

Tyagi, P., D. Singh, S. Mathur, A. Singh, and R. Ranjan. 2022. 'Upcoming progress of transcriptomics studies on plants: An overview', *Front Plant Sci*, 13: 1030890.

Van Emon, J. M. 2016. 'The Omics Revolution in Agricultural Research', *J Agric Food Chem*, 64: 36-44.

Wallace, J. G., S. J. Larsson, and E. S. Buckler. 2014. 'Entering the second century of maize quantitative genetics', *Heredity (Edinb)*, 112: 30-8.

Walley, J. W., R. C. Sartor, Z. Shen, R. J. Schmitz, K. J. Wu, M. A. Urich, J. R. Nery, L. G. Smith, J. C. Schnable, J. R. Ecker, and S. P. Briggs. 2016. 'Integration of omic networks in a developmental atlas of maize', *Science*, 353: 814-8.

Wang, H. V., and J. A. Chekanova. 2019. 'An Overview of Methodologies in Studying lncRNAs in the High-Throughput Era: When Acronyms ATTACK!', *Methods Mol Biol*, 1933: 1-30.

Wang, Z., M. Gerstein, and M. Snyder. 2009. 'RNA-Seq: a revolutionary tool for transcriptomics', *Nat Rev Genet*, 10: 57-63.

Woodhouse, M. R., S. Sen, D. Schott, J. L. Portwood, M. Freeling, J. W. Walley, C. M. Andorf, and J. C. Schnable. 2021. 'qTeller: a tool for comparative multi-genomic gene expression analysis', *Bioinformatics*, 38: 236-42.

Xia, J., J. Wang, and S. Niu. 2020. 'Research challenges and opportunities for using big data in global change biology', *Glob Chang Biol*, 26: 6040-61.

Yang, Y., M. A. Saand, L. Huang, W. B. Abdelaal, J. Zhang, Y. Wu, J. Li, M. H. Sirohi, and F. Wang. 2021. 'Applications of Multi-Omics Technologies for Crop Improvement', *Front Plant Sci*, 12: 563953.

Yang, Z., S. Wang, L. Wei, Y. Huang, D. Liu, Y. Jia, C. Luo, Y. Lin, C. Liang, Y. Hu, C. Dai, L. Guo, Y. Zhou, and Q. Y. Yang. 2023. 'BnIR: A multi-omics database with various tools for Brassica napus research and breeding', *Mol Plant*, 16: 775-89.

You, F. M. 2023. 'Plant Genomics-Advancing Our Understanding of Plants', *Int J Mol Sci*, 24.

Zenda, T., S. Liu, A. Dong, J. Li, Y. Wang, X. Liu, N. Wang, and H. Duan. 2021. 'Omics-Facilitated Crop Improvement for Climate Resilience and Superior Nutritive Value', *Front Plant Sci*, 12: 774994.

Zhao, N., K. Zhang, C. Wang, H. Yan, Y. Liu, W. Xu, and Z. Su. 2020. 'Systematic Analysis of Differential H3K27me3 and H3K4me3 Deposition in Callus and Seedling Reveals the Epigenetic Regulatory Mechanisms Involved in Callus Formation in Rice', *Front Genet*, 11: 766.

Zhou, M., S. Zhu, X. Mo, Q. Guo, Y. Li, J. Tian, and C. Liang. 2022. 'Proteomic Analysis Dissects Molecular Mechanisms Underlying Plant Responses to Phosphorus Deficiency', *Cells*, 11.

Zogli, P., L. Pingault, S. Grover, and J. Louis. 2020. 'Ento(o)mics: the intersection of 'omic' approaches to decipher plant defense against sap-sucking insect pests', *Curr Opin Plant Biol*, 56: 153-61.

CHAPTER 2. qTeller: A TOOL FOR COMPARATIVE MULTI-GENOMIC GENE EXPRESSION ANALYSIS

Margaret R. Woodhouse[1*], Shatabdi Sen[2*], David Schott[3] , John L. Portwood II[1] , Michael Freeling[4], Justin W. Walley[2], Carson M. Andorf[1,3] and James C. Schnable[5]

[1]USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

[2]Department of Plant Pathology & Microbiology, Iowa State University, Ames, IA 50011, USA

[3]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

[4]Department of Plant & Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[5]Center for Plant Science Innovation & Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

Modified from a manuscript published in *Bioinformatics.*

*Shatabdi Sen and Margaret Woodhouse are co-first authors on this publication.

2.1 Abstract

Motivation: Over the last decade, RNA-Seq whole-genome sequencing has become a widely used method for measuring and understanding transcriptome-level changes in gene expression. Since RNA-Seq is relatively inexpensive, it can be used on multiple genomes to evaluate gene expression across many different conditions, tissues, and cell types. Although many tools exist to map and compare RNA-Seq at the genomics level, few web-based tools are dedicated to making data generated for individual genomic analysis accessible and reusable at a gene-level scale for comparative analysis between genes, across different genomes and meta-analyses.

Results: To address this challenge, we revamped the comparative gene expression tool qTeller to take advantage of the growing number of public RNA-Seq datasets. qTeller allows

users to evaluate gene expression data in a defined genomic interval and also perform two-gene comparisons across multiple user-chosen tissues. Though previously unpublished, qTeller has been cited extensively in scientific literature, demonstrating its importance to researchers. Our new version of qTeller now supports multiple genomes for intergenomic comparisons and includes capabilities for both mRNA and protein abundance datasets. Other new features include support for additional data formats, modernized interface and back-end database and an optimized framework for adoption by other organisms' databases.

2.2 Introduction

Since the introduction of RNA-Seq technology over 10 years ago (Wang, Gerstein, and Snyder 2009), the number of available RNA-Seq libraries has increased rapidly (Fig. 2.1). Many software programs, mostly in R, such as EdgeR, ggplot2, WGCNA and DEvis (Langfelder and Horvath 2008; Price et al. 2019; Robinson, McCarthy, and Smyth 2010), have been created to visualize RNA-Seq abundances across different tissues and time points. However, there are few tools that allow users not trained in programming to visualize RNA-Seq expression patterns across multiple genes or genomic intervals, particularly in an interactive way or to compare any given two genes. In 2012, this need was addressed in the creation of qTeller, a web-hosted RNA-Seq visualization platform that allows users to compare RNA-Seq expression across tissues within a genomic interval, across multiple genes or compare expression between any two genes in a given genome (https://github.com/jschnable/qTeller). The platform displays preanalyzed values from publicly available, published datasets. At the time, qTeller hosted instances for Zea mays, Arabidopsis thaliana and Brassica rapa. Although unpublished, qTeller has been used by many researchers and cited extensively, including in the areas of evolution (Man, Gallagher, and Bartlett 2020; Pophaly and Tellier 2015; Wang et al. 2019; Woodhouse et al. 2014), metaanalysis (Hawkins et al. 2015; Jia et al. 2018; Zhang et al. 2018), gene and gene family

identification (Li et al. 2019; Liu, Qu, et al. 2019), quantitative trait and association studies (Liu et al. 2012; Wu, Li, et al. 2016), orthology (Sindhu et al. 2018) and general reviews (Liu, Fernie, and Yan 2020; Wang, Lu, and Deng 2016). qTeller's breadth of use demonstrates its value to the research community. In 2018, the Maize Genetics and Genomics Database (MaizeGDB) (Portwood et al. 2019) released its own version of qTeller for maize (https://qteller.maizegdb.org/). Since then, MaizeGDB has optimized qTeller to host information on protein abundance as well as mRNA abundance (i.e. RNA-Seq data), and to allow comparisons across gene models annotated in different reference genomes. By offering this tool as a web page to the maize community, MaizeGDB helps researchers to quickly compare precomputed gene expression abundances across maize genes and genomes. Here, we present a description of qTeller and its functionality, and how users can download and run the software themselves. The MaizeGDB qTeller is available at https://github.com/MaizeGenetics-and-Genomics-Database/qTeller.

2.3 Material and methods

2.3.1 qTeller basic functionality

There are three main sections of qTeller: Section 2.3.1.1, Section 2.3.1.2 and Section 2.3.1.3. Each section is accessed through a drop-down menu on the web page and presents gene expression information in a different way to meet the needs of users with distinct use cases. Users may investigate expression within one genome, across multiple genomes or compare RNA expression to protein abundances. The three sections are described below.

2.3.1.1 Genes in an interval

Section 2.3.1.1 allows users to select a chromosome and coordinate interval of interest for a given genome (only one genome at a time can be selected) (Fig. 2.2). The primary use case for this section of qTeller is when a researcher has mapped a gene or QTL to a defined interval in

the genome and wishes to use gene expression data within this interval to prioritize among the potential candidate genes within the interval. The interface can be set up for a single reference genome or to have a selectable list of genomes from a set. Next, the user selects the RNA-Seq libraries of interest below the genomic coordinate selection, selects all RNA-Seq libraries under a given set, or selects all RNA-Seq libraries in the database. qTeller then returns the RNA-Seq abundances of all the selected libraries of all the genes within the selected interval for that given genome. A user has the option of selecting 'All Chromosomes' in the dropdown menu and leaving the coordinate boxes blank to return the RNA-Seq abundances for all genes in the genome (excluding unplaced scaffolds). The output is in the format of a table that includes gene model ID, RNA-Seq abundances for each selected library, and a link to visualize the data as a bar chart for every gene model (see Section 2.3.1.3). A user has the option of either viewing the table as a web page or downloading the table as a .csv file. genome start and end position, and check boxes for each of the RNA-Seq experiments (organized by project or paper). Each set of RNA-Seq experiments has an 'All on' and 'All off' option.

2.3.1.2 Genes by name

The Section 2.3.1.2 is similar to Section 2.3.1.1 except that it allows a user to paste a list of gene models of interest instead of selecting genomic coordinates (Fig. 2.3). The use cases for this section of qTeller include users with a set of genes of interest identified via other means (e.g. a set of GWAS hits or a cluster of genes linked by protein interaction data). For a multi-genome instance, a mix of gene models across different genomes is permitted, e.g. allowing users to compare expression from gene models across multiple genomes of a pangenome. Library selection and output tables are the same as for Section 2.3.1.1.

2.3.1.3 Visualize expression

The Section 2.3.1.3 tool draws a bar chart for all libraries for a given input gene model or draws a dot plot of all libraries between two gene model inputs (Fig. 2.4). The latter case is useful for comparing relative expressions between two gene models. The dot plot feature for the multi-genome qTeller instance allows a user to input two gene models from any genome. The use cases for this section of qTeller include comparisons of duplicated genes to identify potential evidence of regulatory subfunctionalization, or comparison of patterns of expression of equivalent gene models between different genetic backgrounds/genomics to identify genotype-specific patterns of regulation as the result of cis- or trans-regulatory divergence. Advanced options for both the bar chart and dot plot allow users to select their libraries of interest instead of visualizing all libraries. Under each visualization output, qTeller generates a shareable link to recreate the bar chart or dot plot images if a user wants to share the data with others or use it in a publication. A user can mouse over the bars in the bar chart, or the dots in the dot plot, to get information about the abundances and how the data were generated experimentally. A user can also change the axes of the dot plot to zoom in on a region of interest for better resolution.

2.3.2 qTeller expanded functionality

2.3.2.1 Protein expression visualization

Gene expression is the measurement of how genes produce functional products used to carry out processes in a cell. There are two primary ways to measure gene expression: mRNA abundance using RNA-Seq, and protein abundance using mass spectrometry [reviewed in (Zhang et al. 2010)]. Gene expression at the mRNA abundance level can be only poorly predicted using data on gene expression at the protein abundance level and vice versa, and both types of data can be used to associate genes with functional characteristics. The functionality of qTeller was expanded to include protein expression as well as RNA-Seq data.

The ‘Compare RNA & Protein’ tool draws four different types of visualization, a bar chart and three different dot plots. The ‘Single-Gene Expression’ tool under ‘Visualize RNA versus Protein’ is similar to the single-genome Section 2.3.1.3 bar chart, except that the user can select either mRNA abundance (FPKM) or protein abundance (NSAF) for all selected libraries for a single input gene model. The first dot plot (Two-Gene Scatterplot, Fig. 2.5A) compares two gene model inputs using the same data type, either mRNA versus mRNA or protein versus protein. This dot plot is useful for comparing relative mRNA or protein abundance between two gene models. The dot plot feature for the multi-genome qTeller instance allows a user to input two gene models from any genome as long as the expression data from both genomes were generated by the same project with a consistent methodology. The second dot plot (‘Single Gene Expression versus Abundance’) is used to make a comparison between mRNA abundance and protein abundance for the single input gene model across different tissues. The third dot plot (‘MultiGene Expression versus Abundance in a Single Tissue’, Fig. 2.5B) is similar to the second dot plot except that it allows a user to select the tissue of interest and enter a list of gene models. This dot plot makes a direct comparison between mRNA and protein abundance for a fixed tissue and set of gene models. The latter two plots also provide a Pearson correlation coefficient that measures linear correlation between two variables and abundance types.

2.3.2.2 Multi-genome functionality

qTeller now offers a multi-genomic option when building and calling a database; this feature is useful for genomes and/or RNA-Seq datasets that were constructed using the same methods across genome assemblies (e.g. NAM founders in maize, doi:10.1101/2021.01.14.426684). This functionality is specific for multiple genomes within the same species, requiring that all genomes have the same number of chromosomes with the same chromosome ID designation (i.e. chr1, chr2, etc. or 1, 2, etc.). The main technical difference between qTeller

and multi-genome qTeller is that an input bed file is required which contains information about each genome ID. The bed file follows the typical structure of a normal bed file, with the gene model ID in Column 4 and the ID of the genome in Column 5 (see Supplementary Table S1). The genome ID can be any alphanumeric string. In order for experiments to be treated as paired data, and thus appropriate for between-genome dot plots and comparisons in multi-genome qTeller, they must be assigned exactly matching values in the 'data_id' column. For instance, if SRR12345 for Genome A is described as 'pollen tube' in the 'data_id' metadata column, then if Genome B's SRR23456 from the same experiment is also from pollen tube tissue, its 'data_id' must also be written as 'pollen tube' exactly, and have the same experiment Source, if the user wishes these experiments to be fetched together.

The biggest difference in the qTeller interface structure for multigenome is under Section 2.3.1.1, where users can select a genome of their choice from the drop-down menu at the top. This reflects a change in the database structure wherein each gene model in the multi-genome database is assigned a genome ID (see Software Usage below). Because Section 2.3.1.1 is based on a genomic coordinate system, more than one genome cannot be fetched at a time. However, under Section 2.3.1.2 or Section 2.3.1.3, gene model IDs from more than one genome can be fetched. Ideally, cross-genome RNA-Seq datasets should be matched with identical descriptions only when the RNA samples used for quantification were collected, processed, and sequenced by the same laboratory, to ensure that any differences observed in relative abundances between genomes are not due to differences in laboratory technique or environment.

2.3.2.3 Expanded qTeller navigation

The expanded qTeller software package includes a reformatted homepage and a menu header on each page for quick access to each of the four tools or links to general information

(contact, data sources and FAQs). Each tool menu item has a submenu listing which genomes are available. There is also a new 'News' item feature on the side of the page.

2.3.3 Basic qTeller software usage

qTeller was originally written in Python 2.7 (http://www.python. org), html and PHP5 for SQlite3 (https://www.sqlite.org/) software. We updated the Python code to Python 3 and ensured that the PHP scripts were PHP7 compatible. Images are drawn using Python Matplotlib (Hunter 2007). Python3 dependencies are listed in qteller_packagelist_python3.txt in our GitHub and can be installed using Python3 pip. The most basic form of qTeller requires only three inputs: a gff or bed file of the gene models of interest; a directory containing files with RNA-Seq and/or protein abundances by experiment; and a metadata file structured as described in Supplementary Table S2. There are two main directories: the build_db directory, where the database is built; and the web_interface directory, which contains the dynamic web pages.

qTeller's basic structure allows for most types of RNA-Seq mapping pipelines to be used, since qTeller accepts fpkm abundances calculated by Cufflinks (genes.fpkm_mapping outputs) from genomic mapping pipelines such as GSNAP (Wu, Reeder, et al. 2016), STAR (Dobin et al. 2013) and TopHat (Kim et al. 2013) or TPM abundances calculated by transcript RNA-Seq mapping programs such as Salmon (Patro et al. 2017) or Kallisto (Bray et al. 2016). However, qTeller's structure is based on gene models, not transcripts; therefore, if a user has Salmon/Kallisto output files that quantify expression at the per-transcript level, it will first be necessary to calculate an aggregate gene-level abundance, whether via averaging or another process, and the resulting gene-level data constructed as .txt file where Column 1 is the Gene ID, and Column 2 is the averaged TPM abundance (see Supplementary Table S3). The qTeller build.py scripts will automatically detect whether the directory containing the RNA-Seq abundances have the .txt or .fpkm_tracking extension and proceed accordingly. The two-column,

gene/abundance .txt file configuration will also work for abundances calculated through EdgeR or some other method. It is important to emphasize that all inputs combined in a single qTeller instance should be mapped using the same pipeline and use the exact same method of counting abundances (either FPKM or TPM or some other method). The combination of datasets generated using different quantification pipelines will generate many apparent differences in expression, resulting from technical differences in quantification rather than biological differences in expression.

The .txt file or .fpkm_mapping outputs must be preprocessed to replace empty abundances as 'Nan' before running them as input files in the qTeller build.py scripts. This preprocessing of the input files enables qTeller to differentiate between 0 abundance value and null abundance value. This preprocessing of the .txt file or .fpkm_mapping outputs can be either done by the user using Excel, or by the qTeller script metadata_NA_to_Nan.py available in our GitHub. One can also customize the code as per the user input file structure.

qTeller will accept multiple biological replicates. However, if a user has a large number of libraries to work with, it is advised to average the biological replicate abundances by gene, and create a text file as described above, for the RNA input; otherwise, visualization of datasets can become crowded and difficult to resolve visually. This method of averaging biological replicates was used to construct the current MaizeGDB qTeller instances.

Because .gff file structures can vary in Column 9 in terms of whether genes are prefaced by 'ID¼' or some other scheme, qTeller allows the user to indicate which identifier is used in the gene model input .gff file by selecting –gff and then –gene_def_tag and typing the identifier afterward. If this option is not selected, qTeller will assume the identifier is 'ID'.

Metadata files need to be organized exactly as the example given in Supplementary Table S2. The qTeller web pages are organized based on the project from which the RNA-Seq data was collected, processed and sequenced (i.e. 'Source' in the metadata files.) This ensures that all the data within a collection has been extracted and sequenced the same way, so as to avoid the issue of artifactual relative abundances due to differences in laboratory technique or environment. Note that abundances in similar tissues across different laboratories (i.e. leaf) may differ somewhat due to differences in laboratory handling and other factors. Metadata is dynamically organized on each web page based on the contents of the database.

qTeller builds a SQLite3 database from the gene model, RNA (and protein) abundance and metadata information. This database is then the source of all data called by qTeller in the web pages for Section 2.3.1.1, Section 2.3.1.2 and Section 2.3.1.3.

Certain files are hard-coded for drop-down menu information and must be manually changed by the user. For instance, in the index_*.php files (index_singlegenome.php, index_multigenome.php and Protein_index.php) that correspond to the Section 2.3.1.1 pages, the chromosome selection drop-down menu must be edited to reflect the number of chromosomes in the target genome(s), and the name of the chromosomes in the target genome. For instance, maize has 10 chromosomes, designated chr1, chr2, etc., whereas Sorghum's chromosomes are designated Chr01, Chr02, etc. in the current Sorghum reference genome release, and Arabidopsis thaliana has only five chromosomes; the drop-down menu in index_*.php will need to be edited to reflect the target genome's configuration (see Supplementary Fig. S1). Also, in the index_multigenome.php file for multiple genomes, the drop-down menu for genome selection will need to be manually changed to reflect the genomes used, based on their designation in

Column 5 of the gene model bed file input. The default index_multigenome.php in our GitHub download is configured for the multigenome test data (see below).

Example databases of a subset of incomplete maize data for single-genome, multi-genome and protein data, as well as example metadata, bed and fpkm files to generate these databases, are included in the build_db directory.

2.4 Results

2.4.1 Use case: MaizeGDB qTeller

There are several features in MaizeGDB's qTeller instance that are uniquely specified for maize, including the maize genomes, maize datasets and MaizeGDB-specific metadata and other information. The homepage for the MaizeGDB qTeller website (https://qteller. maizegdb.org/) has a general description of qTeller, news items, quick links for each of the four tools, two sections for getting started and frequently asked questions (FAQs) and a Contact page linked to a local JIRA (https://www.atlassian.com/software/jira) instance to track errors or issues. The datasets used in MaizeGDB qTeller are described below.

2.4.2 Datasets

2.4.2.1 Maize genomes

MaizeGDB currently hosts three instances of qTeller: RNA-Seq and protein abundance data for the latest two versions (versions 4 and 5) of the reference maize genome B73, and RNA-Seq data for the NAM founder genomes, consisting of the genomes of 26 diverse maize inbred lines (doi:10.1101/2021.01.14.426684, https://namgenomes.org/).

The well-known and most used public founder maize variety B73 was sequenced in 2009 (Schnable et al. 2009). For nearly a decade, B73 was the reference genome for the maize research community, and most of the genomic tools, resources and datasets at MaizeGDB were oriented around this single reference. MaizeGDB's 2018 release of qTeller centered around

version 4 of the B73 genome (RefGen_v4) released in 2017 (Jiao et al. 2017). As sequencing technology became more affordable, additional maize reference-quality genomes were sequenced (Hirsch et al. 2016; Springer et al. 2018; Sun et al. 2018). While these genome assemblies had great potential to advance maize research, the underlying assemblies and supporting datasets (e.g. RNA-Seq) were generated with different methodologies and conditions.

In 2020–2021, the NAM Sequencing Consortium (doi:10.1101/ 2021.01.14.426684, https://nam-genomes.org/) released the first set of maize genomes sequenced, assembled and annotated in a consistent way. The NAM Sequencing Consortium's data release included a new version of B73 (RefGen_v5) and the 25 founder lines of the Nested Association Mapping (NAM) population, which has been used extensively by maize and other researchers to study maize flowering time (Buckler et al. 2009), leaf architecture (Tian et al. 2011), disease resistance (Poland et al. 2011) and other important agronomic traits (Wallace et al. 2014). These assemblies presented the opportunity for constructing pan-genomes and identifying pan-gene sets (genes conserved across the varieties), as well as making it possible to develop pan-genome tools. The multi-genome version of qTeller at MaizeGDB includes these 25 NAM founder lines and B73 RefGen_v5. The MaizeGDB project currently supports 44 maize genomes that could be included into qTeller as additional multigenome gene expression datasets become available.

2.4.2.2 Maize gene expression

The MaizeGDB qTeller has over 200 unique datasets from 12 projects available at MaizeGDB. The B73 version 4 instance of qTeller has RNA-Seq data from six studies (Forestan et al. 2016; Johnston et al. 2014; Kakumanu et al. 2012; Stelpflug et al. 2016; Walley et al. 2016; Waters et al. 2017) covering 158 tissues/conditions. The B73 version 5 instance has data from eight studies (Forestan et al. 2016; Johnston et al. 2014; Kakumanu et al. 2012; Makarevitch et al. 2015; Opitz et al. 2014a; Stelpflug et al. 2016; Walley et al. 2016; Warman et al. 2020)

covering 172 tissues/conditions. The multi-genome set of NAM founders has three studies with 23 tissues/conditions (10.1101/2021.01.14.426684, 10.1105/ tpc.17.00475, 10.1186/s13059-017-1328-6). The ‘Compare RNA & Protein’ tool has data from one mRNA and protein study (Walley et al. 2016) for 23 tissues/conditions; this dataset is currently the only large-scale gene expression atlas that provides both RNA-Seq and protein abundance data.

2.4.2.3 Data processing

All of the RNA-Seq datasets for MaizeGDB qTeller were mapped with a consistent methodology. Fastq files were mapped to either Ensembl AGPv4 B73, Ensembl AGPv5 B73 or the NAM founder genomes using STAR, and abundances calculated using Cufflinks. The protein abundance data was projected to both AGPv4 B73 and AGPv5 B73 based on gene synteny [see methods in (Walsh et al. 2020)].

2.5 Main Figures and table

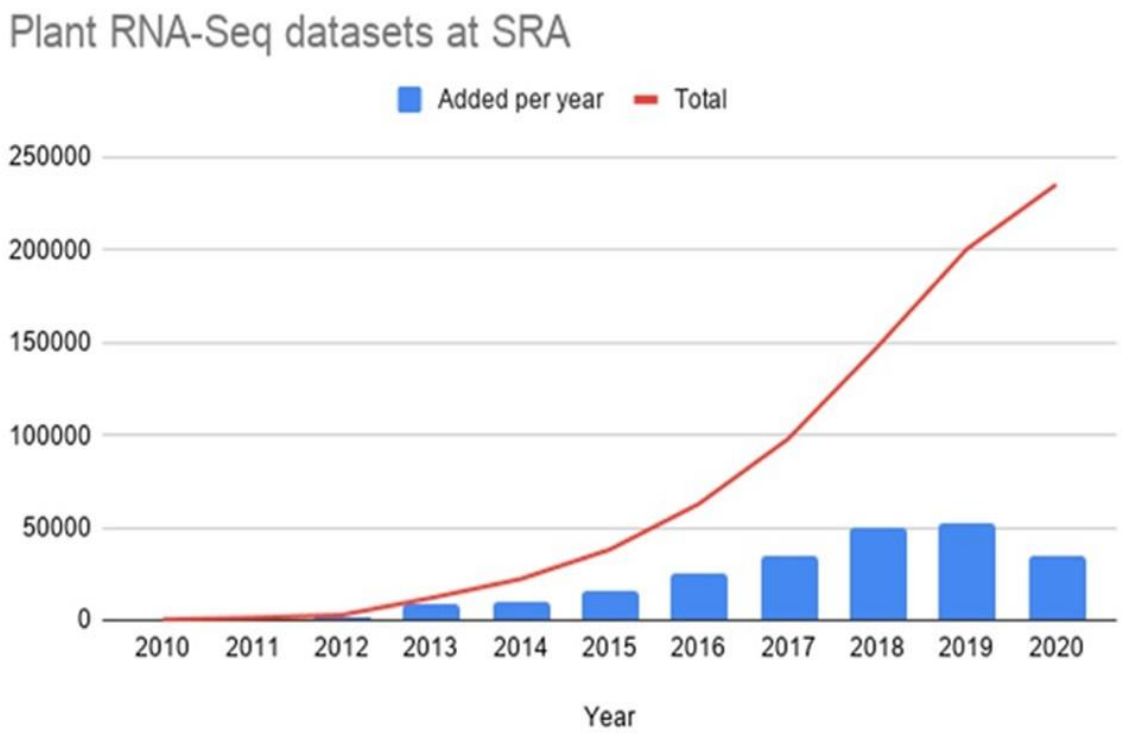

Figure 2.1 Plant RNA-Seq datasets at the NCBI Short Read Archive.

The chart shows the growth of plant RNA-Seq datasets at the GenBank Short Read Archive from 2010 to 2020. The x-axis is labelled by year. The y-axis is labeled by the number of RNA-Seq experiments. The blue bars for each year show the number of experiments added during that calendar year. The red line shows the cumulative number of experiments available during the given year.

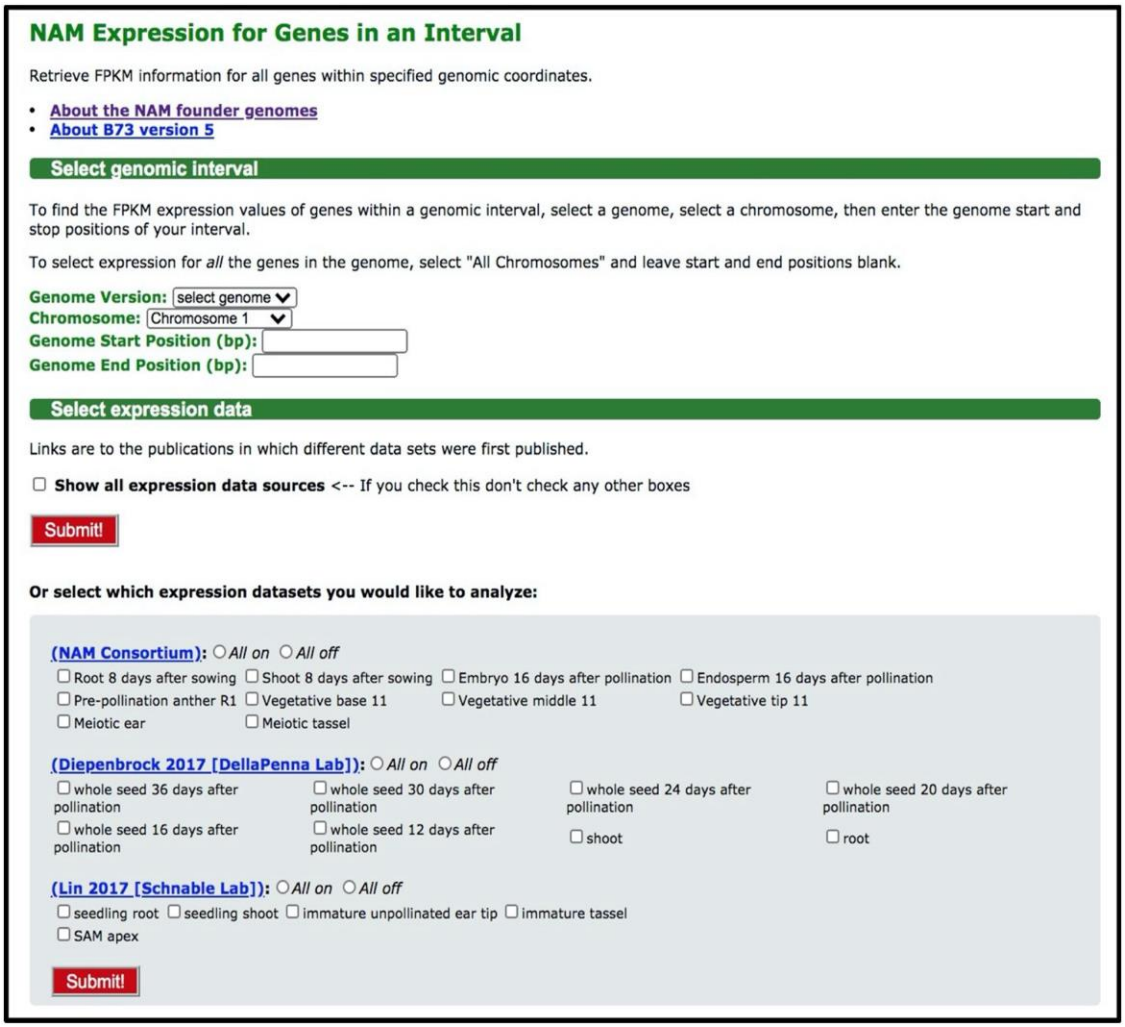

Figure 2.2 ‘Genes in an Interval’ tool.

The screenshot is from MaizeGDB’s qTeller instance for the ‘Genes in an Interval’ tool for a set of maize genomes (NAM founders). The input form has a drop-down menu for the

genomes and chromosomes (including an ‘All Chromosomes’ option), text boxes for the genome start and end position, and check boxes for each of the RNA-Seq experiments (organized by project or paper). Each set of RNA-Seq experiments has an ‘All on’ and ‘All off’ option.

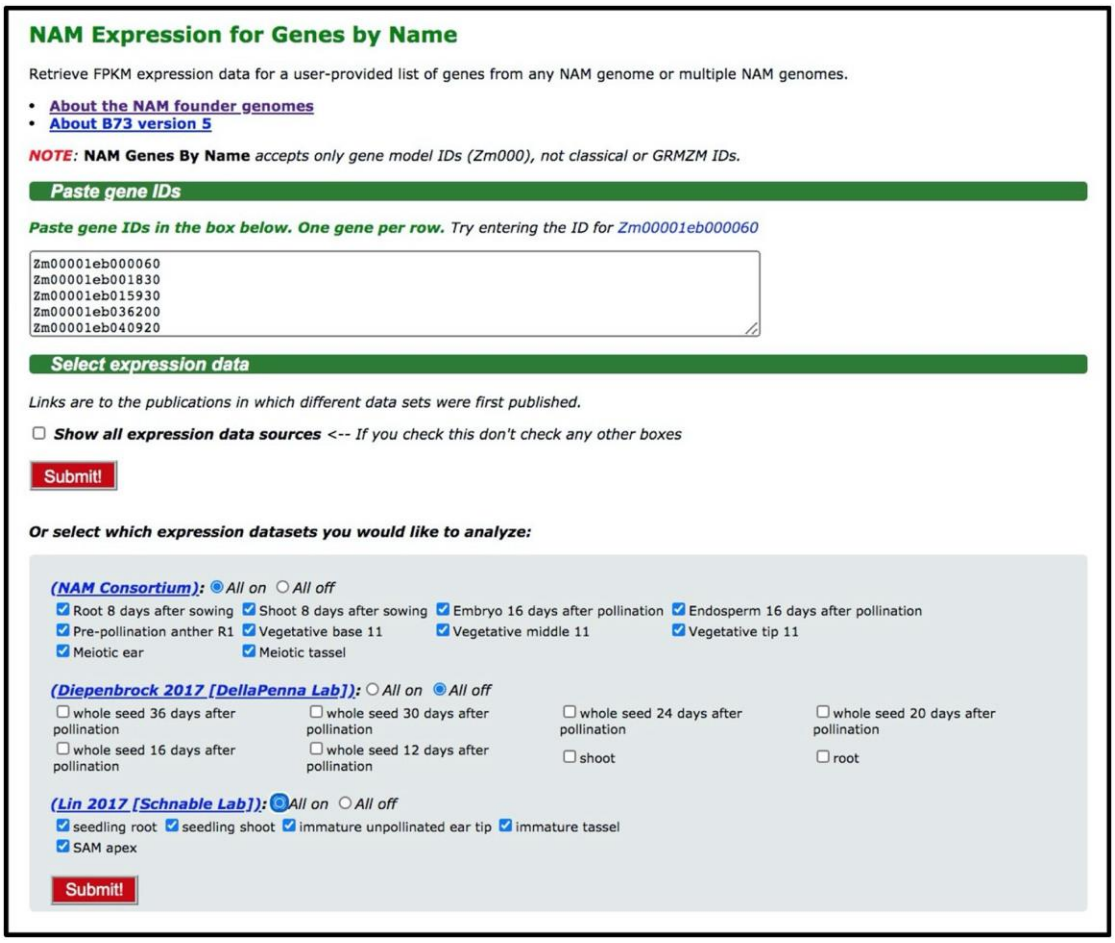

Figure 2.3 ‘Genes by Name’ tool.

The screenshot is from MaizeGDB’s qTeller instance for the ‘Genes by Name’ tool for a set maize genome (NAM founders). The input form takes as input a text box for a list of gene model identifiers and check boxes for each of the RNA-Seq experiments (organized by project or paper). Each set of RNA-Seq experiments has an ‘All on’ and ‘All off’ option.

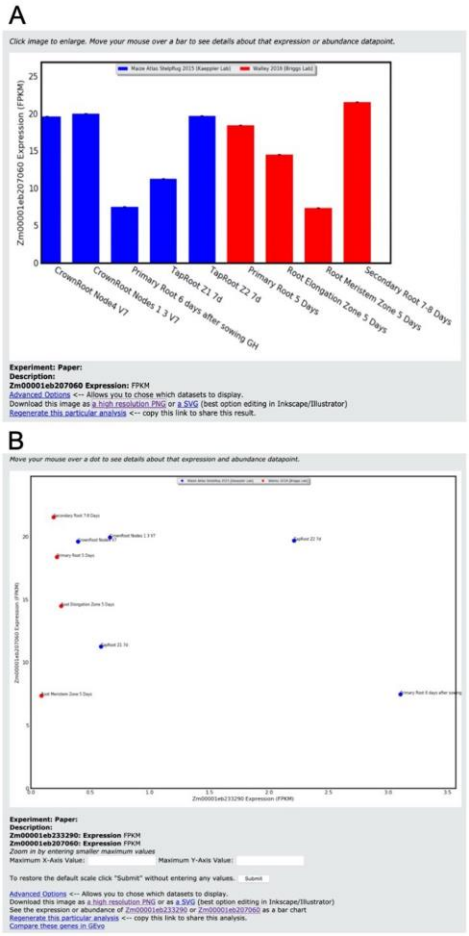

Figure 2.4 ‘Visualize Expression’ tool.

The screenshot is from MaizeGDB’s qTeller instance for the ‘Visualize Expression’ tool for genes in the B73 v5 maize genome. (A) The output of the B73 gene Zm00001eb207060 for a subset of root tissue from the ‘Single-Gene Expression’ tool that creates a bar chart showing the mRNA abundance from selected RNA-Seq experiments. (B) The output from the ‘Two-Gene Scatterplot’ tool which displays a scatter plot comparing the expression for two genes. Zm00001eb207060 from the bar chart image is compared to its retained homeolog Zm00001eb233290. Notice that Zm00001eb207060 is expressed consistently higher in root tissue than Zm00001eb233290.

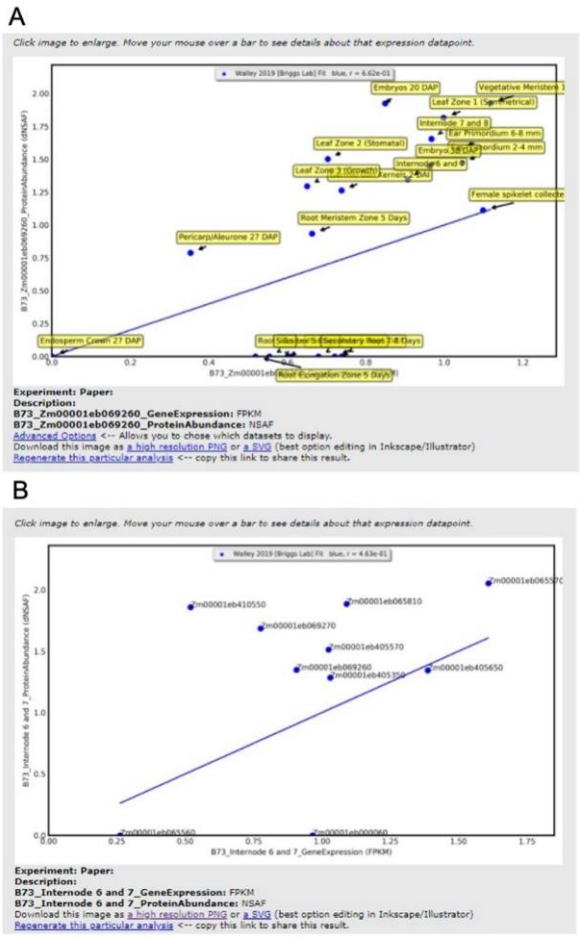

Figure 2.5 ‘Compare RNA and Protein’ tool.

The screenshot is from MaizeGDB’s qTeller instance for the ‘Compare RNA and Protein’ tool for genes in the B73 v5 maize genome. (A) The output from the ‘Single-Gene Expression versus Abundance’ tool that creates a scatter plot comparing mRNA expression and protein abundance from selected RNA-Seq experiments. (B) The output from the ‘Multi-Gene Expression versus Abundance in a single tissue’ tool which displays a scatter plot comparing mRNA expression and protein abundance for the selected tissue and set of gene models.

2.6 References

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter. 2016. 'Near-optimal probabilistic RNA-seq quantification', *Nat Biotechnol*, 34: 525-7.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz, M. M. Goodman, C. Harjes, K. Guill, D. E. Kroon, S. Larsson, N. K. Lepak, H. Li, S. E. Mitchell, G. Pressoir, J. A. Peiffer, M. O. Rosas, T. R. Rocheford, M. C. Romay, S. Romero, S. Salvo, H. Sanchez Villeda, H. S. da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M. D. McMullen. 2009. 'The genetic architecture of maize flowering time', *Science*, 325: 714-8.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Forestan, C., R. Aiese Cigliano, S. Farinati, A. Lunardon, W. Sanseverino, and S. Varotto. 2016. 'Stress-induced and epigenetic-mediated maize transcriptome regulation study by means of transcriptome reannotation and differential expression analysis', *Sci Rep*, 6: 30446.

Hawkins, L. K., J. E. Mylroie, D. A. Oliveira, J. S. Smith, S. Ozkan, G. L. Windham, W. P. Williams, and M. L. Warburton. 2015. 'Characterization of the Maize Chitinase Genes and Their Effect on Aspergillus flavus and Aflatoxin Accumulation Resistance', *PLoS One*, 10: e0126185.

Hirsch, C. N., C. D. Hirsch, A. B. Brohammer, M. J. Bowman, I. Soifer, O. Barad, D. Shem-Tov, K. Baruch, F. Lu, A. G. Hernandez, C. J. Fields, C. L. Wright, K. Koehler, N. M. Springer, E. Buckler, C. R. Buell, N. de Leon, S. M. Kaeppler, K. L. Childs, and M. A. Mikel. 2016. 'Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize', *Plant Cell*, 28: 2700-14.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." In, 90-95. IEEE Computer Society.

Jia, H., W. Sun, M. Li, and Z. Zhang. 2018. 'Integrated Analysis of Protein Abundance, Transcript Level, and Tissue Diversity To Reveal Developmental Regulation of Maize', *J Proteome Res*, 17: 822-33.

Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C. S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, and D. Ware. 2017. 'Improved maize reference genome with single-molecule technologies', *Nature*, 546: 524-27.

Johnston, R., M. Wang, Q. Sun, A. W. Sylvester, S. Hake, and M. J. Scanlon. 2014. 'Transcriptomic analyses indicate that maize ligule development recapitulates gene expression patterns that occur during lateral organ initiation', *Plant Cell*, 26: 4718-32.

Kakumanu, A., M. M. Ambavaram, C. Klumas, A. Krishnan, U. Batlang, E. Myers, R. Grene, and A. Pereira. 2012. 'Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq', *Plant Physiol*, 160: 846-67.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. 2013. 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biol*, 14: R36.

Langfelder, P., and S. Horvath. 2008. 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, 9: 559.

Li, Y., L. Zhai, J. Fan, J. Ren, W. Gong, X. Wang, and J. Huang. 2019. 'Genome-wide identification, phylogenetic and expression analysis of the maize HECT E3 ubiquitin ligase genes', *Genetica*, 147: 391-400.

Liu, J., A. R. Fernie, and J. Yan. 2020. 'The Past, Present, and Future of Maize Improvement: Domestication, Genomics, and Functional Genomic Routes toward Crop Enhancement', *Plant Commun*, 1: 100010.

Liu, R., H. Jia, X. Cao, J. Huang, F. Li, Y. Tao, F. Qiu, Y. Zheng, and Z. Zhang. 2012. 'Fine mapping and candidate gene prediction of a pleiotropic quantitative trait locus for yield-related trait in Zea mays', *PLoS One*, 7: e49836.

Liu, Y., J. Qu, L. Zhang, X. Xu, G. Wei, Z. Zhao, M. Ren, and M. Cao. 2019. 'Identification and characterization of the TCA cycle genes in maize', *BMC Plant Biol*, 19: 592.

Makarevitch, I., A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch, J. Ross-Ibarra, and N. M. Springer. 2015. 'Transposable elements contribute to activation of maize genes in response to abiotic stress', *PLoS Genet*, 11: e1004915.

Man, J., J. P. Gallagher, and M. Bartlett. 2020. 'Structural evolution drives diversification of the large LRR-RLK gene family', *New Phytol*, 226: 1492-505.

Opitz, N., A. Paschold, C. Marcon, W. A. Malik, C. Lanz, H. P. Piepho, and F. Hochholdinger. 2014. 'Transcriptomic complexity in young maize primary roots in response to low water potentials', *BMC Genomics*, 15: 741.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. 2017. 'Salmon provides fast and bias-aware quantification of transcript expression', *Nat Methods*, 14: 417-19.

Poland, J. A., P. J. Bradbury, E. S. Buckler, and R. J. Nelson. 2011. 'Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize', *Proc Natl Acad Sci U S A*, 108: 6893-8.

Pophaly, S. D., and A. Tellier. 2015. 'Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize', *Mol Biol Evol*, 32: 3226-35.

Portwood, J. L., 2nd, M. R. Woodhouse, E. K. Cannon, J. M. Gardiner, L. C. Harper, M. L. Schaeffer, J. R. Walsh, T. Z. Sen, K. T. Cho, D. A. Schott, B. L. Braun, M. Dietze, B. Dunfee, C. G. Elsik, N. Manchanda, E. Coe, M. Sachs, P. Stinard, J. Tolbert, S. Zimmerman, and C. M. Andorf. 2019. 'MaizeGDB 2018: the maize multi-genome genetics and genomics database', *Nucleic Acids Res*, 47: D1146-D54.

Price, A., A. Caciula, C. Guo, B. Lee, J. Morrison, A. Rasmussen, W. I. Lipkin, and K. Jain. 2019. 'DEvis: an R package for aggregation and visualization of differential expression data', *BMC Bioinformatics*, 20: 110.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A. P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J. M. Chia, J. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson. 2009. 'The B73 maize genome: complexity, diversity, and dynamics', *Science*, 326: 1112-5.

Sindhu, A., D. Janick-Buckner, B. Buckner, J. Gray, U. Zehr, B. P. Dilkes, and G. S. Johal. 2018. 'Propagation of cell death in dropdead1, a sorghum ortholog of the maize lls1 mutant', *PLoS One*, 13: e0201359.

Springer, N. M., S. N. Anderson, C. M. Andorf, K. R. Ahern, F. Bai, O. Barad, W. B. Barbazuk, H. W. Bass, K. Baruch, G. Ben-Zvi, E. S. Buckler, R. Bukowski, M. S. Campbell, E. K. S. Cannon, P. Chomet, R. K. Dawe, R. Davenport, H. K. Dooner, L. H. Du, C. Du, K. A. Easterling, C. Gault, J. C. Guan, C. T. Hunter, G. Jander, Y. Jiao, K. E. Koch, G. Kol, T. G. Kollner, T. Kudo, Q. Li, F. Lu, D. Mayfield-Jones, W. Mei, D. R. McCarty, J. M. Noshay, J. L. Portwood, 2nd, G. Ronen, A. M. Settles, D. Shem-Tov, J. Shi, I. Soifer, J. C. Stein, M. C. Stitzer, M. Suzuki, D. L. Vera, E. Vollbrecht, J. T. Vrebalov, D. Ware, S. Wei, K. Wimalanathan, M. R. Woodhouse, W. Xiong, and T. P. Brutnell. 2018. 'The maize W22 genome provides a foundation for functional genomics and transposon biology', *Nat Genet*, 50: 1282-88.

Stelpflug, S. C., R. S. Sekhon, B. Vaillancourt, C. N. Hirsch, C. R. Buell, N. de Leon, and S. M. Kaeppler. 2016. 'An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development', *Plant Genome*, 9.

Sun, S., Y. Zhou, J. Chen, J. Shi, H. Zhao, H. Zhao, W. Song, M. Zhang, Y. Cui, X. Dong, H. Liu, X. Ma, Y. Jiao, B. Wang, X. Wei, J. C. Stein, J. C. Glaubitz, F. Lu, G. Yu, C. Liang, K. Fengler, B. Li, A. Rafalski, P. S. Schnable, D. H. Ware, E. S. Buckler, and J. Lai. 2018. 'Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes', *Nat Genet*, 50: 1289-95.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. 2011. 'Genome-wide association study of leaf architecture in the maize nested association mapping population', *Nat Genet*, 43: 159-62.

Wallace, J. G., P. J. Bradbury, N. Zhang, Y. Gibon, M. Stitt, and E. S. Buckler. 2014. 'Association mapping across numerous traits reveals patterns of functional variation in maize', *PLoS Genet*, 10: e1004845.

Walley, J. W., R. C. Sartor, Z. Shen, R. J. Schmitz, K. J. Wu, M. A. Urich, J. R. Nery, L. G. Smith, J. C. Schnable, J. R. Ecker, and S. P. Briggs. 2016. 'Integration of omic networks in a developmental atlas of maize', *Science*, 353: 814-8.

Walsh, J. R., M. R. Woodhouse, C. M. Andorf, and T. Z. Sen. 2020. 'Tissue-specific gene expression and protein abundance patterns are associated with fractionation bias in maize', *BMC Plant Biol*, 20: 4.

Wang, Y., X. Hua, J. Xu, Z. Chen, T. Fan, Z. Zeng, H. Wang, A. L. Hour, Q. Yu, R. Ming, and J. Zhang. 2019. 'Comparative genomics revealed the gene evolution and functional divergence of magnesium transporter families in Saccharum', *BMC Genomics*, 20: 83.

Wang, Y., W. Lu, and D. Deng. 2016. 'Bioinformatic landscapes for plant transcription factor system research', *Planta*, 243: 297-304.

Wang, Z., M. Gerstein, and M. Snyder. 2009. 'RNA-Seq: a revolutionary tool for transcriptomics', *Nat Rev Genet*, 10: 57-63.

Warman, C., K. Panda, Z. Vejlupkova, S. Hokin, E. Unger-Wallace, R. A. Cole, A. M. Chettoor, D. Jiang, E. Vollbrecht, M. M. S. Evans, R. K. Slotkin, and J. E. Fowler. 2020. 'High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements', *PLoS Genet*, 16: e1008462.

Waters, A. J., I. Makarevitch, J. Noshay, L. T. Burghardt, C. N. Hirsch, C. D. Hirsch, and N. M. Springer. 2017. 'Natural variation for gene expression responses to abiotic stress in maize', *Plant J*, 89: 706-17.

Woodhouse, M. R., F. Cheng, J. C. Pires, D. Lisch, M. Freeling, and X. Wang. 2014. 'Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids', *Proc Natl Acad Sci U S A*, 111: 5283-8.

Wu, T. D., J. Reeder, M. Lawrence, G. Becker, and M. J. Brauer. 2016. 'GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality', *Methods Mol Biol*, 1418: 283-334.

Wu, X., Y. Li, J. Fu, X. Li, C. Li, D. Zhang, Y. Shi, Y. Song, Y. Li, and T. Wang. 2016. 'Exploring Identity-By-Descent Segments and Putative Functions Using Different Foundation Parents in Maize', *PLoS One*, 11: e0168374.

Zhang, G., B. M. Ueberheide, S. Waldemarson, S. Myung, K. Molloy, J. Eriksson, B. T. Chait, T. A. Neubert, and D. Fenyo. 2010. 'Protein quantitation using mass spectrometry', *Methods Mol Biol*, 673: 211-22.

Zhang, X., L. Lei, J. Lai, H. Zhao, and W. Song. 2018. 'Effects of drought stress and water recovery on physiological responses and gene expression in maize seedlings', *BMC Plant Biol*, 18: 68.

2.7 Appendix A. Notes

2.7.1 Data availability statement

The source code for qTeller is open-source and available through GitHub (https://github.com/Maize-Genetics-and-Genomics-Database/qTeller). A maize instance of qTeller is available at the Maize Genetics and Genomics database (MaizeGDB) (https://qteller.maizegdb.org/), where we have mapped over 200 unique datasets from GenBank across 27 maize genomes.

2.7.2 Acknowledgements

We thank the research groups of Iowa State University and USDA-ARS, Corn Insects and Crop Genetics Research Unit.

2.7.3 Author contributions

The authors wish it to be known that, in their opinion, Margaret R. Woodhouse and Shatabdi Sen authors should be regarded as Joint First Authors.

2.7.4 Declaration of interests

The authors declare no competing interests.

2.7.5 Funding

This research was supported by the US. Department of Agriculture, Agricultural Research Service, Project Number [5030-21000-068-00-D] through the Corn Insects and Crop Genetics Research Unit in Ames, Iowa. This material is based upon work supported by the Department of Agriculture, Agricultural Research Service under Agreement No. 58-5030-0- 036 [Iowa State Award: 022172-00001 to J.W.W.]. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer.

2.8 Appendix B: Supplementary tables and figures

All supplementary tables and figures can be found online at Sen et al., 2022: https://academic.oup.com/bioinformatics/article/38/1/236/6354355#supplementary-data

2.9 Appendix C: Consent to include co-authored article in thesis/dissertation

THE PARTIES

Table 2.1 Co-authors

Student Author (Full Name, Major, and Institution)	Shatabdi Sen Bioinformatics and Computational Biology Iowa State University, Ames, IA,50014
List other co-authors and their institutions.	Margaret R. Woodhouse USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA
Title(s) of the co-authored section (chapter, etc.)	qTeller: a tool for comparative multi-genomic gene expression analysis Chapter 2
Journal Name, Book Title, etc. (if applicable)	Bioinformatics

DISTRIBUTION OF TASKS AND RESPONSIBILITIES

In this research publication, I, Shatabdi Sen , was responsible for the following roles: (Select all roles that apply.)

☐Conceptualization
☒Data curation
☒Formal analysis
☐Funding acquisition
☒Investigation
☒Methodology
☐Resources
☒Software
☒Validation
☒Visualization
☐Writing – original draft
☒Writing – review & editing
☐Other: Please describe briefly

CHAPTER 3. MAIZE FEATURE STORE (MFS): A CENTRALIZED RESOURCE TO MANAGE AND ANALYZE CURATED MAIZE MULTI-OMICS FEATURES FOR MACHINE LEARNING APPLICATIONS

Shatabdi Sen[1] , Margaret R. Woodhouse[2], John L. Portwood II[2], and Carson M. Andorf[2,*]

[1]Department of Plant Pathology & Microbiology, Iowa State University, Ames, IA 50011, USA

[2]USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

[3]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript published in *Oxford Database*

3.1 Abstract

The big-data analysis of complex data associated with maize genomes accelerates genetic research and improves agronomic traits. As a result, efforts have increased to integrate diverse datasets and extract meaning from these measurements. Machine learning models are a powerful tool for gaining knowledge from large and complex datasets. However, these models must be trained on high-quality features to succeed. Currently, there are no solutions to host maize multi-omics datasets with end-to-end solutions for evaluating and linking features to target gene annotations. Our work presents the Maize Feature Store (MFS), a versatile application that combines features built on complex data to facilitate exploration, modeling, and analysis. Feature stores allow researchers to rapidly deploy machine learning applications by managing and providing access to frequently used features. We populated the MFS for the maize reference genome with over 14,000 gene-based features based on published genomic, transcriptomic, epigenomic, variomic, and proteomics data sets. Using the MFS, we created an accurate pan-genome classification model with an AUC-ROC score of 0.87. The MFS is publicly available through the maize genetics and genomics database. **Database URL:** https://mfs.maizegdb.org/

3.2 Introduction

The study of cellular, molecular, and genetic interactions in maize generates huge amounts of data. Due to the high dimensionality and heterogeneity of multi-omics data, integrating and analyzing these datasets has proven to be extremely difficult. Recently there has been an increased interest in analyzing large-scale omics data, particularly for predicting genotype-phenotype relationships. Over the last decade, machine learning has found numerous applications in plants, resulting in a slew of papers and reviews (Dai et al. 2020; Lloyd et al. 2015; Singh et al. 2016). There has been particular interest in maize, making it the most studied crop using machine learning (Benos et al. 2021). This interest can be attributed to the fact that it is grown in many parts of the world and has a variety of uses, including direct human consumption, animal feed, the production of ethanol, and other biofuels.

To further advance and facilitate the application of machine learning in crop and plant research, robust analytical methods and tools are required to manage multi-omics data through efficient data management, linkage, and integration strategies. This need is particularly strong for maize research, where a vast amount of data exists. Numerous storage methods have been developed to manage and analyze multi-omics data (Gui et al. 2020), including the Maize Genetics and Genomics Database (MaizeGDB) (https://www.maizegdb.org/), which comprises maize reference sequences, diversity data, expression data, phenotypic data, epigenetic and regulatory data, as well as metabolic pathway data along with multiple tools for genome-wide maize data exploration (Woodhouse, Cannon, et al. 2021); Panzea (https://www.panzea.org/), comprising genotypic and phenotypic data from several maize lines (Zhao et al. 2006); and Phytozome (https://phytozome-next.jgi.doe.gov/) a centralized hub of annotated plant gene families, evolutionary data and functional data (Goodstein et al. 2012). Other comprehensive databases and data repositories such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/)

(Benson et al. 2007), Gramene (http://www.gramene.org/) (Tello-Ruiz et al. 2021), ePlant (http://bar.utoronto.ca/eplant_maize/) (Waese-Perlman et al. 2021), MODEM (http://modem.hzau.edu.cn/) (Liu et al. 2016) and a more recent maize multi-omics database ZEAMAP (http://www.zeamap.com/) (Gui et al. 2020) also collect maize omics data. While these databases are quite useful, they store data in a structured manner using relational databases and require advanced multi-layer data structures to optimize data management and analysis. Additionally, they frequently lack interactive multivariate methods for exploring and integrating datasets. These databases enable users to access data in various file formats, including annotation data in GFF format and SNP datasets in VCF format. Although these datasets are easily accessible via these repositories, they do not come in a format suitable for performing diverse multivariate analyses, particularly at the gene level. Users who wish to apply modeling to these multi-omics datasets must spend considerable time collecting, cleaning, and aggregating before using them for model training.

Regardless of the challenges, omics integration studies have pervaded literature in recent years (Fukushima et al. 2009; Zogli et al. 2020; Deshmukh et al. 2014). As a result, the growing collection of omics data in maize is gaining attention among researchers to carry out systematic integrative analysis and storage of the heterogeneous data (Rajasundaram and Selbig 2016). In response to the challenges of handling heterogeneous data, non-traditional databases (NoSQL) emerged as an alternative, more flexible, and more scalable data store (Gundla and Chen 2016; Wang et al. 2014). Therefore, this paper presents the Maize Feature Store (MFS), a NoSQL-based interactive, modular, and dynamic user interface for systematically integrating and analyzing over 14,407 gene-based features based on the most recent maize multi-omics dataset (version 5 of the B73 reference genome, or B73v5). Feature stores are becoming a powerful

resource for data scientists to have readily available access to high-quality features for rapid deployment of machine learning applications, but feature stores are not available for most model organism databases. We aim to demonstrate how MFS provides a suite of methods and modeling modules, enabling users to find meaningful patterns from the maize omics data.

To demonstrate the utility of the MFS, we discussed the application of MFS in pan-genome analysis using the maize genome (B73v5) as a multi-omics utility case study. The pan-genome represents the entire set of genes within a species (Medini et al. 2005), consisting of a "core" genome, containing gene models shared between all individuals of the species, and the "non-core" genome, made up of near-core, dispensable, and private gene models occurring in most, some, or a single genome, respectively. Plant genomes are highly dynamic, and several challenges remain to be overcome before cost effective and rapid pan-genome construction is possible (Morneau 2021). Therefore, we provide modules aimed at tackling problems associated with pan-genome analysis by applying machine learning algorithms and classifying genes as core or non-core in a new genome using only multi-omics data associated with the genes.

3.3 Materials and Methods

3.3.1 Overview of the Maize Feature Store database

We have created an application that uses a MongoDB database (NoSQL) named "BigFeatureDb". MongoDB is a document-oriented data store that stores data in collections. Collections are made up of documents, and each field in a document is associated with a value. Complex maize omics data has been imported into these embedded data models via the Pymongo library. We stored each omics data type in separate collections for each feature type (e.g. "DNASequenceFeatures"). These collections contain documents corresponding to the gene model set of the B73v5 reference genome (Hufford et al. 2021). The document's key is used as the MongoDB primary key. Within each document, field value pairs are used to hold pairs of

gene model feature names and feature values. This database structuring allows a variety of aggregation operations to process complex queries.

3.3.2 Maize Feature Store Architecture

The Maize Feature Store has three layers, Transform (to ingest and process data and create features), Store (for storing the created features and their metadata), and Serve (to make available the stored features). The data in the Maize Feature Store is stored in the MongoDB database, and the features are extracted and pre-processed from varied sources using customized Python scripts. The front-end application in the Python Flask framework makes the data available to various end-users.

3.3.3 Application Development

We developed an interactive web-based query system to retrieve the desired information from the maize reference genome version B73v5 omics data using Flask, HTML5, JavaScript, and CSS. The server-side scripting uses Python code and Pymongo (v3.11.3) drivers. A sophisticated search query system enables users to conduct multiple searches, data visualization, and modeling.

The graphical user interface is designed to help users conduct an automatic end-to-end analysis of the maize omics data, along with basic exploratory analysis and predictive modeling of the datasets. To do this, the interface is divided into sections and subsections in the form of various menus on the navigation bar. The home page (https://mfs.maizegdb.org/) illustrates the overall functioning of the tool with three major components (“Features and Analysis”, “Models”, and “More”) for getting started with the analyses.

The “Features and Analysis” module (https://mfs.maizegdb.org/features) is divided into three main sections: All data analysis, Downsampled analysis, and User candidate gene analysis. Each of these sections is further subdivided into Sequence Features, Gene Structure Features,

Expression Features, Chromatin Features, Count Features, Correlation Features, and Other Features. These sections have additional subsections with specialized functions that operate dynamically on the selected dataset. Users can select their desired features and labels in each subsection and carry out a wide range of analyses using tables and graphs. Each subsection can analyze either the entire dataset or a randomly down-sampled dataset. The outputs of the selected analysis (tables and graphs) are displayed reactively on a separate webpage. The user can download all the tables (copy or .csv or .xlsx or .pdf) and plots (.png) using specified buttons. Additionally, tables and graphs are interactive, allowing for deeper data exploration. It is crucial to note that some subsections, such as “DNA Sequence” Features, do not display the whole dataset to prevent the complexity of selecting hundreds of features and avoid the visualization becoming unwieldy. However, users can always download the selected subset or the complete dataset via the “Download Source” or “Download All” choices. All the front-end structures were created using Bootstrap (v4.0), jQuery (v3.5.1), and Flask (v1.1.2) Python packages. The plots were built by Dashbio v0.7.1 and plotly (v5.3.1) / matplotlib (v3.4.2), respectively.

The “Predictions” section consists of machine-learning models as a web service. As a use-case, we provide two models: the “Advanced” model (https://mfs.maizegdb.org/model_advanced) and the “Basic” model (https://mfs.maizegdb.org/model_basic), for classifying maize core and noncore genes (Hufford et al. 2021). Two simple forms are built using HTML and CSS to take input from the users on the top 25 features that were highly predictive for differentiating between core and non-core genes. Our application uses a Gradient Boosting Classifier for the "Advanced" model and a Random Forest Classifier for the "Basic" model, both built with scikit-learn (v1.0.2) and wrapped in Flask. The “More” section holds additional information for the smooth functioning of

the interface, such as links to the Data Sources, Tool Sources, Frequently Asked Questions, and Contact Page.

3.3.4 Data acquisition

The central idea behind generating and extracting a broad set of omics data associated with the maize genome is to allow researchers to explore these intrinsic and extrinsic gene features and conclude their research findings linked to any eukaryotic organisms or, more specifically, to maize.

We curated an extensive set of genomics, transcriptomics, epigenomic, variomic, and proteomics data from three major sources: MaizeGDB, peer-reviewed publications, and data generated in other labs (https://mfs.maizegdb.org/data_sources). The B73v5 maize gene models, canonical protein sequences, and coding sequences were collected from the MaizeGDB database. Gene structural features were extracted from the annotation files (GFF) linked to the B73v5 genome. The gene expression (mRNA and protein abundance) datasets across multiple tissue types and conditions were collected from peer-reviewed publications and from other labs. The epigenomic and variomic datasets were gathered from MaizeGDB JBrowse (Woodhouse, Cannon, et al. 2021) and the maize Nested Association Mapping paper (Hufford et al. 2021).

3.3.4.1 Sequence Feature Generation

We used the canonical transcript and protein sequences to generate the sequence features for genes with multiple transcripts. The coding sequence data was used for generating various numerical representation schemes of DNA sequences. Four modules of the rDNAse package (Zhu and Dong 2016), basic tools, nucleic acid composition, autocorrelation, and pseudo nucleotide composition (details on the DNA features can be found here: https://mfs.maizegdb.org/DNAseq) were used to generate DNA sequence features. The genomic sequences were also used to generate various codon and amino acid usage features such as the

codon adaptation index, expected effective number of codons, and stacking energy using the SADEG package (Babak Khorsand 2017).

Numerous structural and physicochemical descriptors, such as amino acid composition, autocorrelation, composition/transition/distribution (CTD), conjoint triad, quasi-sequence order, pseudo amino acid composition, and the amphiphilic pseudo-amino acid composition (details on the protein sequence features can be found here: https://mfs.maizegdb.org/Proteinseq), were extracted from the peptide/protein sequences using the protr package (Xiao et al. 2015). The protein sequences were also used to generate predicted protein subcellular localization features (nucleus, cytoplasm, extracellular, mitochondria, cell membrane, endoplasmic reticulum, plastid, golgi apparatus, lysosome/vacuole, peroxisome) using the WolfPsort (Horton et al. 2007) and Deeploc (Almagro Armenteros et al. 2017) programs. The protein structural features such as coils, hot loops, transmembrane helices, and signal peptides were predicted from the amino acid sequences as an input using DisEMBL (Linding et al. 2003), TMHMM (Krogh et al. 2001), and SignalP (Petersen et al. 2011), respectively.

3.3.4.2 Structure Feature Generation

The gene annotation (GFF) files linked to the B73v5 maize genome were used to extract numerous gene structural features such as the gene length, number of isoforms, exon length, average exon length, number of exons, chromosome associated with each gene, coding sequence length, five-prime untranslated regions (UTR) length and three-prime UTR length using customized Python script. The Python script parses through the GFF file to generate these features.

Distance features such as distance from the chromosome center, distance to the nearest knob, the centromere, and the telomere were also generated for each gene of the B73v5 maize genome. The data was downloaded from MaizeGDB.

3.3.4.3 Expression Feature Collection

The maize transcriptomics and proteomics data consist of expression levels for each gene across multiple tissue types and experimental conditions. The RNA expression features included data from the MaizeGDB qTeller (Woodhouse, Sen, et al. 2021a) B73v5 instance. The MaizeGDB qTeller contains almost 200 unique datasets from 12 projects. Each dataset was mapped with a consistent pipeline to provide fair comparisons. Any future datasets added to the MFS will follow the same pipeline. The B73v5 instance of qTeller contains data from eight studies from multiple labs (Forestan et al. 2016; Warman et al. 2020; Walley et al. 2016; Stelpflug et al. 2016; Opitz et al. 2014b; Makarevitch et al. 2015; Kakumanu et al. 2012; Johnston et al. 2014) covering 172 tissues/conditions. The "Compare RNA & Protein" tool of qTeller incorporates data from a single mRNA and protein study (Walley et al. 2016) spanning 23 tissues/conditions. Apart from gene expression, we estimated the average mRNA abundance level, protein abundance level, maximum mRNA abundance level, maximum protein abundance level, tissue gene abundance breadth, and tissue protein abundance breadth for each gene across all tissues and conditions. The breadth is defined as the number of tissues where the gene or protein showed expression.

3.3.4.4 Chromatin Feature Generation

Chromatin features comprised of chromatin states, three histone modifications (H3K4me3, H3K27me3, H3K27ac), open chromatin as quantified by ATAC-Seq, and DNA methylation (quantified separately in CG, CHG, and CHH contexts) were obtained from the ChromHMM software and Dai, Xiuru et al. (Dai et al. 2020). The chromatin states were generated from ChIP-Seq data (including nine types of histone modifications, H2AZ, H3, H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K56ac) in two tissues, ear and leaf (Ricci et al. 2019). Histone modifications are often found in recurring combinations

at promoters, enhancers, and repressed regions. These combinations are called “chromatin states” and can annotate regulatory regions in genomes. We have included multiple chromatin states features from ChIP-Seq data using the tool ChromHMM (A multivariate HMM for chromatin combinatorics) (Ernst and Kellis 2017).

3.3.4.5 Count Feature Generation

We generated the “Count” features by finding and counting annotations from multiple genome interval files whose genomic coordinates overlapped with the maize gene sites using the bedtools suite (Quinlan and Hall 2010). The genome annotation files included the MaizeGDB B73v5 JBrowse annotations (mutational insertions, transcription factor binding sites, transcription start sites, enhancers, transposable elements, miRNAs) ((Ricci et al. 2019), (Dong et al. 2019; Bolduc et al. 2012; Oka et al. 2017; Vollbrecht et al. 2010; McCarty et al. 2013; Mejia-Guerra et al. 2015)) and G-quadruplexes. The G-quadruplex annotation files were generated using in-house Python scripts from the B73v5 maize genome sequence. Counts were computed for three genomic regions: the first region included the gene body, the second included a 1KB region upstream and downstream of the gene start and end sites, and the third covered a much larger region, comprising 5 KB upstream and downstream of the gene start and end site.

3.3.4.6 Correlation Feature Collection

The correlation features include 12 co-expression modules identified through weighted gene co-expression network analysis. The data comprises 79 tissues, 6-organ developmental gene atlas coupled with five abiotic/biotic stress transcriptome datasets (Hoopes et al. 2019). These topology features were available for B73 AGPv4 gene models; therefore, B73 AGPv4 gene models were converted to B73v5 using a conversion list published on MaizeGDB.

3.3.4.7 Varionomic Feature Generation

Varionomic features included the count of single nucleotide polymorphisms (SNPs) per gene model and the effects of SNPs on the genes. The count of SNPs per gene model was calculated by finding overlapping regions between the SNP data VCF file from (Hufford et al. 2021) and maize gene coordinates using Bedtools, and SnpEff was used to annotate and predict the impact of variations on genes. This tool takes pre-defined variations listed in a VCF file containing the nucleotide change and its location and predicts if the variants are detrimental.

3.3.4.8 Other Feature Generation

The "Other" feature section includes evolutionary gene age (described below) and the total number of presence/absence of associated Pfam-domains per gene model (Hufford et al. 2021; Mistry et al. 2021). The direction and magnitude of natural selection were inferred from the ratio of nonsynonymous substitutions (Kn) / synonymous substitutions (Ks) between Sorghum and maize B73v5 orthologous genes and from the ratio of nonsynonymous substitutions (Kn) / synonymous substitutions (Ks) between maize Tzi8, a tropical maize line (Hufford et al. 2021), and maize B73v5 orthologous genes. Ks and Kn values were derived between syntenic ortholog coding sequences of B73v5 and Sorghum bicolor v3 (https://phytozome-next.jgi.doe.gov/info/Sbicolor_v3_1_1) using the tool CoGe SynMap (Lyons and Freeling 2008) (https://genomevolution.org/coge/SynMap.pl) with the parameters Relative Gene Order; -D 20; - A 5; Quota Align Merge; Syntenic Depth B73:Sorghum 2:1; and CodeML Kn/Ks. Ks and Kn values between B73v5 and the maize tropical cultivar Tzi8 were derived using similar parameters except the Syntenic Depth was set to 1:1.

The evolutionary gene age was calculated by searching for homologs within increasingly broad clades using the phylostratr pipeline (Arendsee et al. 2019). The deepest clade that contains a homolog of the protein(s) encoded by a gene is that gene's age as described by

(Arendsee et al. 2019). The maize gene age is classified into 21 categories based on the presence/absence of the homologs of maize genes in 20 representative eukaryotic species (including cellular organisms, Andropogoneae, commelinids, Embryophyta, Eukaryota, Liliopsida, Magnoliopsida, Mesangiospermae, PACMAD clade, Panicoideae, Petrosaviidae, Poaceae, Poales, Spermatophyta, Streptophyta, Streptophytina, Tracheophyta, Tripsacinae, and Viridiplantae).

3.3.4.9 Label Generation

In addition to the different genomics, proteomics, and transcriptomics features, the Maize Feature Store also includes example biological annotations. They can be used as class labels for users looking to classify their genes of interest to any of the biological annotations or identify relationships between these gene annotations and a variety of features offered through the MFS. These gene annotations are not only meant to act as targets but are also intended to function as features when appropriate. For example, we can use whole-genome duplication (WGD)/tandem gene annotations as features when trying to solve core/non-core gene prediction problems and vice versa. Currently, MFS contains three sample labels: “Classical” (classical/other) genes, “Pangenome” (core/near-core/dispensable/private) genes, “Gene Origin” (WGD/tandem/both) genes, and a “No Label” option. Classical genes are the most well-studied genes in maize, most of which have a visible mutant phenotype (for example, liguleless2) as described by (Schnable and Freeling 2011). We downloaded 430 maize classical genes from MaizeGDB (Classical Genes). The core/near-core/dispensable/private genes and WGD/tandem/both genes were collected from maize pan-genome generated as part of the Nested Association Mapping (NAM) genome sequencing project (Hufford et al. 2021). The “No Label” option lets users view the relationship between the genes independently of any annotations. This selection is provided to enable users to view the properties of all genes without labeling them into different gene

categories or annotations. Using this feature, users can examine the features of multiple genes and can choose to annotate them based on common patterns identified between different genes. As it involves the inspection of all the genes, they work only for the "Submit for analysis" button.

3.3.5 Data Visualization

The MFS user interface is pre-configured with plotly, matplotlib, and Dashbio allowing innovative visualizations such as data distributions, connections between features, and aggregate statistics (minimum, maximum, average, unique categories, outliers, missing values, etc.). This enables researchers to gain rapid insight into the features and make more informed decisions about using specific features. The interface also provides detailed instructions on the usage and interpretation of each plot. Users are given options to conduct each exploratory analysis using the entire omics dataset or the downsampled data using the “Submit analysis” and “Downsampled analysis” buttons.

3.3.6 Downsampled analysis

The ratio of label categories is frequently uneven, resulting in a bias favoring the majority class. For example, seventy-two percent of our genes are marked as core in the maize reference genome version B73v5, and twenty-eight percent are annotated as non-core (near-core, dispensable and private genes). Therefore, we offer the random down-sampling method to address the issue of unbalanced data during exploratory analysis and provide users with the option of “Downsampled analysis”. It is important to note that the size of the downsampled data is different for each label (Classical/Pan-genome/Gene-Origin) selection as the size of the minority class is different in each label.

3.3.7 User candidate gene analysis

The user candidate gene analysis section allows users to do a comparative study on their genes of interest. Users can enter a single gene of interest, or a group of candidate genes linked to specific biological pathways or functions and compare them with other down sampled sets of maize genes. Two types of analyses are possible for the user-defined candidate genes: a) single candidate gene analysis and b) analysis of multiple candidate genes. For single gene analysis, users can enter a single gene of interest and visualize the output for the selected features either in tabular format or graphical format with a marginal plot showing the frequency distribution of the selected gene features for all maize genes along with highlighting the candidate gene (Supplementary Figure S1A). For analysis of multiple candidate genes, users can enter a list of genes and compare their gene list for the selected feature with the other downsampled maize genes in various univariate or multivariate plots. When using multiple candidate genes, it is recommended that a larger gene list be entered (fifty or more) so that a more reliable comparison of the candidate genes and the downsampled other maize genes can be made. The down-sampling is random based on the number of candidate genes; therefore a larger candidate gene list requires more down sampled genes, resulting in a better representation of the population.

To demonstrate the potential use case of the user candidate gene analysis, we gathered a set of fifty stress genes differentially expressed between the control and salt stress samples (Li, Cao, et al. 2017) and used them to identify unique characteristics common among salt stress genes (Supplementary Figure S1B). Using our univariate analysis, we found that the maize B73v5 salt stress genes differed significantly from other downsampled maize genes regarding the gene structural features of isoform count, coding sequence length, three-prime UTR length, and five-prime UTR length. These structural features showed a significantly higher range among the candidate genes. Previous work on stress genes has also discovered that 3'UTR-based mRNA

stability controls are present in stressed cells (Zheng et al. 2018), thereby further supporting our findings from the salt stress genes.

3.3.8 Exploratory analysis

The exploratory analysis module in the Maize Feature Store assists users in visualizing all accessible features and labels in tabular and graphical formats after initial preprocessing, cleaning, and normalization steps. Omics datasets come in diverse scales and follow their own statistical distributions as they are collected from disparate sources; therefore, data standardization becomes crucial for omics datasets. The MFS application allows for the normalization of omics numerical features by centering the features with their mean and the standard deviation between 0 and 1 using the StandardScalar function of Sklearn.

Apart from providing fundamental functionality, high-end modules in MFS calculate and perform various univariate, bivariate, or multivariate analyses such as Histograms, Count and Distribution plots, Pair plots, Box plots, Violin plots, Joint plots, Scatter plots, Correlation plots, Categorical Bar plots, Heatmaps, Clustering plots, and Dimension reductions (PCA) (see Supplementary Methods). However, the “Gene Expression” dataset currently provides a preview of the results by limiting the display of Histograms, Count and Distribution plots, Pair plots, Box plots, Violin plots, Correlation plots, and Heatmaps to five tissues of the selected lab. Since each lab includes multiple tissues, the limit of visualizing five tissues is intended for better analysis and visualization of plots. Users can modify the script to view more than five tissues from a lab. Most of these plots have options to download, zoom-out/zoom in, reset axes, autoscale, toggle spike lines, show the closest data on hover, compare data on hover, box select, pan, and lasso. Users can also select specific legends to view data only for the selected legends. The Histograms, Count and Distribution plots, and the Categorical Bar plots also come with a two-sided p-value analysis displayed at the top of the selected feature chart to determine whether there is enough

statistical evidence in favor of a hypothesis (there is a difference in the selected feature values or frequencies across the different categories of the target variable). For comparing the effect of the selected continuous feature on the classical/other genes target variable (binary), we carry out a two-sample test using the scipy.stats library in Python. For comparing the effect of the selected continuous feature across multiple categories of the target variable such as core/near-core/dispensable/private genes or WGD/tandem/both genes, we carried out a one-way ANOVA test using the Python stats library, and lastly, for comparing the effect of the selected categorical feature across two or more categories of the target variable, we carried out the Chi-square test using the scipy.stats library in Python.

Details on the usage and interpretation of all the plots and tables are also available on the MFS website (https://mfs.maizegdb.org/Structure) and Supplementary Methods. While MFS is intended to facilitate plot generation using a graphical user interface, by hiding sophisticated plotting routines behind MFS modules, users can download the appropriate module Python script for direct replication and transformation of the visualizations.

3.3.9 Data Clustering

The MFS uses advanced functionalities to analyze unlabeled omics data rather than labeled data to overcome the lack of manual annotations. The module can efficiently compute several unsupervised clustering algorithms on downsampled omics data and provides interactive visualization of the results using Dendrograms, Heatmaps, Hierarchical Scatter plots, Hierarchical Heatmaps, and PCA plots (2D, 3D, biplot) (see Supplementary Methods). Different user options are available for some of these modules to dynamically show different results. For example, in the Hierarchical Scatter plots, the “Choose Clusters” option is available where the users can manually enter the number of clusters to visualize in the Pair plot. However, it is recommended that users enter the number of clusters as per the output generated by the

Dendrogram plot. To save time and complexity, we limited the Heatmap plot to only display the relationship between the first hundred down-sampled genes and the selected attributes; however, users with sufficient resources are free to utilize the function and customize it to include as many genes as necessary for their specific analysis.

3.4 Results

3.4.1 Maize Feature Store Workflow

Maize omics data are generally large, complex, and contain a variety of structures. The ability to store and retrieve data effectively is critical in maize research. Historically, huge datasets have been kept as flat files on disk or relational databases. These platforms are difficult to develop, maintain, and adapt to big-data applications because they adhere to inflexible table structures and frequently lack scalability, such as data aggregation. Therefore, we propose to design our application to carry out complex operations, including 1) Flexible, to handle a wide variety of data types. This enables researchers to rapidly evolve data models and conduct customized analyses. 2) Scalable, permitting researchers to easily explore large and complex datasets without waiting long periods for simple queries. 3) Operationally mature, including end-to-end encryption, fine-grained data access control, and operational tooling. These operations can facilitate the management of multi-omics data and the accurate alignment of genes across multiple datasets, thereby increasing the feasibility of multi-omics integrative analysis.

Numerous biological prediction problems (van Dijk et al. 2021) are based on standard feature sets such as gene length, exon number, and gene expression. These conventional feature sets are repeatedly utilized to tackle many different biological problems and to obtain these features from raw data requires users to know bioinformatics, such as annotating gene models from a genomic fasta file, mapping RNA-Seq reads to genomes or extracting counts of exons per gene model. These processes become tedious and repetitive if we use the same features to solve

further biological problems. A feature store allows researchers to overcome this obstacle and improve the usability of the omics in the genotype-to-phenotype context. We developed the Maize Feature Store tool to simplify the management, access, and analysis of omics datasets for a wider range of users.

3.4.2 Application of MFS on pan-genome classification

We illustrate the capability of the Maize Feature Store in applying and analyzing multi-omics data for classifying genes as core or non-core and identifying top omics features that are most helpful in predicting their classification within a pan-genome. As reported in Figure 3.1, several modules were developed to follow a precise exploratory analysis workflow that goes from the data selection to the downstream data analysis and ultimately to modeling. For our case study, we developed two models: one that utilized all omics features (a total of 14,407 features) (https://mfs.maizegdb.org/feature_details) and another that utilized a subset of omics features (a total of 10,271 features) consisting of only the gene structure, gene sequence, and protein sequence data (https://mfs.maizegdb.org/feature_details). The model development lifecycle involved several stages, such as feature engineering, dealing with imbalanced data, feature selection, model building, hyperparameter tuning, and finally selecting the most optimal model (see Supplemental Methods).

An example of the “Data Table” module is shown in Table 3.1. In the “Data Table” module, it is possible to view all the genes and the selected features. Users can sort the table columns and use the search bar to look up specific gene IDs. We used the MFS data exploration and visualization modules to perform several univariate, bivariate, and multivariate analyses of the core and noncore gene structural features (Supplementary Figures S2-S6, see Supplemental Methods). An initial analysis of the data provided a quick visual summary of the potential association between the selected features of interest and the various categories of the “Pan-

genome" label (Figure 3.2 and Supplementary Figures S2-S6). By simultaneously exploring gene structure features, we can observe that several features are significantly correlated in both core and non-core genes. Therefore, the plots can initially demonstrate how the different genomic features can contribute to our understanding of core or non-core genes and highlight the potential for gene structural features in pan-genome classification.

3.4.3 Unified features excel over individual subsets in maize gene classification: core vs. non-core categories

The "Modeling" module of MFS offers an "Advanced Model" form and a "Basic Model" form which allows users to make predictions for their genes based on certain inputs. We trained the "Advanced" model using the top 25 features from a comprehensive set of omics features generated using a Hybrid Feature Selection method and a base Gradient Boosting Classifier with five-fold cross-validation (Supplementary Tables S1-S2 and Figure 3.3A). We built a simplified "Basic" model by training on the top 25 features generated using a similar approach from only the gene structural features and sequence features (Supplementary Tables S3-S4 and Figure 3.3B). To evaluate the specific contributions of each feature type to the overall accuracy of core and noncore gene prediction, we performed individual predictions using the other distinct subsets of features (Expression Features, Chromatin Features, Count Features, Correlation Features, and Other). This involved constructing separate machine-learning models for each feature subset (Supplementary Figure S11-S16). We tested the performance of six machine-learning algorithms for the classification of "Pan-genome" genes on both "Advanced" and "Basic" models, namely: (1. Logistic Regression, 2. Random Forest Classifier, 3. Gradient Boosting Classifier, 4. Extra Trees Classifier, 5. KNeighborsClassifier, and 6. SVM Classifier) and two distinct optimization approaches (1. Random and 2. Grid Search). In general, all five approaches performed well, but Gradient Boosting Classifier performed significantly better in the "Advanced" model with the

area under the Receiver Operating Characteristic Curve (AUC-ROC) = 0.85, Average Precision-Recall (PR) = 0.96, and F1 = 0.92 (Figure 3.3C) and the Random Forest Classifier performed significantly better in the "Basic" model yielding an AUC-ROC = 0.80, Average PR = 0.92 and F1 = 0.89 (Figure 3.3D). We compared the results to random classification to gain a proper perspective on the model performances. Based on random classification, the AUC-ROC would be 0.5. When AUC=0.5, the classifier cannot distinguish between positive (core) and negative (non-core) class points, as the classifier is predicting a random class or a constant class for all the data points. An increase in AUC-ROC and F1 can be seen in the "Advanced" model, especially compared to the "Basic" model. Therefore, our performance increased significantly when we used both intrinsic and extrinsic features, as demonstrated by the "Advanced" model.

Additionally, both of our models: "Basic" and "Advanced", outperformed a previous model in terms of accuracy recently published by Yocca, E, Alan et al. (Yocca and Edger 2021), which predicted core genes of Oryza sativa and Brachypodium distachyon using only intrinsic features such as gene sequence features, evolutionary features, and gene structural features. They achieved an AUC-ROC of approximately 0.77 and an accuracy of approximately 0.71 when trained and tested with the Oryza sativa balanced datasets and an AUC-ROC of approximately 0.86 and an accuracy of approximately 0.80 when trained and tested with the Brachypodium distachyon balanced datasets using a Random Forest method for ML, whereas our "Basic" model (Random Forest Classifier) achieved an AUC-ROC of approximately 0.80 and an accuracy of approximately 0.84 in the testing set, and our "Advanced" model (Gradient Boosting Classifier) achieved an even higher accuracy of approximately 0.89 and AUC-ROC of approximately 0.85. In this way, our models not only classify genes as core or non-core but also challenge the efficacy of current pipelines by comparing model output with pipeline output. Analyses of

complex genomes by pan-genome pipelines often result in the incorrect annotation of genes as core or non-core. Our model can provide extra validation to the pipeline output and identify mis-annotations that may occur in the current pipelines, which are both time-consuming and computationally expensive.

3.4.4 Investigating the features that have strong differentiation powers in both the "Basic" and "Advanced" models

The best performing model (Gradient Boosting Classifier in the "Advanced" model and Random Forest Classifier in the "Basic" model) was used to determine which predictor variables are most significant for prediction performance. In this way, we can gain insights into the biology of core and non-core genes. The 25 most important variables (Figure 3.3A and 3.3B) for training the "Advanced" model and the "Basic" model were generated using a Hybrid Feature Selection method and a base Gradient Boosting Classifier as described in the materials and methods section. A Gradient Boosting Classifier has a built-in variable importance assessment. The Kn/Ks ratio of both sorghum vs. B73 and Tzi8 vs. B73, a measure of evolutionary pressures on protein-coding regions, was among the top five most significant features in the "Advanced" model. There have been previous pan-genome studies that compared synonymous (Ks) and nonsynonymous substitution (Kn) rates (Tao et al. 2019). These studies have indicated that dispensable genes undergo more non-synonymous substitutions, as well as increasing Kn/Ks ratios, implying greater positive selection on dispensable genes (Gordon et al. 2017; Wang et al. 2018; Li et al. 2014). While performing exploratory analysis with the genes, we also observed a difference in the Kn/Ks ratio of Tzi8 vs. B73 among the "Pan-genome" genes with a mean value of 13.90 in the dispensable genes and 4.58 and 4.70 in the near core and core genes, respectively, aligning with results found in the previous studies of greater positive selection on dispensable genes. The two-sided p-value analysis also indicated a significant difference in the Kn/Ks ratio

observed among the “Pan-genome” genes. Other important predictors in our “Advanced” model were the difference in the ratio of the WGD regions among the core or noncore genes, presence, and absence of Pfam domains (protein families, domains, and functional sites extracted from the Pfam database) for coding genes in the core genome set and those in the dispensable genome, Transcription Factor Ethylene Responsive Element Binding Factor domain EREB (stress-responsive transcription factors) and (TE) transposable elements.

Gene duplications play a major role in the evolution of novel traits in eukaryotes (Ohno 1970; Yu et al. 2019). The WGD regions are found to contain a higher ratio of core and near-core genes, whereas non-WGD regions (tandem regions) contain a higher ratio of dispensable and private genes (Liu et al. 2020; Bayer et al. 2020). Additionally, the exploratory analysis also indicated that in our omics dataset, the non-core genes had a higher tandem repeats ratio than the core genes (Supplementary Figure S7). An enrichment of TEs in the vicinity of dispensable genes was reported in B. distachyon (Gordon et al. 2017) and B. oleracea (Golicz et al. 2016). Our model, as well as our exploratory analysis (https://mfs.maizegdb.org/TE), complements the findings of previous studies on transposable elements and Pfam domains (Zhao et al. 2006), as the maize B73 dispensable genes were also found to be enriched with transposable elements around the 1Kb and 5Kb regions upstream and downstream of the gene start site and end site respectively, and the total Pfam domains were also abundant among the maize B73 core genes compared to the dispensable genes. As the EREB transcription factors are involved in plant hormone responses under stress conditions (68), they are more likely to be enriched among dispensable genes than the core genes, and our study confirms this (https://mfs.maizegdb.org/TFbindingSite).

The top features in our "Basic" model having the most influence in the classification of core or non-core genes are the five-prime UTR length, three-prime UTR length, isoforms count, and sequence features such as Composition Transition Distribution (CTDD), pseudo dinucleotide composition (PseDNC) and many more. Most of these features displayed significant differences between the maize B73 core and non-core genes (https://mfs.maizegdb.org/Structure). Earlier studies have also stated that dispensable genes tend to display common features similar to young genes: short gene length, weak homology, low expression, rapid evolution, and turnover (Christine Tranchant-Dubreuil 2019), thereby further supporting our findings on the topological properties of core and non-core genes.

3.5 Discussion

The growing number of omics datasets from diverse sources have highlighted the importance of evaluating specific models and methods for collecting, managing, and analyzing multi-omics data to better explore the interplay between the multiple cellular, molecular, and phenotypic layers. While several multi-layer data structures are available, there is still a need for end-to-end solutions for storing, exploring, and modeling data. To solve this need, we proposed using MFS as a suitable structure to manage commonly used maize omics features. MFS will benefit bioinformaticians, data scientists, and experimental researchers interested in solving complex biological problems. Our tool enables researchers to share and discover features, create more effective machine-learning pipelines, and perform exploratory analyses. It provides users without domain knowledge or modeling experience the ability to identify the most significant factors affecting the target problem. For example, during the exploratory analysis of "Pan-genome" genes (Figure 3.2), we observed that the exon number varied across the pan-genome categories and thus might be a strong predictor of core or non-core genes.

An example of a current application of these models involves classifying genes in a new species closely related to maize as core or non-core without constructing an expensive pan-genome. Our models outperform random assignment for most downstream applications with around 90% accuracy. Our model would also be ideal for newly sequenced or poorly annotated genomes. Where other tools like BLAST could also infer annotation, it does not provide underlying insights for the assignments beyond sequence homology.

Each year, numerous papers and research articles are published on maize, utilizing omics data. However, although data repositories exist, there is a need to extend model organism databases like MaizeGDB to provide end-to-end data analysis. MFS, in this context, provides a central hub of maize omics features with flexible and expandable functionality that enables maize researchers to configure the tool for specific analyses. Additionally, MFS's modeling module utilizes a comprehensive set of omics features to conduct a core/non-core gene classification. Even though several prediction or classification problems have been addressed using a wide range of omics features in mice (Yuan et al. 2012), D. melanogaster (Campos, Korhonen, Hofmann, et al. 2020; Aromolaran et al. 2020), and C. elegans (Campos, Korhonen, Sternberg, et al. 2020), no work on plants, more specifically maize, has been reported. We were able to build a classification model utilizing the comprehensive set of features (“Advanced” model) and perform a comparative study by building another model utilizing just sequence and structural features known as the “Basic” model. Although the “Basic” model was more generalized, the “Advanced” model performed significantly better (Figure 3.3C), thus showing that an elaborate assembly of intrinsic and extrinsic factors from a wide range of sources covering multiple aspects of a gene greatly outperforms the approach based solely on sequence or structural features. We further emphasized the necessity of using both intrinsic and extrinsic

features by comparing our models (both "Basic" and "Advanced") with already existing models by Yocca, E, Alan et al. (Yocca and Edger 2021), which predicted core and non-core genes of Oryza sativa and Brachypodium distachyon, respectively. Our “Advanced” model performed significantly better with an accuracy of almost 25% higher than their same species Oryza sativa model (trained and tested on the Oryza sativa balanced datasets) and almost 11% higher than their Brachypodium distachyon model (trained and tested on the Brachypodium distachyon balanced datasets).

In this work, we aimed at the needs of both experimental and computational researchers. We addressed the need for resources that bridge the gap between the growing number of omics datasets and their potential as training data for modeling and machine learning. We developed a framework that hosts over 14,000 gene-based machine learning features built on multi-omics data to facilitate the exploration and modeling of classification problems. The tool's modularity will allow computational researchers to add additional functionality, fine-tune existing functionalities, and reproduce the entire application for other species of interest.

3.6 Main Figures and Tables

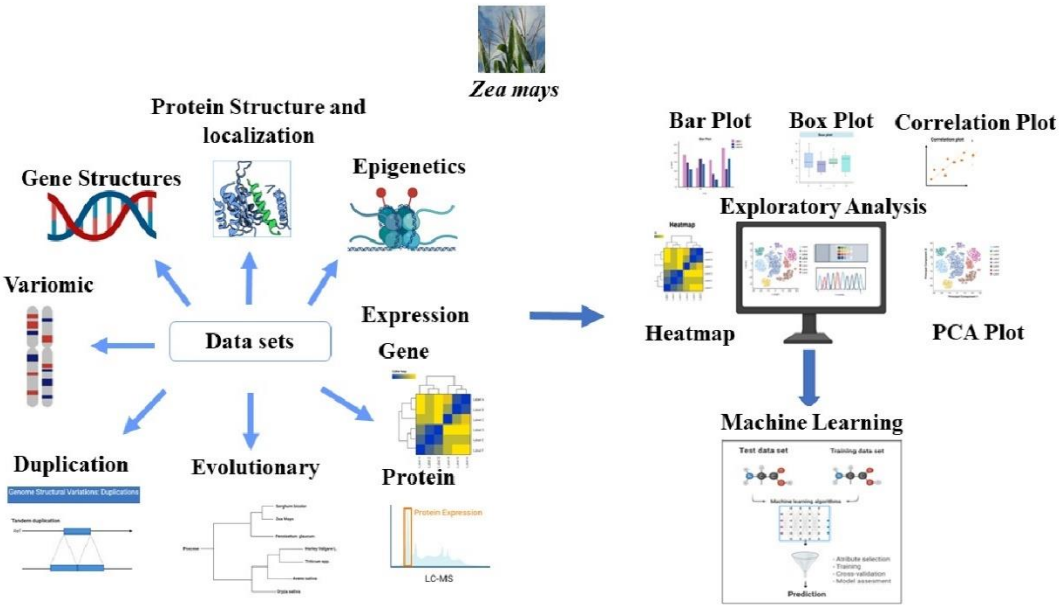

Figure 3.1 Module description: The MFS consists of three main modules: Features, Downstream Analysis, and Modeling.

We assembled the omics features associated with each gene model in Zea mays (B73v5) based on various sources as indicated by the “Data sets” arrows of the figure. Many preliminary and advanced exploratory analyses can be performed on the generated features as indicated by the “Exploratory analysis” module of the figure. Systematic evaluation of machine learning (ML) approaches is used in the Modeling section to solve complex biological problems, such as pan-genome prediction. The Graphical Overview was created using BioRender.com.

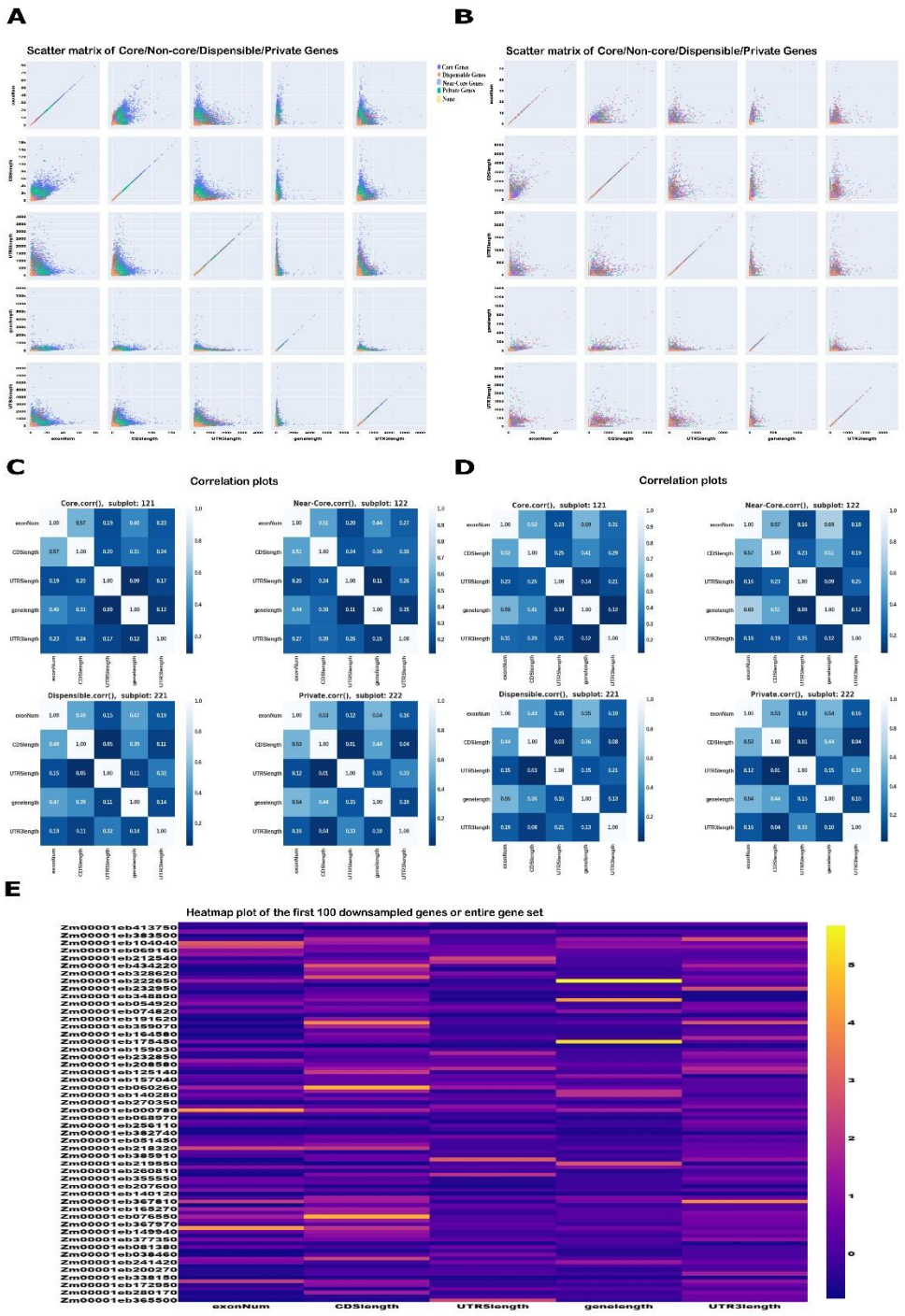

Figure 3.2 Example Maize Feature Store outputs.

The MFS provides users with options to carry out several univariate, bivariate, and multivariate analyses for both the total and downsampled omics data. Univariate analysis example: (A) Total Histogram; (B) Downsampled Histogram; Bivariate analysis example: (C) Total Scatter plot; (D) Downsampled Scatter plot; Multivariate analysis example: (E) Total Correlation plot; (F) Downsampled Correlation plot. These plots were generated from the selected Gene Structures such as Gene length, Exon number, three-prime UTR length, five-prime UTR length, and the selected label ("Pangenome": core/near-core/dispensable/private). The plot's colors and legends indicate the multiple "Pangenome" categories. In addition to the graph, to increase the interpretability of the data, we have also included p-values, mean and standard deviations of the selected datasets. For details on the interpretation of the plots, see (https://mfs.maizegdb.org/Structure).

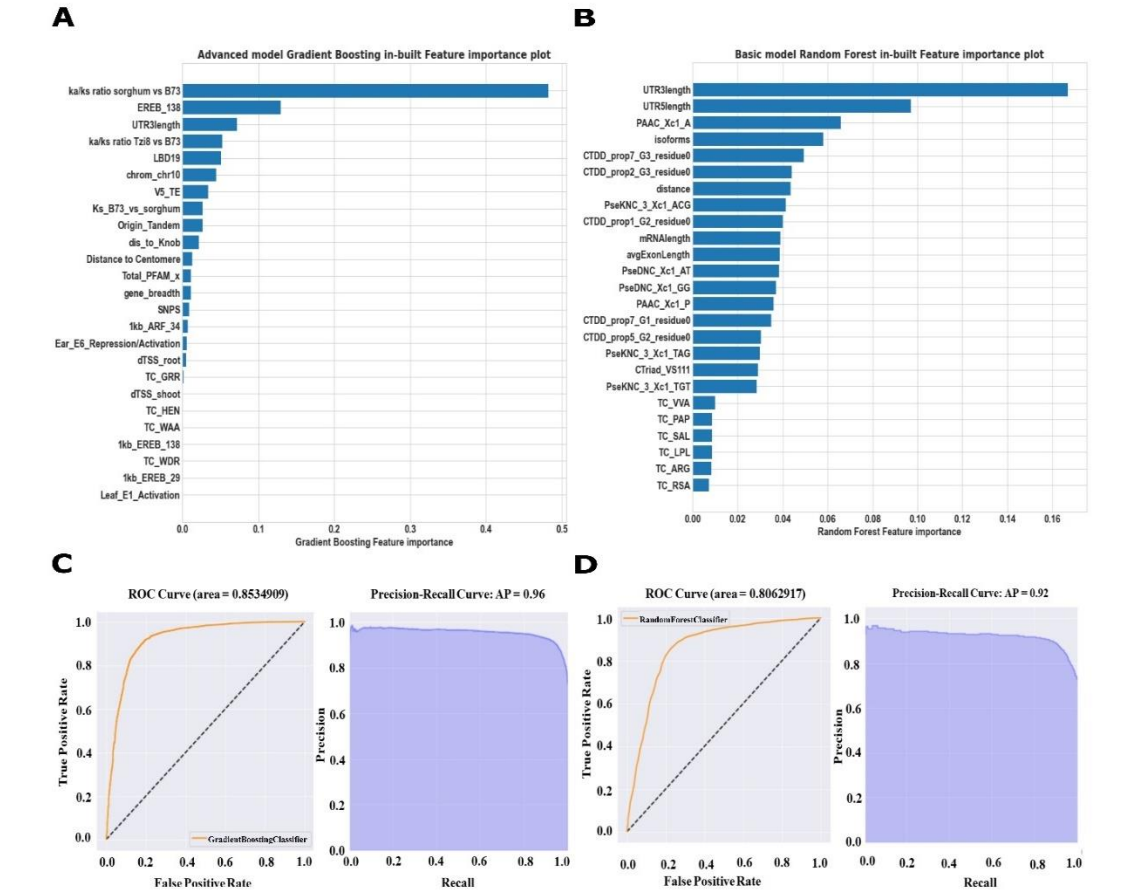

Figure 3.3 Maize Feature Store example Basic and Advanced models.

(A) In our Advanced model, both intrinsic and extrinsic features contributed substantially to the core/non-core gene predictions in maize B73v5. The 25 omics features were ranked based on how useful the model found each feature in predicting the target (core/non-core genes). (B) The Basic model feature importance plot displays only the structural and sequence features most predictive of identifying the core and non-core genes in B73v5. Higher scores indicate that a specific feature has a larger impact on the model used to predict a specific variable (core/non-core). (C, D) The prediction performance of both the “Advanced” model and the “Basic” model was evaluated

across all classifiers on the test set using AUC-ROC (left) and the area under the Precision Recall Curve AUC-PR (right) metrics. For detailed model evaluation and performance analysis, see the Supplementary Figure S17-S18.

Table 3.1 Dynamic visualization of the selected gene structure datasets.

ID	ExonNum	UTR5length (base pairs)	Genelength (base pairs)	UTR3length (base pairs)	PanGenome_label
Zm00001eb000010	9	105	5588	1168	Near-Core Gene
Zm00001eb000020	9	849	5549	313	Core Gene
Zm00001eb000050	7	645	5829	0	Dispensable Gene
Zm00001eb000060	2	299	1023	364	Dispensable Gene
Zm00001eb000070	6	0	8641	0	Dispensable Gene
Zm00001eb000080	9	447	3132	730	Near-Core Gene
Zm00001eb000100	6	82	3105	641	Near-Core Gene
Zm00001eb000110	2	15	821	43	Dispensable Gene
Zm00001eb000120	1	0	628	268	Near-Core Gene

Gene structure datasets (exon number, five-prime UTR length, gene length, three-prime UTR length, and the “Pan-genome” categories) using the MFS’s “Data Table” option. Only ten rows are displayed per page.

3.7 References

Almagro Armenteros, J. J., C. K. Sonderby, S. K. Sonderby, H. Nielsen, and O. Winther. 2017. 'DeepLoc: prediction of protein subcellular localization using deep learning', *Bioinformatics*, 33: 3387-95.

Arendsee, Z., J. Li, U. Singh, A. Seetharam, K. Dorman, and E. S. Wurtele. 2019. 'phylostratr: a framework for phylostratigraphy', *Bioinformatics*, 35: 3617-27.

Aromolaran, O., T. Beder, M. Oswald, J. Oyelade, E. Adebiyi, and R. Koenig. 2020. 'Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features', *Comput Struct Biotechnol J*, 18: 612-21.

Babak Khorsand, Ehsan Sadeghnezhad, Javad Zahiri, Mohsen Sharif, Hassan Zare-mayvan. 2017. 'Stability Analysis in Differentially Expressed Genes'.

Bayer, P. E., A. A. Golicz, A. Scheben, J. Batley, and D. Edwards. 2020. 'Plant pan-genomes are the new reference', *Nat Plants*, 6: 914-20.

Benos, L., A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis. 2021. 'Machine Learning in Agriculture: A Comprehensive Updated Review', *Sensors (Basel)*, 21.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2007. 'GenBank', *Nucleic Acids Res*, 35: D21-5.

Bolduc, N., A. Yilmaz, M. K. Mejia-Guerra, K. Morohashi, D. O'Connor, E. Grotewold, and S. Hake. 2012. 'Unraveling the KNOTTED1 regulatory network in maize meristems', *Genes Dev*, 26: 1685-90.

Campos, T. L., P. K. Korhonen, A. Hofmann, R. B. Gasser, and N. D. Young. 2020. 'Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster', *NAR Genom Bioinform*, 2: lqaa051.

Campos, T. L., P. K. Korhonen, P. W. Sternberg, R. B. Gasser, and N. D. Young. 2020. 'Predicting gene essentiality in Caenorhabditis elegans by feature engineering and machine-learning', *Comput Struct Biotechnol J*, 18: 1093-102.

Christine Tranchant-Dubreuil, Mathieu Rouard, Francois Sabot. 2019. 'Plant Pangenome: Impacts On Phenotypes And Evolution', *Wiley Online Library*.

Dai, X., Z. Xu, Z. Liang, X. Tu, S. Zhong, J. C. Schnable, and P. Li. 2020. 'Non-homology-based prediction of gene functions in maize (Zea mays ssp. mays)', *Plant Genome*, 13: e20015.

Deshmukh, R., H. Sonah, G. Patil, W. Chen, S. Prince, R. Mutava, T. Vuong, B. Valliyodan, and H. T. Nguyen. 2014. 'Integrating omic approaches for abiotic stress tolerance in soybean', *Front Plant Sci*, 5: 244.

Dong, Z., Y. Xiao, R. Govindarajulu, R. Feil, M. L. Siddoway, T. Nielsen, J. E. Lunn, J. Hawkins, C. Whipple, and G. Chuck. 2019. 'The regulatory landscape of a core maize domestication module controlling bud dormancy and growth repression', *Nat Commun*, 10: 3810.

Ernst, J., and M. Kellis. 2017. 'Chromatin-state discovery and genome annotation with ChromHMM', *Nat Protoc*, 12: 2478-92.

Forestan, C., R. Aiese Cigliano, S. Farinati, A. Lunardon, W. Sanseverino, and S. Varotto. 2016. 'Stress-induced and epigenetic-mediated maize transcriptome regulation study by means of transcriptome reannotation and differential expression analysis', *Sci Rep*, 6: 30446.

Fukushima, A., M. Kusano, H. Redestig, M. Arita, and K. Saito. 2009. 'Integrated omics approaches in plant systems biology', *Curr Opin Chem Biol*, 13: 532-8.

Golicz, A. A., P. E. Bayer, G. C. Barker, P. P. Edger, H. Kim, P. A. Martinez, C. K. Chan, A. Severn-Ellis, W. R. McCombie, I. A. Parkin, A. H. Paterson, J. C. Pires, A. G. Sharpe, H. Tang, G. R. Teakle, C. D. Town, J. Batley, and D. Edwards. 2016. 'The pangenome of an agronomically important crop plant Brassica oleracea', *Nat Commun*, 7: 13390.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar. 2012. 'Phytozome: a comparative platform for green plant genomics', *Nucleic Acids Res*, 40: D1178-86.

Gordon, S. P., B. Contreras-Moreira, D. P. Woods, D. L. Des Marais, D. Burgess, S. Shu, C. Stritt, A. C. Roulin, W. Schackwitz, L. Tyler, J. Martin, A. Lipzen, N. Dochy, J. Phillips, K. Barry, K. Geuten, H. Budak, T. E. Juenger, R. Amasino, A. L. Caicedo, D. Goodstein, P. Davidson, L. A. J. Mur, M. Figueroa, M. Freeling, P. Catalan, and J. P. Vogel. 2017. 'Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure', *Nat Commun*, 8: 2184.

Gui, S., L. Yang, J. Li, J. Luo, X. Xu, J. Yuan, L. Chen, W. Li, X. Yang, S. Wu, S. Li, Y. Wang, Y. Zhu, Q. Gao, N. Yang, and J. Yan. 2020. 'ZEAMAP, a Comprehensive Database Adapted to the Maize Multi-Omics Era', *iScience*, 23: 101241.

Gundla, Naresh Kumar, and Zhengxin Chen. 2016. 'Creating NoSQL Biological Databases with Ontologies for Query Relaxation', *Procedia Computer Science*, 91: 460-69.

Hoopes, G. M., J. P. Hamilton, J. C. Wood, E. Esteban, A. Pasha, B. Vaillancourt, N. J. Provart, and C. R. Buell. 2019. 'An updated gene atlas for maize reveals organ-specific and stress-induced genes', *Plant J*, 97: 1154-67.

Horton, P., K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. 2007. 'WoLF PSORT: protein localization predictor', *Nucleic Acids Res*, 35: W585-7.

Hufford, M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, J. Liu, W. A. Ricci, T. Guo, A. Olson, Y. Qiu, R. Della Coletta, S. Tittes, A. I. Hudson, A. P. Marand, S. Wei, Z. Lu, B. Wang, M. K. Tello-Ruiz, R. D. Piri, N. Wang, D. W. Kim, Y. Zeng, C. H. O'Connor, X. Li, A. M. Gilbert, E. Baggs, K. V. Krasileva, J. L. Portwood, 2nd, E. K. S. Cannon, C. M. Andorf, N. Manchanda, S. J. Snodgrass, D. E. Hufnagel, Q. Jiang, S. Pedersen, M. L. Syring, D. A. Kudrna, V. Llaca, K. Fengler, R. J. Schmitz, J. Ross-Ibarra, J. Yu, J. I. Gent, C. N. Hirsch, D. Ware, and R. K. Dawe. 2021. 'De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes', *Science*, 373: 655-62.

Johnston, R., M. Wang, Q. Sun, A. W. Sylvester, S. Hake, and M. J. Scanlon. 2014. 'Transcriptomic analyses indicate that maize ligule development recapitulates gene expression patterns that occur during lateral organ initiation', *Plant Cell*, 26: 4718-32.

Kakumanu, A., M. M. Ambavaram, C. Klumas, A. Krishnan, U. Batlang, E. Myers, R. Grene, and A. Pereira. 2012. 'Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq', *Plant Physiol*, 160: 846-67.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes', *J Mol Biol*, 305: 567-80.

Li, P., W. Cao, H. Fang, S. Xu, S. Yin, Y. Zhang, D. Lin, J. Wang, Y. Chen, C. Xu, and Z. Yang. 2017. 'Transcriptomic Profiling of the Maize (Zea mays L.) Leaf Response to Abiotic Stresses at the Seedling Stage', *Front Plant Sci*, 8: 290.

Li, Y. H., G. Zhou, J. Ma, W. Jiang, L. G. Jin, Z. Zhang, Y. Guo, J. Zhang, Y. Sui, L. Zheng, S. S. Zhang, Q. Zuo, X. H. Shi, Y. F. Li, W. K. Zhang, Y. Hu, G. Kong, H. L. Hong, B. Tan, J. Song, Z. X. Liu, Y. Wang, H. Ruan, C. K. Yeung, J. Liu, H. Wang, L. J. Zhang, R. X. Guan, K. J. Wang, W. B. Li, S. Y. Chen, R. Z. Chang, Z. Jiang, S. A. Jackson, R. Li, and L. J. Qiu. 2014. 'De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits', *Nat Biotechnol*, 32: 1045-52.

Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. 2003. 'Protein disorder prediction: implications for structural proteomics', *Structure*, 11: 1453-9.

Liu, H., F. Wang, Y. Xiao, Z. Tian, W. Wen, X. Zhang, X. Chen, N. Liu, W. Li, L. Liu, J. Liu, J. Yan, and J. Liu. 2016. 'MODEM: multi-omics data envelopment and mining in maize', *Database (Oxford)*, 2016.

Liu, Y., H. Du, P. Li, Y. Shen, H. Peng, S. Liu, G. A. Zhou, H. Zhang, Z. Liu, M. Shi, X. Huang, Y. Li, M. Zhang, Z. Wang, B. Zhu, B. Han, C. Liang, and Z. Tian. 2020. 'Pan-Genome of Wild and Cultivated Soybeans', *Cell*, 182: 162-76 e13.

Lloyd, John P., Alexander E. Seddon, Gaurav D. Moghe, Matthew C. Simenc, and Shin-Han Shiu. 2015. 'Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes', *The Plant Cell*, 27: 2133-47.

Lyons, E., and M. Freeling. 2008. 'How to usefully compare homologous plant genes and chromosomes as DNA sequences', *Plant J*, 53: 661-73.

Makarevitch, I., A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch, J. Ross-Ibarra, and N. M. Springer. 2015. 'Transposable elements contribute to activation of maize genes in response to abiotic stress', *PLoS Genet*, 11: e1004915.

McCarty, D. R., S. Latshaw, S. Wu, M. Suzuki, C. T. Hunter, W. T. Avigne, and K. E. Koch. 2013. 'Mu-seq: sequence-based mapping and identification of transposon induced mutations', *PLoS One*, 8: e77172.

Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli. 2005. 'The microbial pan-genome', *Curr Opin Genet Dev*, 15: 589-94.

Mejia-Guerra, M. K., W. Li, N. F. Galeano, M. Vidal, J. Gray, A. I. Doseff, and E. Grotewold. 2015. 'Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites', *Plant Cell*, 27: 3309-20.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. 2021. 'Pfam: The protein families database in 2021', *Nucleic Acids Res*, 49: D412-D19.

Morneau, Dominique. 2021. 'Pan-genomes: moving beyond the reference'.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*.

Oka, R., J. Zicola, B. Weber, S. N. Anderson, C. Hodgman, J. I. Gent, J. J. Wesselink, N. M. Springer, H. C. J. Hoefsloot, F. Turck, and M. Stam. 2017. 'Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize', *Genome Biol*, 18: 137.

Opitz, N., A. Paschold, C. Marcon, W. A. Malik, C. Lanz, H. P. Piepho, and F. Hochholdinger. 2014. 'Transcriptomic complexity in young maize primary roots in response to low water potentials', *BMC Genomics*, 15: 741.

Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen. 2011. 'SignalP 4.0: discriminating signal peptides from transmembrane regions', *Nat Methods*, 8: 785-6.

Quinlan, A. R., and I. M. Hall. 2010. 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26: 841-2.

Rajasundaram, D., and J. Selbig. 2016. 'More effort - more results: recent advances in integrative 'omics' data analysis', *Curr Opin Plant Biol*, 30: 57-61.

Ricci, W. A., Z. Lu, L. Ji, A. P. Marand, C. L. Ethridge, N. G. Murphy, J. M. Noshay, M. Galli, M. K. Mejia-Guerra, M. Colome-Tatche, F. Johannes, M. J. Rowley, V. G. Corces, J. Zhai, M. J. Scanlon, E. S. Buckler, A. Gallavotti, N. M. Springer, R. J. Schmitz, and X. Zhang. 2019. 'Widespread long-range cis-regulatory elements in the maize genome', *Nat Plants*, 5: 1237-49.

Schnable, J. C., and M. Freeling. 2011. 'Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize', *PLoS One*, 6: e17855.

Singh, A., B. Ganapathysubramanian, A. K. Singh, and S. Sarkar. 2016. 'Machine Learning for High-Throughput Stress Phenotyping in Plants', *Trends Plant Sci*, 21: 110-24.

Stelpflug, S. C., R. S. Sekhon, B. Vaillancourt, C. N. Hirsch, C. R. Buell, N. de Leon, and S. M. Kaeppler. 2016. 'An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development', *Plant Genome*, 9.

Tao, Y., X. Zhao, E. Mace, R. Henry, and D. Jordan. 2019. 'Exploring and Exploiting Pan-genomics for Crop Improvement', *Mol Plant*, 12: 156-69.

Tello-Ruiz, M. K., S. Naithani, P. Gupta, A. Olson, S. Wei, J. Preece, Y. Jiao, B. Wang, K. Chougule, P. Garg, J. Elser, S. Kumari, V. Kumar, B. Contreras-Moreira, G. Naamati, N. George, J. Cook, D. Bolser, P. D'Eustachio, L. D. Stein, A. Gupta, W. Xu, J. Regala, I. Papatheodorou, P. J. Kersey, P. Flicek, C. Taylor, P. Jaiswal, and D. Ware. 2021. 'Gramene 2021: harnessing the power of comparative genomics and pathways for plant research', *Nucleic Acids Res*, 49: D1452-D63.

van Dijk, Aalt Dirk Jan, Gert Kootstra, Willem Kruijer, and Dick de Ridder. 2021. 'Machine learning in plant science and plant breeding', *iScience*, 24: 101890.

Vollbrecht, E., J. Duvick, J. P. Schares, K. R. Ahern, P. Deewatthanawong, L. Xu, L. J. Conrad, K. Kikuchi, T. A. Kubinec, B. D. Hall, R. Weeks, E. Unger-Wallace, M. Muszynski, V. P. Brendel, and T. P. Brutnell. 2010. 'Genome-wide distribution of transposed Dissociation elements in maize', *Plant Cell*, 22: 1667-85.

Waese-Perlman, Ben, Asher Pasha, Chantal Ho, Amirahmad Azhieh, Yushan Liu, Alexander Sullivan, Vincent Lau, Eddi Esteban, Jamie Waese, George Ly, Cornelia Hooper, S. Evan Staton, Nicholas Brereton, Cuong Le, Rex Nelson, Shelley Lumba, David Goodstein, A. Harvey Millar, Isobel Parkin, Lewis Lukens, Juergen Ehlting, Loren Rieseberg, Frédéric Pitre, Anne Brown, and Nicholas J. Provart. 2021. 'ePlant in 2021: New Species, Viewers, Data Sets, and Widgets', *bioRxiv*.

Walley, J. W., R. C. Sartor, Z. Shen, R. J. Schmitz, K. J. Wu, M. A. Urich, J. R. Nery, L. G. Smith, J. C. Schnable, J. R. Ecker, and S. P. Briggs. 2016. 'Integration of omic networks in a developmental atlas of maize', *Science*, 353: 814-8.

Wang, S., I. Pandis, C. Wu, S. He, D. Johnson, I. Emam, F. Guitton, and Y. Guo. 2014. 'High dimensional biological data retrieval optimization with NoSQL technology', *BMC Genomics*, 15 Suppl 8: S3.

Wang, W., R. Mauleon, Z. Hu, D. Chebotarov, S. Tai, Z. Wu, M. Li, T. Zheng, R. R. Fuentes, F. Zhang, L. Mansueto, D. Copetti, M. Sanciangco, K. C. Palis, J. Xu, C. Sun, B. Fu, H. Zhang, Y. Gao, X. Zhao, F. Shen, X. Cui, H. Yu, Z. Li, M. Chen, J. Detras, Y. Zhou, X. Zhang, Y. Zhao, D. Kudrna, C. Wang, R. Li, B. Jia, J. Lu, X. He, Z. Dong, J. Xu, Y. Li, M. Wang, J. Shi, J. Li, D. Zhang, S. Lee, W. Hu, A. Poliakov, I. Dubchak, V. J. Ulat, F. N. Borja, J. R. Mendoza, J. Ali, J. Li, Q. Gao, Y. Niu, Z. Yue, M. E. B. Naredo, J. Talag, X. Wang, J. Li, X. Fang, Y. Yin, J. C. Glaszmann, J. Zhang, J. Li, R. S. Hamilton, R. A. Wing, J. Ruan, G. Zhang, C. Wei, N. Alexandrov, K. L. McNally, Z. Li, and H. Leung. 2018. 'Genomic variation in 3,010 diverse accessions of Asian cultivated rice', *Nature*, 557: 43-49.

Warman, C., K. Panda, Z. Vejlupkova, S. Hokin, E. Unger-Wallace, R. A. Cole, A. M. Chettoor, D. Jiang, E. Vollbrecht, M. M. S. Evans, R. K. Slotkin, and J. E. Fowler. 2020. 'High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements', *PLoS Genet*, 16: e1008462.

Woodhouse, M. R., E. K. Cannon, J. L. Portwood, 2nd, L. C. Harper, J. M. Gardiner, M. L. Schaeffer, and C. M. Andorf. 2021. 'A pan-genomic approach to genome databases using maize as a model system', *BMC Plant Biol*, 21: 385.

Woodhouse, M. R., S. Sen, D. Schott, J. L. Portwood, M. Freeling, J. W. Walley, C. M. Andorf, and J. C. Schnable. 2021. 'qTeller: A tool for comparative multi-genomic gene expression analysis', *Bioinformatics*.

Xiao, N., D. S. Cao, M. F. Zhu, and Q. S. Xu. 2015. 'protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences', *Bioinformatics*, 31: 1857-9.

Yocca, A. E., and P. P. Edger. 2021. 'Machine learning approaches to identify core and dispensable genes in pangenomes', *Plant Genome*: e20135.

Yu, J., A. A. Golicz, K. Lu, K. Dossa, Y. Zhang, J. Chen, L. Wang, J. You, D. Fan, D. Edwards, and X. Zhang. 2019. 'Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars', *Plant Biotechnol J*, 17: 881-92.

Yuan, Y., Y. Xu, J. Xu, R. L. Ball, and H. Liang. 2012. 'Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data', *Bioinformatics*, 28: 1246-52.

Zhao, W., P. Canaran, R. Jurkuta, T. Fulton, J. Glaubitz, E. Buckler, J. Doebley, B. Gaut, M. Goodman, J. Holland, S. Kresovich, M. McMullen, L. Stein, and D. Ware. 2006. 'Panzea: a database and resource for molecular and functional diversity in the maize genome', *Nucleic Acids Res*, 34: D752-7.

Zheng, D., R. Wang, Q. Ding, T. Wang, B. Xie, L. Wei, Z. Zhong, and B. Tian. 2018. 'Cellular stress alters 3'UTR landscape through alternative polyadenylation and isoform-specific degradation', *Nat Commun*, 9: 2268.

Zhu, Minfeng, and Jie Dong. 2016. 'rDNAse : R package for generating various numerical representation schemes of DNA sequences'.

Zogli, P., L. Pingault, S. Grover, and J. Louis. 2020. 'Ento(o)mics: the intersection of 'omic' approaches to decipher plant defense against sap-sucking insect pests', *Curr Opin Plant Biol*, 56: 153-61.

3.8 Appendix A. Notes

3.8.1 Data availability statement

Project name: Maize Feature Store (MFS); Project home page: MFS is freely available on

GitHub at https://github.com/shatabdi123/MFS_Application Web version of MFS is available at

https://mfs.maizegdb.org/. The dataset for MFS can also be accessed on Kaggle: https://kaggle.com/datasets/332177dbd2271966f2291640acf6f7057bde915d939b3bf67545a5f24a0e3fe3. Programming language: Python, R, JavaScript, HTML, CSS; Other requirements: Flask 1.1.2 or higher. The application is platform independent.

3.8.2 Acknowledgements

We thank the research groups of Iowa State University and USDA-ARS, Corn Insects and Crop Genetics Research Unit and Dr. Rita Hayford for their constructive feedback, which has contributed to the improvement of our platform.

3.8.3 Author contributions

Conceptualization, S.S., C.A.; Methodology, S.S.; Data Generation, S.S., M.W.; Formal Analysis, S.S.; Data Interpretation, S.S., C.A.; Software Validation, S.S., J.P; Analysis Validation, S.S.; Writing – Original Draft Preparation, S.S., C.A.; Writing – Review and Editing, S.S., M.W., J.P., and C.A.; Supervision, C.A.; Funding Acquisition, C.A.

3.8.4 Declaration of interests

The authors declared no competing interest.

3.8.5 Funding

This research was supported by the US. Department of Agriculture, Agricultural Research Service, Project Number [5030-21000-068-00-D] through the Corn Insects and Crop Genetics Research Unit in Ames, Iowa. This material is based upon work supported by the Department of Agriculture, Agricultural Research Service under Agreement No. 58-5030-0-036 [Iowa State Award: 022172- 00001 to J.W.W.]. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer. Conflict of Interest: none declared.

3.9 Appendix B: Supplementary tables and figures

All supplementary tables and figures can be found online at Sen et al., 2023: https://mc.manuscriptcentral.com/database.

CHAPTER 4. PREDICTING GENES ASSOCIATED WITH BIOTIC OR ABIOTIC STRESS ACROSS DIFFERENT MAIZE LINES AND RELATED SPECIES

Shatabdi Sen[1] , Rita Hayford[2] , and Carson M. Andorf[2,3*]

[1]Department of Plant Pathology & Microbiology, Iowa State University, Ames, IA 50011, USA

[2]USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

[3]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript to be submitted in *Bioarchive*

4.1 Abstract

Maize is a crop that is highly susceptible to various biotic and abiotic stresses throughout its growth cycle, particularly during the developmental stage and before flowering. To develop stress-resistant and high-yielding crop varieties, it is crucial to understand the molecular mechanisms and identify stress-responsive genes that control plant responses to these stresses. Genome-wide association studies (GWAS) of the maize reference genome, B73, have revealed numerous stress-responsive genes in this model species. However, functional genomics results can sometimes be ambiguous, and sequence similarity alone is not always reliable for identifying stress genes. Transcriptome profiling studies have provided insights into the molecular mechanisms underlying stress response, but these studies have been limited to a subset of species. Additionally, the features underlying stress-associated genes are poorly understood, making computational prediction challenging. To address these challenges, we employed several approaches based on machine learning classification algorithms and evaluated their ability to accurately predict stress-responsive genes based on non-homology gene features. In this study, comprehensive omics datasets from the Maize Feature Store were harnessed, and a machine-learning-based workflow was applied to predict stress-responsive genes in maize. We discovered

strong predictors of stress-responsive genes in the reference genome B73, which can pave the way for computational predictions of stress-responsive genes in other maize accessions and non-model plant species. The models trained on gene expression datasets as target labels outperformed the models trained on GWAS datasets. Furthermore, the results showed that models trained on generic RNA-Seq fold change cutoffs for annotating target labels as stress-responsive or non-responsive genes performed better than models trained on stringent cutoffs. The findings highlight the importance of considering gene expression datasets and using generic cutoffs for accurate classification of stress-responsive genes. This research contributes to the advancement of stress-responsive gene prediction and has implications for improving crop resilience and productivity in the face of biotic and abiotic stresses.

4.2 Introduction

Maize is one of the most widely cultivated crops globally. Its production is not only integral to food security, but it is also pivotal in biofuel production (Gong et al. 2014). It is the third most significant crop grown after wheat and rice (Assem 2015). The demand for maize is increasing, especially in developing nations, due to the rising consumption of animal and human food (Grote et al. 2021), therefore, maize production needs to dramatically increase to meet future demand. However, the yield and growth of maize are influenced by various factors, including changing climate conditions and different types of stress, both biotic and abiotic (Hemathilake and Gunathilake 2022). Under natural conditions, plants undergo different phases to complete their life cycle. In recent years, climatic parameters such as precipitation and temperature have become more unpredictable and extreme, resulting in prolonged droughts and changes in temperature beyond the optimal state (Habib-Ur-Rahman et al. 2022). Such changes have posed significant challenges to crop production. The susceptibility of plants to abiotic stress poses several challenges to sustaining an increase in crop production with changing climatic

patterns (Raza et al. 2019). Abiotic stress refers to the negative impact of non-living factors, such as drought, high temperature, salinity, and nutrient deficiency, on plant growth and development (Gull, Lone, and Wani 2019). Biotic stress on the other hand is imposed by pathogens, including fungal, bacterial, and viral, and can cause heavy damage leading to yield reduction in plants (Nazari et al. 2023). Some examples of prevalent maize diseases are northern corn leaf blight, ear rot, maize rough dwarf disease and sugarcane mosaic disease. Maize is also plagued by pests, including stem borer, pink borer, shoot fly, termites and the storage pest maize weevil (Gong et al. 2014). The origin of new pathogens and insect races due to climatic and genetic factors is a major challenge for plant breeders in breeding biotic stress resistant crops. These stressors can disrupt normal physiological processes in plants, leading to reduced growth, yield, and overall productivity. Approximately 10% of the global maize yield is lost each year as a result of biotic stresses (Jakhar and Singh 2015).

Therefore, the identification of resistant genes paves the way to the development of disease-resistant cultivars and is essential for reliable production in maize and other plant species. The rapid acceleration in genome sequencing is providing complete sequences for dozens of new plant species each year (Henry 2022). Thus, this transformative capability has paved the way for the application of comparative genomics in the identification of stress-responsive genes in plants. This methodology involves the systematic comparison of genomic sequences across different plant species, enabling the identification of conserved regions and variations associated with stress tolerance. However, a significant challenge of homology-based functional annotation is that these annotations are often propagated from one sequence to the next without associated data on provenance (Dai and Shen 2022). Another potential problem is, despite shared evolutionary history, a gene responding transcriptionally to cold stress in one

species may not be a good predictor of whether syntenic orthologous genes in related species would also respond to cold stress in the same treatment at the same developmental stage (Meng et al. 2021). This low conservation of transcriptional responses across conserved genes in related species is consistent with the results of a previous comparison of the transcriptional responses of maize and sorghum to stress (Zhang et al. 2017) and the variation in transcriptional responses to stress between different alleles of the same gene in maize (Gahlaut et al. 2016).

In addition to employing sequence similarity, clustering methodologies also serve as a valuable tool for functional assignment of stress genes (Chiu and Ong 2022). A common strategy involves correlation analysis, utilizing Pearson or Spearman correlation based on gene expression data, coupled with subsequent clustering. This method groups genes with similar expression patterns, allowing for the inference of functional assignment by assessing if annotated genes within a cluster are enriched for recognized stress-related biological functions. The assignment of genes to specific stress-related clusters relies on their correlation strength. However, it's essential to acknowledge that these approaches, grounded in linear or monotonic relationships, may encounter challenges in capturing the intricate spatio-temporal dynamics inherent in the complex relationships among stress-responsive genes (Oyelade et al. 2016).

Traditional wet lab approaches have provided valuable direct insights into stress gene identification. However, these methods have certain limitations that can hinder a comprehensive understanding of stress gene regulation. Direct gene expression measurement and valuable information about stress-responsive genes are possible using wet lab techniques like RNA sequencing and qPCR, but these approaches have certain limitations (Han et al. 2015). Firstly, wet lab approaches may overlook low-abundance genes that are difficult to detect using conventional methods. Secondly, these techniques often lack temporal dynamics, providing a

static picture of gene expression at a specific time point. Additionally, the limited availability of RNA-seq datasets for various stress conditions poses a significant challenge. The scarcity of such datasets stems from the time-consuming and costly nature of RNA-seq experiments, impeding comprehensive exploration and analysis of stress-responsive genes across diverse conditions. Lastly, inherent biases and limitations associated with individual wet lab techniques can affect the accuracy and robustness of stress-responsive gene identification.

To overcome the limitations of existing bioinformatics tools and wet lab approaches, the integration of machine learning with multiomics datasets has emerged as a crucial strategy for stress-responsive gene identification. Machine learning (ML) algorithms can handle large-scale data with diverse molecular information, allowing for the integration of multiple omics datasets, such as transcriptomics, proteomics, and epigenomics (Feldner-Busztin et al. 2023). This integration enables a more comprehensive analysis of stress-responsive gene regulation, capturing the complex interactions and regulatory networks involved. In recent years, researchers have increasingly utilized machine learning algorithms to gain insights into plant phenotyping, gene function prediction, and molecular studies (Mahood, Kruse, and Moghe 2020; Danilevicz et al. 2022). These applications have revolutionized the field by enabling the analysis of large-scale datasets and providing accurate predictions and classifications. ML is making significant strides in plant breeding. Its application in maize, for instance, in classifying DNA sequence regions into active genes and pseudogenes based on features like DNA methylation (Niederhuth and Schmitz 2017). ML is has also been employed to predict crossover regions in the plant genome, where genetic material exchanges between paternal and maternal genomes (van Dijk et al. 2021). Furthermore, ML has gained traction in plant population genetics, exemplified by predicting genomic regions influenced by natural selection (Schrider and Kern 2018). These applications

highlight ML's evolving role in plant breeding, extending beyond traditional gene annotation in newly sequenced genomes. ML, with its emphasis on predictive patterns, serves as a valuable complement to conventional comparative omics approaches in exploring and understanding plant genome function.

In this paper, we explored the integration of machine learning with multiomics datasets as a strategy to address the challenges of wet lab approaches as well as existing bioinformatics tools and pipelines. The characterization of stress-responsive genes in terms of their biophysical and biochemical properties have remained elusive and limited to certain features through direct or indirect observations (Latorre et al. 2022). A systematic and comprehensive study of the features characterizing stress-responsive genes has not yet been performed. This is an important step towards understanding gene expression regulation and improving the tools used in plant biology and biotechnology (Poljsak and Milisav 2012). Hence, here we create a computational model that will predict stress-responsive genes associated with abiotic and biotic stresses in maize (*Zea Mays*) as well as across other maize genome assemblies (e.g., maize inbred lines W22 and Zea mays ssp. mexicana L (TIL-18, TIL-25) (Lu et al. 2017)) by performing meta-analyses of a comprehensive set of multi-omics datasets. Broadly speaking, the workflow will provide a framework that yields insight into the possible characteristics of specific genes and the role they play in response to different environmental stimuli.

4.3 Materials and Methods

4.3.1 Definition of stress-responsive genes

The catalog of stress-responsive genes in Maize B73 was defined using differential expression analysis, performed with DESeq2 (Madzima et al. 2021) (v1.26.0) and genome-wide association studies (GWAS) (Challa and Neelapu 2018).

4.3.1.1 GWAS based definition of target labels

Genome-wide association studies (GWAS) are a powerful tool for investigating multiple or complex traits related to any single/multiple stress. GWAS on various plants/crops have identified novel gene candidates, or genes or quantitative trait loci, responsible for abiotic stress and biotic stress.

The GWAS based labeling of genes was carried out based on clean, processed GWAS dataset from MaizeGDB. The first group of GWAS data from MaizeGDB (Woodhouse, Cannon, et al. 2021) aggregated genome-wide association mapping for 41 different phenotypes to 38,421 SNPs from the Maize Hapmap1 and Hapmap2 projects. Individual SNPs are present with the trait(s) that segregates with that SNP. The original data was from (Wallace et al. 2014).

Association mapping across numerous traits reveals patterns of functional variation in maize. This study used the maize Nested Association Mapping (NAM) population and nearly 30 million segregating variants to identify variants that were significantly associated with at least one phenotype. The phenotypes cover various plant architecture, developmental, disease resistance traits and 12 different metabolites in leaves.

The data was originally mapped to the B73 RefGen_v2 genome. MaizeGDB used the 50bp upstream and downstream flanking sequences for each SNP (in RefGen_v2) from the RefGen_v2 SNP genomic coordinates (retrieved using bedtools getfasta) as query sequences aligned to B73 RefGen_v5 and the 25 NAM founder lines. This data included the top syntenic hit (from blastn, -evalue 0.00001 and dagchainer, -D 10000000 -g 10000 -A 3 -e -0f -x 1 -E 0.1 -M 100) for each of the query sequences.

Reference SNP (RS) identifiers were assigned based on coordinate positions for the B73 RefGen_v4 SNPs downloaded from the European Variation Archive (March 2020). The GWAS SNP data was remapped to RefGen_v4. MaizeGDB using the 100bp upstream and downstream

flanking sequences (retrieved using bedtools getfasta) for each SNP (in RefGen_v2). The data included the top hit (from blastn, -evalue 0.00001) where a GWAS SNP mapped to the same position as an EVA SNP (with a RS ID) in v4.

Another set of GWAS dataset associated with multiple phenotypes were retrieved from MaizeGDB describes genome-wide association mapping (GWAS) from 133 papers covering 531 studies for 279 traits across ~42,000 loci overlapping ~8,400 genes (Portwood et al. 2019). The data was compiled and remapped to B73_v4 at the GWAS Atlas database by the National Genomics Data Center at the Chinese Academy of Sciences (Tian et al. 2011). Each GWAS hit includes trait name, tissue, position, allele, p-value, R^2 value, RS number, RS reference allele, RS alternate allele, publication information, and PubMed link.

All of the original coordinates for B73 RefGen_v2, v3, and v4 have been remapped to B73 RefGen_v5. MaizeGDB used the 100bp upstream and downstream flanking sequences (retrieved using bedtools getfasta) for each SNP (in RefGen_v4) from the RefGen_v4 SNP genomic coordinates from the GWAS Atlas database (retrieved using bedtools getfasta) as query sequences aligned to B73 RefGen_v5 with at least 98% coverage and 98% sequence identity. Only the top hit for each of those query sequences were considered. If multiple top hits exist, the hit nearest the original location is chosen.

BLAST was used for the mapping (example command syntax: "blastn -db Zm-B73-REFERENCE-NAM-5.0.fa -query flanking_sequences.fasta -perc_identity 98 -qcov_hsp_perc 98").

For each GWAS SNP throughout the genome, genes were annotated as stress or non-stress based on the presence or absence of SNPs within three distinct genomic regions: one using the gene body, defined as the region from the annotated transcription start site to the annotated

transcription stop site, a second for the upstream and downstream region, defined as a 1 KB (kilobase) region directly upstream and downstream of the transcription start site and transcription stop site respectively, and a third for the upstream and downstream region, defined as a 5 KB (kilobase) region directly upstream and downstream of the transcription start site and transcription stop site respectively.

4.3.1.2 RNA-Seq based definition of target labels

The RNA-Seq based classification of genes into stress-responsive or non-responsive categories was determined by comparing the expression levels of the genes in treatment versus control samples. This categorization relied on data aligned to high-quality RNA-Seq expression reads, mapped specially to the most recent version of the B73 reference genome (B73v5). Twenty-five (25) quality RNA-Seq datasets from published RNA-Seq studies related to biotic and abiotic stress generated from tissues of the B73 cultivar were used in this analysis (Table 4.1). The RNA-Seq studies collected for this study captured various types of abiotic stress factors including drought, heat, cold, salinity, waterlogging, nitrogen, cadmium, phosphate, nitrate, ammonium and elevated ozone (UV). The biotic datasets were generated from B73 inoculated with pathogens such as *Cercospora Zeina* (causal agent of gray leaf spot), *Fusarium graminearum* (causal agent of Gibberella stalk rot), *Fusarium venenatum*, *Colletotrichum graminearum*, Sugarcane Mosaic Virus, Mites herbivores and Weed stress.

To assess the model's effectiveness on different maize inbred lines, we utilized three additional high-quality RNA-Seq datasets. These datasets were sourced from published studies on abiotic stress (specifically drought and cold stress) and were derived from the tissues of the maize inbred line W22 and the cultivar Zea mays ssp. mexicana L. (specifically, TIL-18 and TIL-25) (Lu et al. 2017).

For our analysis, two sets of RNA-Seq based target labels were generated. Read counts were utilized to discern stress-responsive genes by contrasting gene expression in treatment vs. control samples. In the first set, differentially expressed genes were specified as those with an adjusted P value < 0.05 and an absolute log2 fold change ≥1 and <-1, as determined by EdgeR (Robinson, McCarthy, and Smyth 2010). Non-responsive genes were characterized as those not meeting the expression criteria between treatment and control values at all time points and conditions.

Meanwhile, the second set of RNA-Seq based stress-responsive gene labels employed a more stringent cutoff [|log2(fold change)| ≥ 2 and < -2, p-value < 0.05], ensuring a refined selection of stress-responsive genes based on a higher threshold for fold change and statistical significance.

4.3.1.3 Unified approach of defining target labels

In adopting a unified approach to designate genes as stress-responsive or non-responsive, the synergistic strengths of both GWAS and RNA-Seq studies were harnessed. This strategy aimed to capitalize on the distinct advantages offered by each method. Genes were meticulously annotated as stress-responsive when they were identified as such by either GWAS and DESeq analyses [|log2(fold change)| ≥ 1 and < -1 , p-value < 0.05]. By integrating the results from both genomic and transcriptomic analyses, this unified labeling approach ensured a robust and multi-faceted identification of genes exhibiting stress responsiveness.

4.3.2 Preparing features for predictive analysis

Feature generation is the process of transforming raw, unstructured data into a set of features that describes and represents the diverse attributes of the input data, often for statistical analysis or classification purposes. This process is performed after data collection and integration. In stress-responsive gene prediction, the input data are a set of omics features

associated with the genes transformed into numerical representations (features) and passed to a classifier that is expected to classify as either stress-responsive or non-responsive. This set of features can be broadly categorized as intrinsic and extrinsic features as defined in the Maize Feature Store (Sen et al. 2023). We define intrinsic features as features that can be directly derived from gene and protein sequences without association or comparison with another sequence; examples include gene sequence, protein sequence and codon usage features. Features are extrinsic if they are computed from the sequence's interaction with another sequence or its environment. Examples are localization, which estimates the probability of a gene to reside in a particular compartment within the cell; topology, which computes the degree of interaction among genes or proteins.

The intrinsic and extrinsic set of predictive features were divided into seven categories: Sequence, Gene Model Structure, Gene and Protein Expression, Chromatin, Count, Variomic , and Evolutionary features. All these intrinsic and extrinsic sets of gene features were retrieved from the Maize Feature Store for Maize B73 (reference genome) and were generated for the NAM founder lines as detailed in the Maize Feature Store.

4.3.3 Stress features filtering for modeling

The focus of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results (Chandrashekar and Sahin 2014). The standardized maize multi-omics data can contain thousands of variables of which many of them could be highly correlated with other variables (e.g., when two features are perfectly correlated, only one feature is sufficient to describe the data). The dependent variables provide no extra information about the classes and thus serve as noise for the predictor. This means that the total information content can be obtained from fewer unique features which contain maximum discrimination information

about the classes. Hence by eliminating the dependent variables, the amount of data can be reduced which can improve the stress classification performance. By applying feature selection techniques, we can gain some insight into the process and can improve the computation requirement and prediction accuracy.

There are a lot of ways of performing feature selection, but most feature selection methods can be divided into three major types: Filter-based, Wrapper-based and Embedded. Filter methods act as preprocessing to rank the features wherein the highly ranked features are selected and applied to a predictor. In wrapper methods the feature selection criterion is the performance of the predictor i.e., the predictor is wrapped on a search algorithm which will find a subset which gives the highest predictor performance. Embedded methods (Guyon and Elisseeff 2003) include variable selection as part of the training process without splitting the data into training and testing sets. Therefore, here we proposed to use a vote-based approach to take advantage of the benefits of each one of the feature selection techniques. We applied a variety of feature selection methods from each of the three-feature selection techniques. The chi-squared test was opted for the filter-based feature selection method. The wrapper-based feature selection techniques used were Recursive Feature Elimination, Recursive Feature Elimination with Cross-Validation and finally the embedded methods used were Random Forest-based feature selection, L1-based feature selection, typically associated with LASSO regression and Extra Trees-based feature selection to pick the top variables and assign a vote for each variable chosen. At the end, we calculated the total votes for each variable chosen and then chose the best features based on majority voting.

4.3.4 Stress feature modeling and hyperparameter tuning

For classification methods, we used seven distinct machine-learning algorithms, namely: (1. Logistic Regression, 2. Random Forest Classifier, 3. Gradient Boosting Classifier, 4. Extra

Trees Classifier, 5. KNeighborsClassifier and 6. XGB Classifier). In the initial steps of our methodology, we focused on preparing the dataset for our classification tasks. Data standardization was paramount, involving the scaling of numerical features to a standardized range, with a mean of 0 and a standard deviation of 1. This crucial preprocessing step ensured that each feature contributed fairly to the subsequent machine learning models, preventing biases that might arise from varying feature scales.

Moreover, recognizing the challenges posed by our imbalanced stress gene datasets, where one of our classes (non-responsive genes) was underrepresented compared to the other (stress-responsive genes), we implemented strategies for data balancing. Techniques such as oversampling the minority class and Synthetic Minority Over-sampling Technique (smote) or undersampling the majority class were employed to create a more balanced representation, mitigating potential biases in favor of the majority class.

Furthermore, to enhance the generalizability of our models, we incorporated two different hyperparameter optimization approaches: Random Search and Grid Search. Random Search involves randomly selecting combinations of hyperparameters from a predefined search space, allowing for an efficient exploration of a wide range of possibilities. On the other hand, Grid Search systematically explores a manually specified subset of hyperparameter combinations, providing a more exhaustive evaluation of potential configurations.

4.3.5 Model evaluation

To ascertain stress-responsive genes through computational methods, it is essential to validate the model's predictions and assess its effectiveness. The accuracy of predictions must be verified to ensure the computational approach aligns with experimental methods. Evaluation occurs on unseen data rather than the training set, adhering to standard practices for reliable machine learning model assessments. For our stress-responsive gene classification, we employed

the ten-fold cross-validation technique, a robust method. This process entails randomly dividing the training data into ten groups, constructing the model with nine of them, and evaluating it on the tenth group. The performance is recorded, and this cycle repeats for the remaining groups, providing a comprehensive evaluation of the model's effectiveness.

Since we are addressing a binary classification task, determining whether a gene is stress-responsive or non-responsive, we rely on standard binary classification evaluations metrics to evaluate the model's performance. The metrics included for our analysis are Accuracy, Precision, Sensitivity (Recall), and F1-Score. The evaluation process involved fundamental parameters like True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). TP signifies a correct Positive prediction, FP represents an incorrect Positive prediction, TN corresponds to a correct Negative prediction, and FN indicates an incorrect Negative prediction.

In the context of evaluating binary classification models, precision, recall, and F1-score are metrics that highlight the positive class of interest, often neglecting the negative class. Accuracy, on the other hand, can be deceptive while dealing with imbalanced dataset. We incorporated graphical methods such as the Area Under Receiver Operating Characteristic (ROC-AUC) curve and the Area Under Precision-Recall curves (PR-AUC) for our model assessment. The ROC-AUC curve visually demonstrates the trade-off between the true positive rate (sensitivity) and false positive rate (1—specificity) at different thresholds. It served as a tool to select optimal binary classifiers independently of class distribution, with scores ranging from 0 to 1 across different stress-responsive gene classifiers. The Precision-Recall curves (PR-AUC) depict the trade-off between the true positive rate and positive predictive value, offering more informative insights, particularly in scenarios of imbalanced datasets, as observed in our stress-responsive gene classification.

4.3.6 Deciphering feature significance: Interpretable AI

Understanding the decisions of complex AI models in predicting stress-related genes is vital. To unravel these complexities, we have employed two robust Python frameworks: Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). LIME is an innovative explanation technique that interprets classifier predictions faithfully and understandably by creating a local interpretable model around the prediction. The process involves generating random datapoints near the target data point, with closer points receiving more weight. The outcomes from these datapoints, predicted by the complex model, serve as ground truth, while predictions from simpler models are considered. The final step minimizes differences between complex and simple model outcomes, with a weighted emphasis on instances close to our target data point.

On the other hand, SHAP is a game-theoretic approach based on Shapley values, quantifying each feature's contribution to the model's outcome. Starting with the expected value as the base, SHAP values reveal how each feature contributes to the predicted probability relative to the base value. While initially a method for local interpretation, SHAP can also be employed for global interpretation by aggregating SHAP values across multiple prediction cases.

This comprehensive approach, encapsulating intricate details of the dataset, model architecture, and interpretation results, significantly contributes to the scholarly understanding of key features in stress-responsive gene classification within the expansive domain of AI-driven genomic analysis.

4.4 Results

4.4.1 Comparing model performance across diverse feature combinations

Distinct subsets of features, along with the RNA-Seq determined criteria for target labels (where stress-responsive genes were defined by an adjusted P value < 0.05 and an absolute log2 fold change ≥1 and <-1), were utilized to conduct separate predictions. The goal was to assess the varying contributions of different feature types to the overall accuracy of predicting stress-responsive genes through the construction of diverse machine learning models. The study highlighted the best-performing model for each feature subset, as depicted in Figure 4.1.

Models trained exclusively on gene model structure features, encompassing parameters such as gene length, number of isoforms, exon length, average exon length, number of exons per gene, coding sequence length, five prime untranslated regions (UTR) length, and three-prime UTR length, exhibited strong performance. The most effective model within this structural feature subset was the Gradient Boosting Classifier, achieving notable metrics, including an area under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.81, accuracy of 0.80, Average Precision-Recall (PR) of 0.89, and F1 score of 0.85 (refer to Supplementary Figure 4.5).

In contrast, models trained on genomic count features, such as the count of mutational insertions, transcription factor binding sites, transcription start sites, enhancers, transposable elements, miRNAs, and G-quadruplexes, computed across three genomic regions, demonstrated robust performance. These regions encompassed the gene body, a 1KB region upstream and downstream of gene start and end sites, and a larger region covering 5 KB upstream and downstream of the gene start and end sites. The Gradient Boosting Classifier emerged as the top-performing model among all, achieving an AUC-ROC of 0.84, accuracy of 0.81, Average Precision-Recall (PR) of 0.91, and F1 score of 0.87 (see Supplementary Figure 4.6).

Models trained individually on epigenetic, evolutionary, sequence, and codon-based features exhibited overall similar performance, with an AUC-ROC ranging from 0.77 to 0.79, accuracy around 0.73, Average Precision-Recall (PR) from around 0.85 to 0.87, and F1 from around 0.78 to 0.80 (see Supplementary Figure 4.7, Supplementary Figure 4.8, Supplementary Figure 4.9, and Supplementary Figure 4.10, respectively). Tree-based models such as Gradient Boosting and Random Forest performed optimally when trained on these feature subsets.

Among different subgroups of features, models trained on expression features (with stress-based experiments removed), including gene expression, protein expression, and gene co-expression values, demonstrated the best performance with an AUC-ROC of 0.88, accuracy of 0.83, Average Precision-Recall (PR) of 0.93, and F1 of 0.87 (see Supplementary Figure 4.11). Of all the models, the extra-tree classifier exhibited the most optimal performance when trained exclusively on expression features.

The variomic features-based model exhibited the least favorable performance compared to all other feature subsets, with an AUC-ROC of 0.74, accuracy of 0.73, Average Precision-Recall (PR) of 0.93, and F1 of 0.80 (see Supplementary Figure 4.12).

Ultimately, the combined model, trained using the unified set of features, incorporating intrinsic and extrinsic predictive features such as Sequence, Gene Model Structure, Codon, Gene and Protein Expression, Chromatin, Count, Variomic, and Evolutionary features, emerged as the best-performing model for stress-responsive gene classification. It achieved an AUC-ROC of 0.91, accuracy of 0.86, Average Precision-Recall (PR) of 0.96, and F1 of 0.90 (see Figure 4.2).

Therefore, in terms of performance ranking, the top-performing model was the unified model, followed by the expression-based model, genomic count model, gene structural model,

epigenetic model, evolutionary model, sequence model, codon-based model, and variomic feature-based model (see Figure 4.1).

4.4.2 Model comparison based on different labeling techniques

Distinct models were trained based on the four distinct definitions of stress-responsive genes: the RNA-Seq-based definition employing a generic cutoff [|log2(fold change)| ≥ 1 and < -1, p-value < 0.05], a more stringent criterion [|log2(fold change)| ≥ 2 and <-2, p-value < 0.05], a GWAS-based definition, and a unified approach for defining stress genes.

The model's optimal performance was observed when trained on the RNA-Seq-based definition of stress-responsive genes with a generic cutoff [|log2(fold change)| ≥ 1 and < -1, p-value < 0.05], yielding an impressive area under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.91, accuracy of 0.86, Average Precision-Recall (PR) of 0.96, and F1 of 0.90 (see Figure 4.2).

Models trained on the unified approach for defining stress genes and the more stringent RNA-Seq-based definition of stress-responsive genes [|log2(fold change)| ≥ 2 and < -2, p-value < 0.05] exhibited comparable performance, with AUC-ROC values ranging from 0.83 to 0.86, accuracy of 0.78 to 0.80 , Average Precision-Recall (PR) of 0.85 to 0.90, and F1 of 0.76 to 0.78 (see Supplementary Figure 4.13, Supplementary Figure 4.14, respectively).

In contrast, models trained on the GWAS-based definition of stress-responsive genes as true labels displayed the least favorable performance, with an AUC-ROC of 0.54, accuracy of 0.63, Average Precision-Recall (PR) of 0.36, and F1 of 0.23 (see Supplementary Figure 4.15).

These results strongly indicate that the RNA-Seq-based definition of stress-responsive genes with a generic cutoff [|log2(fold change)| ≥ 1 and < -1, p-value < 0.05] stands out as the gold standard for labeling stress-responsive genes. It is suggested as the most ideal method for

labeling stress-responsive genes when employing machine learning-based predictions or classifications.

4.4.3 Statistical modeling and ranking the most distinctive features of stress-responsive genes

In this study we employed three criteria to identify optimal features. Firstly, the features needed to be easily accessible and available across a wide range of plant organisms. Intrinsic features could be readily extracted as long as an organism possessed a fully sequenced genome. Context-dependent features were considered based on their generation feasibility through existing pipelines, customized Python scripts, or retrieval from major plant databases and peer-reviewed publications.

Furthermore, the chosen features must demonstrate substantial predictive prowess in the context of gene stress classification, aligning with the strengths of gradient boosting. To assess the predictive power of each feature, a bootstrapping approach was employed. This process entailed evaluating the accuracy of each tree in the boosting ensemble using out-of-bag samples for validation. The labels of the feature were permuted, and the resulting average reduction in accuracy was leveraged to ascertain the importance score, utilizing the varImp function from the scikit-learn (v1.0.2) package for gradient boosting.

Thirdly, the features were required to minimize biological redundancy. Biologically redundant features, often derived from a similar source, tend to exhibit high correlations with each other. For instance, CBI (Codon Bias Index) has demonstrated strong Pearson correlation coefficients with CAI (Codon Adaptation Index) and Fop (Frequency of optimal Codons) because these features are all derived from the codon usage of a gene and share similar biological meanings. Including such redundant features not only poses challenges for various machine

learning classifiers but also introduces complexity without necessarily enhancing the inferential and predictive power of the classifier.

Applying the aforementioned criteria, from a myriad of features under consideration, we pinpointed the top 25 features potentially linked to stress genes, displaying relatively mild correlations among themselves (refer to Figure 4.3 and Table 4.2). The selection was based on the top 25 features derived from the most effective model trained on the unified set of features and the RNA-Seq based definition of stress-responsive genes using a generic cutoff [$|\log2(\text{fold change})| \geq 1$ and < -1, p-value < 0.05]. Notably, these features encapsulate diverse aspects ranging from sequence to function.

The predominant factors influencing the classification of stress-responsive or non-responsive genes in our highly effective gradient boosting classifier model include EREB Transcription Factors, specifically EREB 138, EREB 29, and EREB 71. These transcription factors are recognized for their responsiveness to ethylene, a plant hormone integral to diverse stress responses. Additionally, the model gives substantial importance to LBD (LATERAL ORGAN BOUNDARIES DOMAIN) Transcription Factor (LBD 19), PFAM domains, genes categorized as dispensable, gene breadth, gene co-expression features, tandem genes, gene structural features like mRNA length, chromatin features such as activation sites and open chromatin regions, as well as evolutionary features such as gene age. We corroborated these key predictors through validation against previous studies on stress-responsive genes and their attributes. Notably, earlier research has highlighted a robust positive correlation between the number of transcription factors targeting a gene and the likelihood of the gene being responsive to stress (Kimotho, Baillo, and Zhang 2019).

The transcription factors EREB 138, EREB 29, EREB 71, and LBD 19 have been linked to stress responses in maize. EREB10, belonging to the AP2/EREB family, is recognized for its involvement in the maize response to abiotic stress (Fagny et al. 2020). Additionally, the LBD transcription factor ZmLBD5 has been identified as a negative regulator of drought tolerance, impacting abscisic acid synthesis. In Arabidopsis, ZmLBD5 enhances drought sensitivity by suppressing ROS accumulation (Jiao et al. 2022; Xiong et al. 2022). The presence and characterization of the LBD transcription factor family in bread wheat suggests its potential role in stress responses. Screening of LBD transcription factors has revealed their participation in salt stress responses in Rosa rugosa (Wang et al. 2021). Moreover, during shoot-borne root initiation in maize, transcriptome profiling has implicated LBD genes such as rtcs, rtcl, rtcn, and lbd34 in this process (Muthreich et al. 2013). Furthermore, class II LBD genes ZmLBD5 and ZmLBD33 have been identified as regulators of gibberellin and abscisic acid biosynthesis, shedding light on their function in stress responses (Xiong et al. 2021). Collectively, these findings underscore the crucial roles of EREB and LBD transcription factors in maize stress responses.

Previous studies have also corroborated the association between PFAM domains and a spectrum of stress-responsive genes, notably those integral to general stress responses like sigma factors and histidine kinases/phosphatases (Jacob et al. 2014).

Tandem duplicate genes exhibit a heightened presence in co-expression patterns, indicating their potential involvement in stress gene responses (Li et al. 2016). Additionally, dispensable genes, characterized by present/absent variation, are proposed to play a pivotal role in enhancing phenotypic diversity and heterotic performance in hybrids, particularly in the context of stress-responsive genes (Weisweiler et al. 2019).

Chromatin features, including histone modifications and chromatin accessibility, have been correlated with the activation of stress-responsive genes in maize under drought stress conditions (Halder et al. 2022). The role of histone acetylation in the transcriptional regulation of stress-responsive genes underscores the importance of epigenetic regulation in stress gene responses. Moreover, low-temperature stress has been demonstrated to induce genome-wide hypermethylation of transposable elements and centromeres in maize, suggesting the involvement of epigenetic modifications in stress gene responses (Chang et al. 2020).

Evolutionary features, such as gene age and adaptive evolution post-gene duplication, have been identified as influential factors in species adaptability and the regulation of stress-responsive genes (Doughty et al. 2020). Furthermore, the distinct expression patterns of aquaporin genes have been linked to the maintenance of water use efficiency in drought-stressed sorghum compared to maize, providing insights into the genetic basis of drought tolerance in different plant species, specifically in the context of stress gene association (Prasad et al. 2021).

4.4.4 Explainability and Interpretability of the Gradient Boosting model for stress-responsive and non-responsive gene classification

The LIME technique provides a comprehensive explanation and interpretation of the prediction for a specific instance, such as an individual stress-responsive gene (refer to Figure 4.4A). For instance, the predicted outcome for the gene depicted in Figure 4 suggests a high likelihood of being stress-responsive with 100.0% prediction confidence. Furthermore, LIME elucidates the rationale behind this prediction by delineating the contributions of input features (WGCNA Module 11, three-prime UTR length, tandem duplicates, mRNA length, Effective number of codons (Nc)) to the predicted outcome (i.e., a high chance of being a stress-responsive gene).

For instance, Module 11 (Hoopes et al. 2019) exhibits enrichment for leaf-specific genes ($P < 2e-15$), and its association with stress-responsive genes is validated through GO term enrichment, particularly for genes related to 'photosynthesis' (GO:0015979). Similarly, Module 3 (Hoopes et al. 2019), enriched for root-specific and biotic-related DE genes ($P < 0.001$), demonstrates an association with stress-responsive genes, as affirmed by GO enrichment identifying terms like 'response to oxidative stress' (GO:0006979) and 'cell wall organization or biogenesis' (GO:0071554).

In a similar vein, the SHAP technique elucidates the prediction outcome by evaluating the contribution of each feature to the prediction, offering global explanations (refer to Figure 4.4B). The feature importance analysis based on the SHAP technique revealed that, in descending order of significance, PFAM domains, tandem duplicates, gene breadth, Heat treated seeding RNA-Seq expression, SNPs, and three prime UTR length were influential input variables affecting the model's performance in predicting the likelihood of being stress-responsive genes. The detailed contributions of these variables to the prediction of stress-responsive genes are presented in Figure 4.4B. The SHAP beeswarm plot further provides intricate insights into how the parameters within each variable contribute to the desired outcome, offering a global explanation and interpretation.

As depicted in Figure 4.4B, the anticipated outcome can manifest as either non-responsive genes (negative side on the x-axis) or stress-responsive genes (positive side on the x-axis). Consequently, a detailed examination of the impact of each prognostic parameter is presented in Figure 4.4B. The analysis indicates that a higher number of PFAM domains, lower tandem duplicates, genes expressed in a majority of conditions or tissues (gene breadth), elevated gene expression under heat-treated conditions, longer three prime UTR length, a higher number

of genes associated with WGCNA module six (Module 6, jointly enriched for internode-specific genes, biotic-related DE genes, and DE genes under both biotic and abiotic stress, $P < 5e-9$), and a higher number of genes associated with E4 (H3K27ac + H2AZ + H3K4me1 + H3K56ac) histone modification combination are all correlated with an increased likelihood of being stress-responsive genes (Figure 4.4B).

4.4.5 Model performance on other maize lines

We assessed the efficacy of a model trained on widely applicable and generalized features, specifically codon-based sequence features, using independent experimentally validated datasets. Upon training and testing on the maize B73 reference genome, this model displayed satisfactory performance, achieving an area under the Receiver Operating Characteristic Curve (AUC-ROC) within the range of 0.77 to 0.79. The accuracy was approximately 0.73, with Average Precision-Recall (PR) values spanning from 0.85 to 0.87, and F1 scores falling between 0.78 and 0.80 (refer to Supplementary Figure 4.10). Since these same sequence features can be calculated for genes across different maize lines, it becomes feasible to evaluate the predictive capacity of stress-responsive gene expression in one maize line based solely on information about the stress responsiveness of genes in other maize lines.

Hence, it was employed to evaluate the model's performance on datasets generated by different research groups worldwide, specifically focusing on genes designated as stress-responsive in the maize inbred line W22 and the Zea mays ssp. mexicana L. genomes (TIL-18 and TIL-25). The classification was based on the RNA-Seq definition of stress-responsive genes using a generic cutoff [$|\log2(\text{fold change})| \geq 1$ and < -1, p-value < 0.05]. The test dataset exclusively comprised the positive class, representing genes labeled as stress-responsive genes. Due to limited availability of expression datasets for the W22 and mexicana genomes, defining

non-responsive genes was not feasible. Consequently, the codon-based model's performance was assessed solely on the positive test cases for each maize line.

The model demonstrated robust performance in classifying genes as stress-responsive, achieving an accuracy of approximately 93% for TIL-18 by correctly identifying 5,688 stress-responsive genes out of the total 6,147 in TIL-18 (refer to Supplementary Figure 4.16A). Similarly, for TIL-25, the model achieved an accuracy of around 92%, accurately classifying 6,444 genes as stress-responsive genes out of the total 6,982 stress-responsive genes in TIL-18 (Supplementary Figure 4.16B). Additionally, for W22, the model exhibited an accuracy of around 87%, accurately classifying 4033 genes as stress-responsive genes out of the total 4617 stress-responsive genes in W22 (Supplementary Figure 4.16C).

4.5 Discussion

Abiotic and biotic stress responses are traditionally thought to be regulated by discrete signaling mechanisms. Recent experimental evidence revealed a more complex picture where these mechanisms are highly entangled and controlled by a range of cellular, molecular, and genetic mechanisms that act together in a complex regulatory network. Transcription factors, functional domains, gene structures, genome localizations, and expression profiles are key components of this crosstalk, as are heat shock factors and small RNAs. Despite shared evolutionary history, a gene responding transcriptionally to cold stress in one species was not a good predictor of whether syntenic orthologous genes in related species would also respond to cold stress in the same treatment at the same developmental stage (Meng et al. 2021). This low conservation of transcriptional responses across conserved genes in related species is consistent with the results of a previous comparison of the transcriptional responses of maize and sorghum to stress (Zhang et al. 2017) and the variation in transcriptional responses to stress between different alleles of the same gene in maize (Zeng et al. 2021). This suggests that identifying and

characterizing key genes and their potential characteristics, which discriminates abiotic and biotic stress responses, would increase our understanding of plant stress response manyfold and provide targets for genetic manipulation to improve their stress tolerance. Identifying master regulators such as stress-responsive genes that connect both biotic and abiotic stress response pathways is fundamental in providing opportunities for developing broad-spectrum stress-tolerant crop plants.

In summary, our comprehensive methodology, combining standardized and balanced data, a diverse set of classification algorithms, and thoughtful hyperparameter tuning, aims to develop robust models for accurately classifying stress-responsive and non-responsive genes. This multi-faceted approach ensures that our models are well-equipped to handle the intricacies of the classification task and produce reliable predictions.

It is becoming increasingly apparent that genomic sequences represent only one aspect of the complex genetic relationships that have evolved under diverse selection pressures; therefore, it is necessary to consider a variety of features, including both intrinsic and context-dependent features. We hypothesize that a combined or advanced model built utilizing an entire set of omics data will outperform the generalized model build using just the sequence and structural features for maize stress-responsive or non-responsive gene classification.

For the purpose of reproducibility, the supervised classification models trained on just gene features, including sets of features that can be calculated solely from genomic sequence data and gene structural annotation, can provide significant accuracy to predict which genes will transcriptionally respond to a specific abiotic or biotic stress. The success we achieved in prediction based on gene-sequence features greatly expands the potential application of this technique to non-model species—including those adapted to extreme environments—for which a

reference genome sequence has been generated, but substantial functional genomic datasets are lacking. Unlike the combined model, which requires data, the pure genomic feature model can be applied to any species with a sequenced genome and annotated gene models.

While the unified feature model offers superior accuracy and efficiency, obtaining input features for this model necessitates the utilization of specialized sequencing techniques and resequencing data from diverse populations. To establish a highly adaptable prediction platform applicable to any species with a sequenced genome and annotated gene models, we developed a generalized model that exclusively relies on gene and protein sequences along with structural features. These genomic features are conveniently accessible or readily available in the form of GFF files.

Furthermore, we conducted a comprehensive comparison and evaluation of models trained using different labeling techniques for defining genes as stress-responsive and non-responsive. Our findings highlight those genes defined as stress-responsive with a generic cutoff [$|\log2(\text{fold change})| \geq 1$ and < -1, p-value < 0.05] serve as the gold standard for labeling stress-responsive genes, as they lead to optimal performance compared to GWAS-based labeling or a more stringent RNA-Seq-based labeling.

Although the optimal cutoff may vary based on specific analysis goals and dataset characteristics, our validation process, utilizing independent datasets and cross-validation to assess performance across diverse contexts, revealed that a more stringent cutoff heightened specificity (reducing false positives) but concurrently resulted in decreased sensitivity (increasing false negatives). This stricter threshold excluded genes with subtle yet meaningful changes, resulting in information loss and potentially diminishing classification accuracy.

Therefore, a meticulous evaluation of both features and labels is crucial for constructing an optimal model for stress-responsive gene classification. In summary, our exploration involving the integration of various omics datasets, labels, and analyses enhances the annotation of stress-responsive genes, contributing to a deeper understanding of the molecular mechanisms underlying multiple stress responses in plants.

4.6 Main Figures and table

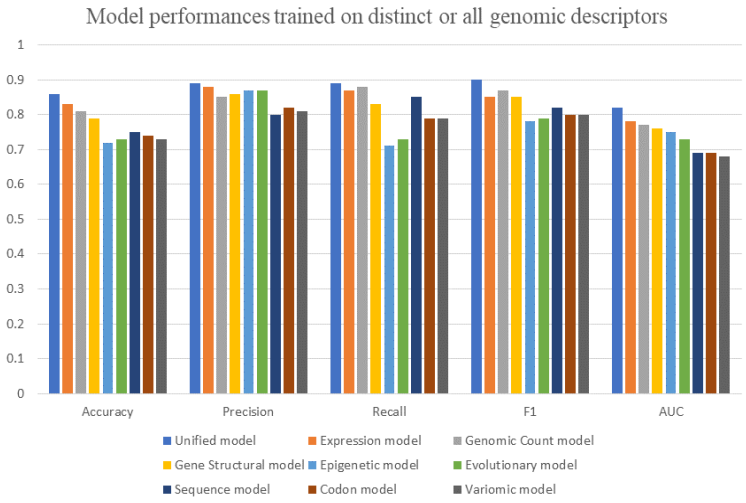

Figure 4.1 Prediction performance chart of the best performing model trained on distinct or all genomic descriptors.

In the chart, we display the prediction performance of the best performing model trained on distinct or all genomic descriptors (Unified, Expression, Count, Structural, Epigenetic, Evolutionary, Sequence, Codon and Variomic). The chart summarizes the models' best scores on the testing dataset for each criterion (accuracy, precision, recall, F1, AUC-ROC)

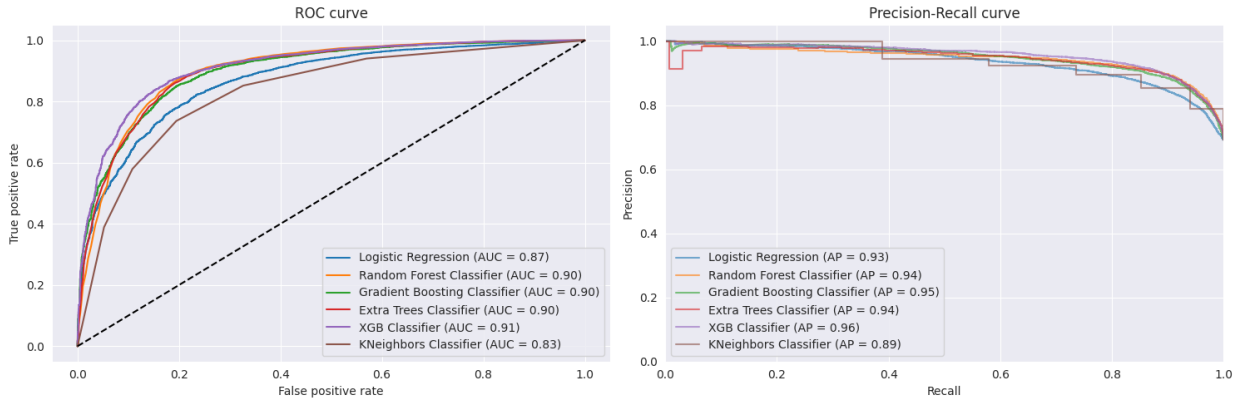

Figure 4.2 Graphical representation of the prediction performance of the "unified" model.

Graphical representation of the prediction performance of the "unified" (structural + sequence + genomic count + epigenetic + expression + variomic + codon + variomic) features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

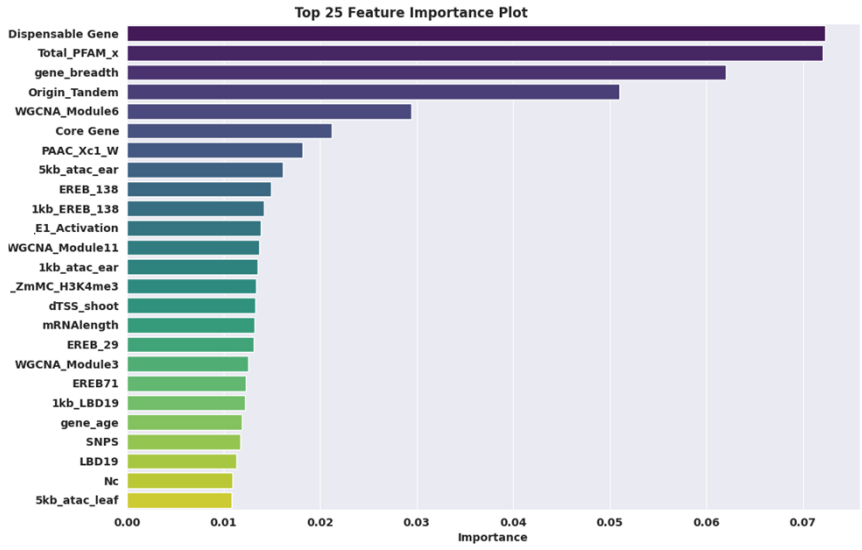

Figure 4.3 Feature Importance plot of the comprehensive unified feature model.

For our comprehensive unified feature model, both intrinsic and extrinsic features contributed substantially to the stress-responsive/non-responsive gene predictions in maize B73v5. The 25 omics features were ranked based on how useful the model found each feature in predicting the target (stress-responsive/non-responsive).

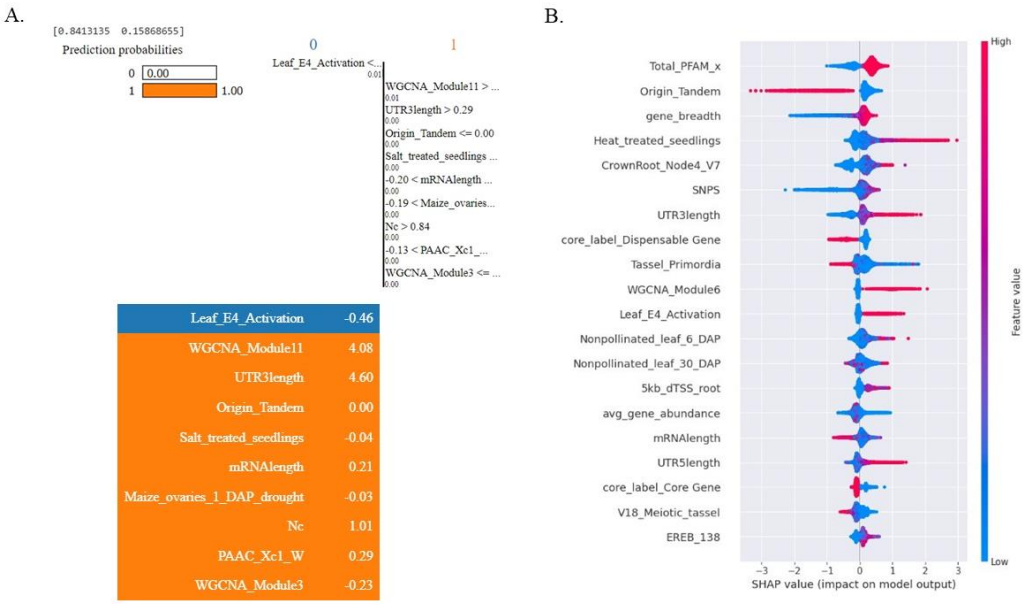

Figure 4.4 LIME and SHAP explainable AI plot.

(A) LIME explainability of a single instance. (B) SHAP beeswarm summary plot on the impact of input variables on the Gradient Boosting classifier model's predictive ability.

Table 4.1 Tabular view of the RNA-Seq data sources.

PMID	DOI/References	Type	Project Accession
28298920	https://doi.org/10.3389/fpls.2017.00290	Abiotic	PRJNA335771
31245722	https://doi.org/10.1002/pld3.57	Biotic	PRJNA325825
34557211	https://doi.org/10.3389/fpls.2021.699146	Biotic	PRJNA730310
24885787	https://doi.org/10.1186/1471-2229-14-141	Abiotic	PRJNA210356
28298916	https://doi.org/10.3389/fpls.2017.00267	Abiotic	PRJNA339768
27461139	https://doi.org/10.1038/srep30446	Abiotic	PRJNA290180
30257650	https://doi.org/10.1186/s12864-018-5109-8	Abiotic	PRJNA398446
26042133	https://doi.org/10.3389/fpls.2015.00341	Abiotic	PRJNA269060
36016425	https://doi.org/10.3390/v14081803	Biotic	PRJNA846583
32756433	https://doi.org/10.3390/genes11080881	Abiotic	PRJNA645274
34502437	https://doi.org/10.3390/ijms22179527	Abiotic	PRJNA723826
33730156	https://doi.org/10.1093/plcell/koab083	Abiotic	PRJNA659061
29206208	https://doi.org/10.3390/ijms18122624	Abiotic	PRJNA420600
31968691	https://doi.org/10.3390/ijms21020686	Abiotic	PRJNA594965
33106639	https://doi.org/10.1038/s41477-020-00787-9	Biotic	PRJNA577898
29426290	https://doi.org/10.1186/s12864-018-4513-4	Biotic	PRJNA357594
28535078	https://doi.org/10.1094/MPMI-03-17-0054-R	Biotic	PRJNA369690
30186298	https://doi.org/10.3389/fpls.2018.01222	Biotic	PRJNA390756
32121334	https://doi.org/10.3390/genes11030267	Abiotic	PRJNA606824

Table 4.1continued

PMID	DOI/References	Type	Project Accession
30537259	https://doi.org/10.1111/tpj.14184	Biotic	PRJEB10574
25569788	https://doi.org/10.1371/journal.pgen.1004915	Abiotic	PRJNA244661
25174417	https://doi.org/10.1186/1471-2164-15-741	Abiotic	PRJNA226757
36345007	https://doi.org/10.1186/s13059-022-02807-7	Abiotic	PRJNA849202
26990640	https://doi.org/10.1371/journal.pone.0151697	Abiotic	PRJNA304223
35579358	https://doi.org/10.1093/genetics/iyac080 Advance	Abiotic	PRJNA604929

The table lists the twenty-five (25) quality RNA-Seq dataset sources with their PMID, Digital Object Identifier (DOI), Type of stress and Project Accession information.

Table 4.2 The top 25 omics features most useful in predicting the target (stress-responsive/non-responsive genes).

Features	Feature Details	Sources
Total_PFAM_x	Total number of Protein domains	https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/
Origin_Tandem	Adjacent sequential genes	https://pubmed.ncbi.nlm.nih.gov/34353948/

Table 4.2 continued

Features	Feature Details	Sources
Gene_Breadth	Variety of expressed genes	https://qteller.maizegdb.org/rna_data_sources.php
Heat_Treated_Seedlings	Expression of genes under heat stress	https://qteller.maizegdb.org/rna_data_sources.php
CrownRoot_Node4_V7	Expression of crown root at node 4	https://qteller.maizegdb.org/rna_data_sources.php
SNPs	count of single nucleotide polymorphisms (SNPs)	https://pubmed.ncbi.nlm.nih.gov/34353948/
UTR3Length	Three prime UTR length	https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/
Dispensable Gene	Genes present across 2 to 23 of the NAM lines	https://pubmed.ncbi.nlm.nih.gov/34353948/
Tassel_Primordia	The expression of tassel primordia	https://qteller.maizegdb.org/rna_data_sources.php
WGCNA_Module6	Weighted gene co-expression network module 6	https://pubmed.ncbi.nlm.nih.gov/30537259/
Leaf_E4_Activation	chromatin states associated with gene activation	https://www.nature.com/articles/s41477-019-0547-0
Nonpollinated_Leaf_6_DAP	Expression of the non-pollinated leaf (6 days after pollination)	https://qteller.maizegdb.org/rna_data_sources.php

Table 4.2 continued

Features	Feature Details	Sources
Nonpollinated_Leaf_30_DAP	Expression of the non-pollinated leaf (30 days after pollination)	https://qteller.maizegdb.org/rna_data_sources.php
5kb_dTSS_Root	Gene counts around 5 kilo base pair of the TSS	B73v5 TSS
Avg_Gene_Abundance	Average gene expression level	https://qteller.maizegdb.org/rna_data_sources.php
mRNA_Length	mRNA length	B73v5 miRNA
UTR5Length	Five prime UTR length	https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/
Core_Gene	Genes present across all the 26 NAM lines	https://pubmed.ncbi.nlm.nih.gov/34353948/
V18_Meiotic_Tassel	Expression of the V18 meiotic tassel	https://qteller.maizegdb.org/rna_data_sources.php
EREB_138	Ethylene-Responsive Element Binding (EREB)Transcription factor	Ricci 2019 TFBS DAP-seq

The table outlines the 25 most valuable omics features for predicting the target (stress-responsive/non-responsive), along with details about their sources or methodologies of data generation.

4.7 References

Assem, S. K. 2015. 'Maize, tropical (Zea mays L.)', *Methods Mol Biol*, 1223: 119-34.

Challa, Surekha, and Nageswara RR Neelapu. 2018. 'Genome-wide association studies (GWAS) for abiotic stress tolerance in plants.' in, *Biochemical, physiological and molecular avenues for combating abiotic stress tolerance in plants* (Elsevier).

Chandrashekar, Girish, and Ferat Sahin. 2014. 'A survey on feature selection methods', *Computers & Electrical Engineering*, 40: 16-28.

Chang, Y. N., C. Zhu, J. Jiang, H. Zhang, J. K. Zhu, and C. G. Duan. 2020. 'Epigenetic regulation in plant abiotic stress responses', *J Integr Plant Biol*, 62: 563-80.

Chiu, J. K. H., and R. T. Ong. 2022. 'Clustering biological sequences with dynamic sequence similarity threshold', *BMC Bioinformatics*, 23: 108.

Dai, X., and L. Shen. 2022. 'Advances and Trends in Omics Technology Development', *Front Med (Lausanne)*, 9: 911861.

Danilevicz, M. F., M. Gill, R. Anderson, J. Batley, M. Bennamoun, P. E. Bayer, and D. Edwards. 2022. 'Plant Genotype to Phenotype Prediction Using Machine Learning', *Front Genet*, 13: 822173.

Doughty, T. W., I. Domenzain, A. Millan-Oropeza, N. Montini, P. A. de Groot, R. Pereira, J. Nielsen, C. Henry, J. G. Daran, V. Siewers, and J. P. Morrissey. 2020. 'Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts', *Nat Commun*, 11: 2144.

Fagny, M., M. L. Kuijjer, M. Stam, J. Joets, O. Turc, J. Roziere, S. Pateyron, A. Venon, and C. Vitte. 2020. 'Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks', *Front Genet*, 11: 606285.

Feldner-Busztin, D., P. Firbas Nisantzis, S. J. Edmunds, G. Boza, F. Racimo, S. Gopalakrishnan, M. T. Limborg, L. Lahti, and G. G. de Polavieja. 2023. 'Dealing with dimensionality: the application of machine learning to multi-omics data', *Bioinformatics*, 39.

Gahlaut, V., V. Jaiswal, A. Kumar, and P. K. Gupta. 2016. 'Transcription factors involved in drought tolerance and their possible role in developing drought tolerant cultivars with emphasis on wheat (Triticum aestivum L.)', *Theor Appl Genet*, 129: 2019-42.

Gong, F., L. Yang, F. Tai, X. Hu, and W. Wang. 2014. '"Omics" of maize stress response for sustainable food production: opportunities and challenges', *OMICS*, 18: 714-32.

Grote, Ulrike, Anja Fasse, Trung Thanh Nguyen, and Olaf Erenstein. 2021. 'Food security and the dynamics of wheat and maize value chains in Africa and Asia', *Frontiers in Sustainable Food Systems*, 4: 617009.

Gull, Audil, Ajaz Ahmad Lone, and Noor Ul Islam Wani. 2019. 'Biotic and abiotic stresses in plants', *Abiotic and biotic stress in plants*: 1-19.

Guyon, Isabelle, and André Elisseeff. 2003. 'An introduction to variable and feature selection', *Journal of machine learning research*, 3: 1157-82.

Habib-Ur-Rahman, M., A. Ahmad, A. Raza, M. U. Hasnain, H. F. Alharby, Y. M. Alzahrani, A. A. Bamagoos, K. R. Hakeem, S. Ahmad, W. Nasim, S. Ali, F. Mansour, and A. El Sabagh. 2022. 'Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia', *Front Plant Sci*, 13: 925548.

Halder, K., A. Chaudhuri, M. Z. Abdin, M. Majee, and A. Datta. 2022. 'Chromatin-Based Transcriptional Reprogramming in Plants under Abiotic Stresses', *Plants (Basel)*, 11.

Han, Y., S. Gao, K. Muegge, W. Zhang, and B. Zhou. 2015. 'Advanced Applications of RNA Sequencing and Challenges', *Bioinform Biol Insights*, 9: 29-46.

Hemathilake, DMKS, and DMCC Gunathilake. 2022. 'Agricultural productivity and food supply to meet increased demands.' in, *Future Foods* (Elsevier).

Henry, Robert J. 2022. 'Progress in plant genome sequencing', *Applied Biosciences*, 1: 113-28.

Hoopes, G. M., J. P. Hamilton, J. C. Wood, E. Esteban, A. Pasha, B. Vaillancourt, N. J. Provart, and C. R. Buell. 2019. 'An updated gene atlas for maize reveals organ-specific and stress-induced genes', *Plant J*, 97: 1154-67.

Jacob, S., A. J. Foster, A. Yemelin, and E. Thines. 2014. 'Histidine kinases mediate differentiation, stress response, and pathogenicity in Magnaporthe oryzae', *Microbiologyopen*, 3: 668-87.

Jakhar, Dan Singh, and Rajesh Singh. 2015. 'Biotic stress response in maize (Zea mays L.)', *Journal of Biotechnology and Crop Science*, 4: 47-51.

Jiao, P., X. Wei, Z. Jiang, S. Liu, S. Guan, and Y. Ma. 2022. 'ZmLBD2 a maize (Zea mays L.) lateral organ boundaries domain (LBD) transcription factor enhances drought tolerance in transgenic Arabidopsis thaliana', *Front Plant Sci*, 13: 1000149.

Kimotho, Roy Njoroge, Elamin Hafiz Baillo, and Zhengbin Zhang. 2019. 'Transcription factors involved in abiotic stress responses in Maize (Zea mays L.) and their roles in enhanced productivity in the post genomics era', *PeerJ*, 7: e7211.

Latorre, P., R. Bottcher, M. Nadal-Ribelles, C. H. Li, C. Sole, G. Martinez-Cebrian, P. C. Boutros, F. Posas, and E. de Nadal. 2022. 'Data-driven identification of inherent features of eukaryotic stress-responsive genes', *NAR Genom Bioinform*, 4: lqac018.

Li, L., R. Briskine, R. Schaefer, P. S. Schnable, C. L. Myers, L. E. Flagel, N. M. Springer, and G. J. Muehlbauer. 2016. 'Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias', *BMC Genomics*, 17: 875.

Lu, X., X. Zhou, Y. Cao, M. Zhou, D. McNeil, S. Liang, and C. Yang. 2017. 'RNA-seq Analysis of Cold and Drought Responsive Transcriptomes of Zea mays ssp. mexicana L', *Front Plant Sci*, 8: 136.

Lundberg, Scott M, and Su-In Lee. 2017. 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, 30.

Madzima, T. F., S. Vendramin, J. S. Lynn, P. Lemert, K. C. Lu, and K. M. McGinnis. 2021. 'Direct and Indirect Transcriptional Effects of Abiotic Stress in Zea mays Plants Defective in RNA-Directed DNA Methylation', *Front Plant Sci*, 12: 694289.

Mahood, E. H., L. H. Kruse, and G. D. Moghe. 2020. 'Machine learning: A powerful tool for gene function prediction in plants', *Appl Plant Sci*, 8: e11376.

Meng, X., Z. Liang, X. Dai, Y. Zhang, S. Mahboub, D. W. Ngu, R. L. Roston, and J. C. Schnable. 2021. 'Predicting transcriptional responses to cold stress across plant species', *Proc Natl Acad Sci U S A*, 118.

Muthreich, N., C. Majer, M. Beatty, A. Paschold, A. Schutzenmeister, Y. Fu, W. A. Malik, P. S. Schnable, H. P. Piepho, H. Sakai, and F. Hochholdinger. 2013. 'Comparative transcriptome profiling of maize coleoptilar nodes during shoot-borne root initiation', *Plant Physiol*, 163: 419-30.

Nazari, L., M. F. Aslan, K. Sabanci, and E. Ropelewska. 2023. 'Integrated transcriptomic meta-analysis and comparative artificial intelligence models in maize under biotic stress', *Sci Rep*, 13: 15899.

Niederhuth, C. E., and R. J. Schmitz. 2017. 'Putting DNA methylation in context: from genomes to gene expression in plants', *Biochim Biophys Acta Gene Regul Mech*, 1860: 149-56.

Oyelade, J., I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi. 2016. 'Clustering Algorithms: Their Application to Gene Expression Data', *Bioinform Biol Insights*, 10: 237-53.

Poljsak, B., and I. Milisav. 2012. 'Clinical implications of cellular stress responses', *Bosn J Basic Med Sci*, 12: 122-6.

Portwood, J. L., 2nd, M. R. Woodhouse, E. K. Cannon, J. M. Gardiner, L. C. Harper, M. L. Schaeffer, J. R. Walsh, T. Z. Sen, K. T. Cho, D. A. Schott, B. L. Braun, M. Dietze, B. Dunfee, C. G. Elsik, N. Manchanda, E. Coe, M. Sachs, P. Stinard, J. Tolbert, S. Zimmerman, and C. M. Andorf. 2019. 'MaizeGDB 2018: the maize multi-genome genetics and genomics database', *Nucleic Acids Res*, 47: D1146-D54.

Prasad, V. B. R., M. Govindaraj, M. Djanaguiraman, I. Djalovic, A. Shailani, N. Rawat, S. L. Singla-Pareek, A. Pareek, and P. V. V. Prasad. 2021. 'Drought and High Temperature Stress in Sorghum: Physiological, Genetic, and Molecular Insights and Breeding Approaches', *Int J Mol Sci*, 22.

Raza, A., A. Razzaq, S. S. Mehmood, X. Zou, X. Zhang, Y. Lv, and J. Xu. 2019. 'Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review', *Plants (Basel)*, 8.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135-44.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Schrider, D. R., and A. D. Kern. 2018. 'Supervised Machine Learning for Population Genetics: A New Paradigm', *Trends Genet*, 34: 301-12.

Sen, S., M. R. Woodhouse, J. L. Portwood, 2nd, and C. M. Andorf. 2023. 'Maize Feature Store: A centralized resource to manage and analyze curated maize multi-omics features for machine learning applications', *Database (Oxford)*, 2023.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. 2011. 'Genome-wide association study of leaf architecture in the maize nested association mapping population', *Nat Genet*, 43: 159-62.

van Dijk, A. D. J., G. Kootstra, W. Kruijer, and D. de Ridder. 2021. 'Machine learning in plant science and plant breeding', *iScience*, 24: 101890.

Wallace, J. G., P. J. Bradbury, N. Zhang, Y. Gibon, M. Stitt, and E. S. Buckler. 2014. 'Association mapping across numerous traits reveals patterns of functional variation in maize', *PLoS Genet*, 10: e1004845.

Wang, Z., R. Zhang, Y. Cheng, P. Lei, W. Song, W. Zheng, and X. Nie. 2021. 'Genome-Wide Identification, Evolution, and Expression Analysis of LBD Transcription Factor Family in Bread Wheat (Triticum aestivum L.)', *Front Plant Sci*, 12: 721253.

Weisweiler, M., A. Montaigu, D. Ries, M. Pfeifer, and B. Stich. 2019. 'Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits', *BMC Genomics*, 20: 787.

Woodhouse, M. R., E. K. Cannon, J. L. Portwood, 2nd, L. C. Harper, J. M. Gardiner, M. L. Schaeffer, and C. M. Andorf. 2021. 'A pan-genomic approach to genome databases using maize as a model system', *BMC Plant Biol*, 21: 385.

Xiong, J., W. Zhang, D. Zheng, H. Xiong, X. Feng, X. Zhang, Q. Wang, F. Wu, J. Xu, and Y. Lu. 2022. 'ZmLBD5 Increases Drought Sensitivity by Suppressing ROS Accumulation in Arabidopsis', *Plants (Basel)*, 11.

Xiong, Jing, Xuanjun Feng, Weixiao Zhang, Xianqiu Wang, Yue Hu, Xuemei Zhang, Fengkai Wu, Wei Guo, Wubing Xie, and Qingjun Wang. 2021. 'Class II LBD genes ZmLBD5 and ZmLBD33 regulate gibberellin and abscisic acid biosynthesis', *bioRxiv*: 2021.04. 08.439062.

Zeng, R., Z. Li, Y. Shi, D. Fu, P. Yin, J. Cheng, C. Jiang, and S. Yang. 2021. 'Natural variation in a type-A response regulator confers maize chilling tolerance', *Nat Commun*, 12: 4713.

Zhang, Y., D. W. Ngu, D. Carvalho, Z. Liang, Y. Qiu, R. L. Roston, and J. C. Schnable. 2017. 'Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize', *Plant Cell*, 29: 1938-51.

Appendix A. Notes

4.8.1 Data availability statement

Project name: Predicting genes associated with biotic or abiotic stress across different maize lines and related species at https://github.com/shatabdi123/MFS_Application. The dataset can also be accessed on Kaggle: https://kaggle.com/datasets/332177dbd2271966f2291640acf6f7057bde915d939b3bf67545a5f24a0e3fe3. Programming language: Python, R.

4.8.2 Acknowledgements

We thank the research groups of Iowa State University and USDA-ARS, Corn Insects and Crop Genetics Research Unit.

4.8.3 Author contributions

Conceptualization, S.S., C.A.; Methodology, S.S.; Data Generation, S.S., R.H., M.W.; Formal Analysis, S.S.; Data Interpretation, S.S., C.A.; Software Validation, S.S.; Analysis Validation, S.S.; Writing – Original Draft Preparation, S.S., C.A.; Writing – Review and Editing, S.S., R.H., and C.A.; Supervision, C.A.; Funding Acquisition, C.A.

4.8.4 Declaration of interests

The authors declared no competing interest.

4.8.5 Funding

This research was supported by the US. Department of Agriculture, Agricultural Research Service, Project Number [5030-21000-068-00-D] through the Corn Insects and Crop Genetics Research Unit in Ames, Iowa. This material is based upon work supported by the Department of Agriculture, Agricultural Research Service under Agreement No. 58-5030-0-036 [Iowa State Award: 022172- 00001 to J.W.W.]. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer. Conflict of Interest: none declared.

4.9 Appendix B: Supplementary tables and figures

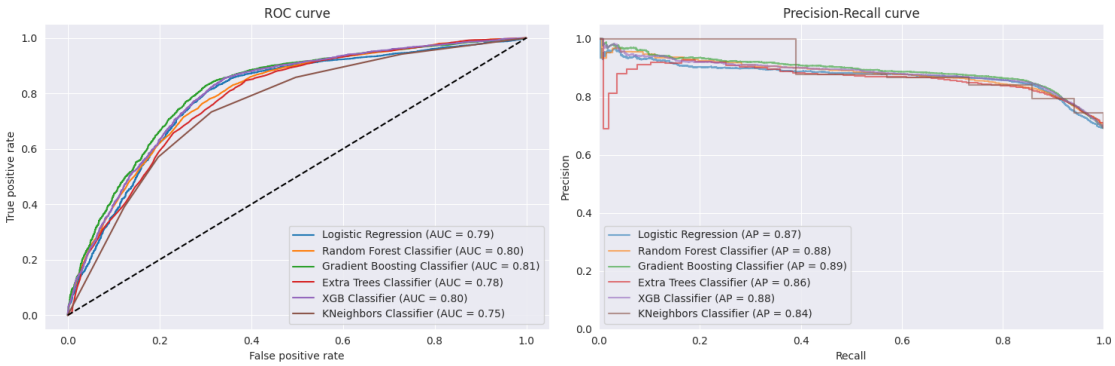

Figure 4.5 Graphical representation of the prediction performance of the “Gene Structure” model.

The Graphical representation of the prediction performance of the “Gene Structural” features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

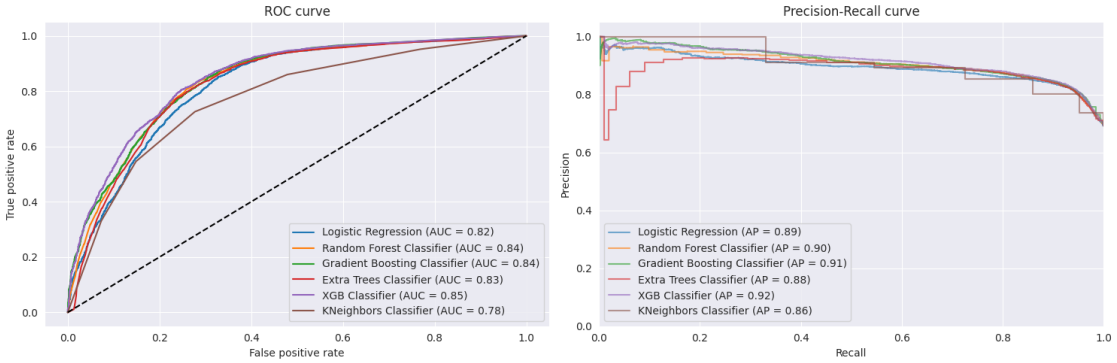

Figure 4.6 Graphical representation of the prediction performance of the "Genomic Count" model.

The Graphical representation of the prediction performance of the "Genomic Count" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

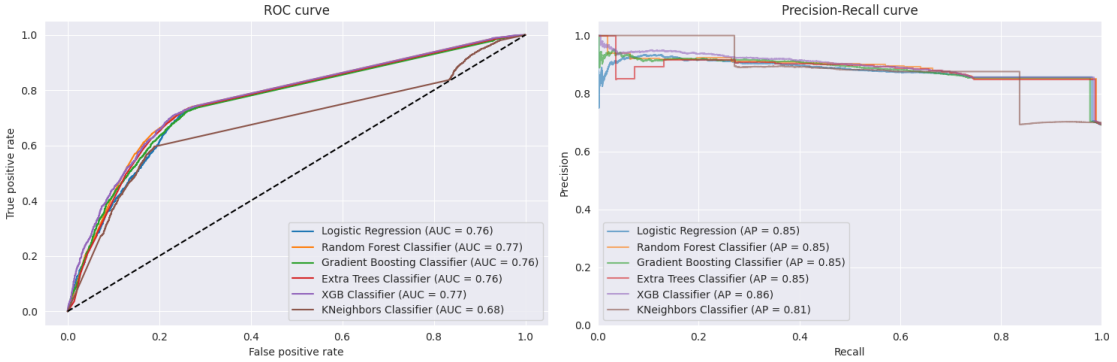

Figure 4.7 Graphical representation of the prediction performance of the "Epigenetic" model.

The Graphical representation of the prediction performance of the "Epigenetic" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

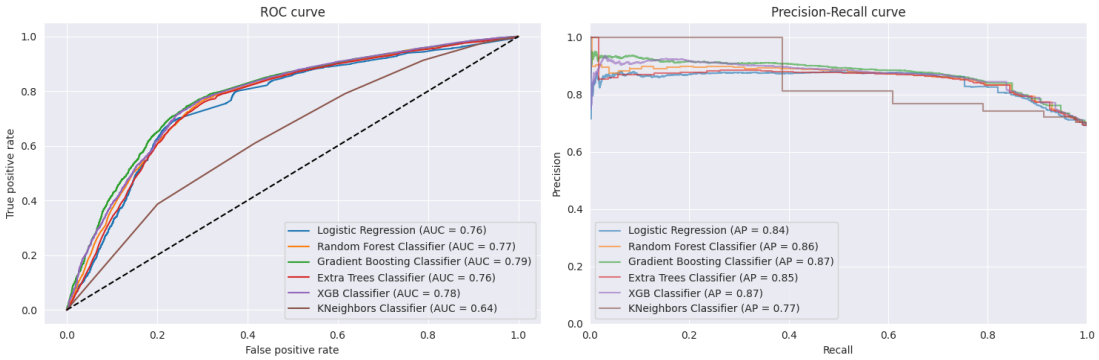

Figure 4.8 Graphical representation of the prediction performance of the "Evolutionary" model.

The Graphical representation of the prediction performance of the "Evolutionary" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

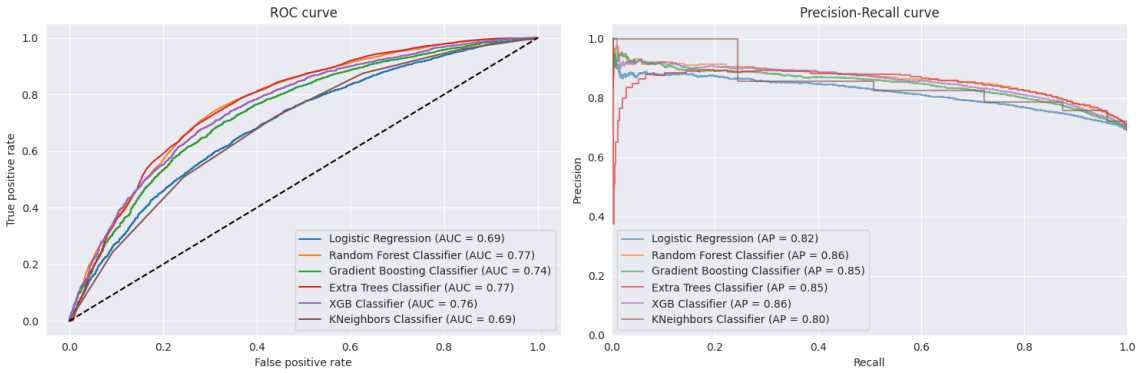

Figure 4.9 Graphical representation of the prediction performance of the "Sequence" model.

Graphical representation of the prediction performance of the "Sequence" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

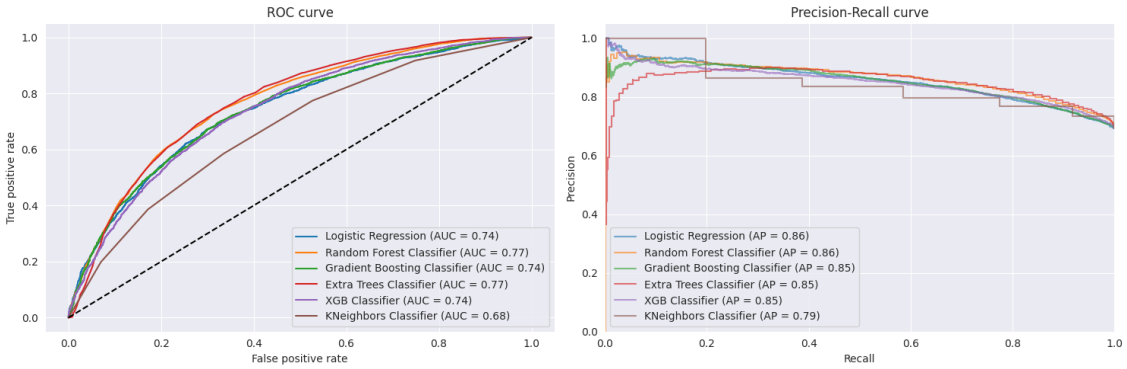

Figure 4.10 Graphical representation of the prediction performance of the "Codon" model.

Graphical representation of the prediction performance of the "Codon" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

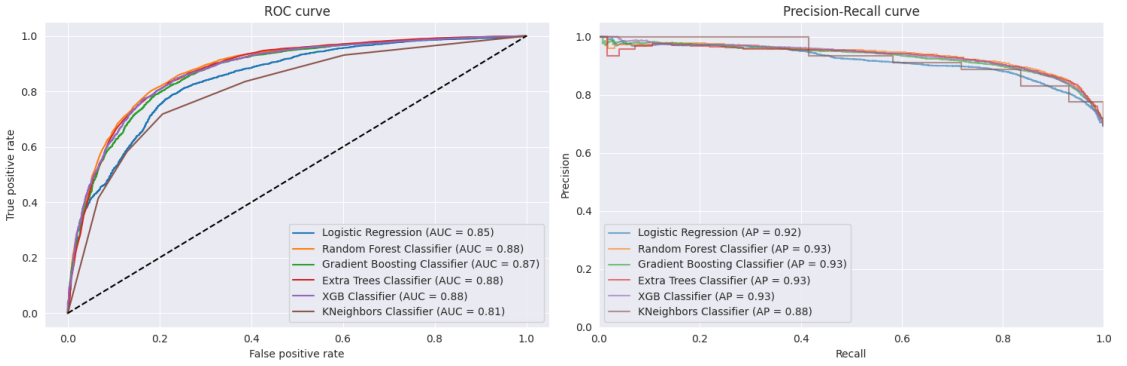

Figure 4.11 Graphical representation of the prediction performance of the "Expression" model.

The Graphical representation of the prediction performance of the "Expression" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

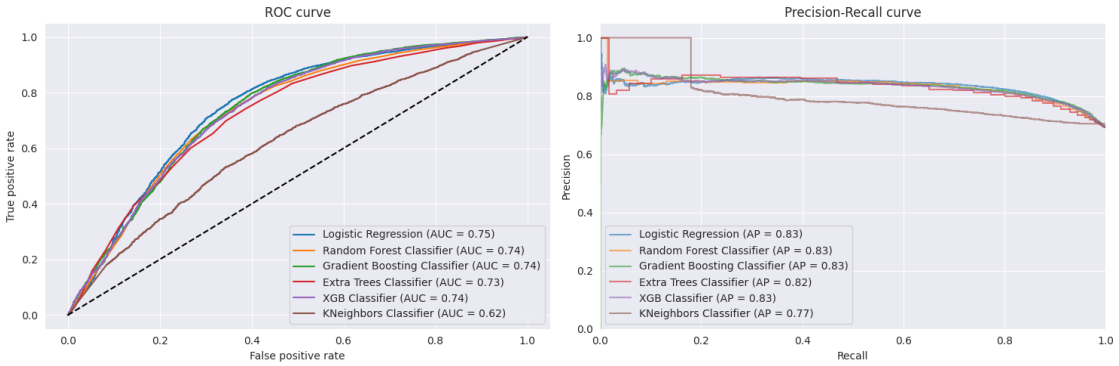

Figure 4.12 Graphical representation of the prediction performance of the "Variomic" model.

The Graphical representation of the prediction performance of the "Variomic" features based model evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

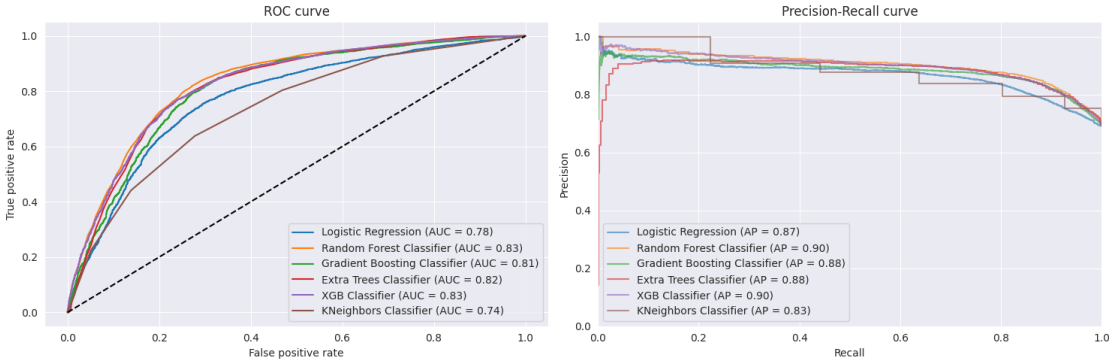

Figure 4.13 Graphical representation of the prediction performance of the model trained using the unified approach of defining the stress-responsive genes (GWAS + RNA-seq cut off).

The Graphical representation of the prediction performance of the model trained using the unified approach of defining the stress-responsive genes (GWAS + RNA-seq cut off

[|log2(fold change)| ≥ 1 and < -1, p-value < 0.05]) and on the comprehensive set of features . The performance is being evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

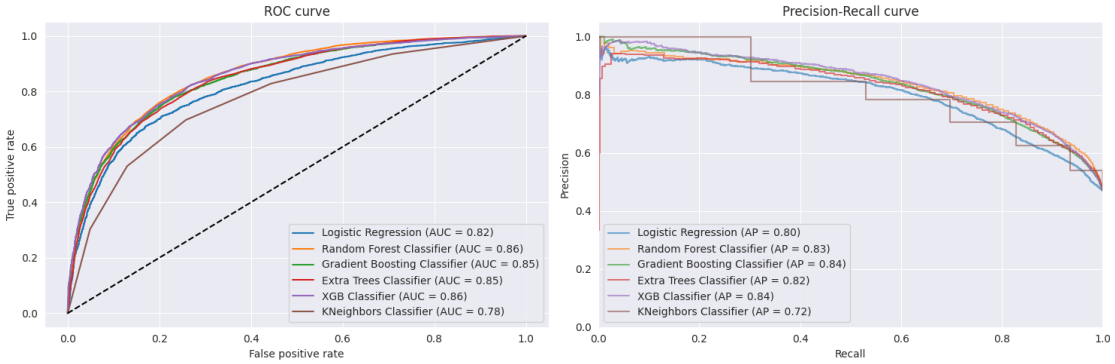

Figure 4.14 Graphical representation of the prediction performance of the model trained using the stringent RNA-seq cut off.

The Graphical representation of the prediction performance of the model trained using the stringent RNA-seq cut off [|log2(fold change)| ≥ 2 and < -2, p-value < 0.05] and on the comprehensive set of features . The performance is being evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

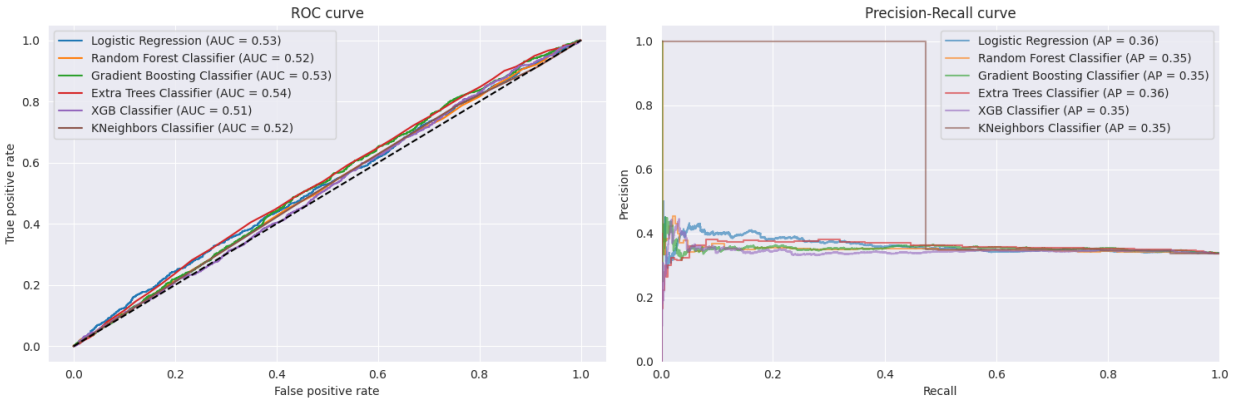

Figure 4.15 Graphical representation of the prediction performance of the model trained using GWAS based definition of stress-responsive genes.

The Graphical representation of the prediction performance of the model trained using GWAS based definition of stress-responsive genes as true labels and on the comprehensive set of features . The performance is being evaluated on the test set using AUC-ROC (left) and Precision Recall Curve PR metrics (right).

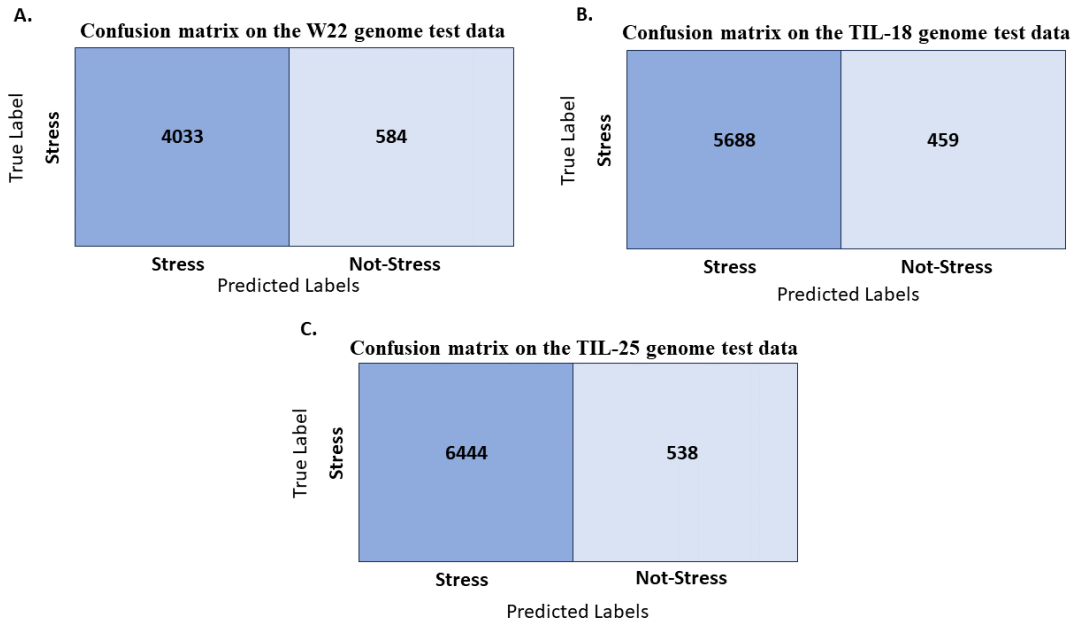

Figure 4.16 Confusion matrix displaying for evaluating W22, TIL-18, TIL-25 genome test datasets.

(A) Confusion matrix displaying the true positive and false negative instances on W22 genome test data. (B) Confusion matrix displaying the true positive and false negative instances on TIL-18 genome test data. (C) Confusion matrix displaying the true positive and false negative instances on TIL-25 genome test data.

CHAPTER 5. GENERAL CONCLUSION

The surge in sequencing capacity and cost reduction has facilitated the accumulation of extensive omics datasets, notably within repositories such as MaizeGDB (https://www.maizegdb.org/) , PMBB (Panzea Maize Bioinformatics Database) (https://www.panzea.org/), and MaizeCODE (http://www.maizecode.org/) . This reservoir of public omics data presents an opportunity for comprehensive meta and predictive analyses. A substantial portion of plant genes, particularly in maize, remains underexplored, exhibiting dissimilarity to known gene sequences and posing challenges in deciphering their functions. Homology-based approaches run the risk of introducing inaccuracies and misleading functional annotations. The strategic use of meta-analysis, leveraging the abundant omics data available, proves indispensable in understanding the contextual nuances for studying these genes (Bhandary et al. 2018).

However, despite the wealth of omics data, the shortage of reusable datasets is evident. Many laboratories worldwide contribute data to these repositories, yet the lack of user-friendly explanations for their datasets poses a hurdle for scientists seeking to repurpose the information (Gomez-Cabrero et al. 2014). This renders a substantial portion of deposited omics data inaccessible and complicates endeavors in meta-analysis.

Therefore, in this thesis, our primary objective was to address the challenges posed by vast and diverse omics datasets in terms of their accessibility and interpretability for addressing intricate biological issues related to maize genomes. Initially, we harnessed transcriptome data from a multitude of sources, incorporating over 200 unique datasets from 12 projects accessible through MaizeGDB for 26 Nested Association Mapping founder lines (Hufford et al. 2021). To enhance user interaction, we developed an interactive interface facilitating the comparison of

RNA-seq expressions across various data sources for a user-defined gene list or genomic interval. Additionally, users can visually compare expressions between two genes, capitalizing on the expanding pool of public RNA-Seq datasets.

Subsequently, we introduced a framework, the Maize Feature Store (MFS), housing gene-based machine learning features derived from multi-omics data to aid in exploring and modeling classification problems. The MFS incorporates over 14,000 gene-based features sourced from published genomic, transcriptomic, epigenomic, variomic, and proteomics datasets. Furthermore, the MFS integrates supervised and unsupervised machine-learning algorithms, streamlining the analysis and prediction of intricate genome annotations. A practical application of the MFS showcased its effectiveness in achieving high classification accuracy when distinguishing core and non-core genes within the maize pan-genome.

Lastly, we harnessed the comprehensive array of omics features to unravel the intricacies of stress-responsive genes, pinpointing key factors associated with these genes. This chapter serves as a synthesis of critical findings and outlines potential avenues for future research arising from the investigations conducted in this dissertation.

5.1 Specific findings and contributions

5.1.1 qTeller: a tool for comparative multi-genomic gene expression analysis

qTeller was developed to address the need for an accessible tool to organize, integrate, access, compare and visualize gene expression data. Though it was previously unpublished, the tool has been used and cited broadly by the plant research community in the study of evolution, meta-analyses, gene and gene family identification, quantitative trait and association studies and ontology. MaizeGDB expanded qTeller’s functionality to include multiple genomes and protein abundance data and enhanced the website layout to make qTeller even easier to use. qTeller was

designed for plant species but is broadly extendable to any species with a sequenced genome and RNA-Seq or protein abundance data.

5.1.2 Maize Feature Store (MFS): A centralized resource to manage and analyze curated maize multi-omics features for machine learning applications

In this work, we aimed at the needs of both experimental and computational researchers by providing them with two instances of the tool: a user-friendly instance and a modular instance. The user-friendly side of the tool will assist researchers without any prior computational background, by providing them a convenient one-click interface to retrieve, analyze and model heterogeneous maize data. In contrast, the tool's modularity will allow computational researchers to add additional functionality, fine-tune the existing functionalities, and model or perhaps even reproduce the entire application for the species of interest. Therefore, our approach enables both experimental and computational researchers to perform comprehensive analyses of maize multi-omics data, including methods to analyze the relationship between gene phenotypes and gene length, copy numbers and expression levels, epigenetic markers, cross-species conservation, and SNP densities, thereby covering a significantly larger portion of the maize genome and phenome.

Our models outperform random assignment for most downstream applications, but their accuracy rates are not high enough to replace pan-genomes altogether. However, if 89% accuracy is a satisfactory trade-off between complexity and ease in capturing preliminary variation in genes without comprehensive genome resources then our model can be the most optimal approach. Our model would also be ideal for newly sequenced or poorly annotated genomes. Where other tools like BLAST could also infer annotation, it does not provide underlying insights for the assignments beyond sequence homology.

5.1.3 Predicting genes associated with biotic or abiotic stress across different maize lines and related species

This study aimed to develop a more efficient framework for narrowing down the search space for stress-responsive genes, facilitating experimental validation by focusing efforts on a smaller, highly relevant subset of genes. A comprehensive examination of the top-ranked gene characteristics identified by the model, particularly those discriminating between the stress responsive and non-responsive genes, was conducted. These characteristics played a crucial role in advancing our understanding of the intricate nature of plant responses to multiple stressors.

The simplified models, trained solely on sequence or gene structural features, can be applied to predict stress-responsive genes in less-studied species with newly sequenced genomes. This research marks a significant stride in unraveling gene expression regulation, enhancing tools in synthetic biology and biotechnology for stress gene identification. Additionally, it addresses the lack in systematically characterizing stress-gene features, with particular implications for plant science research, given the intrinsic connection between stress responses and sustainability or productivity.

5.2 References

Bhandary, P., A. S. Seetharam, Z. W. Arendsee, M. Hur, and E. S. Wurtele. 2018. 'Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data', *Plant Sci*, 267: 32-47.

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner. 2014. 'Data integration in the era of omics: current and future challenges', *BMC Syst Biol*, 8 Suppl 2: I1.

Hufford, M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, J. Liu, W. A. Ricci, T. Guo, A. Olson, Y. Qiu, R. Della Coletta, S. Tittes, A. I. Hudson, A. P. Marand, S. Wei, Z. Lu, B. Wang, M. K. Tello-Ruiz, R. D. Piri, N. Wang, D. W. Kim, Y. Zeng, C. H. O'Connor, X. Li, A. M. Gilbert, E. Baggs, K. V. Krasileva, J. L. Portwood, 2nd, E. K. S. Cannon, C. M. Andorf, N. Manchanda, S. J. Snodgrass, D. E. Hufnagel, Q. Jiang, S. Pedersen, M. L. Syring, D. A. Kudrna, V. Llaca, K. Fengler, R. J. Schmitz, J. Ross-Ibarra, J. Yu, J. I. Gent, C. N. Hirsch, D. Ware, and R. K. Dawe. 2021. 'De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes', *Science*, 373: 655-62.